Université du Québec
Institut national de la recherche scientifique
Centre Énergie Matériaux Télécommunications

Physiological features for mental state monitoring in real life conditions

By

Abhishek Tiwari

A thesis submitted in fulfillment of the requirements for the degree of
Doctorate of Sciences, Ph.D
in Telecommunications

Evaluation Committee

Internal evaluator and committee president:	Prof. Sofiène Affes
External evaluator 1:	Prof. Behnaz Ghoraani Florida Atlantic University
External evaluator 2:	Prof. Rajkumar Elagiri Ramalingam Indian Institute of Technology Madras
Research advisor:	Prof. Tiago H. Falk

"The road goes ever on and on, down from the door where it began, now how far ahead the road has gone, and I must follow it if I can"

Bilbo

Acknowledgements

My deepest gratitude to my supervisor Dr. Tiago H. Falk for the support, opportunities and challenges provided throughout the PhD program, his role as supervisor surpasses all expectations, as such, his mentorship and attitude have been vital for the completion of this thesis, and for my personal development, it is great pleasure to work with him.

I would like also to thank all the members and ex-members of the MuSAE lab for their incredible help, advice, company, and for the fun we have had while discussing ideas, running experiments, rushing for deadlines, debating and so on. In particular, I would like to thank Raymundo Cassani, Shruti Kshirsagar and Anderson Avila for their support and camaraderie.

Without my family, I would not be the person that I am today, I would like to largely thank my family, my parents Suresh and Leela, and my brother Vivek and sister-in-law Anupama; who despite the distance are always in my mind, thanks for giving your love, support, advice and care.

There are not words to describe how grateful I am with every single person who helped to make this work possible. This thesis dedicated to all of you.

Abstract

Mental states refer to psychological states and emotions of an individual at any given instant. Constant negative mental states (e.g. stress, anxiety) can lead to a decline in job performance and efficiency. Such states can also cause job burnout and lead to various mental health disorders. Over the long term, they can also cause cardiovascular diseases in individuals. As a result, the overall cost associated with mental health far exceeds that of physical ailments. Currently, mental state monitoring is done using subjective questionnaires, audio-visual monitoring, and physiological signals. Of these, physiological signals allow for continuous, unobtrusive, and objective measurement of mental states while having very few privacy concerns. However, mental state monitoring models using physiological signals have traditionally been developed for data collected in controlled laboratory conditions using bulky data acquisition equipment. Such experiments try to control for noise and other confounding factors such as physical activity, circadian rhythm to name to few. As a result, models developed for such data cannot be used in real life a.k.a "in-the-wild" conditions. In this doctoral thesis, we present the steps towards the development of mental state monitoring models for data collected in highly ecological conditions. To achieve this goal, three main tools have been explored. These include (i) non-linear physiological features, (ii) building noise robustness in the features, and (iii) using multiple physiological modalities.

First, we explored noise-robust motif features for EEG based affect recognition. These features focus on the shape of the time series while ignoring the amplitude. The outcome of this exploration showed that motif based features can outperform traditionally used power spectral density and asymmetry features while being robust to artefacts. The motif based features show further improvement on fusion with the benchmark features thus suggesting complementarity of the features. Secondly, we evaluated various multi-scale entropy features for heart rate variability based mental workload assessment in ambulatory conditions. Multi-scale permutation entropy outperforms other entropy based methods. Additionally, comparable performance for different physical activity levels was achieved using the proposed features. Next, we proposed subband - complexity and -spectral descriptor heart rate variability features for stress and anxiety evaluation for in-the-wild conditions. These feature try to separate the effects of confounding factors by separately characterizing the high and low frequency behavior of the inter-beat interval series. The proposed features outperform the benchmark and commonly used non-linear feature sets in ambulatory and in-the-wild conditions. Lastly, we showed that multi-modal systems help improve performance for mental workload and stress assessment by building noise robustness as well as improving performance over short term windows. It is hoped that the insights presented herein help in the further development of methods for assessment of mental states in highly ecological conditions.

Keywords: Mental states, physiological signals, EEG, heart rate variability, stress, anxiety, wearables

Résumé

Les états mentaux désignent les états psychologiques et les émotions d'un individu à un instant donné. Des états mentaux négatifs constants, comme le stress et l'anxiété, peuvent entraîner une baisse des performances et de l'efficacité au travail. Ces états peuvent également provoquer un épuisement professionnel et entraîner divers troubles de la santé mentale. À long terme, ils peuvent également provoquer des maladies cardiovasculaires chez les individus. Par conséquent, le coût global associé à la santé mentale dépasse de loin celui des maladies physiques. Actuellement, la surveillance de l'état mental se fait à l'aide de questionnaires subjectifs, d'un suivi audio-vidéo et de signaux physiologiques. Parmi ces derniers, les signaux physiologiques permettent une mesure continue, discrète et objective des états mentaux tout en posant très peu de problèmes de confidentialité. Cependant, les modèles de surveillance de l'état mental utilisant des signaux physiologiques ont traditionnellement été collectées dans des conditions de laboratoire contrôlées, à l'aide d'un équipement d'acquisition de données encombrant. Ces expériences tentent de contrôler le bruit et d'autres facteurs de confusion tels que l'activité physique et le rythme circadien, pour n'en citer que quelques-uns. Par conséquent, les modèles développés pour de telles données ne peuvent pas être utilisés dans des conditions réelles. Dans ce doctorat, nous présentons les étapes vers le développement de modèles de suivi de l'état mental pour des données collectées dans des conditions réelles. Pour atteindre cet objectif, trois outils principaux ont été explorés. Il s'agit (i) des caractéristiques non linéaires, (ii) de la robustesse au bruit des caractéristiques, et (iii) de l'utilisation de modalités multiples physiologiques.

Tout d'abord, nous avons exploré les caractéristiques de motifs robustes au bruit pour la reconnaissance des emotions à partir de l'EEG. Le résultat de cette exploration a montré que les caractéristiques basées sur les motifs peuvent surpasser la densité spectrale de puissance et les caractéristiques d'asymétrie traditionnellement utilisées tout en étant robustes aux artefacts. Les caractéristiques basées sur les motifs montrent une amélioration supplémentaire lors de la fusion avec les caractéristiques de référence, ce qui suggère une complémentarité des caractéristiques. Deuxièmement, nous avons évalué diverses caractéristiques d'entropie multi-échelle pour l'évaluation de la charge mentale dans des conditions ambulatoires. L'entropie de permutation multi-échelle est plus performante que les autres méthodes basées sur l'entropie. En outre, les caractéristiques proposées ont permis d'obtenir des performances comparables pour différents niveaux d'activité physique. Ensuite, nous avons proposé des caractéristiques de descripteur de sous-bande -complexité et -spectral pour l'évaluation du stress et de l'anxiété dans des conditions réelles. Ces caractéristiques tentent de séparer les effets des facteurs de confusion en caractérisant séparément le comportement des hautes et basses fréquences des séries d'intervalles entre les battements. Les caractéristiques proposées sont plus performantes que les ensembles de caractéristiques non linéaires de référence qui sont couramment utilisés dans des conditions ambulatoires et en conditions réelles. Enfin, nous avons montré que les systèmes multimodaux contribuent à améliorer les performances de l'évaluation de la charge mentale et du stress en renforçant la robustesse au bruit et en améliorant les performances sur des fenêtres à court terme. Nous espérons que les résultats présentés ici contribueront au développement de méthodes d'évaluation des états mentaux dans des conditions réelles.

Mots-clés: États mentaux, signaux physiologiques, EEG, variabilité de la fréquence cardiaque, stress, anxiété, wearable

Contents

List of Figures

List of Tables

List of Abbreviations

$meanRR$	Mean of RR intervals
$pNNx$	Proportion of RR values with successive difference greater than x milliseconds
$rmssd$	Root Mean Square of Successive Difference of RR intervals
$sdRR$	Standard Deviation of RR intervals
ACC	Accuracy
ANS	Autonomic Nervous system
AV	Atrioventricular Node
BACC	Balanced Accuracy
C	Clustering coefficient
$\mathbf{C}_s$	Small world clustering coefficient
CMSE	Composite Multi-scale entropy
CNS	Central Nervous system
COPD	Chronic Obstructive Pulmonary Disease
CorrDim	Correlation Dimension
CV	Cross-Validation
DEAP	Dataset for Emotion Analysis using EEG, Physiological signals
DFA	Detrend Fluctuation Analysis
DL	Deep Learning
DNN	Deep Neural Network
DoF	Direction of flow
dRR	Magnitude of the difference intervals of the RR series
ECG	Electrocardiogram
EEG	Electroencephalogram
EMD	Empirical Mode Decomposition
ENPQ	Ecole Nationale de Police du Québec
F1	F1-score
FT	Fourier Transform
G	Global efficiency
GSR	Galvanic Skin Response
HCI	Human computer interaction
HF	HRV High Frequency

HRV	Heart Rate Variability
IADS	International Affective Digitized Sounds
IAPS	International Affective Picture System
IC	Independent Component
Isod	Inter-Scale Ordinal Distance
k	Degree of connectivity
L	Characteristic path length
L$_s$	Small world characteristic path length
LE	Lyapunov Exponent
LF	HRV Low Frequency
LOSO	Leave-One-Subject-Out
LOsO	Leave-One-Sample-Out
MATB-II	Multi-Attribute Task Battery-II
MCC	Matthews Correlation Coefficient
MIST	Montreal Imaging Stress Task
ML	Machine Learning
mPE	modified Permutation Entropy
mPE_wt	Weighted Modified Permutation Entropy
mRMR	Minimum redundancy maximum relevance
MSE	Multi-scale entropy
MuLES	MuSAE Lab EEG Server
NASA-TLX	NASA Task Load Index
PASS	Physical Activity under StreS
PE	Permutation Entropy
PNS	Parasympathetic Nervous system
PPG	Photoplethysmography
PSD	Power Spectral Density
PSD	Power Spectral Density
RBF	Radial Basis Function
RFE	Recursive feature elimination
RR	Inter-beat Interval
RSA	Respiratory Sinus Arrhythmia
RVM	Relevance Vector Machines
S	Small worldness of a network
SA	Sinoatrial Node
SAM	Self Assessment Manikin
SampEn	Sample Entropy
SE	Shannon Entropy
SNS	Sympathetic Nervous system

STAI	State-Trait Anxiety Inventory
SVM	Support vector machine
SWAT	Subjective Workload Assessment Technique
SWELL	Smart Reasoning for Well-being at Home and at Work Dataset
TILES	(Tracking IndividuaL performancE with Sensors
Tr	Transitivity
TSST	Trier Social Stress Test
VAS	Visual Analog Scale
VLF	HRV Very Low Frequency
WAUC	Workload Assessment Under physical aCtivity
wICA	wavelet enhanced Independent Component Analysis

Synopsis

0.1 Introduction

L'état mental désigne l'ensemble des émotions et des états psychologiques actuels d'une personne. Les états mentaux positifs peuvent accroître la satisfaction et l'épanouissement dans la vie [1], alors que les états mentaux négatifs entraînent une baisse des performances [2] et peuvent provoquer divers troubles de santé [3, 4, 5]. Une mauvaise santé mentale, en plus de diminuer la qualité de vie, a également un coût économique important en termes de perte de productivité et de frais de santé [6]. La détection et l'intervention précoces de la détérioration des états mentaux peuvent être des moyens efficaces pour prévenir les burnouts dans différentes professions [7, 8]. Les états mentaux sont également directement liés à la performance des tâches. Plus précisément, la charge mentale définie comme "le coût encouru par un opérateur humain pour atteindre un niveau de performance particulier" [9] est un indicateur important de la capacité d'un individu à réaliser une tâche avec succès. L'évaluation de la charge mentale est un facteur important dans l'optimisation de la performance d'une tâche pour un individu [10], en particulier dans les emplois critiques pour la sécurité tels que les premiers intervenants. Un niveau optimal de charge mentale est souhaitable pour réussir à accomplir les tâches, car des niveaux plus élevés peuvent provoquer du stress [11, 12]. L'évaluation de la charge mentale peut également aider à prendre des décisions en matière de conception ergonomique, par exemple, la conception d'un cockpit pour les pilotes ou bien le tableau de bord du conducteur dans les voitures [10]. L'interaction homme-machine (IHM) peut également être améliorée en intégrant l'intelligence émotionnelle dans les systèmes [13]. Par exemple, des systèmes d'apprentissage conscients de l'alerte et de l'ennui pour aider les éducateurs [13], et des systèmes de détection de la fatigue mentale pour les conducteurs ou les contrôleurs aériens [10]. Parmi les autres domaines où l'intelligence émotionnelle est utile, il y a la maison intelligente [14], les jeux [15] et les systèmes conversationnels intelligents [16]. Globalement, la surveillance de la santé mentale peut grandement contribuer à améliorer la qualité de vie et la qualité d'expérience des individus.

0.1.1 Méthodes actuelles de surveillance des états mentaux

Les états mentaux ont traditionnellement été evalué à l'aide de questionnaires subjectifs. Divers questionnaires standardisés ont été testés et utilisés pour différents états mentaux. Ces questionnaires s'enquièrent généralement de l'état mental d'un individu de manière périodique et les réponses sont données par une note (échelle continue) ou par plusieurs choix prédéfinis (échelle discrète). Pour la reconnaissance des émotions, les exemples incluent les six émotions de base présentées par [17] ainsi que le modèle arousal-valence présenté par [18]. La figure 1.1 illustre quelques émotions représentatives et leurs positions sur cette échelle. Pour des états mentaux plus spécifiques, plusieurs questionnaires ont été proposés. Il s'agit notamment des questionnaires NASA-TLX [9] et SWAT [19] pour la charge mentale, STAI [20] pour l'évaluation de l'anxiété, et l'échelle de Borg [21] pour la fatigue et l'effort perçu. De plus, les échelles visuelles analogiques (par exemple, des échelles de 5 et 10 points) ont également été utilisées pour évaluer les états mentaux [22, 23]. Les questionnaires subjectifs sont faciles à utiliser mais présentent plusieurs limites, notamment les biais psychologiques tels que la règle pic-fin et l'effet de récence [24, 25], la négligence et la complaisance dans les réponses conduisant à des évaluations erronées [26], ainsi que l'interruption et la frustration supplémentaires causées par le remplissage des questionnaires lorsque la tâche est en cours. Une alternative à la surveillance de l'état mental par le biais de questionnaires consiste à utiliser les données audio-visuelles générées par les individus. En général, les informations audio et vidéo sont faciles à collecter à l'aide de microphones et de caméras. Cependant, de tels systèmes sont confrontés à trois défis majeurs : i) le contrôle volontaire, ii) le fait que les modèles audio-visuels ne soient pas universels et que les émotions puissent changer d'une langue et d'une culture à une autre ; et iii) les problèmes de confidentialité liés au partage de ces données.

Une autre façon d'évaluer l'état mental consiste à utiliser des signaux neuro-physiologiques [13, 10]. Ces signaux reflètent l'activité du système nerveux central (SNC) et du système nerveux autonome (SNA). L'EEG capte l'activité électrique du cerveau (SNC) à l'aide d'électrodes placées à la surface du cuir chevelu. L'EEG a une haute résolution temporelle, est non invasif, rapide et peu coûteux, ce qui en fait une méthode privilégiée pour étudier les réponses du cerveau aux stimuli émotionnels [27, 28] aussi bien en laboratoire que dans des applications mobiles réelles [29, 30, 31]. L'activité du SNA régule à son tour les fonctions corporelles et peut être saisie à l'aide de signaux physiologiques tels que l'ECG, le signal respiratoire, la température de la peau et le GSR, pour en citer quelques-uns. Le SNA peut être divisé en deux systèmes : Le SNP, qui détend le corps, et le SNS, qui est associé à la réaction de combat-fuite. Ces deux systèmes ont un impact différent sur les signaux physiologiques et peuvent aider à distinguer différents états mentaux [32] individuellement ou de manière combinés dans un cadre multimodal [13]. Ces signaux peuvent être contrôlés en continu, ce qui rend possible leur utilisation en temps réel. Ces signaux sont également indépendants du biais lié au sujet. Les réponses physiologiques sont involontaires et universelles, et souffrent donc comparativement de moins variabilité entre les cultures. Enfin, elles suscitent moins d'inquiétudes quant au respect de la vie privée que leur équivalent audio-vidéo.

0.1.2 Passer du laboratoire au cas réel

Les études de surveillance de l'état mental ont traditionnellement été menées dans des conditions contrôlées en laboratoire. Ces expériences exigent généralement que le sujet reste immobile afin de minimiser le bruit des mouvements et de fournir un stimulus fixe pour évoquer un état mental donné pendant une petite période de temps. La surveillance des états mentaux dans des conditions "sauvages", c'est-à-dire dans des environnements hautement écologiques et quotidiens, à l'aide de dispositifs portables, présente une série de défis. Ces défis sont les suivants : (i) Bruit : Les signaux physiologiques peuvent être corrompus par différents types de bruit, par exemple le mouvement des muscles faciaux pour l'EEG et les artefacts de mouvement des électrodes pour l'ECG, pour n'en citer que quelques-uns. Bien que des algorithmes d'amélioration puissent être utilisés, ils nécessitent souvent de grandes quantités de ressources informatiques. Ces algorithmes ne sont pas en temps réel et peuvent supprimer des informations de signal pertinentes qui sont nécessaires à la prédiction de l'état mental [33]. (ii) Facteurs de confusion : Plusieurs facteurs de confusion peuvent interférer avec l'état mental mesuré. Il s'agit notamment des effets du rythme circadien, de l'activité physique et de la fatigue, pour n'en citer que quelques-uns. Pour minimiser leurs effets, les expériences en laboratoire tentent d'enregistrer les données au même moment de la journée pour tenir compte de la variabilité circadienne et éviter que le sujet ne se déplace [34]. Cependant, les expériences dans en cas réel impliqueront probablement de multiples stimuli émotionnels et physiques en même temps, ce qui nuit à la précision et à la fiabilité de la prédiction de l'état mental. (iii) Disponibilité des données : La plupart des bases de données existantes pour la surveillance de l'état mental ont été collectées dans des conditions de laboratoire [32, 28] avec des appareils de mesure encombrants, et avec seulement un petit nombre de participants enregistrant des données pendant une durée limitée. Ces données de courte durée ne permettent pas d'explorer les effets à long terme des stimuli émotionnels sur la physiologie. Par conséquent, le manque de bases de données publiques dans un cas réel rend difficile le développement de solutions fiables dans des environnements hautement écologiques.

0.1.3 Outils pour la surveillance de l'état mental en cas réel

Trois grandes familles d'innovations pour la surveillance de l'état mental dans des cas réels ont été proposées dans cette thèse. Ce sont : (i) Les caractéristiques non linéaires : Il existe de plus en plus de preuves de l'utilité des caractéristiques non linéaires dans différents contextes cliniques [35, 36] et pour les tâches de surveillance de l'état mental en laboratoire [37, 38]. Cependant, leur utilité dans un cas réel doit encore être évaluée. (ii) Caractéristiques robustes au bruit : Les méthodes d'extraction de caractéristiques qui intègrent la robustesse directement dans les caractéristiques peuvent minimiser les effets du bruit [39]. (iii) Fusion multimodale : Les systèmes multimodaux sont plus robustes au bruit et peuvent encore améliorer les performances en raison de la présence d'informations complémentaires dans diverses modalités. Ces caractéristiques rendent les systèmes multimodaux idéaux pour les applications en cas réels.

Ces trois innovations ont été décrites dans plusieurs manuscrits, dont la liste détaillée se trouve dans la section 1.5.

0.1.4 Organisation de la thèse

Alors que ce chapitre d'introduction (Chapitre 1) a présenté les défis du suivi de l'état mental dans des cas réels et a présenté les bases des contributions décrites ci-dessous, le reste de cette thèse est structuré de la manière suivante : Le chapitre 2 fournit une vue d'ensemble des méthodes de pointe en matière de surveillance de l'état mental, ainsi qu'une liste des bases de données actuellement disponibles. Le chapitre 3 présente l'utilisation de caractéristiques EEG basées sur des motifs pour améliorer la reconnaissance des émotions sans artefact. Ensuite, le chapitre 4 combine les motifs avec des caractéristiques non linéaires multi-échelles pour le signal ECG afin de prédire la charge mentale dans des conditions ambulatoires. Ensuite, le chapitre 5 traite des facteurs de confusion affectant les signaux physiologiques en utilisant des caractéristiques sous-bande pour l'ECG. Dans le chapitre 6, des systèmes multimodaux sont présentés pour améliorer la robustesse au bruit et les performances de la surveillance de l'état mental en utilisant plusieurs modalités physiologiques en même temps. Enfin, le chapitre **??** présente les conclusions générales de cette thèse, ainsi que les domaines de recherche futurs.

0.2 Chapitre 2: Surveillance de l'état mental basée sur la physiologie : état de l'art

Divers signaux neurophysiologiques ont été utilisés pour la surveillance de l'état mental. Ces signaux reflètent les changements dans le SNC ou le SNA et partagent certaines propriétés et défis de base, (i) Ils sont collectés de manière non invasive à partir de diverses parties du corps en utilisant différentes techniques de détection, par exemple, des électrodes de cuir chevelu pour l'EEG. (ii) la plupart de ces signaux présentent des propriétés non linéaires, (iii) ils sont caractérisés par un ensemble de caractéristiques de référence permettant de saisir les changements du SNC ou du SNA, (iv) ils sont sensibles à différents types d'artefacts physiologiques ou non physiologiques qui peuvent modifier leurs caractéristiques. Une fois les signaux acquis et traités pour améliorer la qualité du signal, les caractéristiques peuvent être extraites. Ces caractéristiques peuvent ensuite être combinées ou introduites individuellement dans un pipeline d'apprentissage machine (ML) et utilisées pour prédire les états mentaux. Les sections suivantes couvrent d'abord les signaux neurophysiologiques pertinents utilisés dans cette thèse, en décrivant leurs propriétés, les sources d'artefacts et leur suppression, et enfin les caractéristiques de référence traditionnelles utilisées. Ensuite, les différents types de systèmes multimodaux sont présentés, suivis par les différents composants du pipeline ML. Ensuite, certains des ensembles de données disponibles publiquement et leurs limites sont décrits. Enfin, les ensembles de données collectés dans des cas réels pour cette thèse sont présentés.

0.2.1 Electroencéphalographie

0.2.1.1 Source du signal et acquisition

Les signaux EEG capturent l'activité électrique spontanée du cerveau et sont devenus un outil fiable pour le suivi de l'état mental des individus [28]. Les signaux sont capturés en plaçant de multiples électrodes sur le cuir chevelu à l'aide d'un capuchon d'électrode, comme le montre la Fig. 2.1. Ces électrodes, dont le nombre varie de 20 à 256, sont placées sur le cuir chevelu en fonction de normes spécifiques visant différentes régions du cerveau ; le placement de 10 à 20 électrodes [40], par exemple, est le montage le plus populaire. Le choix du nombre d'électrodes est généralement basé sur le compromis entre la résolution spatiale souhaitée et le confort du sujet. Ces dernières années, des travaux ont exploré l'utilisation d'un plus petit nombre d'électrodes placées sur des régions spécifiques du cuir chevelu en fonction de la tâche à explorer. Des résultats prometteurs ont été trouvés pour diverses applications allant du diagnostic de la maladie d'Alzheimer [41] à la classification des états mentaux [42].

0.2.1.2 Propriétés du signal

Les électrodes tentent de capter des informations provenant de différentes régions du cerveau. Ces régions sont globalement divisées en quatre zones principales, à savoir les lobes temporal, occipital, frontal et pariétal, sur la base de l'anatomie du cerveau [43]. Il est également démontré que ces régions sont (en gros) responsables de fonctions cérébrales spécifiques. En ce qui concerne les émotions, deux régions, à savoir l'amygdale (dans la partie frontale du lobe temporal) et le cortex préfrontal (situé dans le lobe frontal), présentent des modifications lorsqu'on leur présente des stimuli émotionnels. En général, l'activation de l'amygdale est liée aux émotions négatives. Les signaux EEG d'individus sains ont également été divisés en plusieurs bandes sur la base d'une analyse visuelle d'individus sains [44]. Ces régions comprennent les bandes delta (δ : $1 - 4Hz$), thêta (θ : $4 - 8Hz$), alpha (α : $8 - 12Hz$), bêta (β : $12 - 32Hz$) et gamma (γ : $32 - 45Hz$). La quantification des propriétés des signaux EEG en fonction des différents rythmes cérébraux s'est avérée utile pour la reconnaissance des émotions [13] et d'autres tâches de surveillance de l'état mental [45, 32]. On a observé que l'activité cérébrale est fonction de divers processus allant d'une activité électrique rapide à des réactions chimiques plus lentes et à des processus diffusifs. Ces interactions non linéaires ont été étudiées à l'aide de la dynamique non linéaire. La quantification des signaux EEG à l'aide de telles mesures a montré qu'elle permettait de capturer des informations non redondantes par rapport aux mesures (linéaires) traditionnellement utilisées [36].

0.2.1.3 Suppression des artefacts

Les signaux EEG sont très sensibles au bruit et sont affectés, par exemple, par les mouvements oculaires, les clignements d'yeux et les interférences des lignes électriques, pour n'en citer que quelques-uns. Ces artefacts nécessitent généralement une détection et une suppression manuelles. Récemment, les algorithmes de suppression automatique des artefacts ont fait l'objet d'un grand intérêt. Ces méthodes reposent généralement sur des techniques de séparation aveugle des sources et sont coûteuses en termes de calcul. De plus, les algorithmes peuvent supprimer des informations pertinentes du signal EEG pour le problème de prédiction, ce qui peut entraîner une diminution des performances. De tels effets ont été précédemment démontrés pour la prédiction de la gravité de la maladie d'Alzheimer [33], par exemple. D'autres étapes de prétraitement comprennent le filtrage dans le domaine temporel, avec des filtres coupe-bande pour éliminer les interférences du réseau électrique (50 ou 60 Hz, selon le pays), et le filtrage passe-bande pour améliorer uniquement les composantes spectrales liées à l'EEG. Ces méthodes permettent d'améliorer la qualité du signal avant une analyse plus approfondie.

0.2.1.4 Fonctionnalités utilisées

Les caractéristiques les plus couramment utilisées pour les applications de surveillance de l'état mental sont liées à la puissance des différentes bandes d'énergie EEG et à leurs asymétries inter-hémisphériques. Plus précisément, (i) les caractéristiques de puissance spectrale mesurent la puissance de l'EEG dans les différentes bandes prédéfinies et sont les caractéristiques les plus couramment utilisées dans divers domaines, y compris la classification des états émotionnels. Le calcul de la puissance dans le domaine temporel pour une bande d'énergie donnée est décrit dans l'équation (ii) L'asymétrie de la puissance de l'EEG entre les deux hémisphères s'est également avérée être une caractéristique utile pour la détection de l'état émotionnel [28] et l'évaluation de l'humeur générale [46]. L'asymétrie est calculée en prenant le rapport de la DSP entre l'électrode gauche et sa paire d'électrodes droite correspondante. Les paires d'électrodes couramment utilisées sont décrites dans la section 2.2.4. En général, l'asymétrie EEG peut donc refléter la dimension de la valence [28, 46]. D'autre part, l'asymétrie pré-frontale en alpha et l'asymétrie temporelle dans la bande gamma sont observables pour la reconnaissance de l'éveil [47]. Récemment, d'autres propriétés du signal EEG ont également été utilisées pour la détection des états mentaux. Il s'agit notamment de métriques dérivées de l'analyse de la connectivité fonctionnelle [48], de caractéristiques de modulation d'amplitude [**?**] et de caractéristiques non linéaires [28, 49].

La figure 2.2 représente le pipeline général d'analyse des données EEG montrant les différents composants décrits dans les sections précédentes.

0.2.2 Electrocardiogramme

0.2.2.1 Source du signal et acquisition

L'ECG est une mesure non invasive de l'activité, de la structure et de la fonction cardiaques. Il enregistre les différents potentiels électriques produits par le cœur à la surface du corps. Cette activité électrique rythmique est basée sur une stimulation bioélectrique répétitive des muscles cardiaques : l'électrocardiogramme. Cette activité électrique peut être captée en plaçant des paires d'électrodes à la surface du corps et est appelée ECG. L'évaluation clinique de l'ECG nécessite l'utilisation de positions standard pour le placement des électrodes. Le placement standard de l'ECG à 12 dérivations a été largement adapté comme norme pour les enregistrements cliniques. Les dispositifs portables, cependant, utilisent des placements d'ECG à une seule dérivation pour des raisons pratiques. Ces dispositifs ont été largement comparés aux placements ECG à 12 dérivations pour diverses applications [50, 51, 52] et présentent en général une corrélation élevée avec les mesures cliniques. La figure 2.3 représente un ECG typique composé de différentes formes d'onde générées à partir d'un cycle cardiaque pour un individu sain, à savoir : l'onde P, le segment PR, le complexe QRS, le segment ST et l'onde T. Les intervalles de temps entre ces différents pics et les différences d'amplitudes sont connus pour contenir des informations cliniques importantes [53].

La variabilité de la série RR représente l'activité du SNS et est appelée variabilité de la fréquence cardiaque (VFC). Plus précisément, le SNS est responsable de la diminution de la valeur RR et de sa variabilité tandis que le PNS augmente les valeurs RR ainsi que leur variabilité. Le SNP domine généralement l'activité de la fréquence cardiaque au repos et pendant la récupération, tandis que le SNS est actif pendant la réaction "combat-fuite" du corps. Bien que ces deux systèmes modifient la VRC de manière contradictoire, leur interaction est de nature non linéaire. Si, en général, l'activation du SNS peut supprimer le PNS, elle peut même augmenter l'activité du PNS ou au contraire ne pas la modifier. La durée recommandée d'enregistrement de l'ECG pour l'analyse du VRC est divisée en deux catégories : l'analyse à long terme et l'analyse à court terme, la première reposant sur des fenêtres de 24 heures et étant généralement utilisée en milieu clinique. Pour l'analyse à court terme, une fenêtre de 5 minutes est recommandée, car elle permet de mesurer clairement les éléments périodiques de basse fréquence du système sympathique. La mesure du VRC à court terme est le résultat de deux sources distinctes. Premièrement, la relation complexe et dynamique entre le SNP et le SNS. Deuxièmement, les mécanismes de régulation qui contrôlent la VRC par le biais de l'arythmie du sinus respiratoire (ARS) : [54, 55]. L'ARS désigne l'accélération et le ralentissement du cœur par la respiration via le nerf vague. Récemment, des fenêtres inférieures à 5 minutes (offrant une résolution temporelle accrue) ont été explorées pour diverses applications de surveillance de l'état mental. Cependant, la diminution de la fenêtre affecte la fiabilité de ces mesures par rapport aux analyses standard de 5 minutes [56] et leur utilisation reste controversée [54].

0.2.2.2 Suppression des artefacts

Le signal ECG brut est généralement corrompu par diverses sources de bruit, notamment le mouvement des électrodes, les interférences des lignes électriques et le bruit des instruments, pour n'en citer que quelques-unes. Ces bruits peuvent rendre difficile la détection de la forme d'onde de l'ECG. Par conséquent, l'analyse du VRC ne peut pas être effectuée sur les signaux bruts et nécessite une certaine forme de filtrage. De nombreux algorithmes de filtrage différents (généralement basés sur le débruitage des ondelettes ou l'EMD) ont été proposés pour filtrer les signaux ECG afin d'améliorer la qualité du signal : [57, 58, **?**, 59]. Pour l'analyse du VRC, les méthodes de filtrage n'ont pas besoin de perversion de la morphologie des ondes, car nous devons uniquement détecter les pics QRS. Ainsi, un simple filtrage passe-bande dans la gamme 5-25 Hz est couramment utilisé pour améliorer les pics du QRS. Un tel filtrage passe-bande a été intégré dans un grand nombre d'algorithmes de détection des QRS [60]. Une fois le signal ECG filtré, les pics du QRS sont détectés à l'aide d'un algorithme de détection des pics. L'algorithme de Pam-Tompkins [61] est l'une des méthodes les plus utilisées. Au fil des années, plusieurs modifications de l'algorithme original ont également été proposées pour augmenter la robustesse aux artefacts [62, 63, 64].

Un ECG bruité peut altérer la série RR en raison de pics QRS manquants ou faux. Les artefacts physiologiques constituent une autre source de bruit des séries RR sous la forme de plusieurs battements anormaux. Ces battements anormaux peuvent fausser les mesures du VRC calculées et sont connus pour introduire des erreurs dans plusieurs mesures du VRC [39]. Traditionnellement, l'analyse visuelle de l'ECG et des séries RR a été effectuée pour exclure les régions présentant un signal de mauvaise qualité ou pour modifier les séries RR détectées. Ces dernières années, des algorithmes automatisés de suppression, d'interpolation ou de filtrage des signaux RR ont été proposés. Chacune de ces méthodes peut avoir un impact différent sur les mesures de la VRC et il n'existe pas de normes actuelles pour le traitement des séries chronologiques RR bruyantes. Les méthodes existantes tentent d'abord de détecter les segments anormaux, puis exécutent divers algorithmes de filtrage et de détection des aberrations. Le pipeline global d'analyse de la VRC est illustré dans la Fig 2.4. Après le filtrage des séries RR, les caractéristiques VRC sont calculées à partir des séries. La section suivante décrit ces caractéristiques VRC couramment utilisées.

0.2.2.3 Caractéristiques de référence du VRC

Les caractéristiques des domaines temporel et fréquentiel sont couramment recommandées comme caractéristiques de référence pour l'analyse du VRC [55]. Comme l'EEG, la série RR présente également un comportement complexe de type fractal [65]. Au cours des dernières décennies, ces caractéristiques ont été étudiées à l'aide de caractéristiques basées sur la dynamique non linéaire et la théorie du chaos. Ces caractéristiques ont généralement montré une meilleure performance par rapport à l'ensemble de caractéristiques de référence et ont été recommandées pour diverses applications cliniques [35]. Ces différents ensembles de caractéristiques de référence sont décrits ci-dessous.

Les caractéristiques temporelles sont directement calculées sur la série RR. Les caractéristiques couramment utilisées comprennent la moyenne, l'écart-type, la racine carrée moyenne des différences successives de la série RR ainsi que le rapport des intervalles RR consécutifs différant de plus de "x" millisecondes, divisé par le nombre total d'intervalles RR (appelé pNNx). Ces caractéristiques représentent l'activité du SNS ou du PNS et sont souvent corrélées les unes aux autres [54]. L'analyse dans le domaine fréquentiel doit prendre en compte l'échantillonnage non uniforme des séries RR. Habituellement, cela se fait en rééchantillonnant la série temporelle RR à l'aide de méthodes d'interpolation linéaire ou cubique afin d'obtenir des échantillons équidistants [55]. Pour ce faire, on utilise l'indice temporel du pic QRS comme indice temporel de l'intervalle RR, puis on rééchantillonne pour obtenir la série dite de tachogrammes RR. Ensuite, la DSP du tachogramme est calculée en utilisant différentes méthodes, à savoir le FT, le périodogramme de Welch ou les modèles auto-régressifs. Comme pour l'analyse EEG, la DSP du VRC a été divisée en plusieurs bandes. Pour les segments de série RR de 5 minutes, ces bandes sont : très basse fréquence (VLF) (0,0033 Hz-0,04 Hz), basse fréquence (LF) (0,04 Hz-0,15 Hz) et haute fréquence (HF) (0,15 Hz-0,4 Hz). La puissance dans chacune de ces bandes peut être utilisée comme caractéristique. Les émotions négatives telles que le stress, l'anxiété et, en général, une charge mentale plus élevée sont associées à une réponse accrue du SNS avec un retrait du PNS. Cela correspond à une diminution de la puissance HF avec une augmentation de la puissance LF conduisant à une augmentation du rapport LF/HF ; [32, 66, 67, 68].

La série RR présente un comportement non linéaire complexe [65]. On a observé que ce comportement change en fonction de différentes conditions physiques et psychologiques [35]. La non-linéarité de la série chronologique RR a été quantifiée à l'aide de différentes mesures, telles que l'entropie, les fractales et les systèmes chaotiques/dynamiques. Les mesures basées sur l'entropie pour quantifier cette complexité comprennent (i) l'entropie d'échantillon (SE, définie dans l'équation 2.2) et (ii) l'entropie de permutation (PE, définie dans l'équation 2.3). PE calcule la distribution des motifs dans une série temporelle. Les motifs sont des configurations récurrentes dans une série temporelle définie par un degré (n) et un retard (λ), comme le montre la Fig. 2.5. Les motifs sont robustes au bruit car ils ne prennent en compte que la forme sous-jacente de la série temporelle. Ces entropies ont été utilisées pour diverses applications de surveillance de l'état mental [69, 70]. Les mesures fractales couramment utilisées comprennent (i) l'analyse de la fluctuation de la tendance, (ii) l'exposant de Lyapunov et (iii) la dimension de corrélation. Ces mesures ont été utilisées pour un grand nombre d'applications de surveillance de l'état mental [32, 70, 37]. Cependant, la performance de ces caractéristiques pour les applications de surveillance de l'état mental dans des cas réels reste encore à voir.

0.2.3 Autres modalités physiologiques

D'autres signaux physiologiques ont été utilisés pour le suivi de l'état mental [13]. Il s'agit notamment de :

Photopléthysmographe (PPG): La PPG est une technique optique non invasive et peu coûteuse qui détecte les changements pulsatiles du volume sanguin. Le signal PPG peut être mesuré à partir de différents endroits, comme le bout des doigts, le poignet et le lobe de l'oreille. La PPG peut être utilisée pour mesurer la fréquence cardiaque et le VRC. Par conséquent, la VRC basée sur la PPG a été utilisée pour la surveillance de l'état émotionnel dans un cadre multimodal [13], ainsi que pour la détection du stress [71], pour n'en citer que quelques-uns.

Signal de respiration: La respiration peut être mesurer par des dispositifs tels que les smart-shirts ou des sangles au niveau de la poitrine. La respiration peut être considérée comme un mélange de deux processus : la respiration métabolique et la respiration comportementale [72], cette dernière étant affectée par des stimuli internes et externes. En général, ces changements sont observables dans le rythme et la profondeur de la respiration. Il a été démontré que le stress mental et l'anxiété augmentent à la fois la fréquence et la profondeur respiratoires [73, 74]. Cette variabilité respiratoire a été quantifiée à l'aide de fonctions statistiques simples, telles que la moyenne et l'autocorrélation [75]. D'autres caractéristiques respiratoires couramment utilisées comprennent les puissances et les rapports de bande pour différentes régions de fréquence [13].

La reponse electrodermale (RED): Le signal RED capte les changements d'activité des glandes sudoripares et se compose de deux éléments : une composante tonique lente et une composante phasique plus rapide. Cette composante phasique est directement liée au niveau d'excitation émotionnelle. Par conséquent, les descripteurs statistiques ainsi que les puissances de bande des différentes régions de fréquence du RED ont été utilisés comme caractéristiques pour diverses applications de surveillance de l'état mental [13, 76, 77].

Température de la peau: On sait que la température de la peau varie en fonction des changements d'états émotionnels [13]. Généralement, des descripteurs statistiques sont utilisés comme caractéristiques pour les séries de température.

Nous abordons, ensuite, le développement de systèmes de reconnaissance multimodale qui s'appuient sur plusieurs de ces modalités.

0.2.4 Surveillance multimodal de l'état mental

Il est bien connu que les signaux physiologiques véhiculent des informations complémentaires pour la surveillance de l'état mental [13, 78, 79]. Les systèmes multimodaux offrent également une robustesse accrue contre les défaillances des capteurs, car la mauvaise qualité du signal dans une modalité peut être compensée par d'autres flux de signaux. Les informations multimodales peuvent être combinées à différents niveaux, comme l'illustre la figure 2.6. Il s'agit notamment (i) du niveau du signal (Fig. 2.6-a) : Les différents signaux physiologiques présentent souvent un comportement mixte car ils peuvent être régis par des mécanismes sous-jacents similaires. Par conséquent, des caractéristiques permettant de quantifier ce couplage ont été utilisées. Diverses caractéristiques de couplage comme le couplage cardio-respiratoire [80, 81], le couplage VRC-pression sanguine [82] et le couplage RED-EEG [83] ont été utilisées pour la détection de l'état mental, (ii) Niveau des caractéristiques (Fig. 2.6-b) : Pour cette fusion, les caractéristiques des différentes modalités sont extraites séparément et combinées avant de les introduire dans un pipeline ML. Il s'agit de l'une des approches les plus couramment utilisées [13, 78]. La fusion de caractéristiques nécessite de synchroniser les différentes modalités en faisant la moyenne sur les époques, par exemple. Habituellement, les caractéristiques de couplage, dans le cas de la fusion au niveau du signal, sont également fusionnées avec les caractéristiques individuelles, et (iii) le niveau de décision (Fig. 2.6-c) : La fusion au niveau décisionnel combine la sortie des pipelines ML de chacune des modalités en utilisant le vote majoritaire ou des méthodes pondérées. Cette méthode offre une certaine robustesse contre la corruption des modalités uniques par le bruit [84] et peut prendre en compte la qualité du signal [85].

La fusion au niveau du signal et des caractéristiques peut être effectuée avant le pipeline ML. En général, ces deux schémas sont capables d'apprendre l'interaction entre les signaux et les caractéristiques et peuvent fournir une plus grande amélioration de performance. La fusion au niveau des décisions, quant à elle, peut contribuer à la robustesse face à la défaillance d'un ou de plusieurs capteurs, en contrepartie de la performance. Dans la section suivante, les différents composants du pipeline d'apprentissage automatique (ML) sont décrits.

0.2.5 Canal d'apprentissage automatique

Une fois les caractéristiques extraites, la prédiction de l'état mental est réalisée à l'aide d'algorithmes d'apprentissage automatique. Les algorithmes d'apprentissage automatique supervisé utilisent les caractéristiques d'entrée ainsi que les classes cibles d'état mental correspondantes (binaires ou continues) pour apprendre la relation entre les deux. La classification binaire (par exemple, stress vs pas de stress) a été couramment utilisée pour la prédiction de l'état mental [48, 32]. Dans ce cas, les classes cibles des ensembles de données qui sont continues sont binarisées en fonction d'une valeur seuil spécifique. La donnée réelle du stimulus expérimental peut également être utilisée comme classe cible (par exemple, conditions de jeu vidéo stressantes ou relaxantes). En général, le pipeline ML implique une réduction de la dimensionnalité pour éviter un ajustement excessif, suivie de l'entraînement d'un modèle ML tout en utilisant une méthode d'évaluation spécifique. Enfin, les résultats sont évalués en utilisant certains chiffres de mérite. Dans cette section, nous décrivons les différents aspects du pipeline.

Sélection des caractéristiques: permet de réduire la dimensionnalité de l'ensemble de données afin d'éviter le surapprentissage. En outre, elle élimine plusieurs caractéristiques qui pourraient être fortement corrélées et ne pas fournir d'informations supplémentaires au classificateur. Plusieurs approches différentes de sélection de caractéristiques ont été proposées dans la littérature [86]. La sélection des caractéristiques permet également de mieux comprendre le processus de prise de décision du modèle en révélant les principales caractéristiques utilisées pour faire la prédiction. Les méthodes de sélection des caractéristiques comprennent : (i) ANOVA, (ii) mRMR et (iii) RFE. Ces méthodes prennent souvent en compte la corrélation avec les classes cibles ainsi que l'interaction avec d'autres caractéristiques.

Évaluation: Les algorithmes ML doivent être entraînés puis évalués sur différents échantillons de données afin d'éviter tout biais optimiste dans les résultats. La validation croisée est une stratégie utilisée pour créer différents fractionnements d'apprentissage et de test pour l'évaluation afin de réduire le biais en cas de données limitées. Dans la validation croisée standard, appelée validation croisée k-fold, l'ensemble de formation est divisé en k ensembles plus petits. Ensuite, pour chacun des "parties", l'étape suivante est répétée. Un modèle est formé en utilisant k-1 des parties comme données de formation et validé sur la partie restante des données (c'est-à-dire le k-ième partie retenu). Ce processus est répété jusqu'à ce que tous les k parties aient été utilisés pour la validation, ce qui conduit à k résultats différents. La moyenne de ces k-résultats constitue la performance finale. Idéalement, les modèles développés devraient pouvoir être généralisés aux données de nouveaux individus ou être personnalisés pour des individus spécifiques. C'est pourquoi différentes stratégies d'évaluation prenant en compte les données spécifiques au sujet ont été créées. Il peut s'agir de : *(i)* modèles adaptés au sujet, *(ii)* le test de modèle LOSO (Leave-one-subject-out), où les modèles sont formés sur les données de N-1 sujets et l'évaluation est effectuée sur le Nième sujet retenu jusqu'à ce que tous les sujets soient utilisés comme ensembles de test. Cependant, pour les ensembles de données comportant un petit nombre de participants (comme c'est généralement le cas dans les études sur les émotions), il est difficile d'assurer la généralité entre les sujets. Enfin, le test inter-sujets suppose que les données de chaque sujet sont disponibles à la fois dans les échantillons d'apprentissage et de test ; cette méthode est largement utilisée dans la littérature. Le paramètre K-fold CV couramment utilisé avec des données mélangées est un exemple de modèle inter-sujets.

Classification et facteur de qualité: Les machines à vecteurs de support (SVM) ont été largement utilisées dans la littérature [38, 71, 87, 88] en raison de leur capacité à traiter des données non linéaires. Une étude récente sur la reconnaissance des émotions à l'aide de signaux EEG [28] a révélé que la majorité des travaux utilisaient des SVM avec différents noyaux ; le noyau de la fonction de base radiale (RBF) a été largement employé pour aider à gérer ces non-linéarités. De plus, il a été démontré que les SVM ont une stabilité et une capacité de généralisation élevées et qu'ils ne sont pas affectés par un ajustement excessif. Pour évaluer les performances des classifieurs, la précision (ACC) et le score F1 (F1) sont les mesures les plus couramment utilisées [38, 48]. Ces métriques sont fiables lorsque l'ensemble de données est équilibré. Cependant, les données collectées en conditions réelles sont souvent déséquilibrées en raison du manque de contrôle sur les stimuli expérimentaux ainsi que des données manquantes ou bruitées. Dans de tels cas, il est nécessaire d'utiliser des mesures qui sont robustes aux déséquilibres. Les mesures couramment utilisées pour les ensembles de données déséquilibrés comprennent : (i) la précision équilibrée (BACC) : définie comme la moyenne de la sensibilité et de la spécificité (Eq. 2.5), et (ii) le coefficient de corrélation de Matthews (MCC) : Récemment, le MCC a été utilisé pour le calcul des performances pour les données déséquilibrées [89]. Le MCC prend en compte les quatre valeurs de la matrice de confusion pour renforcer la robustesse face au déséquilibre (Eq. 2.6).

La section suivante décrit les ensembles de données physiologiques disponibles pour la surveillance de l'état mental, ainsi que leurs limites.

0.2.6 Ensembles de données disponibles

Les données recueillies pour la surveillance de l'état mental à l'aide de signaux physiologiques reposent sur l'obtention d'états mentaux spécifiques tout en contrôlant simultanément les facteurs de confusion physiques ou mentaux. Il existe plusieurs ensembles de données accessibles au public pour la surveillance de l'état mental. Cependant, ces ensembles de données sont limités dans leur portée par deux facteurs : *(i)* le type de stimulus émotionnel utilisé dans ces expériences pourrait ne pas représenter correctement les situations de la vie réelle, et *(ii)* le type de conditions de contrôle utilisées pour garantir une collecte de données de haute qualité pourrait rendre l'utilisation difficile dans des conditions réelles en raison de la variabilité introduite dans les signaux. La présente section décrit les types de stimuli émotionnels fournis, certains des ensembles de données accessibles au public ainsi que leurs limites.

0.2.6.1 Stimuli émotionnels couramment utilisés

Les méthodes habituelles pour susciter ces états sont les suivantes : (i) Utilisation de stimuli multimédias : Un grand nombre d'études s'appuient sur l'utilisation de contenus multimédia, tels que la musique et les clips audio [90, 91, 92, 93], les images [94, 95] et les vidéos [13, 96, 84] pour susciter les états mentaux souhaités. Parmi les exemples de bases de données multimédias normalisées, citons l'International Affective Picture System (IAPS) [97], et l'International Affective Digitized Sounds (IADS) [98], (ii) Utilisation d'environnements simulés : De telles stratégies sont utilisées pour induire des états mentaux spécifiques par opposition à un large éventail d'émotions. Parmi les exemples les plus connus, citons le Trier Social Stress Test (TSST) [99] et le Montreal Imaging Stress Task (MIST) pour provoquer le stress [100], (iii) Utilisation de tests standardisés : Il existe plusieurs tests permettant d'évoquer la charge mentale, et par la suite, le stress mental tout en employant des ressources mentales spécifiques. Les plus populaires sont le test de Stroop [101], la recherche visuelle [102], et le calcul mental [103], pour n'en citer que quelques-uns. (iv) Utilisation de jeux vidéo/simulés : Les jeux de simulation peuvent aider à susciter des états mentaux dans des conditions qui reproduisent la vie réelle et sont plus immersives. Ainsi, les simulateurs de pilotage [104, 10] et de conduite [105, 106, 10] ont été utilisés pour étudier l'attention, la charge mentale et la fatigue. Les jeux vidéo ont également été utilisés pour fournir divers stimuli mentaux [107] dans des environnements normaux et virtuels [108]. Un autre simulateur populaire utilisé pour la charge mentale est la Multi-Attribute Task Battery-II (MATB-II) [109]. Nous décrivons ensuite les caractéristiques de certaines bases de données accessibles au public pour la recherche sur le suivi de l'état mental.

0.2.6.2 Ensembles de données accessibles au public

Il existe plusieurs ensembles de données accessibles au public pour les applications de surveillance de l'état mental. Souvent, ces ensembles de données sont multimodaux et mesurent diverses modalités de signaux physiologiques et autres. Ces bases de données permettent de comparer et de normaliser les méthodes dans la littérature. Pour la détection de l'état émotionnel, la base de données pour l'analyse des émotions à l'aide de signaux EEG et physiologiques (DEAP) et la base de données Mahnob HCI sont les bases de données les plus couramment utilisés, la grande majorité des études existantes s'appuyant sur l'un ou l'autre de ces bases de données, voire les deux [28]. Pour la base de données DEAP, 32 participants en bonne santé (50% de femmes, âge moyen = 26,9 ans) ont regardé 40 vidéos d'une minute. Des données EEG à 32 canaux (système de placement 10-20, taux d'échantillonnage : 512 Hz) et d'autres signaux physiologiques périphériques ont été enregistrés à l'aide d'un système Biosemi ActiveTwo (Amsterdam, Pays-Bas). On a présenté aux participants 40 vidéos musicales d'une minute au contenu émotionnel variable. Après la présentation de chaque vidéo, les participants ont été invités à évaluer les vidéos musicales sur des échelles discrètes de 9 points pour la valence et l'arousal en utilisant le questionnaire SAM. Les données EEG prétraitées accessibles au public ont été utilisées dans le chapitre 3. Un pipeline de prétraitement standard comprenant le référencement commun, le sous-échantillonnage à 128 Hz, le filtrage passe-bande (4-45 Hz) et la suppression des artefacts liés au clignement des yeux a été utilisé. MAHNOB-HCI [78], quant à lui, est une base de données multimodale enregistrée en réponse à des stimuli affectifs. Les signaux comprennent des signaux EEG et des signaux physiologiques périphériques, ainsi que des enregistrements de visages et de sons, et des données sur le regard. Vingt-sept participants ont pris part à deux expériences. Pour la première expérience, ils ont regardé 20 vidéos émotionnelles et ont indiqué eux-mêmes leur degré d'arousal, de valence, ainsi que des mots-clés émotionnels. Dans la deuxième expérience, de courtes vidéos et images ont été montrées une première fois sans aucune classe cible, puis avec des classes cibles correctes ou incorrectes. Les participants ont évalué leur accord ou leur désaccord avec les classes cibles affichées. Afin de contrôler tout bruit, les participants étaient assis pendant les expériences et ne parlaient pas et ne pratiquaient aucune autre activité physique dans les deux bases de données. Ces deux bases de données ont été largement utilisées. Cependant, elles ont recueilli des signaux physiologiques à l'aide d'équipements spécialisés et non portables avec des taux d'échantillonnage plus élevés. Ces appareils limitent l'utilisation des méthodes développées à des environnements de laboratoire contrôlés. Par conséquent, l'ensemble de données DREAMER [110] a récemment été publié. Le dispositif expérimental est similaire à celui du jeu de données DEAP, mais il fait appel à des dispositifs portables disponibles dans le commerce pour la collecte de données physiologiques.

Pour des états mentaux plus spécifiques tels que le stress, l'anxiété et la charge mentale, il n'existe qu'un petit nombre de jeux de données accessibles au public. Le plus populaire est celui décrit dans [111]. Il a recueilli les signaux physiologiques de participants conduisant un itinéraire fixe dans des conditions d'autoroute (stress moyen) et de ville (stress élevé). Les niveaux de base sans stress ont été recueillis en demandant au conducteur de rester assis dans la voiture. Les enregistrements comprennent l'ECG, le RED, l'électromyogramme (utilisé pour mesurer la tension du haut du dos (trapèze)) et les schémas respiratoires. Un autre ensemble de données récemment collectées pour l'évaluation du stress est le SWELL (Smart Reasoning for Well-being at Home and at Work). Cet ensemble de données est constitué de données multimodales provenant de 25 participants effectuant un travail de connaissance typique. Des facteurs de stress, sous forme d'interruptions de courrier électronique et de pression temporelle, ont été utilisés pour manipuler les niveaux de stress. Plusieurs évaluations subjectives ont été recueillies à la fin de la session d'une heure et les facteurs confondants ont été contrôlés en incluant une période de repos de 8 minutes avant la ligne de base. Les signaux recueillis à l'aide d'appareils disponibles dans le commerce comprennent l'ECG et le RED (enregistrés avec l'appareil Mobi, fréquence d'échantillonnage : 2048 Hz) ainsi que la vidéo et la posture du corps (à l'aide d'un kinect 3D).

0.2.6.3 Limitations des ensembles de données existants

Les ensembles de données actuellement disponibles présentent les limites suivantes lorsqu'il s'agit de les utiliser pour le développement d'applications de surveillance de l'état mental dans des conditions réelles. Ces limites sont les suivantes : (i) Mouvement : Les études de surveillance de l'état mental exigent généralement que le participant reste immobile. Ceci est fait pour minimiser les artefacts de mouvement dans les signaux collectés. De plus, l'activité physique peut modifier la dynamique des signaux physiologiques étudiés et accroître les ressources mentales utilisées. ii) Stimuli utilisés : Une variété de stimuli a été utilisée pour moduler divers états mentaux. Cependant, ces conditions ne sont pas nécessairement représentatives des conditions de vie réelles, (iii) Contrôle des facteurs de confusion : En dehors du mouvement, un grand nombre d'études utilisent des étapes supplémentaires pour contrôler d'autres effets. Cela se fait généralement en introduisant une période de base de repos et/ou de respiration profonde, et (iv) la durée de l'expérience : La plupart des expériences ont été menées pendant une petite durée, les stimuli durant moins de 5 minutes. De telles durées peuvent ne pas rendre compte des effets à long terme du stress et ne pas être suffisamment longues pour extraire des caractéristiques pertinentes.

0.2.7 Ensembles de données recueillies

Les limitations décrites dans la section précédente empêchent l'utilisation des méthodes développées dans des conditions de laboratoire contrôlées d'être utilisées de manière fiable dans des conditions réelles. Il est donc nécessaire de collecter des données qui reflètent les situations ambulatoires et réelles. Les bases de données suivantes ont été collectées dans le cadre de ce doctorat et utilisées pour les recherches décrites ici.

0.2.7.1 Données WAUC

Pour le jeu de données Workload Assessment Under physical aCtivity (WAUC) [**?**], des données ont été collectées auprès de 48 participants (23 femmes, 27.4 ± 6.6 ans). Vingt-deux participants ont réalisé l'expérience sur un tapis roulant et vingt-six sur un vélo stationnaire (comme indiqué sur la Fig. 2.7). La charge de travail a été modulée sur la tâche MATB-II (illustrée à la Fig. 2.8) en modifiant la difficulté de la tâche (facile ou difficile, correspondant à une charge de travail faible et élevée, respectivement). Chaque niveau de difficulté a été réalisé au repos (aucun mouvement), à une activité moyenne (3km/h : tapis roulant, 50 rpm : vélo) ou à une activité élevée (5 km/h, 70 rpm). On a ainsi obtenu un total de 6 combinaisons de charge de travail mental et physique. Ces 6 sessions duraient 10 minutes chacune. L'ordre des sessions était contrebalancé. L'expérience était précédée d'un tutoriel de 10 minutes. De plus, avant chaque session, deux lignes de base (aucune charge de travail physique/aucune charge de travail mental ; uniquement une charge de travail physique) de 1 et 2 minutes, respectivement, ont été recueillies. À la fin de chaque session, le sujet a rempli le questionnaire TLX (NASA-Task Load Index) [9] pour évaluer les différentes dimensions de la charge de travail pendant une période de pause de 5 minutes entre les sessions. Divers signaux physiologiques ont été recueillis, notamment l'EEG (500 Hz, Enobio 8 canaux), l'ECG (250 Hz), la respiration (18 Hz), l'accéléromètre (18 Hz) provenant du Bioharness 3, le PPG (64 Hz), la température cutanée (4 Hz), la reponse electrodermale (4 Hz) et l'accélération provenant de l'Empatica E4. L'acquisition et la synchronisation des signaux ont été effectuées à l'aide du serveur open-source MuSAE Lab EEG Server (MuLES) : [112].

0.2.7.2 Données PASS

Pour la base de données Physical Activity and StresS (PASS) [113], des données ont été recueillies auprès de 48 participants qui ont pratiqué une activité physique (niveau nul, moyen et élevé) sur un vélo stationnaire. Le stress a été modulé en passant d'un jeu vidéo à l'autre : Timeframe (jeu sans stress basé sur la collecte et l'exploration d'objets) et Outlast (un jeu d'horreur à la première personne). Pour les deux jeux, les sessions de jeu ont été divisées en trois segments prédéterminés choisis parmi les jeux pour assurer l'uniformité entre les participants. Chaque niveau de difficulté a été exécuté au repos (aucun mouvement), à une activité moyenne (50 tours par minute) ou à une activité élevée (70 tours par minute). On a ainsi obtenu un total de 6 combinaisons de stress et d'activité physique. Ces 6 sessions ont duré 10 minutes chacune. L'ordre des sessions était contrebalancé, les mêmes sessions de jeu étant effectuées en séquence. Le signal EEG (220 Hz) a été recueilli à l'aide de l'appareil Muse à 8 canaux, tandis que les mêmes signaux périphériques ont été recueillis comme pour l'ensemble de données WAUC. À la fin de chaque session, les participants ont rempli le questionnaire NASA-TLX modifié avec une dimension de stress ajoutée (échelle de 21 points). La figure 2.9 montre le dispositif expérimental utilisé.

0.2.7.3 ENPQ dataset

Pour cet ensemble de données, les données ont été collectées auprès de 27 (6 femmes) stagiaires policiers suivant un cours de 15 semaines à l'École nationale de police du Québec (ENPQ, Nicolet, Canada). Le cours comprenait la formation et l'évaluation de diverses compétences liées à la police (par exemple, la conduite d'une voiture de police, le combat à mains nues, les scénarios d'arrestation, les enquêtes criminelles, l'utilisation d'une arme à feu). Les données ont été recueillies en trois vagues. La première vague consistait en une collecte de données de 15 semaines sur la fréquence cardiaque des participants à l'aide d'un Fitbit pendant toute la durée du cours. La deuxième vague, quant à elle, a permis de recueillir des données au cours de cinq sessions (de 3 heures chacune) de divers exercices de tir où les participants ont été formés au maniement des armes à feu et aux compétences connexes. Plus de détails sur les cinq sessions peuvent être consultés dans le tableau 2.1. L'ECG (250 Hz) et les courbes respiratoires (18 Hz), ont été collectés pendant cette vague à l'aide du Bioharness Bh3. A la fin de chaque session, les stagiaires ont rapporté leurs évaluations de stress et de charge mentale à l'aide du questionnaire NASA-TLX [9] et de fatigue à l'aide de l'échelle de Borg [21]. La troisième vague a été réalisée lors de l'exercice du simulateur d'intervention. Cet exercice prolonge celui du stand de tir en se concentrant sur les aspects de la prise de décision et de l'intervention lorsqu'une arme à feu est impliquée. Ces exercices suivent le format standard d'une salle de classe avec une participation pratique des stagiaires dans un environnement de simulation impliquant des équipes de 1 à 4 personnes. La simulation a été réalisée à l'aide de l'installation décrite par la Fig. 2.10. Le scénario de simulation était affiché via un vidéoprojecteur (en bleu sur la figure) et les stagiaires étaient libres d'interagir avec les personnages à l'écran (par exemple, les suspects, les victimes, les témoins) et de se déplacer dans la zone de simulation comme ils le souhaitaient, ainsi que de se mettre à l'abri derrière différentes barricades (en rouge). Pendant la simulation, l'instructeur pouvait manipuler le comportement des suspects, allant de coopératif à hostile, modulant ainsi directement et/ou répondant aux actions et décisions des stagiaires. Chaque simulation durait plusieurs minutes. Comme pour l'exercice au stand de tir, les stagiaires portaient le dispositif Bioharness. Cette base de données permet de collecter des données physiologiques à long terme qui comprennent divers aspects sociaux (coordination et discussions), physiques (exercices de simulation et de manipulation) et mentaux (les niveaux de difficulté des diverses sessions sont différents). Cette base de données permet de tester les méthodes développées pour les données de longue durée qui peuvent comporter plusieurs facteurs de confusion inconnus.

0.2.7.4 TILES dataset

Les données ont été recueillies auprès de 200 participants (66 hommes, âge $38,6 \pm 9,8$ ans) provenant d'un groupe d'employés d'un grand hôpital urbain de Californie [114]. Deux tiers des participants étaient des infirmières et un tiers des employés de l'hôpital. Les données ont été recueillies pendant une durée de 10 semaines. Les participants ont effectué leur journée de travail comme d'habitude mais ont été invités à remplir une brève enquête quotidienne par téléphone qui comprenait des informations sur les niveaux d'anxiété et de stress sur une échelle de 5 points. Les participants ont été équipés de plusieurs capteurs portables permettant de recueillir diverses données biométriques, notamment les caractéristiques audiométriques, la fréquence cardiaque, la fréquence respiratoire et la qualité du sommeil. Un badge audiométrique personnalisé a été utilisé, comme détaillé dans [115], ainsi qu'un Fitbit Charge 2 et un OMsignal smartshirt. Le capteur basé sur le vêtement mesure une série RR à 4 Hz obtenue à partir de capteurs d'électrocardiographie (ECG) textiles intégrés dans la chemise. Le capteur fournit un paramètre de qualité interne appelé *RRPeakCoverage*, les valeurs les plus élevées indiquant une meilleure qualité. La distribution de la métrique de qualité est illustrée à la Fig. 2.12. Près de 27% des segments étaient en dessous des bons niveaux de qualité *RRPeakCoverage*, montrant ainsi l'impact du bruit sur les paramètres de collecte de données dans la nature. La collecte des données TILES a été effectuée dans des conditions totalement naturelles avec différents types de personnel hospitalier (infirmières, techniciens de laboratoire) pendant plusieurs quarts de travail d'une journée entière (environ 8 heures par quart). Par conséquent, cet ensemble de données permet de tester des caractéristiques qui pourraient saisir les effets à long terme des états mentaux. Comme pour d'autres ensembles de données, il n'existe pas de données de base/de repos et l'analyse se fait entre les états faibles/élevés.

0.2.8 Discussion

Ce chapitre a présenté le contexte de la surveillance de l'état mental à l'aide de signaux physiologiques. Tout d'abord, les différentes modalités physiologiques ont été abordées. Leurs méthodes d'acquisition, leurs propriétés, leur prétraitement et leurs caractéristiques de référence, ainsi que leur relation avec les états mentaux. Cette présentation a été suivie d'une brève description des différents types de systèmes multimodaux et de leurs avantages. Ensuite, les différents composants du pipeline d'apprentissage automatique ont été discutés, allant de la sélection et de l'évaluation des caractéristiques aux mesures de performance utilisées. Les bases de données disponibles ont ensuite été examinés, ainsi que leurs limites. Enfin, nous avons décrit certains des nouveaux ensembles de données collectés et utilisés pour cette thèse et leurs avantages. Dans l'ensemble, ce chapitre aborde le contexte de la littérature sur la surveillance de l'état mental tout en développant les différents outils nécessaires pour faire passer la recherche du laboratoire aux conditions réelles.

0.3 Chapitre 3 : Analyse basée sur les motifs des signaux EEG

Comme décrit dans le chapitre 2, les signaux EEG sont très sensibles aux artefacts, tels que les clignements des yeux et les mouvements musculaires [116]. Pour surmonter ces problèmes, des algorithmes de suppression des artefacts peuvent être utilisés. Il est également possible de développer de nouvelles caractéristiques résistantes au bruit et/ou d'explorer des stratégies de fusion multimodale [79]. Dans ce chapitre, l'accent est mis sur cette dernière solution et des caractéristiques basées sur les motifs sont proposées et testées seules ou avec d'autres caractéristiques complémentaires pour la reconnaissance des emotions. Si le PE est la métrique basée sur les motifs la plus populaire, d'autres approches ont également été développées. Celles-ci incluent le calcul d'une métrique de similarité entre deux séries temporelles basée sur des distributions de motifs [117], ainsi qu'une synchronisation entre deux séries temporelles [118].

0.3.1 Caractéristiques basées sur les motifs

PE est la caractéristique basée sur le motif la plus couramment utilisée et a été étudié ici. En outre, d'autres caractéristiques basées sur les statistiques des motifs récurrents dans la série de motifs peuvent être extraites. Les caractéristiques proposées ici ne prennent en compte que les motifs de degré $n = 3$ et de valeur de décalage $\lambda = 1$. Ces paramètres ont été suggérés dans le passé pour des tâches connexes [69, 118]. Les autres caractéristiques proposées comprennent : (i) la distance ordinale de dissimilarité : Cette [117] est une métrique très proche de l'indice d'asymétrie de référence et mesure la dissimilarité entre deux séries de motifs pour différentes paires d'électrodes (Eq. 3.1). Les paires d'électrodes pour le calcul des caractéristiques ont été décrites dans la section 2.2.4, et (ii) la synchronisation des motifs : La synchronisation des motifs a été proposée comme outil d'analyse de la connectivité fonctionnelle [118] et mesure l'apparition simultanée de motifs dans deux séries temporelles (Eq. 3.2). Une analyse grapho-théorique est ensuite menée sur les connexions fonctionnelles obtenues. Pour cette analyse, chaque électrode du cuir chevelu représente un nœud du réseau cérébral. Par conséquent, les graphes pondérés obtenus ont des poids qui représentent le niveau d'interaction entre les deux nœuds. Les réseaux non pondérés, obtenus par seuillage par la valeur moyenne du réseau pondéré, sont plus robustes aux connexions fonctionnelles bruyantes : [119]. Pour cette analyse, les réseaux pondérés et non pondérés ont été utilisés pour extraire divers paramètres de réseau, à savoir le degré de connectivité, le coefficient de regroupement, la transitivité, la longueur du chemin caractéristique, l'efficacité globale, et la direction du flux.

0.3.2 Mise en place expérimentale

Les données EEG prétraitées de la base de données DEAP ont été utilisées pour cette analyse. Les signaux ont été décomposés en bandes thêta, alpha, bêta et gamma. Ensuite, les caractéristiques de puissance spectrale pour les quatre bandes ainsi que les caractéristiques d'asymétrie de référence (pour les paires d'électrodes mentionnées dans la section 3.3) ont été calculées comme référence. De plus, les ratios des sous-bandes de l'EEG sont également inclus comme caractéristiques de référence. Les rapports calculés comprennent : $\frac{\gamma}{\beta}, \frac{\beta}{\theta}, \frac{\alpha}{\theta}, \frac{(\alpha+\beta)}{\gamma}$, and $\frac{(\gamma+\beta)}{\theta}$. Enfin, l'entropie de Shannon [120] a été utilisée comme caractéristique pour mesurer la complexité des séries temporelles EEG. Le tableau 3.1 fournit un résumé du nombre de caractéristiques extraites pour chaque groupe et sous-groupe de caractéristiques.

Nous explorons les effets de la combinaison des caractéristiques basées sur les motifs proposés avec les caractéristiques de référence. Cependant, étant donné la petite taille de l'ensemble de données, il est important d'éviter les problèmes de la dimentionnalité et de surapprentissage, ce qui nécessite une sélection des caractéristiques. Nous avons exploré ici trois approches différentes de sélection de caractéristiques : ANOVA, RFE, mRMR. Pour les expériences présentées ici, 90$ des données sont réservées à la sélection des caractéristiques et à l'apprentissage du classificateur, et les 10% restants sont réservés aux tests. Le meilleur algorithme de sélection des caractéristiques et le nombre optimal correspondant de caractéristiques sont ensuite sélectionnés. Ici, les classificateurs SVM (avec des noyaux RBF) sont entraînés sur deux problèmes de classification binaire différents, à savoir la discrimination entre les états de valence faible et élevée, ainsi que les états d'arousal faible et élevée. Comme nous souhaitons explorer les avantages des caractéristiques de motifs proposées et les comparer aux caractéristiques de référence, nous n'effectuons pas d'optimisation des hyperparamètres du classificateur et utilisons les paramètres par défaut donnés dans la bibliothèque Scikit-learn [121]. Les classes cibles binaires élevées/faibles pour la classification de la valence et de l'arousal ont été obtenues en utilisant un seuil propre au sujet afin de supprimer le biais subjectif des évaluations, comme le montre la figure 3.1. Pour prédire correctement la performance de l'ensemble de données déséquilibré résultant, une précision équilibrée (BACC) a été utilisée comme figure de mérite. Pour tester la signification des performances atteintes par rapport au hasard, un test t indépendant à un échantillon contre un classificateur à vote aléatoire a été utilisé ($p \leq 0.05$), comme suggéré dans [13]. Afin d'obtenir une performance plus généralisée du classificateur, une fois l'étape de sélection des caractéristiques effectuée, l'entraînement et le test du classificateur sont réalisés 100 fois avec différentes partitions entraînement/test. Les valeurs de BACC indiquées dans le tableau correspondent à la moyenne ± l'écart type de toutes les valeurs de BACC atteintes sur l'ensemble de test sur l'ensemble des 100 itérations. Plusieurs stratégies de fusion de caractéristiques ont été testées pour combiner les différents ensembles de caractéristiques. Elles comprennent (i) la fusion de caractéristiques : Pour cette méthode, les vecteurs de caractéristiques du motif et du repère sont combinés avant la sélection des caractéristiques, (ii) fusion au niveau du score : La méthode de fusion par décision pondérée proposée dans [122] a été utilisée (Eq. 3.17), et (iii) la fusion associative de sortie : Des preuves psychologiques ont suggéré une forte intercorrélation entre les dimensions de valence et d'éveil [123, 124, 125, 126]. Le cadre de l'OAF a été exploré ici et est représenté par le schéma fonctionnel de la Fig. 3.2. Comme on peut le voir, les classificateurs individuels font d'abord les prédictions de valence et d'arousal pour chaque groupe de caractéristiques individuelles. Cette étape est ensuite suivie d'une étape de prédiction finale qui prend en compte les dimensions de la valence et de l'arousal afin de mieux prédire chacun des deux résultats.

0.3.3 Résultats expérimentaux

La sélection des caractéristiques a été mise en œuvre dans les caractéristiques de référence seules, dans la caractéristique de motif proposée seule, et dans l'ensemble combiné référence-motif. Les valeurs optimales de BACC obtenues sont présentées dans les Tableaux 3.2-3.4, respectivement, ainsi que le nombre final de caractéristiques (nof) utilisées dans les modèles. La sélection RFE a généralement donné lieu à la précision la plus élevée avec le meilleur compromis entre $BACC$ et nof. Globalement, la meilleure précision a été obtenue avec l'ensemble combiné, suivi de près par les modèles formés sur les caractéristiques de motifs proposées. Les tableaux 3.5 et 3.6 présentent les 20 caractéristiques les plus utilisées dans les modèles qui ont obtenu le meilleur $BACC$ pour la valence et l'excitation, respectivement. Pour évaluer la contribution de chaque sous-groupe de caractéristiques individuelles à la reconnaissance des états affectifs. Le tableau 3.7 indique la précision équilibrée pour chaque sous-groupe de caractéristiques individuelles pour le modèle le plus performant trouvé après la sélection des caractéristiques RFE. Le tableau 3.8 compare les différentes méthodes de fusion pour les ensembles de caractéristiques de référence et de motifs. Alors que la fusion des caractéristiques donne les meilleures performances pour l'estimation de la valence, la fusion des niveaux de score est la méthode la plus performante pour l'arousal.

0.3.4 Discussion

D'après les tableaux 3.2-3.4, on peut voir que pour la valence, l'ensemble fusionné de caractéristiques de référence et de motifs avec la sélection de caractéristiques RFE est la méthode la plus performante avec une valeur BACC de $0,6010$, tandis que pour l'excitation, l'ensemble fusionné de caractéristiques avec la sélection de caractéristiques mRMR donne la meilleure performance de $0,5645$. Si l'on examine les performances des différents groupes de caractéristiques dans le tableau 3.7, pour la valence, toutes les caractéristiques basées sur les motifs ont obtenu des performances similaires, les caractéristiques "small world" étant les seules à ne pas être significativement meilleures que le point de référence (c'est-à-dire $p < 0,01$ et indiquées par un astérisque dans le tableau). Pour l'excitation, on observe que les sous-groupes de caractéristiques du graphe et "small world" ne sont pas significativement meilleurs que le point de référence, alors que d'autres caractéristiques de motifs, comme l'entropie de permutation et la dissimilarité de distance ordinale, le sont. Dans l'ensemble, les modèles reposant sur ces deux sous-groupes de caractéristiques se sont avérés fournir les informations les plus discriminantes pour les modèles de valence et d'éveil. Les tableaux 3.2-3.4 montrent les effets de la fusion de caractéristiques et les gains obtenus avec l'ensemble combiné par rapport à l'utilisation d'un groupe de caractéristiques individuellement. Pour la dimension de la valence, par exemple, des gains de $8,6\%$ et de $2,4\%$ ont été obtenus avec la fusion de caractéristiques par rapport à l'utilisation du repère et de la caractéristique du motif seuls, respectivement. Comme le montre le tableau 3.5, le modèle basé sur l'ensemble combiné s'est appuyé sur les caractéristiques des deux groupes de caractéristiques, soulignant ainsi leur complémentarité pour la prédiction de la valence. Pour l'arousal, la fusion de caractéristiques a donné lieu à des gains plus modestes par rapport aux caractéristiques de référence $(6,1\%)$ et aux caractéristiques de motifs $(2,6\%)$, le modèle mRMR le plus performant ne s'appuyant pas du tout sur les caractéristiques de référence. Pour la fusion des décisions, l'espace de pondération a été exploré par pas de 0,1 et il a été constaté que pour la valence, l'ensemble de caractéristiques de référence a donné un poids de 0,2 (c'est-à-dire 0,8 pour les caractéristiques des motifs), tandis qu'un poids de 0,3 a été trouvé pour l'arousal (c'est-à-dire un poids de 0,7 pour les motifs). Ces résultats soulignent l'importance des caractéristiques de motifs par rapport aux caractéristiques de référence pour la prédiction de la valence et de l'arousal. Enfin, la méthode de fusion associative de sortie a été surclassée par toutes les autres méthodes de fusion, même si elle s'est révélée nettement supérieure au hasard. Néanmoins, pour la dimension de la valence, elle a obtenu des résultats similaires à la fusion au niveau du score sans avoir besoin d'une recherche exhaustive des poids. Dans l'ensemble, la fusion au niveau des caractéristiques s'est avérée être la meilleure stratégie pour la valence et a été observée comme étant significativement meilleure que la fusion associative au niveau des scores $(p_{val} \approx 0,01)$ et des sorties $(p_{val} \approx 0,01)$, tandis que la fusion au niveau des scores pour l'excitation était significativement meilleure que la fusion associative au niveau des caractéristiques $(p_{val} < 0,01)$ et des sorties $(p_{val} < 0,01)$.

Ce chapitre a fait les premiers pas pour évaluer les avantages des caractéristiques basées sur les motifs par rapport aux repères existants basés sur le spectre. À cette fin, aucune optimisation n'a été effectuée sur les classificateurs en tant que tels afin de comparer directement les performances obtenues avec la même configuration de classificateur mais avec des entrées de caractéristiques variables. En tant que tel, on s'attend à ce que des gains supplémentaires puissent être observés non seulement avec l'optimisation des hyperparamètres du classificateur, mais aussi avec des méthodes de classification plus complexes ou des schémas de fusion alternatifs.

0.4 Chapitre 4 : Caractéristiques ECG non linéaires et multi-échelles pour une évaluation robuste de l'état mental

Dans ce chapitre, nous explorons les propriétés non linéaires de l'ECG en utilisant des caractéristiques d'entropie multi-échelles et nous testons leur robustesse pour l'évaluation de l'état mental dans des cas réels. L'entropie multi-échelle (MSE) [127] a été proposée pour caractériser la complexité des séries chronologiques physiologiques à plusieurs échelles. L'algorithme est basé sur l'obtention de l'entropie à différentes échelles de temps en utilisant un algorithme de mise à l'échelle. Alors que l'on utilise traditionnellement des échantillons d'entropie avec une mise à l'échelle à gros grains, plusieurs variantes [128] de l'algorithme d'entropie [129, **?**] ainsi que la méthode de mise à l'échelle [130, 131] ont été proposées ces dernières années. Les variantes de l'entropie comprennent la PE [132] l'entropie de permutation modifiée (mPE) [133], pour tenir compte des mêmes battements consécutifs, et l'entropie de permutation pondérée [134], pour intégrer l'amplitude du signal dans la mesure de l'entropie. L'EQM a également été calculée sur la volatilité (variance) [135] ainsi que sur l'amplitude des intervalles de différence (c'est-à-dire $dRR_i = abs(RR_{i+1} - RR_i)$ de la série RR, qui présentent tous deux un comportement non linéaire. Dans ce chapitre, nous nous intéressons particulièrement à l'évaluation des états mentaux de l'utilisateur dans un contexte ambulatoire, dans lequel le mouvement peut non seulement introduire des artefacts qui jouent un rôle néfaste sur la qualité du signal, mais aussi provoquer des changements dans la dynamique cardiaque qui peuvent altérer la mesure du VRC.

0.4.1 Caractéristiques entropiques multi-échelles

Les méthodes d'entropie multi-échelles reposent sur deux étapes : i) la mise à l'échelle et ii) le calcul de l'entropie sur les différentes échelles. Ici, nous avons exploré différents algorithmes pour ces deux étapes, comme le résume le Tableau 4.1. Bien que la mise à l'échelle à grain grossier (*cg*) (Eq. 4.1) ait été la méthode proposée par défaut, son remplacement par une mise à l'échelle par moyenne mobile (*mavg*) (Eq. 4.2, pour les séries temporelles courtes ou une procédure composite (*comp_cg*)(Eq. 4.3) qui réduit la variance de l'entropie aux échelles supérieures, ont également été proposés. De plus, le moment 2^{nd} à gros grain (*mom*) (Eq. 4.5) pour calculer la série de volatilité a été proposé. Nous avons également proposé l'échelonnement du moment de la moyenne mobile (*mavg_mom*) pour obtenir des performances plus élevées pour les séries temporelles plus courtes. Les algorithmes d'entropie testés comprennent (i) l'entropie d'échantillon, (ii) l'entropie de permutation (PE), (iii) l'entropie de permutation modifiée (mPE) et (iv) l'entropie de permutation pondérée (mPE_wt). Les valeurs d'entropie ont été calculées à la fois pour la série RR et la série dRR pour une échelle de $s = 1 \dots 10$. De plus, la moyenne et l'écart-type des mesures d'entropie sur toutes les échelles ont été calculés, ce qui donne un total de 24 caractéristiques pour chaque type de mesure d'entropie.

Comme la série RR présente des propriétés statistiques fractales, la distance ordinale (décrite au chapitre 3) a été calculée sur les différentes séries d'échelles (appelée distance ordinale inter-échelle (Isod)) pour les séries RR et dRR. Pour limiter la taille de l'espace des caractéristiques, nous avons calculé les distances pour des échelles limitées. En outre, nous calculons les statistiques d'Isod, à savoir la moyenne, l'écart type et la première différence, ce qui donne un nombre total de 66 caractéristiques de distance ordinale. Pour des raisons de simplicité, seules les structures de motifs originales basées sur le PE sont considérées et la mise à l'échelle est effectuée par l'algorithme de mise à l'échelle de la moyenne mobile. Les figures 4.1 et 4.2 montrent la série RR et la série de volatilité RR (en utilisant $mavg\_mom$) pour les échelles $s = 1$ à $s = 3$ et les échelles $s = 2$ à $s = 4$, respectivement, pour un segment ECG de cinq minutes. Comme on peut le constater, la mise à l'échelle supprime certaines informations de haute fréquence de la série, généralement associées à des artefacts.

0.4.2 Mise en place experimentale

Les fonctions proposées ont été testées pour l'évaluation de la charge mentale en utilisant les données ECG (250 Hz) de l'ensemble de données WAUC. Tout d'abord, le signal ECG de tous les sujets a été inspecté visuellement et deux sujets ont été retirés car les données étaient corrompues en raison d'un dysfonctionnement du capteur. Pour les autres sujets, la série d'intervalles entre les battements a été extraite comme suit. Tout d'abord, l'ECG a été filtré à l'aide d'un filtre passe-bande d'une largeur de bande de 4 à 40 Hz pour améliorer le complexe QRS. Ce filtrage a été suivi d'un algorithme de détection des QRS basé sur l'énergie [136]. La série RR a ensuite été filtrée pour éliminer les valeurs aberrantes à l'aide d'un algorithme de détection basé sur l'intervalle ($\geq 280ms$ et $\leq 1500ms$), d'un algorithme de détection des valeurs aberrantes par moyenne mobile et d'un filtre basé sur le pourcentage de changement dans les valeurs RR consécutives ($\leq 20\%$). Les mesures standard de la VRC dans les domaines temporel et fréquentiel ont été extraites et utilisées comme mesures de référence. Une liste complète de ces mesures conventionnelles est présentée dans le tableau 6.6. Les caractéristiques ont été extraites sur des segments de 5 minutes de séries RR avec un chevauchement de 4 minutes, ce qui donne six séries RR pour chacune des sessions expérimentales de 10 minutes. Pour l'évaluation, une validation croisée à cinq reprises a été utilisée. L'évaluation de la charge de travail est effectuée comme une tâche de classification binaire, où les classes cibles de charge mentale élevée et faible sont tirées de la tâche MATB-II. Un classificateur à vecteur de support (SVM) avec un noyau RBF est utilisé. Pour explorer la performance de la généralisation, la procédure mentionnée ci-dessus est répétée 50 fois avec différentes graines aléatoires. Cela conduit à 250 (5 fois 50 répétitions avec différentes graines aléatoires) ensembles d'entraînement et de test et classifications. Pour évaluer l'importance des caractéristiques, nous utilisons la sélection de caractéristiques RFE et examinons la fréquence des caractéristiques apparaissant dans les 20 premiers ensembles pour les 250 combinaisons possibles. La précision (Acc) et le score F1 (F1) ont été utilisés comme chiffres de mérite.

0.4.3 Résultats expérimentaux

Les figures 4.4, 4.5, et 4.6 montrent les performances des algorithmes pour des niveaux d'activité physique respectivement nul, moyen et élevé. En outre, la fusion des méthodes de mise à l'échelle basées sur $comp\_cg$ et $mavg\_mom$ à l'aide de l'algorithme mPE a été étudiée, car elle a permis d'obtenir des performances constamment meilleures dans les trois conditions d'activité physique. Le tableau 4.3 montre les résultats de la fusion pour les différents niveaux d'activité physique. Les tableaux 4.4, 4.5 et 4.6 montrent les méthodes de référence, de distance ordinale inter-échelle et d'entropie multi-échelle les plus performantes ainsi que leur fusion pour les niveaux d'activité physique nul, moyen et élevé, respectivement. Dans les tableaux, "nof" indique le nombre de caractéristiques utilisées dans chaque cas. Les tableaux 4.7, 4.8, et 4.9 montrent les caractéristiques (apparaissant plus de 70% du temps) classées selon leur fréquence d'occurrence ($freq$) pour des niveaux d'activité physique nul, moyen et élevé.

0.4.4 Discussion

D'une manière générale, pour tous les cas d'activité physique et pour tous les algorithmes d'entropie, les méthodes d'échelonnement basées sur la moyenne mobile (*mov_avg* et *mavg_mom*) et l'échelonnement composite (*comp_cg*) surpassent les approches basées sur le grainage grossier (*cg*, *mom*). On constate également que les méthodes d'échelonnement basées sur l'entropie de permutation modifiée basée sur le moment 2^{nd} (*mom* et *mavg_mom*) obtiennent généralement un pouvoir prédictif plus élevé dans tous les cas de charge de travail physique, ce qui indique l'importance des séries de volatilité des séries RR, ainsi que des séries dRR. Enfin, les algorithmes basés sur l'entropie de permutation modifiée sont plus performants que les méthodes basées sur l'entropie d'échantillon. La fusion des méthodes d'échelonnement basées sur *comp_cg* et *mavg_mom* avec mPE donne une amélioration significative ($p < 0,01$) de 3,53$\% enprcisionetde 3,30\%$ en score F1 et de 1,90$\% enprcisionetde 1,63\%$ en score F1, respectivement, par rapport à l'algorithme *comp_cg* + mPE le plus performant. Lors de la comparaison avec l'ensemble de référence, l'entropie multi-échelle et les caractéristiques Isod sont nettement plus performantes ($p < 0,01$) que l'ensemble de référence pour tous les niveaux d'activité physique. La fusion avec l'ensemble de référence améliore encore les performances. Dans l'ensemble, des performances similaires dans l'évaluation de la charge mentale peuvent être obtenues avec l'ensemble de caractéristiques proposé pour tous les niveaux d'activité physique. Avec des gains par rapport à la référence, des mesures VRC de $24,41\%$ en Acc et de $27,97\%$ en F1 peuvent être obtenues même à des niveaux d'activité élevés.

0.5 Chapitre 5 : Séparation des facteurs de confusion à l'aide des caractéristiques des sous-bandes du VRC pour un meilleur suivi de l'état mental " dans la nature ".

Bien que les mécanismes exacts à l'origine des changements observés dans la complexité des séries RR soient encore inconnus, des résultats récents [137, 138, 139] ont suggéré une influence des systèmes SNS et PNS dans différentes conditions cliniques. Par exemple, le comportement chaotique de la composante HF a été lié à la variabilité circadienne (cycle veille/sommeil) qui est indépendante des changements de la puissance de la bande HF liés à l'âge [140]. De plus, la synchronisation des caractéristiques de la bande HF avec la respiration et la pression artérielle a été observée lors de l'exécution de tâches liées à une charge mentale plus élevée [82]. En fait, il a été démontré que l'interaction entre le SNS et le PNS (c'est-à-dire les bandes LF et HF) suit un comportement de couplage non linéaire [54]. Des mesures quantifiant cette interaction non linéaire ont été proposées pour distinguer les individus souffrant d'insuffisance cardiaque congestive [141] et d'apnée obstructive du sommeil [142] des témoins sains. La plupart de ces caractéristiques basées sur la complexité ont été calculées sur l'ensemble du spectre de la série RR. On s'attend à ce que les mesures de la complexité basées sur les sous-bandes puissent fournir des informations supplémentaires [128]. Par exemple, les changements de fréquence de crête de la bande HF ont été liés à l'activité physique [143] et l'entropie des bandes LF et HF individuellement s'est avérée utile pour la détection de l'apnée obstructive du sommeil [144]. Dans ce chapitre, nous proposons des mesures de complexité des sous-bandes et de nouvelles caractéristiques de descripteurs spectraux afin de mieux caractériser le stress et l'anxiété dans des conditions ambulatoires et " dans la nature ".

0.5.1 Caractéristiques proposées

L'ensemble de caractéristiques proposé peut être divisé en deux ensembles : celui basé sur la complexité des sous-bandes et celui basé sur les descripteurs spectraux des sous-bandes. Les caractéristiques de complexité de sous-bande nécessitent une séparation des séries temporelles LF et HF. Cela a été fait en créant la série de tachogrammes (échantillonnée à 4 Hz) à partir de la série RR échantillonnée de manière non uniforme. Ensuite, deux filtres passe-bande dans la gamme 0,04-0,15 Hz et 0,15-0,4 Hz ont été utilisés pour séparer les composantes LF et HF du tachogramme, générant ainsi deux nouvelles séries temporelles, à savoir rr_{lf} et rr_{hf}, respectivement. Une série de tachogrammes représentative (en haut) ainsi que les séries rr_{lf} (au milieu) et rr_{hf} (en bas) sont illustrées à la Fig. 5.1. Enfin, les caractéristiques non linéaires (telles que décrites dans la section 5.4.2.2) sont extraites à la fois de rr_{lf} et de rr_{hf}. De plus, l'interaction non linéaire entre les deux séries a été quantifiée à l'aide de la métrique de l'entropie de transfert (Eq. 5.1) [145]. Pour calculer les caractéristiques des descripteurs sous-bande-spectraux, la FFT du tachogramme est d'abord calculée et la densité spectrale de puissance des composantes de fréquence LF et HF a été extraite. Plusieurs descripteurs spectraux ont ensuite été calculés pour chaque région. Les descripteurs spectraux comprennent (i) le centroïde (Eq 5.2), (ii) l'étalement (Eq 5.3), (iii) l'asymétrie (Eq 5.4), (iv) l'aplatissement (Eq 5.5), (v) la crête et (Eq 5.6) (vi) l'entropie spectrale (Eq 5.7).

0.5.2 Mise en place expérimentale

Les jeux de données PASS et TILES décrits dans le chapitre 2 ont été utilisés pour cette analyse. Pour le jeu de données PASS, le signal ECG brut collecté à partir du BH3 et échantillonné à 250 Hz a été utilisé pour évaluer la VRC. Le signal a d'abord été nettoyé en utilisant le filtre passe-bande 5-25 Hz. Ensuite, la détection du complexe QRS a été effectuée à l'aide d'un détecteur de QRS basé sur l'énergie [146] pour créer la série temporelle RR. Comme les artefacts (par exemple, les artefacts musculaires, le mouvement des électrodes, les battements ectopiques) peuvent provoquer des erreurs dans les séries RR, un filtre supplémentaire pour éliminer les aberrations RR a été utilisé. En revanche, pour l'ensemble de données TILES, le smart-shirt fournit directement les valeurs RR. Un maximum de quatre intervalles RR sont détectés par le smart-shirt par seconde. La série RR est reconstruite à partir des valeurs RR fournies et est passée à travers le filtre des aberrations RR comme ci-dessus. Après le prétraitement, les caractéristiques de référence du domaine temporel et fréquentiel (décrites dans le tableau 6.6) et les caractéristiques non linéaires ainsi que l'ensemble de caractéristiques proposé sont extraits. Les caractéristiques non linéaires comprennent (i) les caractéristiques basées sur l'entropie (SE et PE), et (ii) les caractéristiques basées sur les fractales (DFA, LE et CorrDim). Pour la base de données PASS, toutes les caractéristiques sont calculées en utilisant des fenêtres de 240 secondes avec un chevauchement de 120 secondes pour chaque session. Au total, 42 caractéristiques VRC sont disponibles pour l'analyse. Pour la base de données TILES, les caractéristiques ont d'abord été extraites sur des fenêtres non chevauchantes de 5 minutes pour chaque jour afin de tenir compte de la variabilité à court terme du VRC, comme cela a été fait dans [147, 148]. Comme les données peuvent être bruitées dans certaines fenêtres, les caractéristiques ont été extraites uniquement pour les fenêtres où la métrique de qualité RRPeakCoverage est > 0.3. Ensuite, les caractéristiques ont été agrégées sur une journée entière en utilisant les 11 fonctions statistiques suivantes. En outre, le *RRPeakCoverage* a été utilisé pour créer trois nouvelles fonctions tenant compte de la qualité. Au total, nous disposons de 588 caractéristiques pour l'analyse.

Une classification binaire a été effectuée pour l'évaluation du stress et de l'anxiété. Pour la base de données PASS, les données réelles des ensembles ont été utilisée comme classes cibles de stress. Pour la base de données TILES, en revanche, un seuil global a été utilisé pour binariser les évaluations du stress et de l'anxiété. Une procédure CV 5 fois est répétée 10 fois (5X10 exécutions) avec différentes graines aléatoires. Les résultats de classification indiqués sont la moyenne et l'écart type sur les 50 passages. Pour évaluer l'importance des caractéristiques, nous utilisons la sélection des caractéristiques et examinons les caractéristiques qui se classent dans le premier ensemble plus de 70% des 50 essais. Un classificateur SVM (noyau RBF) et un poids de classe "équilibré" ont été utilisés. BACC, F1 et MCC ont été utilisés comme chiffres de mérite. La sélection de caractéristiques RFE est utilisée pour sélectionner les 13 meilleures caractéristiques pour la base de données PASS et les 100 meilleures caractéristiques pour la base de données TILES pour chaque essai. L'importance des caractéristiques a été calculée sur la base des résultats de la sélection des caractéristiques sur les

50 itérations. Les caractéristiques apparaissant à plus de 70% ont été classées en fonction de leur fréquence d'apparition ($freq$) pour obtenir l'ensemble de caractéristiques le plus performant pour les jeux de données PASS. Pour le jeu de données TILES, en raison de l'ajout de fonctions calculées en plus de la série de caractéristiques quotidiennes, les caractéristiques les plus importantes (70%) ont d'abord été séparées de leurs fonctions correspondantes et les fréquences des caractéristiques identiques avec des fonctions différentes ont été agrégées et renormalisées. De plus, la fréquence des fonctionnelles des principales caractéristiques a également été notée.

0.5.3 Résultats expérimentaux

Les performances de classification des contraintes pour les bases de données PASS et TILES sont présentées dans les tableaux 5.2 et 5.3, respectivement. La performance de classification sur l'anxiété pour la base de données TILES est disponible dans le Tableau 5.4. Les deux premières lignes présentent les résultats pour les caractéristiques de référence, tandis que les caractéristiques proposées sont données dans les lignes 3 et 4. Les six lignes suivantes représentent la fusion de différents ensembles de caractéristiques. En particulier, 'Band-All' correspond à la fusion des caractéristiques de complexité et de bande-spectrale, 'Fuse-Complexity' à la fusion de l'ensemble de référence avec les caractéristiques de complexité de bande, 'Fuse spectral' à la fusion de l'ensemble de référence avec les caractéristiques de bande-spectrale, 'Fuse RR-Complexity' à la fusion de l'ensemble de référence avec les caractéristiques de complexité RR, 'Fuse-Band-All' à la fusion de l'ensemble de référence avec les caractéristiques 'Band-All', et 'Fuse-All' à la fusion de tous les ensembles de caractéristiques extraits. Les caractéristiques mises en évidence en gras dans chaque tableau montrent l'ensemble des caractéristiques les plus performantes (basé sur la valeur MCC). Les caractéristiques les plus performantes pour le stress et leur fréquence d'occurrence pour la base de données PASS sont présentées dans le Tableau 5.5. Le tableau 5.6, quant à lui, montre les principales caractéristiques du stress et de l'anxiété pour TILES.

0.5.4 Discussion

Pour la prédiction du stress avec PASS (Tableau 5.2), la meilleure performance est obtenue par la fusion des caractéristiques de référence avec les caractéristiques proposées de complexité de bande et de descripteur spectral (Fuse-Band-All) avec des améliorations significatives ($p < 0,01$) de $4,64\%$ en BACC, $14,7\%$ en F1, et $24,2\%$ en MCC par rapport au jeu de caractéristiques de référence seul. Cette combinaison de caractéristiques présente également une amélioration significative ($p < 0,01$) de $7,66\%$ en F1 par rapport à la fusion de l'ensemble de référence avec les caractéristiques de complexité RR couramment utilisées (Fuse-RR-Complexity). Quant à TILES, pour la prédiction du stress (Tableau 5.3), la meilleure performance est obtenue avec la fusion de tous les ensembles de caractéristiques (Fuse-All) avec une amélioration significative de $6,13\%$ en BACC, $5,5\%$ en F1 et $31,6\%$ en MCC par rapport à l'ensemble de caractéristiques de référence seul. Cette performance est comparable à celle de la fusion du benchmark avec les caractéristiques proposées, ainsi qu'à celle de la fusion du benchmark avec les seules caractéristiques spectrales de la bande. Enfin, pour la prédiction de l'anxiété (Tableau 5.4), la meilleure performance est à nouveau obtenue par la combinaison de tous les ensembles de caractéristiques avec des améliorations significatives de $6,45\%$ en BACC, $9,89\%$ en F1, et $36,4\%$ en MCC par rapport à l'ensemble de caractéristiques de référence seul. Dans l'ensemble, sur les deux ensembles de données et les états mentaux, les caractéristiques de complexité de bande et de descripteur spectral montrent un comportement complémentaire non seulement entre elles, mais aussi avec les caractéristiques de référence existantes.

Si l'on examine les caractéristiques les plus performantes, pour PASS, dans l'ensemble, 4 des 11 caractéristiques proviennent de l'ensemble de caractéristiques proposé, dont trois caractéristiques de la bande HF (2 spectrales et 1 de complexité) et une de la bande LF (spectrale). Pour TILES, en ce qui concerne le stress, neuf des 15 premières caractéristiques sont issues de l'ensemble de caractéristiques proposé. Parmi les principales caractéristiques proposées, six des neuf caractéristiques proviennent de la bande HF (2 complexes, 4 spectrales), une de la bande LF (spectrale), ainsi que les caractéristiques d'entropie de transfert LF vers HF et HF vers LF. Pour la prédiction de l'anxiété, les principales caractéristiques se chevauchent largement avec les principales caractéristiques du stress, 10 des 15 principales caractéristiques apparaissant dans les deux ensembles de caractéristiques. Cela pourrait être dû au fait que les deux états mentaux sont fortement corrélés, car un stress continuellement élevé peut conduire à l'anxiété [149]. Enfin, parmi les fonctions utilisées pour l'agrégation, la moyenne pondérée par la qualité et l'écart type sont les fonctions les plus utilisées pour le stress et l'anxiété dans le premier ensemble de caractéristiques. Cela montre l'importance de la qualité du signal dans la prédiction de l'état mental, corroborant ainsi les résultats de [147]. Ce chapitre nous permet de conclure que la séparation du signal en composantes de sous-bandes permet de traiter les facteurs de confusion qui affectent la série RR en raison de la socialisation, de l'activité physique et du rythme circadien.

0.6 Chapitre 6 : Systèmes multimodaux pour le suivi de l'état mental en conditions réelles

Les systèmes multimodaux de surveillance de l'état mental sont connus pour offrir de meilleures performances que les systèmes reposant sur une seule modalité. Cependant, la plupart des études ont été réalisées dans des conditions de laboratoire contrôlées. Dans ce chapitre, nous explorons le potentiel des systèmes multimodaux pour la surveillance de l'état mental dans des conditions réelles, en présence de facteurs de confusion, tels que l'activité physique et l'interaction sociale. Cette étude est réalisée en deux parties. Dans la première partie, nous explorons un système multimodal pour le suivi de la charge mentale sur des données collectées dans des conditions ambulatoires. Dans la deuxième partie, nous explorons la prédiction multimodale en utilisant le VRC et la respiration à très court terme.

0.6.1 Évaluation multimodale de la charge de travail mental dans des conditions ambulatoires

0.6.1.1 Mise en place expérimentale

L'ensemble de données WAUC (Section 2.8.1)) a été utilisé pour cette analyse. Les signaux EEG y ont d'abord été filtrés avec un filtre passe-bande (1-45 Hz). Ensuite, les artefacts ont été filtrés à l'aide de l'algorithme wICA [150, 151]. Le signal a ensuite été décomposé dans les bandes de fréquences conventionnelles suivantes : delta (δ, 1-4 Hz), thêta (θ, 4-8 Hz), alpha (α, 8-12 Hz), bêta (β, 12-30 Hz) et low-gamma (γ_1, 30-45 Hz). Ensuite, la série d'intervalles entre les battements (RR) a été extraite du signal ECG. Tout d'abord, l'ECG a été soumis à un filtrage passe-bande (4-40 Hz) suivi d'un algorithme de détection des QRS basé sur l'énergie [136]. La série RR a ensuite été filtrée pour éliminer les valeurs aberrantes [146]. Pour les autres signaux périphériques, un filtrage dans différentes plages est effectué. Ici, l'accent a été mis uniquement sur les caractéristiques de référence largement utilisées pour diverses modalités physiologiques afin d'évaluer les améliorations obtenues uniquement par les techniques multimodales. Des caractéristiques de référence de puissance de bande et de descripteur statistique ont été extraites pour tous les signaux. Les 94 caractéristiques extraites sont résumées dans le tableau 6.1. Sept ensembles de caractéristiques ont été explorés pour l'évaluation de la charge mentale dans différentes conditions d'activité physique pour les conditions de vélo et de tapis roulant séparément ; ces ensembles de caractéristiques correspondaient à la combinaison des caractéristiques extraites de divers capteurs. Dans tous les cas, le niveau de charge mentale (contrôlé par les paramètres MATB-II) a été utilisé comme étiquettes cibles. Un modèle de régression logistique a été utilisé pour la classification. Le paramètre " Leave-one-subject-out " (LOSO) a été utilisé pour l'évaluation avec la métrique BACC comme facteur de qualité. Les évaluations ont été réalisées à la fois sur les époques et par sessions. Pour l'évaluation basée sur les époques, les résultats individuels pour chaque session ont été considérés comme le résultat final. Pour l'évaluation par session, un vote majoritaire des époques pour une session donnée a été effectué pour obtenir un résultat par session robuste au niveau du bruit.

0.6.1.2 Résultats expérimentaux

Les performances au niveau de l'époque et de la session pour les conditions de vélo et de tapis roulant pour différents niveaux d'activité physique sont présentées dans les tableaux 6.2 et 6.3, respectivement. Les tableaux 6.4 et 6.5 montrent les trois meilleures caractéristiques de chaque capteur ainsi que leur rang relatif dans le classement de l'ensemble de caractéristiques fusionnées pour les cas d'activité physique nulle, moyenne et élevée pour les conditions de tapis roulant et de vélo, respectivement.

0.6.1.3 Discussion

Dans l'ensemble, les performances du tapis roulant étaient supérieures à celles du vélo pour les trois niveaux d'activité physique, avec des différences de $27,3\%$, $3,9\%$ et $18,3\%$ entre les modèles les plus performants pour les niveaux d'activité physique nul, moyen et élevé, respectivement. Nous observons que pour la plupart des ensembles de caractéristiques, la performance par session est supérieure à la performance par époché dans le cas du tapis roulant ; elles sont comparables l'une à l'autre dans le cas du vélo. La performance par session est plus résistante aux artefacts physiologiques de courte durée ainsi qu'aux états mentaux du sujet qui peuvent corrompre les résultats, donc une sortie par session est plus utile dans un cadre réaliste. La variabilité des performances dans les deux conditions pourrait être due aux différences expérimentales entre le vélo et le tapis roulant. L'activité sur tapis roulant impliquant un mouvement de la tête dû à la marche et à la course, il aurait pu être plus difficile de se concentrer sur un écran fixe, alors que la position plus stable du haut du corps sur le vélo a ajouté à la demande de la tâche.

0.6.2 Surveillance de l'état mental basé sur la respiration et le VRC à très court terme

.

0.6.2.1 Mise en place expérimentale

L'ENPQ (Section 2.8.3) a été utilisé pour cette analyse. Les données ECG ont été inspectées visuellement et les sujets dont le QRS clair était absent ont été supprimés ainsi que les données respiratoires correspondantes. Ensuite, un filtrage passe-bande (5-25 Hz) suivi d'une détection des QRS a été effectué sur le signal ECG. La série a ensuite été filtrée pour éliminer les valeurs aberrantes. Le signal brut de la respiration a quant à lui été sous-échantillonné à 6 Hz, puis filtré en passe-bas (2 Hz) pour éliminer le bruit. Enfin, plusieurs caractéristiques de référence dans les domaines temporel et fréquentiel ont été extraites de la série RR filtrée et de la courbe de respiration améliorée (Tableau 6.6). Au total, 15 VHR et 14 courbes respiratoires ont été calculées pour différentes tailles de fenêtres (60, 90, 120, 180, 240 et 300 s), sans chevauchement entre fenêtres consécutives. Pour l'évaluation, un CV 5 fois plus élevé (répété 50 fois) a été utilisé. Les classes cibles de stress et de charge mentale binarisées par sujet (à partir des évaluations NASA-TLX) ont été utilisées pour la classification à l'aide d'un SVM (noyau RBF). Le taux de précision équilibré (BAC) a été utilisé comme chiffre de mérite. Dans l'ensemble, les classificateurs ont été formés avec les 15 caractéristiques les plus importantes (en utilisant la sélection de caractéristiques RFE) de la VRC, de la respiration, ainsi que des ensembles de caractéristiques fusionnées. Nous analysons la liste des caractéristiques les plus fréquemment sélectionnées (caractéristiques apparaissant au moins 80% du temps sur les 250 essais). De plus, pour évaluer la stabilité à travers la durée de la fenêtre, nous analysons également les caractéristiques qui se sont avérées importantes à travers toutes les durées de fenêtre.

0.6.2.2 Résultats expérimentaux

Les résultats de la classification pour le stress et la charge mentale sont présentés dans les tableaux 6.7 et 6.8, respectivement. Le tableau 6.9 répertorie les principales caractéristiques cohérentes pour le stress et la charge de travail mental.

0.6.2.3 Discussion

Pour le stress, nous pouvons observer une augmentation des performances à mesure que la durée du segment augmente, passant d'un taux d'alcoolémie de 0,557 pour des segments de 60 secondes à un taux d'alcoolémie de 0,594 pour des segments de 5 minutes. Pour la respiration, par contre, les changements sont plus faibles et passent de BAC=0.541 (60 s) à 0.560 (300 s). Pour l'évaluation de la charge mentale, l'impact de l'analyse à très court terme semble être encore moins prononcé que pour le stress. Un BAC=0.561 est atteint pour les segments VRC de 60 s, alors qu'un BAC=0.579 pour 300 s, avec des résultats similaires pour les caractéristiques de la respiration, où une baisse de 1.8% est observée entre les segments de 1 et 5 minutes. Il est important de noter que cette analyse à très court terme a permis d'obtenir une précision supérieure à celle des classificateurs de VRC et de respiration utilisant des durées de 5 minutes, ce qui corrobore davantage l'utilité de la méthode proposée dans des contextes opérationnels adaptatifs.

Nous montrons ici l'importance de la fusion de plusieurs modalités de signaux non seulement pour améliorer la performance des modèles d'évaluation de l'état mental des utilisateurs ambulants, mais aussi pour assurer la robustesse contre les artefacts de mouvement tout en sacrifiant une certaine résolution temporelle. En outre, certaines applications peuvent nécessiter une résolution temporelle rapide. Cependant, les caractéristiques extraites sur des époques plus courtes peuvent être plus sensibles aux artefacts et aux facteurs de confusion. En utilisant la fusion multimodale de caractéristiques de VRC et de respiration à très court terme, nous montrons que la perte de performance due à l'utilisation de caractéristiques de VRC à très court terme peut être compensée.

0.7 Conclusion

Dans cette thèse de doctorat, nous avons étudié les défis que pose le suivi de l'état mental dans des conditions réelles. Ces défis comprennent : (i) le bruit et les artefacts, et (ii) les facteurs confondants, dans les signaux physiologiques. Ces limitations entravent les performances des systèmes de surveillance automatisés. Pour relever ces défis, nous avons proposé plusieurs méthodes, mentionnées ci-dessous.

Caractéristiques de motifs EEG robustes et bruyantes: Les signaux physiologiques sont très sensibles aux artefacts bruités. Une solution à ce problème réside dans l'utilisation de méthodes de filtrage des artefacts avant l'étape d'extraction des caractéristiques. Cependant, ces méthodes exigent généralement des ressources informatiques élevées et fonctionnent en mode hors ligne avec une certaine intervention humaine. Pour surmonter cette limitation, nous avons proposé l'utilisation de représentations de motifs robustes au bruit pour l'extraction de caractéristiques. Les caractéristiques basées sur les motifs intègrent la robustesse directement dans les caractéristiques en ne considérant que la forme de la série chronologique physiologique pour l'extraction des caractéristiques. Nous avons étudié l'utilisation de diverses caractéristiques basées sur les motifs pour une application de reconnaissance des émotions basée sur l'EEG. Les résultats expérimentaux montrent que ces méthodes sont plus performantes que les caractéristiques EEG de référence pour la reconnaissance de la valence et de l'arousal dans le cadre d'une classification binaire. De plus, une amélioration des performances est observée par la fusion de ces caractéristiques proposées avec les caractéristiques de référence, montrant ainsi la complémentarité avec les ensembles de caractéristiques existants. Enfin, nous espérons que ces caractéristiques robustes au bruit et à faible complexité de calcul pourront être utilisées pour la reconnaissance des émotions basée sur l'EEG dans les environnements bruités dans des conditions réelles.

Fonctions ECG non linéaires: Les caractéristiques de pointe utilisées pour diverses applications de surveillance de l'état mental sont incapables de saisir le comportement complexe non linéaire observé dans les séries chronologiques physiologiques. Au cours de la dernière décennie, diverses mesures non linéaires ont surpassé les méthodes standard des domaines temporel et fréquentiel, en particulier pour les applications cliniques. Nous avons proposé ici l'utilisation de caractéristiques non linéaires multi-échelles du VRC pour l'évaluation de la charge mentale dans des applications ambulatoires. Nous avons montré que ces caractéristiques surpassent l'ensemble des caractéristiques de référence dans la prédiction de la charge mentale pour différents niveaux d'activité physique. De plus, nous montrons que l'utilisation de caractéristiques d'entropie de permutation multi-échelle surpasse les mesures d'entropie multi-échelle standard, démontrant ainsi les propriétés de robustesse au bruit des méthodes basées sur les motifs. L'entropie de permutation multi-échelle est présentée comme une caractéristique non linéaire robuste au bruit et simple à calculer pour l'évaluation de la charge mentale dans la nature.

Complexité de la bande et caractéristiques des descripteurs spectraux pour les tachogrammes: Le SNP et le SNS ont un impact différent sur le VRC. Cette connaissance a été utilisée pour dériver des caractéristiques de VRC dans le domaine fréquentiel pour diverses applications de surveillance de l'état mental. Plus précisément, les énergies des bandes LF et HF du tachogramme VRC sont principalement influencées par les systèmes nerveux sympathique et parasympathique, respectivement. Ces bandes sont toutefois affectées différemment par les divers facteurs de confusion, tels que l'activité physique, la parole et le rythme circadien. De plus, les caractéristiques spectrales du spectre de puissance du VRC pour les bandes individuelles sont peu connues. Pour surmonter cette limitation, nous avons proposé une analyse basée sur la complexité et les descripteurs spectraux des séries temporelles des bandes LF et HF pour la surveillance de l'état mental. Les caractéristiques proposées montrent une meilleure performance et une complémentarité avec les caractéristiques de référence pour la prédiction du stress et de l'anxiété sur différents ensembles de données. Les résultats obtenus suggèrent que l'analyse de la complexité des bandes individuelles du VRC, ainsi que la caractérisation des composantes du spectre du VRC, aident à traiter les facteurs de confusion qui ne sont pas possibles avec les caractéristiques traditionnelles du VRC.

Systèmes multimodaux: Dans la recherche sur le contrôle de l'état mental en laboratoire, les propriétés complémentaires de divers signaux physiologiques ont été bien documentées, ce qui a donné lieu à divers systèmes et applications multimodaux plus performants que les systèmes à modalité unique. Les systèmes multimodaux peuvent également devenir un outil utile pour les expériences en conditions réelles afin de fournir une robustesse supplémentaire au bruit, où la sensibilité au bruit de différents capteurs et modalités peut être exploitée à notre avantage. Nous évaluons ici cette hypothèse pour deux applications, à savoir l'évaluation de la charge mentale dans des conditions ambulatoires, ainsi que le suivi du stress et de l'anxiété des étudiants de l'académie de police pendant leurs exercices sur le terrain. Nous montrons que la fusion de capteurs et la prédiction par session peuvent améliorer les performances du système dans un contexte de LOSO. Nous montrons également que les configurations multimodales sont essentielles pour les analyses à très court terme requises pour les applications sensibles au temps, telles que la surveillance des premiers intervenants. Les résultats présentés ici démontrent l'utilité des systèmes multimodaux pour traiter les données bruitées et les facteurs de confusion que l'on trouve souvent dans des conditions réelles.

Chapter 1

Introduction

1.1 Mental State Monitoring

Mental state refers to the current set of emotions and psychological states of a person. Consistent positive mental states, such as calmness and joy, can help increase the feeling of fulfillment and life satisfaction of individuals [1]. On the other hand, constant negative states, such as stress and anxiety, are directly related to a decline in the quality-of-life of individuals and their performance in everyday tasks [1, 2]. Further, persistent negative mental states can lead to severe health problems, such as cardiovascular diseases [3], cognitive dysfunctions [5], and mental health disorders such as depression [4, 5], ultimately leading to chronic illness or even death [152]. The past decades have seen an increasing amount of stress in the workforce with 50% of employees reporting suffering from "work stress" [153]. This has been further aggravated by the COVID-19 pandemic [154] with people reporting declining mental health around the world.

Poor mental health, caused due to excessive stress and other negative emotions, leads to an added emotional and economic burden on individuals suffering from mental health problems and their families [155, 156]. In the past decade, the cost of poor mental health on individuals and society has come to the forefront. In 2017, an estimated 792 million people lived with a mental health disorder, making it the leading cause of disability. The economic burden due to poor mental health worldwide is expected to grow six-fold to A$2.8 trillion dollars in the next 30 years including health care costs, lost productivity, and reductions in health-related quality of life [6]. This number far exceeds the cost due to physical problems in the population. Due to these concerning trends, mental health has been acknowledged as one of the most important factors leading to stronger communities, hence its mention in the World Health Organization (WHO) goals for sustainable development.

A possible way to improve mental health is through early intervention methods [157, 158], which have shown great success in allied domains, such as management of cardiovascular diseases and cancer. Early detection of deteriorating mental health (e.g., due to excessive stress and anxiety) can prevent conditions such as burnout and depression in highly demanding professions, e.g., nurses [7] and surgeons [8]. This is particularly true now with the worldwide COVID-19 pandemic that has put an immense burden on these jobs. For example, reports of record-high burnout cases within healthcare workers have been reported [159, 160, 161].

Mental states are also directly related to task performance. More specifically, mental workload defined as “the cost incurred by a human operator to achieve a particular level of performance” [9] is an important indicator of the ability of an individual to successfully perform a task. Assessment of mental workload is an important factor in optimizing the task performance for an individual [10]. Too little mental workload may cause boredom and lack of engagement, while high mental workload requirements for an extended duration of time may lead to increased fatigue [162] and poor performance. Therefore an optimal level of mental workload is desirable for successfully performing tasks, as higher than optimal levels can cause stress [11, 12].

Monitoring and predicting such states is specifically relevant for jobs where individuals have to deal with high stress situations on a day-to-day basis and where a decrease in performance may lead to safety-critical situations, such as police officers and first responders. Additionally, mental workload assessment can help make useful design decisions and work routines. Representative examples include cockpit design for pilots and driver dashboards, where mental resource allocation can be optimized to prevent unnecessary strain [10].

Moreover, unobtrusive monitoring of the emotional states of individuals can help improve human computer interaction (HCI) by building emotional intelligence into the systems [13]. Monitoring states such as boredom or loss of attention can help educators improve learning resources and personalize training to different students [29]. This has become specially relevant with the surge in online learning due to the pandemic, which limits the teacher-student interaction.Mental fatigue monitoring, in turn, can play a crucial role in ensuring driver safety by alerting drivers when they are too tired, as well as air traffic controllers [10]. Mental state monitoring is also an important aspect in measuring a user's perceived quality-of-experience with real [163] and virtual reality systems [164]. Other domains where the development of emotional intelligence for systems can be relevant include smart home [14], gaming [15], intelligent conversational systems [16], and neuromarketing [165], to name a few. Overall, monitoring of mental states can play a key role in the mental health of individuals, their performance, and their overall quality-of-life. It can also be useful in improving human-computer interaction, allowing technologies to become more intuitive and personalized, ultimately leading to improved learning and other day-to-day experiences.

1.2 Current Methods to Monitor Mental States

Mental states have traditionally been monitored using subjective questionnaires. Various standardized questionnaires have been tested and used for different mental states. These questionnaires usually inquire about the mental state of an individual periodically and are answered by a rating (continuous scale) or one of several predefined choices (discrete scale). Discrete categorization of emotions include the six basic emotions: happiness, sadness, fear, anger, disgust and surprise, as introduced by [17]. Other more complex emotions (such as fatigue or anxiety) are considered to be a combination of these basic emotions under this framework. Several multi-dimensional continuous scales for emotion classification have been introduced [18, 166, 167]. The most popular among these has been the arousal-valence model introduced by [18]. In this scale, each emotional state can be placed on a 2D plane with valence and arousal as the horizontal and vertical axes, respectively. Arousal can range from inactive (e.g., uninterested, bored) to active (e.g., alert,excited), whereas valence ranges from unpleasant (e.g., sad, stressed) to pleasant (e.g., happy, elated). Figure 1.1 depicts some representative emotions and where they are placed on this scale. While most variations in emotional states can be explained on this 2D plane, one more dimension has been added - dominance [18]. Dominance ranges from a helpless and weak feeling (without control) to an empowered feeling (in control of everything). Typically, the self assessment manikin (SAM) [168] is used to assess these three different dimensions. SAM is a picture-oriented way to assess the different dimensions and can be used independent of language, thus making it possible to use across cultures and countries. The SAMs for valence, arousal, and dominance are shown in Fig. 1.2

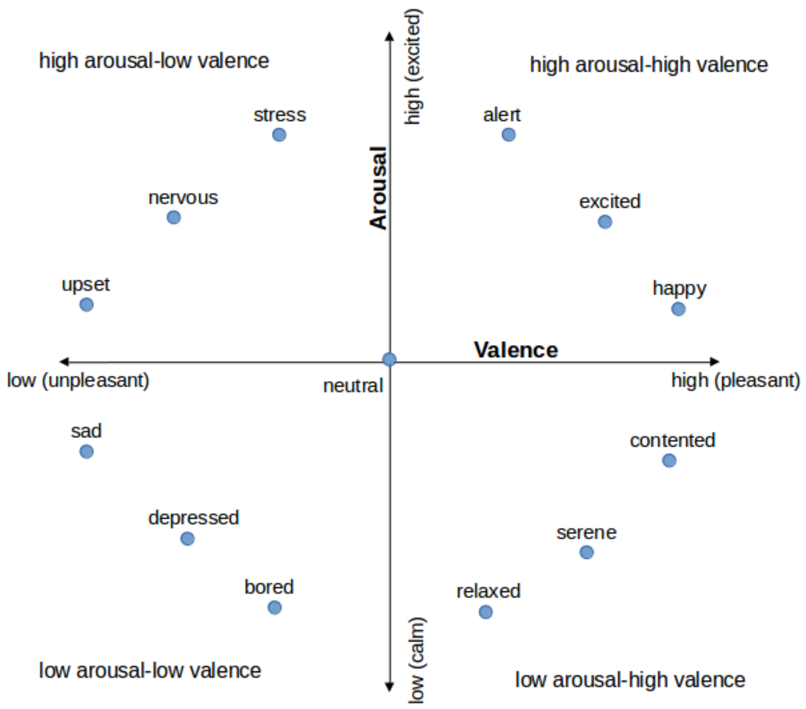


Figure 1.1 – Valance and arousal scale and mapping of different emotions

Other questionnaires to study the degree of specific mental states have also been developed. Popular tools for mental workload assessment include the NASA Task Load Index (NASA-TLX) [9] and the Subjective Workload Assessment Technique (SWAT) [19] questionnaires. Both use multi-dimensional scales to capture different aspects of mental workload. The NASA-TLX contains six dimensions, namely: mental demand, physical demand, temporal demand, performance, effort, and frustration. All these dimensions are rated on a 21-point scale. The overall mental workload measure is often calculated as the average value of the these dimensions. Stress is very close to the concept of mental workload [12] and has been interchangeably used in the literature [37]. However, several questionnaires have been developed for stress assessment [169, 170]. The State-Trait Anxiety Inventory (STAI) [20] has been one of the most popular tools for anxiety measurement. It measures the current state of anxiety by using items that measure subjective feelings of apprehension, tension, nervousness and worry. Finally, the Borg fatigue scale is used to measure the perceived exertion for a given task [21]. Recently, these questionnaire have been deployed using smartphone applications making it possible for their use outside of laboratory conditions and without any supervision [171, 172]. Apart from these standard questionnaires, the visual analog scale (VAS) with specific ranges (e.g., 5- or 10-point scales) has also been used to rate mental states [22, 23].

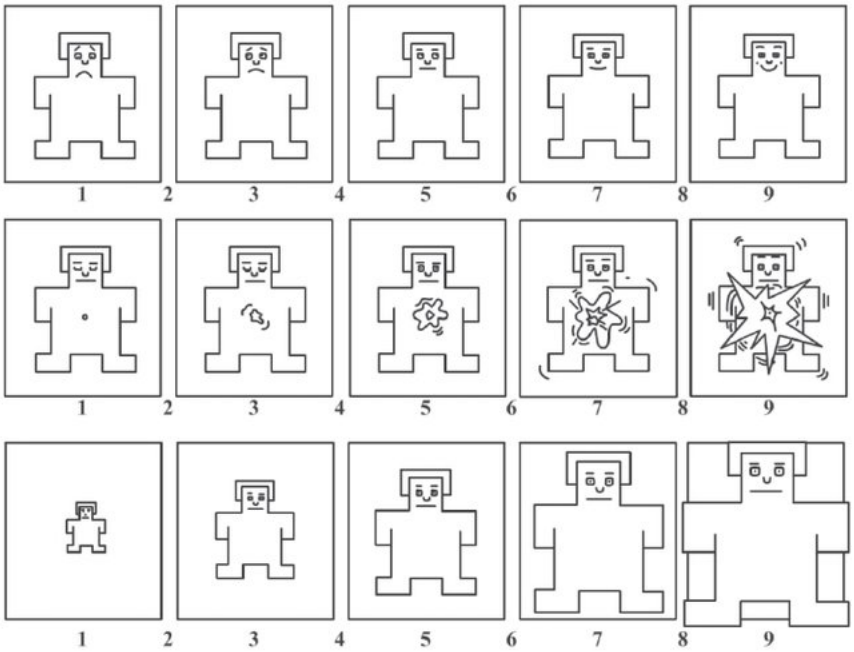


Figure 1.2 – SAMs for (top to bottom) valance, arousal and dominance dimensions

Subjective questionnaires are easy to use, however, they suffer from various limitations. One such limitation includes the presence of psychological biases, such as peak-end and recency biases. These relate to the fact that individuals tend to use the most intense region and the final region of an experience with a forgetting "recency" factor that modulates the contribution of the stimulus in the final rating [24, 25]. Such problems have been specially noted in the assessment of negative mental states [173]. Additionally, the length of the questionnaire may cause lack of compliance and careless responding, thus also leading to erroneous ratings [26]. Moreover, questionnaires are usually administered after a task is completed, thus does not capture the real-time experience of the task. While one may increase the sampling rate of the questions to provide real time feedback, this may hinder the performance of the task itself and/or increase mental workload, thus negatively affecting the experiment [10]. Increased frustration due to such hindrances has also been linked to subject-dropout from studies. Lastly, questionnaires related to negative mental states, such as anxiety or depression, require increased self-focus and introspection, which could further exacerbate mental health conditions [174].

An alternative to mental state monitoring via questionnaires is by using audio-visual data generated by individuals. Humans communicate emotions with one another using speech, facial expressions, and body gestures. Hence, using these modalities has been a popular method for emotion state monitoring. As a result, a large number of audio-visual databases for emotion recognition have been established [175, 176, 177]. Typically, audio and video information is easy to collect with microphones and cameras and a large amount of research has been conducted in this field with different affect recognition challenges, such as the Audio-Visual Emotion Challenge (AVEC) [178] and the Interspeech COMPARE challenge series [179]. However, such systems face three major challenges: i) both speech and facial expressions can be controlled voluntarily and individuals can hide their real mental states or fake a given state, ii) both modalities are not universal and emotional audio-visual patterns can changes across languages and cultures [180, 181], and iii) due to privacy concerns, many users have hesitations in sharing and recording their everyday conversations.

Another way to evaluate mental state is by using neuro-physiological signals [13, 10]. These signals reflect the activity of the central and autonomic nervous systems. Central nervous system (CNS) activity can be captured by capturing the electrical, magnetic or hemodynamic activity of the brain. Electroencephalography (EEG) is one such method that captures the electrical activity of the brain using electrodes placed on the scalp surface. EEG has high temporal resolution compared to other neuroimaging approaches, thus can be used for real-time mental state monitoring. Furthermore, the use of EEG is noninvasive, fast, and inexpensive, making it a preferred method in studying the brain responses to emotional stimuli [27, 28] both in the lab as well as in real-life mobile applications [29, 30, 31].

The autonomic nervous system (ANS) activity, in turn, regulates bodily functions and can be captured using physiological signals such as electrocardiogram (ECG), respiration signal, skin temperature, and galvanic skin response (GSR), to name a few. The ANS can be further divided into two systems: the parasympathetic nervous system (PNS), which relaxes the body, and the sympathetic nervous system (SNS), which is associated with the flight-or-fight response. These two systems impact physiological signals differently and can help distinguish between different mental states [32]. Further, combining the different signals to create a multi-modal system for mental state monitoring has shown to improve performance [13]. This is due to the different signals providing complementary information, as well as added robustness to sensor failure and high noise levels that could be present in individual modalities.

Overall, neurophysiological signals provide several advantages over questionnaire and audio-visual based methods. Compared to subjective questionnaires, these signals can be monitored continuously without any interruption to the task being performed. This allows for real-time monitoring of mental states due to high temporal resolution provided by some of the collected signals (e.g., EEG). These signals are also independent from subject biases which can often lead to ratings which are not reliable. In comparison to audio-visual mental state monitoring, physiological responses are involuntary, hence cannot be faked. The CNS and ANS activities are also universal across people, thus suffer from comparatively less variability across cultures. Lastly, while physiological monitoring also requires care with how human data is collected and stored for privacy reasons, it conveys fewer concerns around privacy then its audio-video counterpart.

1.3 Moving from the lab to "in-the-wild"

A large number of studies using various physiological signals have been conducted over the past years for mental state monitoring. The majority of these studies, however, have been conducted in controlled laboratory settings with bulky sensors and data acquisition setups [10, 32]. These experiments typically require the subject to remain stationary in order to minimize movement noise and provide a fixed stimulus to evoke a given mental state for a small duration of time. Recent developments in wearable sensor technologies, however, have allowed for commercially-available wearable devices to be used to monitor mental states in "in-the-wild" conditions, i.e., in highly ecological, everyday settings.

In-the-wild conditions, however, face several different challenges, such as:

1. Noise: Physiological signals can be corrupted with different types of noise. For EEG, for example, noise sources can include eye blinks, facial muscle artefacts, and power line interference [28]. Similarly, for ECG typical noise sources can include electrode movement, muscle artefact noise, and baseline wander. However, when measured in-the-wild, other noise sources may also contaminate the signals. In the case of EEG signals, this can include electrode movement, head movement, and even gait artefacts due to physical activity [182]. With ECG signals, in turn, physical activity can cause severe electrode movements, thus causing outliers in the measured data which may prevent the signal from any further analysis [183]. While enhancement algorithms can be used, these often take large amounts of computational resources, are not real-time, or may remove relevant signal information

needed for mental state prediction [33]. Another physiological artefact during physical activity is the so-called ectopic beat which is generated by impulses originating from outside the sinoatrial node [184].

2. Confounding factors: Physiological signals are the result of various biological processes interacting with one other at different time scales [185]. These can include effects of circadian rhythm, physical activity, and fatigue, to name a few. These confounding factors can be visible in physiological signals and interfere with analyses. For example, ECG signal analysis is sensitive to the time of day due to the effects of the circadian rhythm [186], as well as with posture, with differences seen in supine vs sitting positions [187, 188]. Additionally, heart rate may show similar changes due to different stimuli; for example, both mental stress and physical activity are responsible for increased heart rate. For EEG signals, in turn, different brain regions are involved with different activities, such as prefrontal cortex with mental workload [189] and the temporal lobe with movement [190]. A combination of these activities can change the measured EEG signal characteristics. Similarly, breathing changes could be caused by both physical activity and talking [191]. Often experiments try to control for such factors by restricting the other confounding stimulus. This is why often subjects are asked to stay still or to record data at the same time of the day to account for circadian variability [34]. In-the-wild experiments, however, will likely involve multiple emotional and physical stimuli at the same time, thus hampering mental state prediction accuracy and reliability.
3. Data availability: Most existing databases for mental state monitoring have been collected in laboratory conditions [32, 28] where the experiment controls for the above mentioned confounding factors and noise. Additionally, by relying on bulky measurement devices, proprietary software, and specialized manpower, laboratory experimentation can be costly and time consuming. As a result, a large number of available datasets only contain data collected from a small number of participants and for small time durations, ranging from 1-minute segments as in [13] to 5-minute sessions in [32]. These short-duration data make it impossible to explore the long-term effects of emotional stimuli on physiology. Therefore, a lack of publicly-available in-the-wild databases makes it difficult to develop solutions that can be used reliably in highly-ecological settings.

1.4 Tools for In-the-Wild Mental State Monitoring

The overarching goal of this doctoral research is to develop various tools that allow for mental state assessment in-the-wild, where noise and confounding effects are present and contaminating the recorded physiological signals. More specifically, three main families of innovations have been proposed, namely:

1. Non-linear features: Physiological signals are a result of non-linear processes [185] and there is growing evidence of the usefulness of non-linear features in different clinical settings [35, 36]. Non-linear features have shown to outperform traditional linear ones for several mental state monitoring applications in the lab [37, 38]. However, their usefulness in-the-wild still needs to be assessed.
2. Noise-robust features: Noise can severely deteriorate features extracted from physiological signals [39]. This problem can be solved using feature extraction methods which build robustness directly into the features while capturing the most important signal characteristics.
3. Multi-modal fusion: Different modalities may provide complementary information for mental state monitoring. Additionally, multi-modal systems are more robust to noise compared to systems relying on a single modality, as systems can learn to rely on certain modalities only if signal quality surpasses certain thresholds. These characteristics make multi-modal systems ideal for in-the-wild applications.

These three innovations have been described in several manuscripts, as listed below in chronological order. Where appropriate, the chapters of this thesis in which these publications appear are also specified.

1.5 Publications derived from the thesis

Publications included in the thesis

Articles published in refereed journals

- "Fusion of Motif-and spectrum-related features for improved EEG-based emotion recognition", **Tiwari, A.**, Falk, T. H. (2019), *Computational intelligence and neuroscience, 2019.* [192] [Chapter 3]

- "Multi-scale heart beat entropy measures for mental workload assessment of ambulant users", **Tiwari, A.**, Albuquerque, I., Parent, M., Gagnon, J. F., Lafond, D., Tremblay, S., H Falk, T. (2019), *Entropy, 21(8), 783*, [193] [Chapter 4]

Under review

- "New Measures of Heart Rate Variability based on Subband Tachogram Complexity and Spectral Characteristics for Improved Stress and Anxiety Monitoring in Highly Ecological Settings", **Tiwari, A.**, Falk, T. H. (2021), *Journal of Biomedical And Health Informatics* [194] [Chapter 5]

Conference proceedings and abstracts

- "Prediction of Stress and Mental workload During Police Academy Training Using Ultra-short-term Heart Rate Variability and Breathing Analysis", **Tiwari, A.**, Cassani, R., Gagnon, J. F., Lafond, D., Tremblay, S., Falk, T. H. (2020, July), *In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC) (pp. 4530-4533). IEEE.* [195] [Chapter 6]

- "Movement Artefact-Robust Mental Workload Assessment During Physical Activity Using Multi-Sensor Fusion", **Tiwari, A.**, Cassani, R., Gagnon, J. F., Lafond, D., Tremblay, S., Falk, T. H. (2020, October), *In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 3471-3477). IEEE.* [196] [Chapter 6]

Other publications

Articles published in refereed journals

- "Electroencephalography Amplitude Modulation Analysis for Automated Affective Tagging of Music Video Clips", Clerico, A., **Tiwari, A.**, Gupta, R., Jayaraman, S., Falk, T. H. (2018), *Frontiers in computational neuroscience, 11, 115.* [197]

- "Lossless Electrocardiogram Signal Compression: A Review of Existing Methods", **Tiwari, A.**, Falk, T. H. (2019), *Biomedical Signal Processing and Control, 51, 338-346.* [198]

- "WAUC: A Multi-modal Database for Mental Workload Assessment under Physical Activity", Albuquerque, I., **Tiwari, A.**, Parent, M., Cassani, R., Gagnon, J. F., Lafond, D., Falk, T. H. (2020), *Frontiers in Neuroscience, 14.* [199]

- "PASS: A Multimodal Database of Physical Activity and Stress for Mobile Passive Body/Brain-Computer Interface Research", Parent, M., Albuquerque, I., **Tiwari, A.**, Cassani, R., Gagnon, J. F., Lafond, D., Falk, T. H. (2020), *Frontiers in Neuroscience, 14, 1274.* [113]

Conference proceedings and abstracts

- "On the Analysis of EEG Features for Mental Workload Assessment During Physical Activity", Albuquerque, I., **Tiwari, A.**, Gagnon, J. F., Lafond, D., Parent, M., Tremblay, S., Falk, T. (2018, October), *In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 538-543). IEEE.* [200]

- "A Comparison of Two ECG Inter-beat Interval Measurement Methods for HRV-Based Mental Workload Prediction of Ambulant Users", **Tiwari, A.**, Albuquerque, I., Parent, M., Gagnon, J. F., Lafond, D., Tremblay, S., Falk, T. H. (2019), *CMBES Proceedings, 42.* [201]

- "Stress and Anxiety Measurement "in-the-wild" Using Quality-aware Multi-scale HRV Features", **Tiwari, A.**, Narayanan, S., Falk, T. H. (2019, July), *In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 7056-7059). IEEE.* [147]

- “A Comparative Study of Stress and Anxiety Estimation in Ecological Settings Using a Smart-shirt and a Smart-bracelet”, **Tiwari, A.**, Cassani, R., Narayanan, S., Falk, T. H. (2019, July), *In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 2213-2216). IEEE.* [202]
- “Breathing Rate Complexity Features for “in-the-wild” Stress and Anxiety Measurement”, **Tiwari, A.**, Narayanan, S., Falk, T. H. (2019, September), *In 2019 27th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.* [148]
- “Mental Workload Assessment During Physical Activity Using Non-linear Movement Artefact Robust Electroencephalography Features”, **Tiwari, A.**, Albuquerque, I., Gagnon, J. F., Lafond, D., Parent, M., Tremblay, S., Falk, T. H. (2019, October), *In 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) (pp. 4149-4154). IEEE.* [203]
- “Speech-based Stress Classification based on Modulation Spectral Features and Convolutional Neural Networks”, Avila, A. R., Kshirsagar, S. R., **Tiwari, A.**, Lafond, D., O’Shaughnessy, D., Falk, T. H. (2019, September), *In 2019 27th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.* [204]
- “Prediction of Psychological Flexibility with Multi-scale Heart Rate Variability and Breathing Features in an “in-the-wild” Setting”, **Tiwari, A.**, Villatte, J. L., Narayanan, S., Falk, T. H. (2019, September) ,*In 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 297-303). IEEE.* [205]
- “Cross-Subject Statistical Shift Estimation for Generalized Electroencephalography-based Mental Workload Assessment”, Albuquerque, I., Monteiro, J., Rosanne, O., **Tiwari, A.**, Gagnon, J. F., Falk, T. H. (2019, October). *In 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) (pp. 3647-3653). IEEE.* [206]
- “A Multimodal Approach to Improve the Robustness of Physiological Stress Prediction During Physical Sctivity”, Parent, M., **Tiwari, A.**, Albuquerque, I., Gagnon, J. F., Lafond, D., Tremblay, S., Falk, T. H. (2019, October), *In 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) (pp. 4131-4136). IEEE.* [207]
- “Optimal Filter Characterization for Photoplethysmography-based Pulse Rate and Pulse Power Spectrum Estimation”, Cassani, R., **Tiwari, A.**, Falk, T. H. (2020, July) ,*In 2020*

42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC) (pp. 914-917). IEEE. [208]

- "Initial Investigation into Neurophysiological Correlates of Argentine Tango Flow States: a Case Study", Cassani, R., **Tiwari, A.**, Posner, I., Afonso, B., Falk, T. H. (2020, October) *,In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 3478-3483). IEEE.* [209]
- "Human mental state monitoring in the wild: Are we better off with deeper neural networks or improved input features?", Pimentel, A. , **Tiwari ,A.** , Falk, T. H. (2021, May) *CMBES Proceedings,vol. 44, 2021* [210]
- "Remote COPD severity and exacerbation detection using heart rate and activity data measured from a wearable device", Tiwari, A. , Liaqat, S. , Liaqat, D. , Gabel, M., E. de Lara, and T. Falk,, *In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2021* [211]
- "Context-aware speech stress detection in hospital workers using Bi-LSTM classifiers", Gaballah, A. , **Tiwari, A.**, Narayanan, S. , Falk, T. , *In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021*

1.6 Thesis organization

While this introductory chapter has presented the challenges with mental state monitoring in-the-wild and laid out the foundation for the contributions described herein, the remainder of this dissertation is structured as follows: Chapter 2 provides an overview of the state-of-the-art methods in mental state monitoring, as well as lists currently available databases. Chapter 3 presents the use of motif based EEG features for improved artefact-robust emotion recognition. Next, motifs are combined with non-linear multi-scale features for ECG signal for prediction of mental workload in ambulatory conditions in Chapter 4. Following this, Chapter 5 deals with confounding factors affecting physiological signals using subband features for ECG. In Chapter 6, multi-modal systems are presented to improve noise robustness and performance for mental state monitoring by using multiple physiological modalities at the same time. Lastly, Chapter 7 provides the general conclusions of this thesis, as well as future research areas.

Chapter 2

Physiology-based mental state monitoring: a review of the state-of-the-art

2.1 Introduction

Various neurophysiological signals have been used for mental state monitoring. These signals reflect the changes in the CNS or ANS and share some basic properties and challenges:

1. They are collected non-invasively from various parts of the body using different sensing techniques. For example, while EEG is collected using wet or dry electrodes placed on the scalp, photoplethysmography (PPG) makes use of infrared light to capture the information about blood flow.
2. Physiological signals are caused by the interaction of various processes occurring over multiple time scales. As a result, most of these signals have shown to exhibit non-linear properties.
3. They have been characterized by sets of traditional benchmark features that capture relevant signal characteristics associated with CNS or ANS changes.
4. They are susceptible to different types of artefacts. These artefacts are either physiological or non-physiological in nature and can lead to changes in the signal characteristics being

estimated. For example, while artefacts in EEG might be caused by eye blinks or facial muscle movement, ECG artefacts may be caused by physical activity causing sensor movement.

Once the signals are acquired and processed to improve signal quality, features can then be extracted. These features may then be combined or input individually into a machine learning (ML) pipeline and used for prediction of mental states. The following sections will first cover the relevant neurophysiological signals used in this thesis, describing their properties, artefact sources and their removal, and finally the traditional benchmark features used to set up a baseline system with which we can compare against. Next, the different types of multi-modal systems are covered, followed by the different components of the ML pipeline. Next, some of the publicly available datasets and their limitations are described. Finally, the in-the-wild datasets collected for the thesis are introduced.

2.2 Electroencephalography

2.2.1 Signal Source and Acquisition

EEG signals capture the spontaneous electrical activity of the brain and have become a reliable tool in mental state monitoring of individuals [28]. The signals are captured by putting multiple electrodes on the scalp using an electrode cap, as shown in Fig. 2.1. The electric potentials are measured between electrodes placed on various points on the scalp and a reference electrode. The reference is located relatively far away from the scalp electrodes of interest, such as average mastoids, average earlobes and neck [44, 212]. These electrodes range from 20 to 256 in number and are placed on the scalp based on specific standards aiming at different regions of the brain; the 10-20 electrode placement [40], for example, is the most popular montage. These standards help avoid variations between subjects with different head sizes and shapes. The choice of the number of electrodes is typically based on the tradeoff between the desired spatial resolution and the comfort of the subject. A large number of electrodes requires a long time to set up, which could lead to unwanted effects, such as drowsiness or frustration. When assessing mental states, these could serve as confounding factors. In recent years, work has explored the use of a smaller number of electrodes that are placed on specific regions of the scalp based on the task being explored. Promising results have been found for various applications ranging from Alzheimer's disease diagnosis [41] to mental state classification [42].

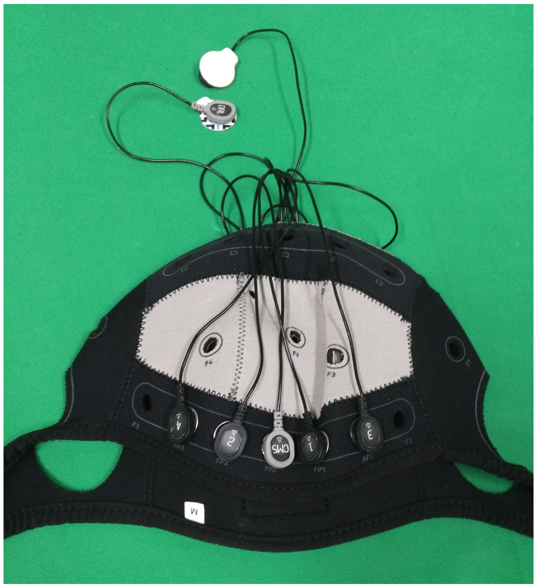


Figure 2.1 – Portable EEG cap (Enobio) using both wet and dry electrodes

2.2.2 Signal Properties

The electrodes try to capture information from different brain regions. These regions are broadly divided into four main areas, namely, the temporal, occipital, frontal, and parietal lobes, based on the anatomy of the brain [43]. These regions are also shown to be (broadly) responsible for specific brain functions. The frontal lobe is responsible for the conscious thought. The temporal lobe for the senses of sound and smell, and the processing of complex visual stimuli such as faces and scenes. The parietal lobe is responsible for integrating sensory information from various senses, as well as the manipulation of objects. Finally, the occipital lobe is responsible for the sense of sight [28]. More specifically for emotions, two regions, namely the amygdala (in the frontal portion of temporal lobe) and the prefrontal cortex (located in the frontal lobe) show changes when presented with emotional stimuli. In general, the activation of the amygdala has been related to negative emotions [213].

The EEG signals for healthy individuals have also been divided in several bands based on visual analysis of healthy individuals [44]. These regions include the so-called delta (δ: $1-4Hz$) , theta ($\theta : 4-8Hz$) , alpha ($\alpha : 8-12Hz$), beta ($\beta : 12-32Hz$), and gamma ($\gamma : 32-45Hz$) bands. The quantification of EEG signal properties based on different brain rhythms has been useful for a wide variety of applications including emotion recognition [13] and other mental state monitoring tasks [45, 32].

Brain activity has been observed to be a function of various processes ranging from fast electrical activity to slower chemical reactions and diffusive processes. Such non-linear interactions often possess “scale free” dynamics. These dynamics represent signal properties that are preserved across various time scales. Traditionally, frameworks of fractal geometry and non-linear dynamics has been used to describe such properties. The quantification of EEG signals using such metrics has shown to capture non-redundant information compared to traditionally used (linear) metrics [36].

2.2.3 Artefact removal

EEG signals are highly susceptible to noise and are impacted by e.g., eye movements, blinks, facial muscle movement, and ECG activity, to name a few. Further, environmental noise such as power line interference can also observed in EEG signals. These artefacts usually require manual detection and removal from individuals trained in EEG analysis. In general, these artefacts are controlled in laboratory experiments by asking the subjects to stay still and avoid any facial movements or speaking. However, these limitations are not possible in real-life. Recently, automated artefact removal algorithms have received a large amount of focus. These methods usually rely on blind source separation techniques and are computationally expensive [214]. Moreover, the algorithms might remove relevant EEG signal information for the prediction problem and may result in a decrease in performance. Such effects have been previously shown for Alzheimer’s disease severity prediction [33], for example. Other pre-processing steps include time-domain filtering, with notch filters to remove power grid interference (50 or 60 Hz, depending on the country), and bandpass filtering to enhance only EEG-related spectral components. Other techniques such as resampling and re-referencing may also be employed. These methods help improve the quality of the signal before further analysis.

2.2.4 Features used

The most commonly used features for mental state monitoring applications are related to the power of the different EEG energy bands and their inter-hemispheric asymmetries.

1. Spectral Power Features: Spectral power features measure the EEG power in the different pre-defined bands and are the most commonly used features across various domains [215] including emotional state classification [28]. These features are calculated either in the time or frequency domain [215]. For the frequency domain, the power spectrum density (PSD) is first estimated using either the Fourier transform (FT) or the Welch's periodogram. Following this, the power of a given frequency band is computed as the integral of the PSD over the band's frequency range. In the time domain, the EEG signal is decomposed by bandpass filters into the time series of each frequency bands of interest. The power is then computed as the average of the samples squared, as in Eq 2.1.

$$power = \frac{\sum_{t=1}^{m} x^2(t)}{m}, \tag{2.1}$$

where $x(t)$ represents the bandpassed EEG series for a given band and m represents the length of the time series. To allow for comparison within and across different subjects, the power of each individual band is typically normalized by the total EEG power.
2. Asymmetry index features: Asymmetry of the EEG power between the two hemispheres has also shown to be a useful features for emotional state detection [28] and overall mood assessment [46]. The asymmetry is calculated by taking the ratio of the PSD between the left electrode and its corresponding right electrode pair. The electrode pairs for the 10-20 standard system are: Fp1-Fp2, F7-F8, F3-F4, T3-T4, C3-C4, P3-P4, T5-T6, and O1-O2.

Changes in alpha power and its asymmetry between the left and right hemisphere have been linked to emotions. With a relative right frontal activation being related to negative states, such as fear and disgust, while increased left frontal activity associated with positive mental states (e.g., joy and happiness). In general, the EEG asymmetry may therefore reflect the valance dimension [28, 46]. On the other hand, pre-frontal asymmetry in alpha and temporal asymmetry in the gamma band are observable for arousal recognition [47]. For more specific mental states, task demand, and temporal pressure related to mental workload (and therefore stress) these are often associated with a decrease in the alpha band power in various cerebral regions, including frontal, central and parietal regions, combined with an increase in theta power at frontal and parietal regions [10, 216].

Recently, other properties of the EEG signal have also been used for detection of mental states. Quantifying such properties has provided complementary information to the commonly used PSD and asymmetry features. Functional connectivity quantifies the flow of neural information between different brain regions. As cognitive function requires the coordinated flow of information between different brain regions, such metrics have shown usefulness in detection of emotional states [48]. Amplitude modulation features which quantify the rate-of-change of spectral components have also been linked to emotional changes [197]. Further, the non-linearity of the EEG signal quantified by metrics such as fractal dimension and entropy have also been used for emotion detection [28]. Multi-scale entropy, which is a measure of complexity in a time series at different scales, has been shown to detect visually elicited mental stress [49].

Figure 2.2 depicts the general EEG data analysis pipeline consisting of data acquisition using an standard electrode placement system, followed by artefact and band-pass filtering to keep relevant information, feature extraction (e.g., band energy and asymmetry) and classification to estimate the current mental state.

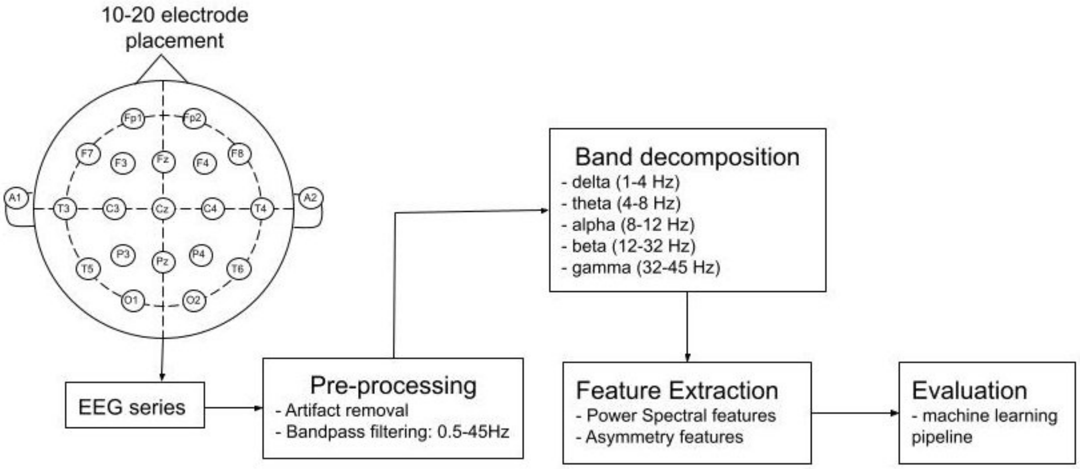


Figure 2.2 – EEG analysis pipeline

2.3 Electrocardiogram

2.3.1 Signal Source and Acquisition

The ECG is a non-invasive measure of heart activity, structure and function. It records the different electrical potentials produced by the heart on the surface of the body. This rhythmic electrical activity is based on repetitive bio-electrical stimulation to the heart muscles. The rhythm of the heart is set by a small region of cardiac muscle cells in the right atrium called the sinoatrial node (SA) that acts as a spontaneous pacemaker, but is under the control of various endogenous and exogenous factors, such as psycho-physiological stressors, respiration, blood pressure, and thermo-regulation through parasympathetic and sympathetic nerve systems [217].

The heart is composed of two upper chambers, the atria and two lower chambers, the ventricles. During normal heart rhythm (termed sinus rhythm) each beat spontaneously generated from the SA produces a propagating wave of electrical activity that spreads throughout the four chambers of the heart in a coordinated fashion. Each impulse propagates throughout the atria before being channeled through the atrioventricular (AV) node to the ventricles. The dipole cardiac vector created by this activity forms a varying electric field on the body surface. This electrical activity can be captured by putting electrode pairs on the surface of the body and is referred to as an ECG. However, electrode pairs placed at different positions generally yield different potential gradients because of the spatial dependence of the electric field generated by the cardiac vector [217, 218]. As a result, clinical evaluation of ECG requires the use of standard positions. The standard 12-lead ECG placing has widely adapted as the standard for clinical recordings. However, the 12-lead system is not practical for everyday wearable devices. As a result, these devices make use of single-lead ECG placements. Such devices have been extensively compared to 12-lead ECG placements for various applications [50, 51, 52] and in general show high correlation with clinical metrics.

Figure 2.3 depicts a typical ECG comprised of different waveforms generated from one cardiac cycle for a healthy individual, namely:

1. P wave: This represents atrial depolarization. The depolarization is not particularly powerful, resulting in a small amplitude for this wave. It usually lasts between 60 to 120 ms and represents the time taken by the pulse to propagate both atria. The spectral characteristics of this waveform are in the low frequency range ≤ 10 Hz.
2. PR segment: This segments is free from any waves and represents the time between the activation of atria to the start of activation of the ventricles.
3. QRS complex: This region is composed of Q, R and S waves. The Q wave is short, downward and negative corresponding to the depolarization of the septum. The R wave following it is long and narrow, representing the depolarization of the left ventricle apex. Finally, the S wave is small, downward and negative corresponding to the depolarization of the basal and

rear regions of the left ventricle. The complex lasts between 60 to 90 ms. The time interval between two QRS peaks, called the inter-beat (RR) interval, represents ANS activity on the heart rate. The spectral information for the complex lies in the 10-50 Hz range.

4. ST segment: This represents the time between the S and T waves when the ventricles contract and return to rest. This segments also represents the baseline level of the ECG signal.
5. T wave: This wave represents the repolarization of the ventricles (the time when the ventricles have finished their activation stage and they are ready for a new contraction) and its duration ranges between 100 to 250 ms. Similar to the P wave, the spectral characteristics of the T waves are are in the low frequency region.

The time intervals between these various peaks and differences in amplitudes are known to contain important clinical information [53]. As such, long-term ECG monitoring remains the go-to method of detecting, diagnosing, and monitoring various heart diseases [219] [220] [221].

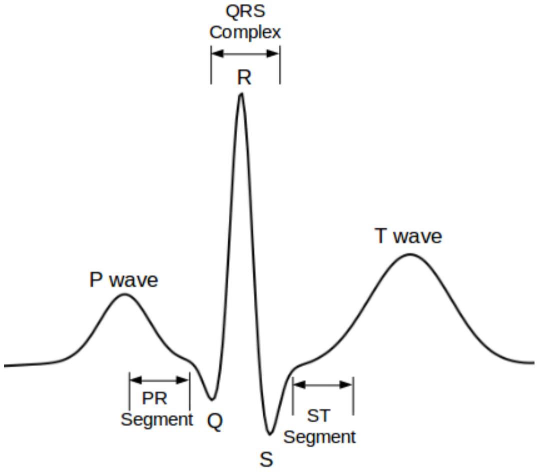


Figure 2.3 – Single ECG waveform showing the QRS complex and other segments

Since its inception by Holter [222] in the early 1960's, long-term ECG monitoring has been used to monitor patients in critical conditions in order to gather information about pathological events that may occur only sporadically over long periods of time. Typical monitors record data from up to three ECG leads during 24-48 hours on a single battery charge. The data is stored on the device and has to be downloaded by the physician at the hospital for analysis and offline diagnosis. The advances in data sensing and acquisition technologies have allowed for development of portable wireless data collection methods which can collect and transmit ECG data using single-lead ECG typically at a lower sampling rate. This data has been used for various applications, such as monitoring physical activity, health and stress [223].

The variability of the RR series represents ANS activity and is referred to as the heart rate variability (HRV). More specifically, the SNS is responsible for decreasing the RR value and its variability while the PNS increase RR values along with its variability. PNS typically dominates the heart rate activity at rest and during recovery while the SNS is active during the “fight-or-flight” response of the body. These two systems operate at different time scales, with PNS nerves being faster ($< 1s$) in comparison to SNS nerve activity ($> 5s$). While these two systems change the HRV in a contradictory manner, their interaction is non-linear in nature. While, generally SNS activation can suppress the PNS it may even increase or cause no change in the PNS activity [54].

The recommended length of ECG recording for HRV analysis is divided in long- and short-term analysis, with the former relying on windows of 24h and are typically used in clinical settings. Long-term HRV analysis captures various slow varying aspects of HR variability including circadian rhythm, core body temperature, and sleep cycle. Generally, for clinical applications, these recordings achieve better predictive results compared to short-term measurements [224, 225]. For short term analysis, in turn, a window of 5 minutes is recommended, as it allows for clear measurements of the low frequency periodic elements of the sympathetic system. Short-term HRV measurement is the result of two distinct sources. First, the complex and dynamic relationship between the PNS and SNS. Second, includes the regulatory mechanisms that control HR via respiratory sinus arrhythmia (RSA) [54, 55]. RSA refers to the respiration-driven speeding and slowing of the heart via the vagus nerve.

Recently, windows shorter than 5 minutes have been explored for various mental state monitoring applications. These windows allow for monitoring of individual’s mental state at an improved temporal resolution. Windows as small as 30 seconds have been explored. However, decreasing the window affects the reliability of these metrics in comparison with standard 5-minute analyses [56] and their use remains controversial [54]. While, the features calculated for these different window lengths use the same mathematical formula, their physiological significance and meaning changes based on window length and should not be used interchangeably [226].

2.3.2 Artefact removal

The raw ECG signal is usually corrupted by various sources of noise. These can include instrument and measurement noise, power line interference, electrode movement, and movement artefacts, which can make the detection of the ECG waveform difficult. Therefore, HRV analysis, which relies on the clear detection of QRS peaks, cannot be performed on the raw signals and require some form of filtering. Many different filtering algorithms have been proposed for filtering ECG signals to improve signal quality. A large number of these algorithms focus on preserving all the various waves in the signal for clinical analysis [57, 58, 227]. Wavelet and empirical mode decomposition (EMD) based denoising methods are among the most commonly used methods. Wavelet denoising is based on thresholding the wavelet coefficients before reconstructing the signal. EMD based methods, in turn, decompose the signal and remove the mode most correlated with noise. More recently, modulation spectrum based denoising has been proposed which outperforms standard wavelet and EMD based methods for different types of noise and levels [59]. For HRV analysis, filtering methods do not need to perverse the wave morphology as we only need to detect the QRS peaks. As such, simple bandpass filtering in 5-25 Hz range is commonly used to enhance the QRS peaks. Such band-passing has been built into a large number of QRS detection algorithms [60].

Once the ECG signal has been filtered, the QRS peaks are detected using a peak detection algorithm. The Pam-Tompkins algorithm [61] is one of the most widely used methods. The method first enhances the QRS complex using bandpass filters which also remove any other noise in the signal. Then, it squares the signal to further amplify the QRS peak. Finally, it applies adaptive thresholds to detect the peaks of the filtered signal. Once the QRS peaks are detected, the RR series is created from the time difference between consecutive peaks. Over the years, several modification to the original algorithm have also been proposed to increase robustness to artefacts [62, 63, 64].

Noisy ECG may corrupt the RR series due to missing or false QRS peaks. Physiological artefacts form another source of RR series noise in the form of several abnormal beats. These include premature beats (ectopic beats), caused by electrical impulses generated from outside the SV node, or atrial fibrillation, often caused due to underlying cardiovascular conditions. Ectopic beats can occur in perfectly healthy individuals [228] and may even increase with physical activity [184]. These abnormal beats can distort the HRV metrics calculated and are known to introduce errors in several HRV metrics [39]. Traditionally, visual analysis of the ECG and the RR series have been done to either exclude regions with poor signal quality or edit the RR series detected. In recent years, automated algorithms for deleting, interpolating or filtering of RR signals have been proposed. Each of these methods might impact the HRV metrics differently and no current standards exist for treating noisy RR time series [39]. Existing methods first attempt to detect abnormal segments and then perform various filtering and outlier detection algorithms. These include:

1. Physiological range based detection of abnormal beats: The range for healthy adult humans is around $\geq$ 320 ms and $\leq$ 1400 ms. Anything outside this range is considered to be an abnormal beat.
2. Moving average filter based detection of abnormal beats: A moving average filter is calculated on the RR series and any value outside a specified range from the moving average value is considered as abnormal.
3. Quotient based filter: Quotient based filtering involves detecting abnormal beats based on the rate-of-change of RR value from one value to the next. The acceptable rate-of-change between two beats is considered to be around 20%.

Following the filtering of RR series, HRV features are calculated from the series. The overall HRV analysis pipeline is shown in Fig 2.4. First, the ECG is filtered by band-passing between 5-25 Hz range to enhance the QRS complex peaks. Following this, the RR series is extracted from ECG signal using automated QRS complex detection algorithms. Then, RR filtering is used to remove abnormal beats. Finally, HRV is calculated on the filtered RR series using standard time- and frequency- features. These features have been widely recommended for HRV analysis [55] and try to quantify the overall ANS activity. The next section describes these commonly-used HRV features.

2.3.3 Benchmark HRV features

As mentioned above, the filtered RR series is used to extract several HRV features. Time- and frequency- domain features are commonly recommended as benchmark features for HRV [55] analysis. Similar to EEG, the RR series also exhibits complex fractal like behavior [65]. In the past few decades, these characteristics have been studied using non-linear dynamics and chaos theory based features. These features have generally shown an improved performance over the

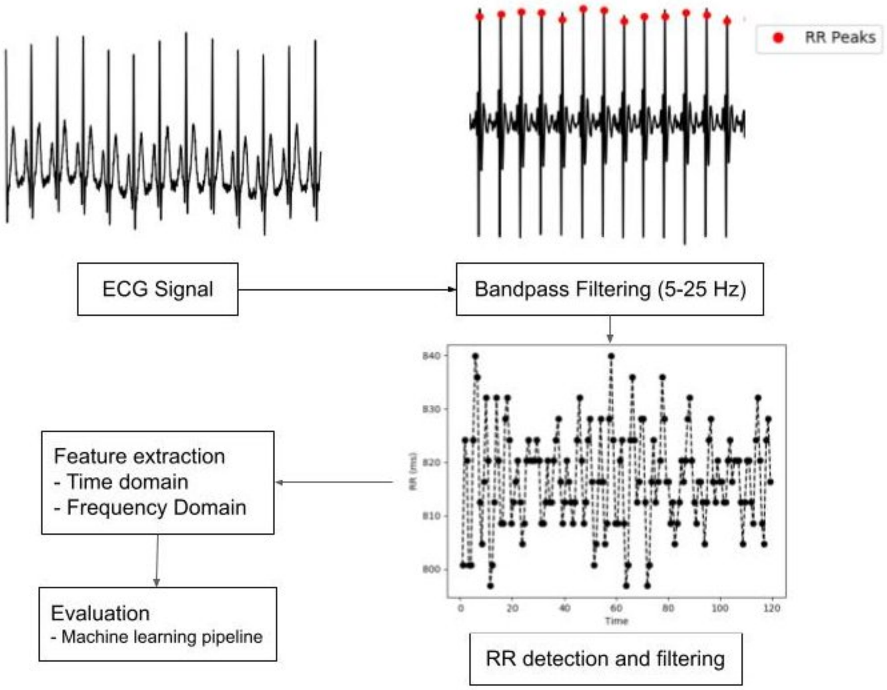


Figure 2.4 – ECG analysis pipeline showing bandpass filtering followed by RR detection and feature extraction

benchmark feature set and have been recommended for various clinical applications [35]. These different benchmark feature sets are described below.

2.3.3.1 Time Domain Features

Time domain features are directly calculated over the RR series. The commonly used features include [55, 54].

1. Mean of RR ($meanRR$): It is one of the most commonly used features. A decreased $meanRR$ represents sympathetic dominance while an increase in its value is caused in activation of the PNS. Its value is closely related to arousal, with high arousal states such as excitement, anger and anxiety decreasing its value while low arousal states, such as boredom and fatigue increasing $meanRR$.
2. Standard deviation of RR ($sdRR$): While the $meanRR$ captures the overall level of the heart rate, $sdRR$ is commonly used metric for capturing its variability. It is closely related to the SNS and a decrease in $sdRR$ is an indicator of sympathetic activation. As mentioned earlier,

sdRR is sensitive to RR series outliers [39]. The mean normalized version of *sdRR* called coefficient of variation has also shown usefulness in the literature [229, 230].

3. Root mean square of successive difference (*rmssd*): This feature has been linked to PNS response. An increased *rmssd* value is an indicator of PNS activation while a decrease suggests its withdrawal. It's relation to PNS withdrawal makes it a useful feature for assessing negative mental states, which usually involve an SNS activation with a corresponding PNS withdrawal.
4. $pNNx$: It is defined as the ratio of the consecutive RR intervals differing by more than "x" milliseconds, divided by the total number of RR intervals. The most commonly used value for x is $50ms$ with $20ms$ also being used in certain cases. Similar to *rmssd*, it is an indicator of PNS activity and is highly correlated with *rmssd* value. Both $pNN50$ and *rmssd* have been shown to be severely impacted by ectopic beats.

While these features are the most commonly used, several other time domain features have also been used as benchmark features for various mental state monitoring applications. The interested reader is referred to [38, 37] for more examples.

2.3.3.2 Frequency Domain Features

The RR series is not sampled at uniform intervals due to differences in the duration of adjacent heart beats. Therefore, a frequency domain analysis needs to take the non-uniform sampling into account. One such approach directly calculates the frequency domain transform from the non-uniformly sampled RR series as a function of beat index. In such a case, the frequency domain axis of the spectrum is a function of cycles per beat [231]. However, a more popular approach is to resample the RR time series using linear or cubic interpolation methods in order to obtain equally spaced samples [55]. This is done by using the time index of the QRS peak as the time index for the RR interval followed by resampling to obtain the so-called RR tachogram series. Following this, the PSD of the tachogram is calculated using different methods, namely, FT, Welch's periodogram or auto-regressive models. Similar to EEG analysis, HRV PSD has been divided into various bands. For 5-minute RR series segments, these bands are: very low frequency (VLF) (0.0033 Hz-0.04 Hz), low frequency (LF) (0.04 Hz-0.15 Hz), and high frequency (HF) (0.15 Hz-0.4 Hz). The power in each of these bands can be used as features, as described below:

1. VLF power: VLF power is usually related to overall health of an individual [54] and requires a full 5-minute recording to be captured correctly for HRV analysis.
2. LF power: LF power is associated with both SNS and PNS activity and is correlated with $sdRR$. A minimum segment length of 2 minutes is required to accurately calculate LF power.
3. HF power: HF power is modulated by the PNS system as well as the RSA. HF power is highly correlated with $pNN50$ and $rmssd$, both of which also indicate PNS activity. Both HF and LF powers have been measured in absolute and relative terms. Relative power (refereed as LFnu and HFnu) is calculated by dividing the band powers to the total power minus the VLF power. HF power is highly susceptible to ectopic beats which introduce high frequency components into the RR tachogram [39]. HF power can be inaccurately estimated with a minimum segment length of 1 minute.
4. LF-to-HF ratio: The LF-to-HF (LF/HF) ratio has originally been used to calculate the balance between SNS and PNS activity (sympatho-vagal balance). The idea being that LF and HF powers represent SNS and PNS activities, respectively. However, recent discoveries of PNS contribution to the LF and RSA contributing to the HF bands have made this a controversial metric [232, 54].

Negative emotions such as stress, anxiety, and in general, higher mental workload are associated with an increased SNS response with a PNS withdrawal. This corresponds to a decrease in HF power with an increase in LF power further leading to an increased LF/HF ratio. A meta-analysis [32] recognized three of the time- and frequency- domain features that consistently decreased with increased stress; these include, the mean of the RR series, $rmsdd$, and $pNN50$. Additionally, a majority of the studies reported that $sdRR$ and HF decreased during stress, while showing increases in LF/HF and LF [32, 66, 67, 68]. Similar changes are brought on to the HRV features by an SNS increase with mental fatigue [10].

2.3.3.3 Non-linear features

The RR series exhibits complex non-linear behavior [65]. This behavior has been observed to change based on different physical and psychological conditions [35]. Non-linearity in the RR time series has been quantified using different measures, such as entropy, fractal and chaotic/dynamical system measures. The ANS adapts the heart rate based on the current demands of the body.

Such an adaptation process happens continuously, leading to irregularity in the RR series. This irregularity can be quantified using sample and permutation entropy for RR time series, as detailed next:

1. Sample Entropy: Sample entropy (SampEn) is the negative natural logarithm of an estimate of the conditional probability that if two sets of vectors ($X_m(i)$ and $X_m(j)$) of length m have a distance $< r$, then two sets of vectors ($X_{m+1}(i)$ and $X_{m+1}(j)$) of length $m+1$ also have a distance $< r$, based on some distance metric $d_m(X, Y)$. It is formally defined as:

$$SampEn = -\log \frac{N_{m+1}}{N_m}, \tag{2.2}$$

   where N_m is number of vector pairs with $d_m(X_m(i), X_m(j)) < r$ and N_{m+1} is number of vector pairs with $d_m(X_{m+1}(i), X_{m+1}(j)) < r$. SE has shown to be an important predictor of various mental states such as stress [32] and anxiety [70] as well as mental fatigue [233].

2. Permutation Entropy: The permutation entropy (PE) algorithm quantifies the occurrence of motifs in the series. Motifs are defined as recurring patterns in the time series with a degree m and lag λ. Based on the rank ordering of the motif pattern we assign it a specific symbol j. Representative motifs of degree ($m = 3$) and lag ($\lambda = 1$) are shown in Fig. 4.3. Each motif can be represented as an alphabet or a number. The time series ($X(t)$) is first converted to the ordinal series ($X^{m,\lambda}(j)$), where $1 \geq j \leq N - m$ where N is the size of the time series using the following relations (for degree $m = 3$):

$$X(i)^{m,\lambda} = \begin{cases} 1 & \text{if } X(i) < X(i+\lambda) \text{ \& } X(i+\lambda) < X(i+2\lambda) \text{ \& } X(i) < X(i+2\lambda), \\ 2 & \text{if } X(i) < X(i+\lambda) \text{ \& } X(i+\lambda) > X(i+2\lambda) \text{ \& } X(i) < X(i+2\lambda), \\ 3 & \text{if } X(i) > X(i+\lambda) \text{ \& } X(i+\lambda) < X(i+2\lambda) \text{ \& } X(i) < X(i+2\lambda), \\ 4 & \text{if } X(i) < X(i+\lambda) \text{ \& } X(i+\lambda) > X(i+2\lambda) \text{ \& } X(i) > X(i+2\lambda), \\ 5 & \text{if } X(i) > X(i+\lambda) \text{ \& } X(i+\lambda) > X(i+2\lambda) \text{ \& } X(i) > X(i+2\lambda), \\ 6 & \text{if } X(i) > X(i+\lambda) \text{ \& } X(i+\lambda) < X(i+2\lambda) \text{ \& } X(i) > X(i+2\lambda). \end{cases}$$

   The PE is then given by:

$$PE = -\sum_{j}^{m!} p(\pi_j^{m,\lambda}) \cdot \log(p(\pi_j^{m,\lambda})), \tag{2.3}$$

where $p(\pi_j^{m,\lambda})$ is the relative frequency of the motif pattern represented by $\pi_j^{m,\lambda}$ and calculated as:

$$p(\pi_j^{m,\lambda}) = \frac{\sum_{j \leq m!} \nVdash_{u:type(u)=\pi_j}(X_j^{m,\lambda})}{\sum_{j \leq m!} \nVdash_{u:type(u)\in\Pi}(X_j^{m,\lambda})}, \tag{2.4}$$

where $\nVdash_A(u)$ denotes the indicator function of set A defined as $\nVdash_A(u) = 1$ if $u \in A$ and $\nVdash_A(u) = 0$ otherwise and $type(.)$ denotes the map from pattern space to symbol space.

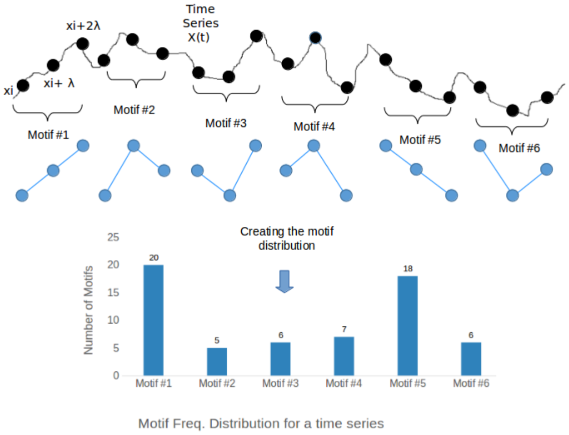


Figure 2.5 – Motifs for degree 3 and lag 1 along with the motif distribution

The robustness of motif features comes from the fact that they only consider the underlying shape of the time series and not the amplitude. As most artefacts for physiological signals manifest as short-term amplitude fluctuations in the signal, PE is more robust to outliers in the time series as it removes the amplitude information from the signal [234]. Previously it has been shown to be robust to eye blink artefacts for EEG data [235]. Additionally, PE also holds advantage over traditionally used Shannon entropy measure, which is based on calculation of entropy based on distribution of data, as motif distribution takes into account the time ordering of the time series [234]. Due to these advantages, PE of RR series has been previously used for prediction of different emotional states [236]. PE has in the past been used to recognize sleep state [237], as well as the effects of anesthesia [235], to detect seizures [238, 117], and to measure alertness [69].

Moreover, the RR series complexity can be quantified using fractal measures. The concept of fractals comes from geometry and is used to quantify objects which are too irregular for traditional geometrical descriptions and also exhibit some self-similar behavior, i.e., they look similar at different scales. Temporal processes are said to exhibit fractal behavior when the fluctuations in time series at smaller and larger scales are statistically equivalent. Such signals are characterised by various properties, including power law spectral density with the slope (α) for the $1/f$ behavior representing the scaling factor. These fractal properties can also be studied with phase space reconstruction and quantifying the phase trajectory of the non-linear systems [239]. The following measures have been widely used to quantify fractal behaviors in time series:

1. Detrend fluctuation analysis (DFA) quantifies fractal-like correlation properties of the time series data. The root mean square fluctuations of the integrated and detrended data are measured within the observation windows of various sizes and then plotted against the size of the window on a log—log scale. The scaling exponent represents the slope of this line. It has been used for stress prediction [32].
2. Lyapunov exponent (LE): The behavior of dynamical systems can be studied as an evolution of phase space trajectory. The trajectories can either diverge or converge from the initial state of the system representing increasing chaotic behavior or predictability. The Lyapunov exponent is an estimate of the duration for which the dynamical system behaves predictably before it becomes chaotic. The largest Lyapunov exponents usually govern the overall behavior of the system and can be predicted using phase space reconstruction for time series [240]. It has previously been used to predict anxiety [70].
3. Correlation dimension: An attractor is defined as a region of the phase space describing the steady state of the system where various initial states of the system converge to. A chaotic (also called strange) attractor is fractal in nature and its fractal dimension is an estimate of the complexity of the system. The correlation dimension (CorrDim) [241] is used a describe the fractal geometry of a chaotic attractor and signifies the number of independent variables required to describe a dynamical system. It has been used previously for mental workload prediction [37].

These non-linear approaches have been used in the literature to estimate stress [32] , mental workload [37] and anxiety [70]. More specifically, work in [37] showed CorrDim as being a robust metric for predicting mental workload when the stimulus is provided for a long duration (∼1 hour). While more traditional metrics go back to their original resting state values after high mental workload is maintained for more than half an hour, correlation dimension remains significantly lower than its resting state value for the entire duration of the high mental workload condition. Work from [32] also recognized CorrDim as one most consistent metrics for recognizing high mental stress in a meta-analysis. Other non-linear metrics such as Hurst exponents have also shown to predict social anxiety [38]. While PE has been used widely due to its robustness, modulation spectrogram based HRV [242] measurement has also shown robustness to different types of noise compared to traditional HRV metric. However, the performance of these features for in-the-wild mental state monitoring applications still remains to be seen.

2.4 Other Physiological Modalities

Various other physiological signals have been used for mental state monitoring [13]. These include:

1. Photoplethysmographam (PPG): PPG is a non-invasive, low-cost optical technique that detects the pulsatile changes in blood volume. A PPG device is comprised of two main components: a light source that sheds light on the skin and a detector that measures small variations in the intensity of the back-scattered light. These variations are related to the changes in blood volume that occur with each cardiac pulse. The PPG signal can be measured from various locations, such as finger tips, wrist, and ear lobes. Given the fact that PPG measures the blood volume changes at each pulse, PPG can be used for measurement of heart rate and HRV. As a result, PPG-based HRV has been used for emotional state monitoring in a multimodal setting [13], as well as for stress detection [71], to name a few.
2. Respiration signal: Respiration is an easy modality to be monitored by smart-garment based devices, such as smart-shirts or chest straps place around the chest. Typically, breathing rate based measures have been used as correlates of mental states such as anxiety, amusement, and boredom [243]. Respiration can be considered as a mixture of two processes of metabolic and behavioral breathing originating from different parts of the brain [72], with the latter being affected by internal and external stimuli. Usually, these changes are observable in two different aspects of breathing: (i) respiration or breathing rate and (ii) tidal volume or breathing depth. Mental stress has been shown to increase both respiration rate and breathing depth [73]. Similarly, anticipatory anxiety is associated with an increase in respiration rate [74]. This respiratory variability has been quantified with simple statistical functionals, such as mean, coefficient of variation, and auto-correlation [75] of the breathing curve. Other commonly used respiration features include band powers and ratios for different frequency

regions [13]. Similar to other physiological signals, breathing signals are known to possess complex fractal behavior [244].

3. Galvanic skin response (GSR): The GSR signal captures changes in sweat gland activity and consists of two components: a slow tonic component which changes very slowly and a faster phasic component. This faster phasic component has been linked to levels of emotional arousal [245]. As a result, GSR has been collected for various mental state monitoring applications [13, 76, 77]. Typically, various statistical descriptors along with band powers of different frequency regions have been used [13].
4. Skin temperature: Skin temperature is known to change with changes in emotional states [13]. Typically, statistical descriptors are used as features for the temperature series.

Next, we discuss the development of multi-modal recognition systems that rely on multiple such modalities. These systems are known to generally provide improved performance and enhanced robustness to artefacts when compared to single modality methods.

2.5 Multi-modal mental state monitoring

It is well known that different physiological signals carry complementary information for mental state monitoring [13, 78, 79]. For example, EEG signals can capture the CNS activity while the peripheral signals can reflect ANS activity. These different signals may even capture information about specific emotional dimensions, e.g., GSR may respond to arousal while EEG may reflect both valence and arousal. This behavior can also help compensate the limitations of one modality with another. Multi-modal systems also provide added robustness against sensor failures, as poor signal quality in one modality may be compensated by other signal streams. For example, EEG signals are known to be contaminated with ocular, heart, and breathing rate artefacts. Such details may be used in scenarios where e.g., heart rate data may be lost. Multi-modal information can be combined at different levels, as shown in Fig. 2.6:

1. Signal level (Fig. 2.6-a): The different physiological signals often show coupling behavior as they may be governed by similar underlying mechanisms. One of the most well known mechanisms for such coupling is via RSA, where respiration drives heart rate. Such coupling may be impacted by mental states. As a result, features to quantify such coupling have been

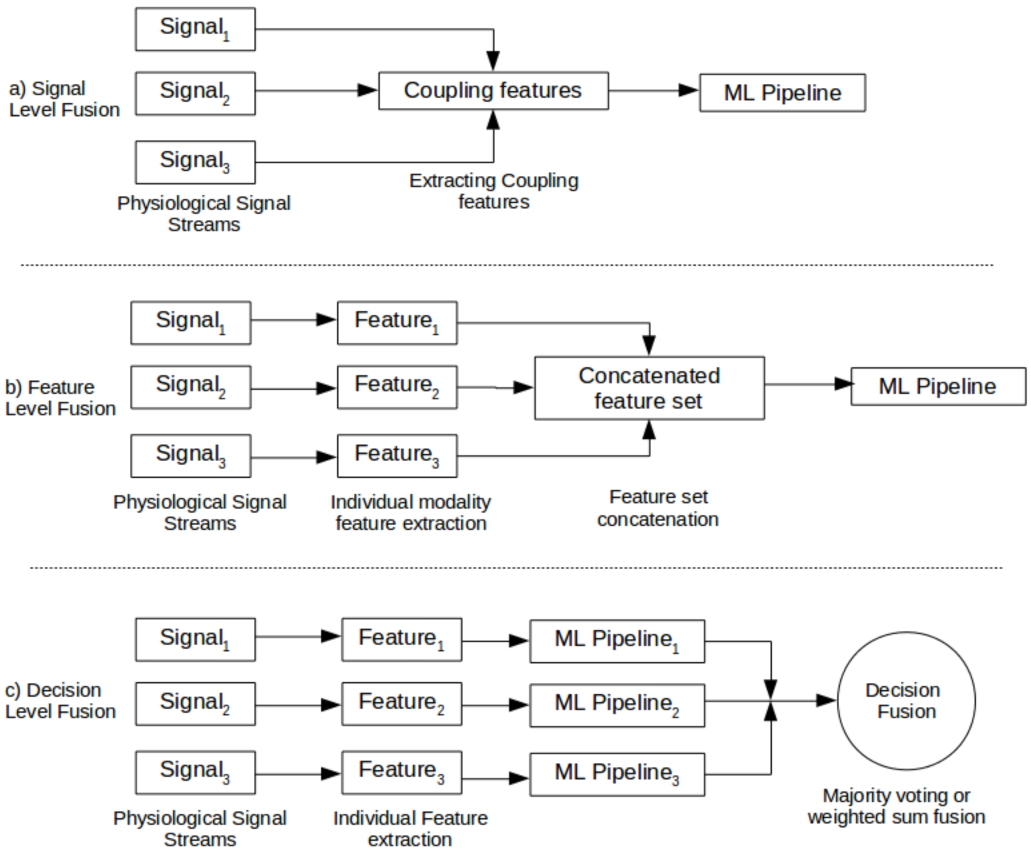


Figure 2.6 – Various multi-modal fusion strategies a) Signal level fusion, b) Feature level fusion, and c) Decision level fusion

used. The works from [80, 81] used cardio-respiratory coupling features to improve emotion recognition. Similarly, [82] observed changes in the phase synchronization of HF component of HRV with respiration and blood pressure during mental stress tasks. Coupling between GSR and EEG signal has also been used for emotional state detection [83].

2. Feature level (Fig. 2.6-b): For this fusion, the features from the different modalities are separately extracted and combined before inputting them into a ML pipeline. This is one of the most commonly used approaches [13, 78]. Different physiological signals often have distinct temporal dynamics which requires separate epoch lengths for feature extraction, e.g., 1-3 seconds for EEG and 5 minutes for HRV. As a result, feature level fusion requires synchronizing the different modalities by averaging over epochs or other strategies. Usually, the coupling features derived in the case of signal level fusion are fused with individual features too.

3. Decision level (Fig. 2.6-c): Decision level fusion combines the output of the ML pipelines from each of the modalities using majority voting or weighted methods. This method provides robustness against corruption of single modalities by noise [84] and may take signal quality into account [85].

As described above, signal and feature level fusion can be done prior to the ML pipeline. In general, these two schemes are able to learn the interaction between signals and features and may provide the largest amount of performance improvement. Decision level fusion, in turn, may help with robustness to failure of one or multiple sensors as a trade-off for performance. It is also important to note that while these strategies are widely used for combining features from different modalities, they can also be used to combine different feature sets from the same modality [48]. In the next section, different components of the machine learning (ML) pipeline are described.

2.6 Machine learning pipeline

Once the features are extracted, mental state prediction is done using machine learning algorithms. Supervised ML algorithms make use of the input features along with the corresponding mental state labels to learn the relationship between the two. The labels can be binary or continuous. Binary classification (e.g., stress vs no stress, low vs high valence/arousal) has been commonly used for mental state prediction [48, 32]. In this case, dataset labels that are continuous are binarized based on a specific threshold value. Alternatively, the ground truth of the experimental stimulus can also be used as a label (e.g., stressful videogame vs relaxing videogame conditions). Typically, the ML pipeline involves dimensionality reduction to prevent over-fitting, followed by training a ML model while using a specific evaluation method. The evaluation method ensures the training and test samples remain separate from one another in order to prevent optimistic bias in the results. Finally, the results are evaluated using certain figures-of-merit. In this section, we describe the different aspects of the pipeline.

2.6.1 Feature selection

Feature selection helps reduce the dimensionality of the dataset to help avoid overfitting. Additionally, it eliminates several features that might be highly correlated and may not provide any additional information to the classifier. Several different feature selection approaches have been proposed in the literature [86]. Feature selection also helps provide insight into the decision making process of the model by revealing the top features used to make the prediction. Some of the commonly used feature selection methods are described below.

1. ANOVA (analysis of variance) based feature ranking and selection: This selection method is based on calculating the significance of the input features with respect to the output values and return the ranked features according to their obtained p-values.
2. Minimum redundancy maximum relevance (mRMR) feature selection: mRMR is a mutual information based algorithm that optimizes two criteria simultaneously: the maximum-relevance criterion (i.e., maximizes the average mutual information between each feature and the target vector) and the minimum-redundancy criterion (i.e., minimizes the average mutual information between two chosen features). The algorithm finds near-optimal features using forward selection with the chosen features maximizing the combined max-min criteria.
3. Recursive feature elimination (RFE): Given an external estimator that assigns weights to features, the least important features are pruned from the current set of features. The procedure is recursively repeated on the pruned set until the desired number of features to select is reached. This technique considers the interaction of features with the learning algorithm to give the optimal subset of features. Since recursive training and feature elimination is required, this method takes a significant amount of runtime as the number of features increase.

2.6.2 Evaluation

ML algorithms are expected to be trained and then evaluated on different samples of data. Doing so ensures no information leakage in present in the model, hence prevent optimistic biases in the results. While larger datasets can divide the data into separate training and test subsets, for smaller datasets doing so can drastically reduce the training data size and introduce variability into the results. Therefore, for such cases cross-validation (CV) is an appropriate approach.

In the standard CV, called k-fold CV, the training set is split into k smaller sets. Following this, for each of the “folds”, the following step is repeated. A model is trained using k-1 of the folds as training data. This trained model is then validated on the remaining part of the data (i.e., the k-th held-out fold). This process is repeated until all the k folds have been used for validation leading to k different results. The performance measure reported by k-fold cross-validation is then the average of the values computed for each loop. This approach can be computationally expensive, but assures that all data is used for training whilst minimizing any biases in the results. K-fold cross validation has been used for various mental state monitoring applications.

Ideally, the models developed are expected to generalize over data of new individuals or be personalized for specific individuals. Therefore, different evaluation strategies that take subject-specific data into consideration have been created. These can include *(i)* subject-wise models, usually developed for EEG signals due to their smaller epoch size allowing for large number of data samples collected for each subject. *(ii)* Leave-one-subject-out (LOSO) model testing, where models are trained on the data of N-1 subjects and evaluation is done on the held-out Nth subject. This is repeated across all subjects in the dataset. However, for datasets with small number of participants (as is typically the case with emotion studies [28]) it is hard to ensure generalizability across subjects. In recent years, deep learning based approaches have shown some improvement in generalization with EEG data for mental workload detection [246], especially when domain generalization tools are used [206]. *(iii)* Finally, cross-subject testing assumes data from each subject is available in both the training and test samples; this method is widely used in the literature. The commonly used K-fold CV setting with shuffled data is an example of a cross-subject model.

2.6.3 Classifiers

Support vector machines (SVM) have been widely used in the literature [38, 71, 87, 88] due to its ability of handle non-linear data. A recent survey on emotion recognition using EEG signals [28] found that the majority of works used SVM with different kernels; the radial basis function (RBF) kernel was widely employed to help with such non-linearities. Additionally, SVM has been shown to have high stability and generalization ability and is not as affected by overfitting [71]. Previous work showed that mRMR feature selection paired with a SVM classifier [247] achieved the best performance in EEG-based emotion recognition tasks [248]. More recently, deep learning models that rely on large datasets as well as computational resources have also gained popularity [249]

2.6.4 Figures-of-merit

To assess classifier performance, accuracy (ACC) and F1-score (F1) are the most commonly used measures [38, 48]. These metrics are reliable when the dataset is balanced. This is usually the case with experiments conducted in controlled laboratory conditions. In such experiments, the experimental stimulus is provided in a balanced manner while collecting data with high signal quality. However, experiments conducted in-the-wild do not have controlled stimuli. Additionally, noisy and missing data may remove certain instances from the dataset, hence further exacerbating the imbalance. Such cases require the use of metrics that are robust to imbalances. Commonly used metrics for imbalanced datasets include:

1. Balanced accuracy (BACC): is defined as the average of the sensitivity and specificity metrics. The minimum value is 0 with a maximum value of 1. In terms of values of the confusion matrix, it is defined as:
$$BACC = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2}, \tag{2.5}$$
where TP represents the true positives, TN the true negatives, FP represents the false positives and FN the false negatives.
2. Matthews correlation coefficient (MCC): Recently, MCC has been used for performance calculation for imbalanced data [89]. MCC takes into account all four values in the confusion matrix, and a high value (close to 1) means that both classes are predicted well, even if one class is disproportionately under- (or over-) represented. It is defined as:
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{2.6}$$
MCC has shown robustness in predicting imbalanced data over traditionally used metrics [250].

The next section describes the available physiological datasets for mental state monitoring along with their limitations.

2.7 Available Datasets

The data collected for mental state monitoring using physiological signals relies on eliciting specific mental states while simultaneously controlling for any physical or mental confounding factors. Several publicly available datasets exist for mental state monitoring. However, these datasets are limited in their scope by two factors: *(i)* the type of emotional stimulus used in these experiments might not represent real life situations correctly, and *(ii)* the type of control conditions used to ensure high quality data collection might make in-the-wild use difficult due to the variability introduced in the signals. The current section describes the types of emotional stimuli provided, some of the publicly available datasets, and their limitations.

2.7.1 Commonly used emotional stimuli

In order to study and model the physiological responses to various mental states, these states need to be evoked in individuals. Studies have tried to evoke these states using different strategies that attempt to make the emotional response as real as possible within a controlled laboratory environment, hence assuring high quality data collection. The usual methods for eliciting these states include:

1. Using multimedia stimuli: A large number of studies rely on the use of multimedia content, such as music and audio clips [90, 91, 92, 93], pictures [94, 95] and videos [13, 96, 84] to elicit desired mental states. These methods typically use multimedia divided in categories to evoke a range of varying emotional responses. The International Affective Picture System (IAPS) [97], for example, is one such standardized database consisting of pictures that have been widely used for eliciting mental states. The labelling for each image was done by asking subjects of different ages, genders, and ethnicities to score the images on the three dimensions of valence, arousal, and dominance. Similarly, the International Affective Digitized Sounds (IADS) [98] consists of affective sounds characterized in terms of valence and arousal dimensions. Such databases and the use of standardized stimuli allow for replicating studies in a reliable manner while also allowing for comparison across studies.
2. Using simulated environments: Such strategies are used for inducing specific mental states in contrast to a wide range of emotions. The Trier Social Stress Test (TSST) [99] is one

of the most popular methods for eliciting stress. The test uses social evaluative situations to induce stress. For this test, the participants are brought to a room with judges and/or audience and are asked to prepare a five minute presentation. This first phase of the experiment induces anticipatory stress in individuals. Following this, the participant gives the presentation in front of the audience for 5 minutes while the audience observes them without comment. Immediately following the presentation, the participants are asked to perform a mental asthmatic task (counting back from 1022 in steps of 13) for 5 minutes which they restart each time a mistake is made. Several different modifications to this protocol have been used [38, 251]. A popular modification is the Montreal Imaging Stress Task (MIST) [100], which consists of a series of computerized mental arithmetic challenges, along with social evaluative threat components that are built into the program or presented by the investigator.

3. Using standardized tests: Several tests exist for evoking mental workload, and subsequently, mental stress while employing specific mental resources. The most popular ones include the Stroop test [101], visual search [102], mental arithmetic [103] , mental rotation [252], as well as the n-back task [253]. These tests have been shown to engage various mental resources such as working memory, visual attention, and spatial processing. As a result they have been widely used for modulating stress [32] and mental workload [254] levels. A long duration use of such tasks has also been used for mental fatigue research [69].
4. Using video/simulation games: Simulated games can help elicit mental states in conditions which replicate real life and are more immersive. As a result, pilot [104, 10] and driving simulators [105, 106, 10] have been used to study attention, mental workload, and fatigue. Video games have also been used to provide various mental stimuli [107] in normal and virtual [108] environments. Another popular simulator used for mental workload is the Multi-Attribute Task Battery-II (MATB-II) [109]. It consists of four different tasks, namely, system monitoring, tracking, communications monitoring, and resource management. These tasks need to be performed simultaneously employing various mental resources. The difficulty level of all of the four tasks can be modulated to control mental workload levels. MATB-II has shown to be better at evoking mental workload than typically used standardized tests, such as n-back and visual search [255]. The tasks also represent real life work conditions more closely, as they engage multiple mental resources simultaneously, something not achieved with other types of tests.

Next we describe the characteristics of some publicly available databases for mental state monitoring research.

2.7.2 Publicly Available Datasets

Several publicly available datasets exist for mental state monitoring applications. Often, these datasets are multi-modal and measure various physiological and other signal modalities. These datasets allow for comparison and standardization of methods across the literature.

For emotional state detection, the Dataset for Emotion Analysis using EEG, Physiological signals (DEAP) and Mahnob HCI dataset are the most commonly used datasets with a large majority of existing studies relying on one or both of these datasets [28]. For the DEAP dataset, 32 healthy participants (50% females, average age = 26.9 years) watched 40 one-minute long videos. Thirty-two channel EEG data (10-20 placement system, sampling rate: 512 Hz) and other peripheral physiological signals were recorded using a Biosemi ActiveTwo system (Amsterdam, Netherlands). Participants were presented with 40 one-minute long music videos with varying emotional content. These video clips were selected based on a previous analysis of several hundred videos as they were shown to elicit the strongest reactions across the four quadrants in the valence-arousal space (i.e., low valence, low arousal; low valence, high arousal; high valence, low arousal; and high valence, high arousal). Prior to each video there was a baseline period of five seconds where the participants were asked to fixate at a cross in the middle of the screen. Following the presentation of each video, participants were asked to rate the music videos on discrete 9-point scales for valence and arousal using SAM. While other dimensional ratings, such as dominance and liking were also collected, these have not been explored in this thesis. To control for any noise the participants were sitting during the experiments and were not talking or performing any other physical activity. The dataset has been widely used for affect recognition using EEG signals [256, 197] and its combination with other physiological signals [257, 258]. The EEG data is available for public download in raw format or in pre-processed format, which includes common referencing, down-sampling to 128 Hz, bandpass filtering between 4-45 Hz, and eye blink artefact removal via independent component analysis. Since this is a standard pipeline for EEG processing, the analysis reported in Chapter 3 is done on the pre-processed data. Data per subject was epoched into forty 60 s long trials with a 3 s long pre-stimulus baseline. The pre-stimulus baseline was then subtracted from the preprocessed data.

MAHNOB-HCI [78], in turn, is a multimodal database recorded in response to affective stimuli. A large number of physiological and other signals were recorded. These include both EEG and peripheral (ECG, GSR, respiration amplitude, and skin temperature) physiological signals along with face and audio recordings, as well as eye gaze data. Similar to the DEAP dataset, a Biosemi ActiveTwo system (Amsterdam, Netherlands) was used for recording physiological signals. Twenty-seven participants participated in two experiments. For the first experiment, they watched 20 emotional videos and self-reported their arousal, valence, dominance, and predictability ratings, as well as emotional keywords. In the second experiment, short videos and images were shown once without any tag and then with correct or incorrect tags. Agreement or disagreement with the displayed tags was assessed by the participants. Noise was controlled by asking the participants to remain still during the course of the experiment.

Both of the above mentioned databases have been widely used. However, they made use of specialised, non-portable and expensive equipment with higher sampling rates for collecting the physiological signals. Such devices restrict the use of the developed methods to controlled laboratory environments. As a result, recently the DREAMER dataset [110] was published. Data comprises EEG (10-20 placement, 128 Hz sampling rate) and ECG (256 Hz sampling rate) signals recorded from 23 participants along with their affective ratings (i.e., valence, arousal, and dominance) after watching film clips of varying lengths. All the signals were captured using portable, wearable, wireless, low-cost, and off-the-shelf equipment that can be used in everyday applications. Similar to other databases, however, the subjects were asked to minimize their physical movement during experimental sessions to minimize artefacts.

For more specific mental states such as stress, anxiety and mental workload, only a small number of publicly available datasets exist. The most popular being the one described in [111]. Data was recorded from 24 participants driving around the Boston area and rating the stress level induced by driving. Participants drove on a fixed route of around 20 miles. While stressful events could not be controlled in such a setting, the route was chosen to include highway and city driving to emulate medium and high levels of stress. Baseline levels of no stress were collected by asking the driver to sit still in the car. The recordings include ECG, galvanic skin response (GSR), electromyogram (used to measure upper back (trapezius) tension), and respiration patterns. The signals were collected using an embedded computer in a modified car.

Another popular database for stress assessment is the Smart Reasoning for Well-being at Home and at Work (SWELL) dataset [259]. This dataset consists of multimodal data from 25 participants performing typical knowledge work. This includes tasks such as making presentations, writing reports, searching for information and reading e-mails for a 1 hour period. Stressors, in the form of email interruptions and time pressure, were used to manipulate stress levels. The participants were asked to not smoke or drink caffeine 3 hours prior to the experiment to remove any confounding factors. To further remove any other effects an 8-minute rest period was included before the start of the work session. Several subjective ratings were used including NASA-TLX for mental effort, as well as valence, arousal and dominance emotional ratings. Additionally, stress was also assessed with a visual analog scale (VAS) from not stressed (0) to very stressed (10). Signals collected include ECG and GSR (recorded with Mobi device, sampling frequency: 2048 Hz) as well as video, body posture (using a kinect 3D). This dataset provides two distinct advantages over a large number of other stress studies [32]. First, the stressors used correspond to real work compared to artificial tasks (e.g., n-back or mental arithmetic). Also, the data has been collected with off-the-shelf sensors which can be used in real life settings.

In the next section, the limitations of the currently available datasets are described.

2.7.3 Limitations of Existing Datasets

The currently available datasets have the following limitations when it comes to using them for development of in-the-wild mental state monitoring applications.

1. Movement: Mental state monitoring studies have typically required the participant to remain stationary. This is done to minimize any movement artefacts in the signals collected. Additionally, physical activity may change the dynamics of the physiological signals being studied, e.g., physical movement and exercise can lead to increased SNS activation [260] which may mask effects of mental states, which can also be associated with SNS activation [32]. Physical activity also requires a large number of mental resources, thus adding to mental workload for individuals. This is specially relevant in mental state monitoring of jobs such as first responders that involve physical activity along with the need for quick decision making. Hence, by excluding physical movement, currently available databases do not account for noise as well as physiological effects of activity found in real life conditions.

2. Stimuli used: As described in Section 2.7.1, a variety of stimuli have been used to modulate various mental states. However, these conditions may not necessarily be a representation of real life conditions. For example, [261] reported higher ratings for mental workload when comparing real life vs simulated flight tasks. As a result, focus has been shifted towards experimental conditions which more closely represent real life conditions, such as TSST and MATB-II assessments.
3. Controlling for confounding factors: Apart from movement, a large number of studies use additional steps to control for other effects. This is usually done by introducing a baseline period of rest and/or deep breathing. Stress studies using HRV additionally control for effects of circadian rhythm by having subjects record data at the same time of day [32]. Additional precautions include ensuring the subjects did not consume any caffeine, alcohol, and/or drugs some hours before the experimental period [34], along with performing no physical activity or having any sleep deprivation [37]. All such confounding variables are impossible to avoid in real life situations and hence models developed for such conditions may not translate well to real life situations.
4. Experiment duration: Most experiments have been conducted for a small duration of time with the stimuli lasting less than 5 minutes. Such durations may not capture the long term effects of stress and may not be long enough to extract relevant features (e.g., long-term HRV). For example, [37] showed that most HRV features go back to their baseline levels from initial changes when a mental workload task is performed for a long duration. Therefore, features showing relevance for small duration stimuli may not remain important when the same stimulus is used for a longer time period.

2.8 Datasets Collected

The limitations described in the previous section prevent the use of methods developed in controlled laboratory conditions to be used reliably in real life conditions. Therefore, there is a need for collection of data which reflects ambulatory and real life situations. The following databases were collected as part of this doctoral thesis and used for the research described herein.

2.8.1 WAUC dataset

The Workload Assessment Under physical aCtivity (WAUC) dataset [199] collected multi-modal data from 48 participants (23 female, 27.4 ± 6.6 years old). Screening was performed in order to prevent any potential risk to the participants during the experiments. Candidates with cardiovascular diseases, neurological disorders, history of feeling dizzy or fainting were excluded from the experiment. The participants were asked to wear comfortable sportswear. Twenty-two participants utilized a treadmill during the experiment and 26 a stationary bike. Participants consented to participating in the experiment and were remunerated for their time. The experimental protocol was approved by the Ethics Review Boards of INRS, Université Laval and the PERFORM Centre (Concordia University), the latter being the location in which data was collected.

Before starting the data collection, a tutorial explaining the experimental procedure and task to be executed was shown to the participants. Next, various sensors were placed on the subject. These included a portable 8-channel wireless EEG headset (500 Hz, Enobio, Neuroelectrics), a portable chest strap (Bioharness 3, Zephyr) for ECG (250 Hz), respiration (18 Hz), and accelerometer (18 Hz) signal recording, and an E4 (Empatica) wristwatch which measures PPG (64 Hz), skin temperature (4 Hz), galvanic skin response (4 Hz) and acceleration. Signal acquisition was done using the open-source MuSAE Lab EEG Server (MuLES) [112], which was also used to send triggers marking the beginning and end of trials.

Following this step, in the case of participants using the treadmill, a safety harness was placed on the participant's chest to avoid falls. For participants using the stationary bike, they were asked to adjust the seat according to their preference. The height of the screen was then adjusted to provide a more comfortable set-up. Three levels of physical activity were considered: no movement, medium (treadmill: 3km/h, bike: 50 rpm), and high (treadmill: 5km/h, bike: 70 rpm). Figure 2.7 shows the experimental setup for both the bike and treadmill conditions.

In order to elicit high and low mental workload levels, the MATB-II was used. It is a computer-based task designed to evaluate operator performance and mental workload. Figure 2.8 depicts the MATB-II interface seen by the participant [109]. As mentioned before, MATB-II encompasses four tasks, all grouped under a single-window interface. These four tasks include system monitoring (top-left), tracking (top-center), communications (bottom-left) and resource management (bottom-center). In this study, the communication task was not used. The interface also includes

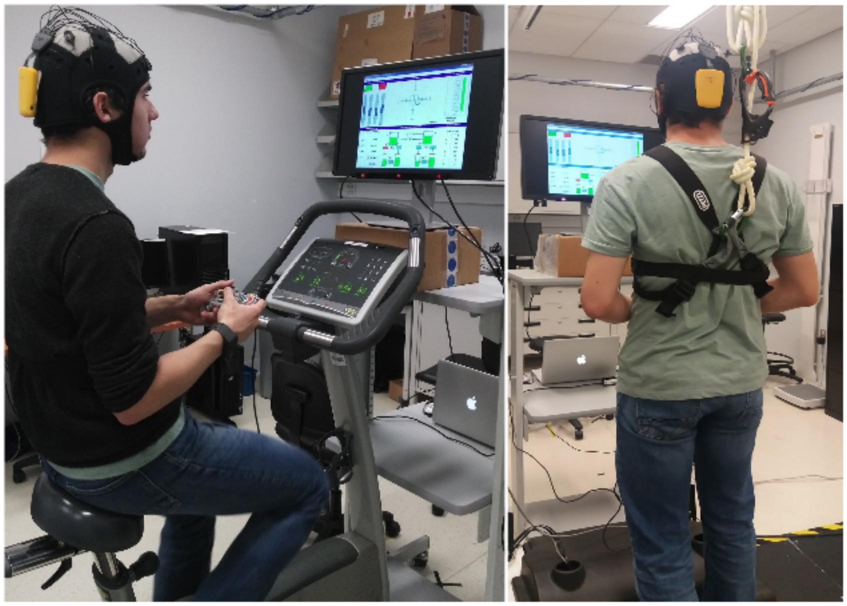
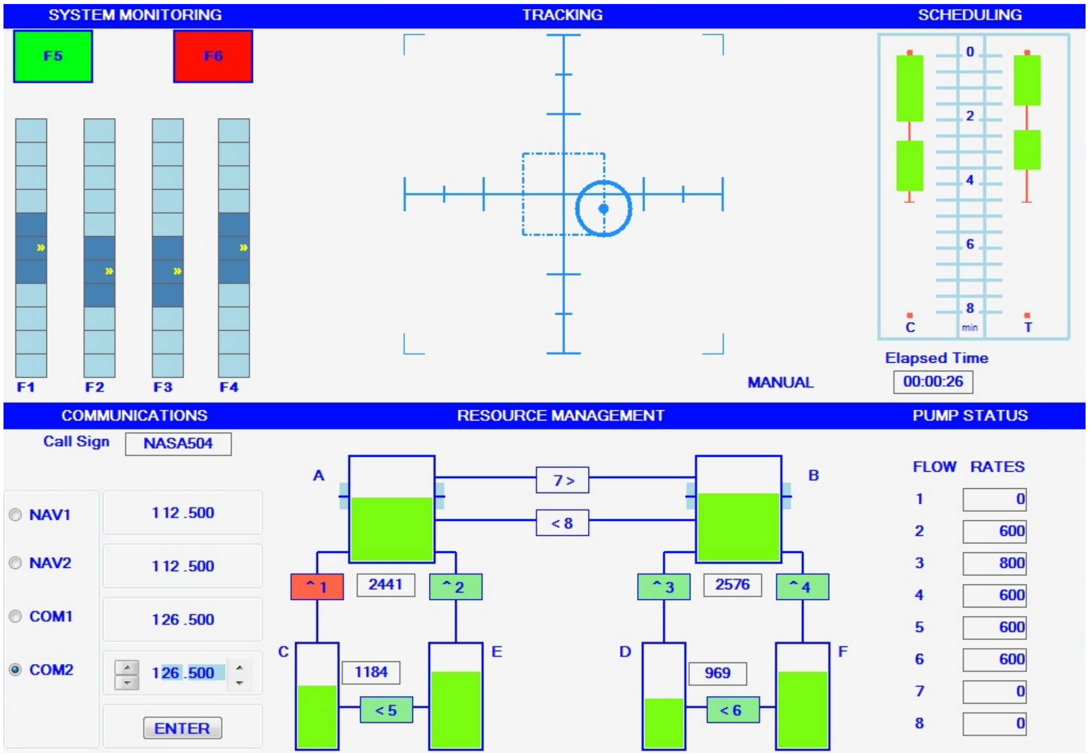


Figure 2.7 – Experimental Setup for both bike and treadmill conditions.

a scheduler (top-right), which only showed the time remaining in each trial, as well as the pump status (bottom-right), which complemented information on the resource management task (see below). Since participants were simultaneously doing a physical task, a mouse could not be used to interact with MATB-II. Instead, participants were instructed to use an xBox One controller.

The system monitoring task requires the participant to monitor four sliders and report deviations from their normal state. The two warning lights (see F5 and F6 on Fig 2.8) were not used in this study. In their normal states, sliders were oscillating around the middle position. In their deviation state, sliders started oscillating around the top or the bottom of the sliders. Participants had to use the directional pad of the controller to report deviations. The tracking task requires the participant to keep a target (the circle) within a box (the square). As the trials went on, the target started to move randomly. Participants had to use the joystick of their controller to brink it back near the center of the square. The resource management task requires the participant to balance a network of fuel tanks. Participants were instructed to keep the level of tanks A and B as close as possible to 2500 units (this level is indicated by ticks on tanks A and B, see Fig 2.8). Fuel gradually depleted from tanks A and B. To keep the tanks at level, participants could use eight pumps (labeled 1 to 8, between the tanks) to move fuel between tanks. To activate pumps, participants had to use the other joystick of the controller to move the cursor and "click" on the pumps. Pumps were configured to fail from time to time. When a pump failed, it turned red and became unusable.



Elapsed Time 00:00:26

Call Sign NASA504

○ NAV1 112 .500
○ NAV2 112 .500
○ COM1 126 .500
● COM2 126 .500

FLOW	RATES
1	0
2	600
3	800
4	600
5	600
6	600
7	0
8	0

Figure 2.8 – MATB-II game for eliciting different levels of mental workload

Pumps were "repaired" automatically after a while and the participant could resume using it if needed. Two levels of workload were used for the mental task (i.e., low and high). Compared to the low workload condition, the high workload condition had more frequent sliders deviations (for system monitoring), faster random oscillations (for the tracking task) and more frequent pump failures (for resource management monitoring).

In total, six combinations of combined mental workload and physical activity were tested (2 levels of mental workload × 3 levels of physical activity). The experiment was then split into six sessions, each one corresponding to one of the six combinations previously described. The order in which the combinations were done was counterbalanced to avoid ordering effects. Before each session, two baseline sessions were performed. During the first session, there were neither physical or mental activity and the participants were asked to close their eyes and relax for 60 seconds. Then the subject was asked to open their eyes and start moving according to the corresponding session physical activity until reaching the desired level. After reaching a stable activity level, the second baseline was collected while the participant kept the pace for 2 minutes without any mental effort involved. This baseline has been added to ensure the stationarity of the heart rate dynamics for a given activity level prior to introducing the mental workload condition. Lastly, the experimenter gave the joystick to the participant who then started the first experimental session for a duration of 10 minutes. After each session, a 5-minute break was given. After each of the experimental sessions, participants were asked to fill the NASA-TLX questionnaire and report their perceived fatigue level based on the Borg scale. Overall, the experimental protocol lasted roughly two hours.

2.8.2 PASS dataset

The Physical Activity and StresS (PASS) database [113] includes multimodal physiological signals from 48 participants while they performed stress inducing tasks in ambulatory conditions. Participants consented to participate in the study, which received ethical board approval from the affiliated institutions. The physical activity was modulated at three different levels using a stationary bike. More specifically, in the 'no physical activity' state, participants sat on the bike without pedaling, whereas in the 'medium' and 'high physical activity' states they were told to maintain speeds of 50 and 70 rpm, respectively. Stress levels, in turn, were modulated by two different videogames.

For low stress condition, the subjects played the game Timeframe, a first-person game centred on exploration and collection of artefacts in an abandoned city. It is designed to be a peaceful, relaxing experience with a bright environment and calming music. The players cannot die in the game. Additionally, participants were told that the number of artefacts collected would not matter, further reducing any stress-inducing factors. Play sessions were divided in three segments. There were no differences in these segments, except that participants were instructed to not seek the same artefacts as before to avoid repeating the same gameplay.

For the high stress condition, on the other hand, subjects played the game Outlast, a first-person horror game where players have to survive while navigating an eerie asylum and evade being attacked by its scary inmates. Players in this game can have the options of avoiding, escaping and hiding from the attackers with no option for fighting back. This made individual skills less relevant for the experiment. It is also deterministic and features a linear story, increasing the similitude of experience between participants. To further increase the stressful state, the game was played with dimmed lights. Similar to Timeframe, the play sessions were divided into three predetermined segments selected from the game. Both games were played with a Xbox One controller. Figure 2.9 shows the experimental setup used.

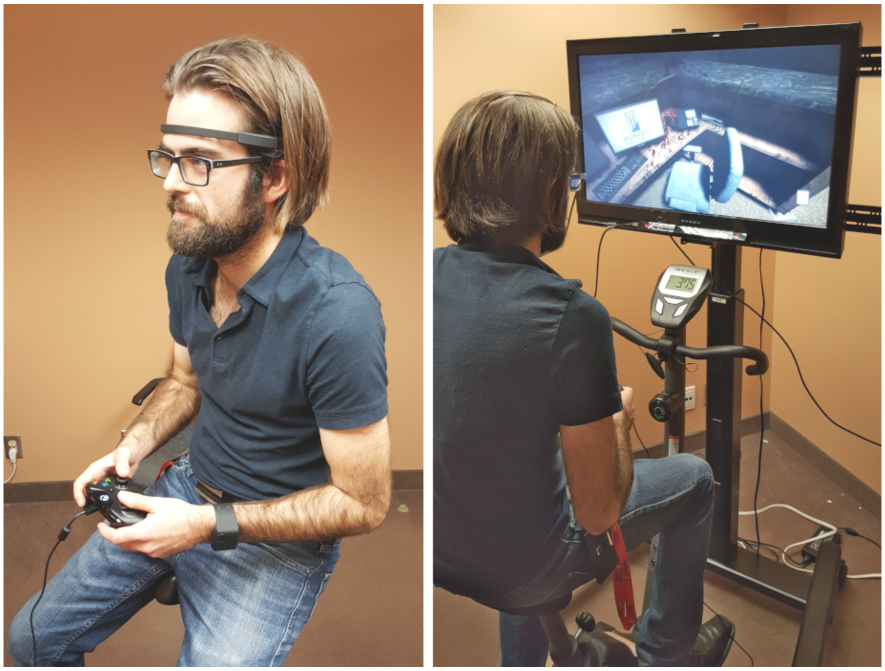


Figure 2.9 – Experimental setup from the front (left). Experimental setup from the back (right). BioHarness 3 not shown since worn under the shirt.

All participants completed six conditions (2 × 3) including the combination of two stress levels (low/high) and three physical activity levels (no, medium, and high) in 10-minute counterbalanced sessions with 5 minute breaks. All conditions for the same game were performed in sequence. This was done to avoid constant swing between calm and stressful psychological states. At the end of each session, subjects rated their stress level using a 21-point scale.

Physiological data was collected using three off-the-shelf devices. ECG (250 Hz) and breathing (18 Hz) were collected using a Bioharness 3 (BH3, Zephyr) chest strap. The Empatica E4 (Empatica) wristband was used to collect skin temperature (4 Hz), galvanic skin response (4 Hz), and PPG (64 Hz) signals. Finally, a Muse headband (Muse) was used to collect 8-channel electroencephalography (EEG, 220 Hz) data. Signal acquisition was done using the open-source MuSAE Lab EEG Server (MuLES) [112], which was also used to send triggers marking the beginning and end of trials.

Both the WAUC and PASS databases evoke different mental states in ambulatory conditions. Therefore, these databases provide opportunities to study the effect of noise and physiological confounders due to different levels of physical activity. Additionally, as mentioned previously, mental workload and stress studies usually compare these mental states against rest conditions while these

Table 2.1 – Session description of the shooting range exercise

Session number	Description
1	Firearm handling & blank cartridge
2	Basic stance
3	While covered & kneeling stance
4	Displacement, rapid, flashlight, decision making
5	Recap

databases try to compare different levels of mental workload and stress. Additionally, circadian rhythm based variability has not been controlled with the experimental session either happening in the morning or afternoon hours for the different subjects based on availability.

2.8.3 ENPQ dataset

For this dataset, data was collected from 27 (6 females) police trainees taking a 15-week course at the Quebec National Police Academy (École Nationale de Police du Québec, ENPQ, Nicolet, Canada). The course included training and evaluation of various police-related skills (e.g., police car driving, hand-to-hand combat, arrest scenarios, criminal investigation, firearm usage). Participants were recruited at the beginning of the semester and the experimental protocol was approved by the Ethics Review Boards of the Institut National de la Recherche Scientifique (INRS), Université Laval, and ENPQ.

Data was collected in three waves. The first wave consisted of a longitudinal data collection in which participants were asked to wear a heart rate wrist monitor (Fitbit) at all possible times throughout the duration of the 15 weeks of the course. This was used to gauge a baseline HRV for each participant. The second wave, in turn, consisted of a shooting range exercise where participants were trained on the handling of firearms and related skills. In total five sessions (3 hour each) were held, each detailing a separate aspect of the task at hand. More details about the five sessions can be seen in Table 2.1. Participants in the second wave were also asked to use a wearable device (BioHarness 3, Zephyr) that collected ECG (250 Hz) and breathing curves (18 Hz), which were all time aligned and streamed using the MuLES data acquisition software [112]. The exercise proceeded as it normally would, with the exception of an additional questionnaire at the end of each session in which the trainees reported their stress and mental workload ratings using a French version of the NASA-TLX [9] and fatigue using the Borg scale [21].

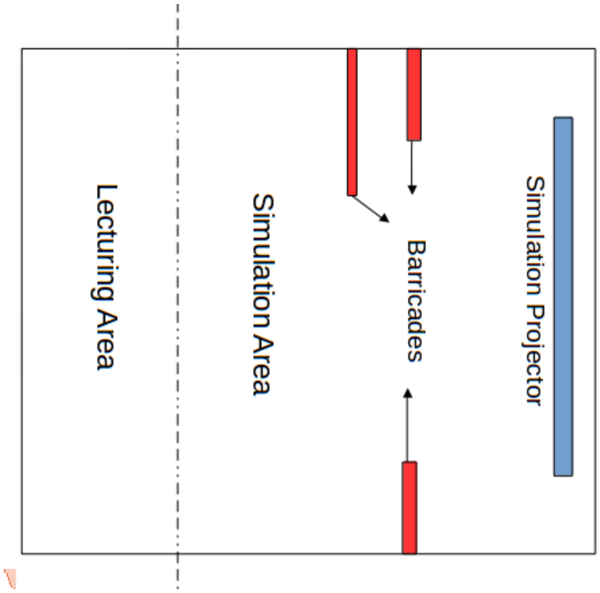


Figure 2.10 – Setup for the intervention simulation area at ENPQ

The third and final data collection wave was performed during the intervention simulator exercise. This exercise extends the shooting range one by focusing on aspects of decision making and intervention when a firearm is involved. These exercises follow standard classroom format along with hands-on trainee participation in a simulation environment involving teams of 1-4 people. The simulation was carried out using the setup depicted by Fig. 2.10. The simulation scenario was displayed via a video projector (shown in blue in the figure) and the trainees were free to interact with characters on the screen (e.g., suspects, victims, witnesses) and to move around in the simulation area as they wish, as well as to take cover behind different barricades (shown in red). To make the exercise more realistic, low velocity automated guns with non-harmful bullets were used during the simulation. Sensors embedded into the simulation area were used to record the timing and location of the shots taken by each trainee. During the simulation, the instructor could manipulate suspect behavior ranging from cooperative to hostile, thus directly modulating and/or responding to trainee actions and decisions. Each simulation lasted for several minutes. Similar to the shooting range exercise, trainees wore the Bioharness device, thus ECG and breathing curves were collected.

This database collects long term physiological data (approx 3 hours per session) in a police course which consists of various social (coordination between subjects, discussions with instructor), physical (simulation and handling exercises) and mental aspects (difficultly levels of various sessions is different). This database allows for testing of the developed methods for long duration data which may have several unknown confounding factors.

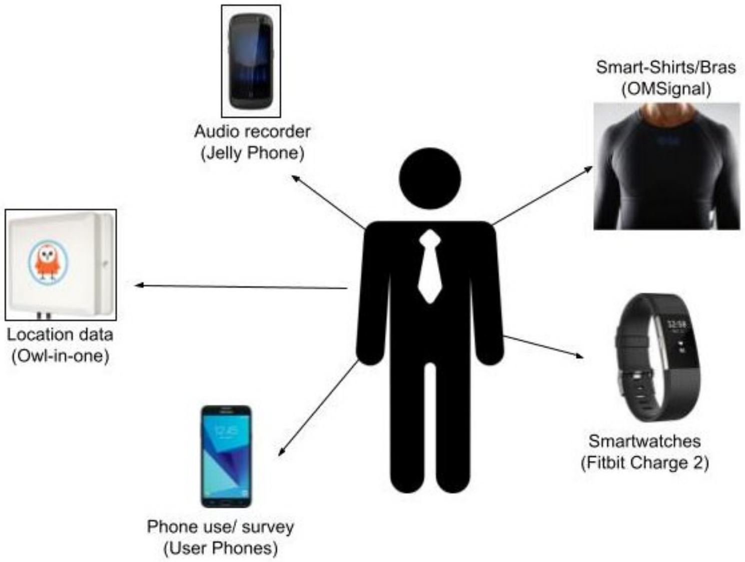


Figure 2.11 – Sensors used in the TILES study

2.8.4 TILES dataset

The TILES (Tracking IndividuaL performancE with Sensors,) dataset was collected from 200 participants (66 male, 134 female; age 38.6 ± 9.8 years) from a pool of employees (nurses and staff) of University of Southern California's Keck Hospital. Two-thirds of the participants were nurses while one-third were hospital staff. Data was collected for a continuous duration of 10 weeks. Participants consented to participate in the study, which received ethical board approval from the affiliated institutions. Complete details about this publicly available dataset can be found in [114].

Participants carried out their work day as usual but were asked to fill out a brief phone-based daily survey that included information on levels of anxiety and stress on a 5-point scale. Participants were outfitted with multiple wearable sensors to collect a variety of biometric data, including audio features, heart rate, respiratory rate, and sleep quality. More specifically, a custom audiometric badge, a Fitbit Charge 2, and an OMsignal smartshirt were used. Additionally, mobile phone use and location data were also collected. The different sensors used as shown in Fig 2.11

Regarding the physiological signals, the OMsignal shirt collects the RR interval (via internal QRS-peak peak detection algorithm of the shirt) and breathing peak amplitude information. Additionally, the Fitbit Charge 2 collects heart rate information sampled every 5 seconds. A total of four RR intervals can be recorded within each second for the OMSignal shirt. The *RRPeakCoverage*

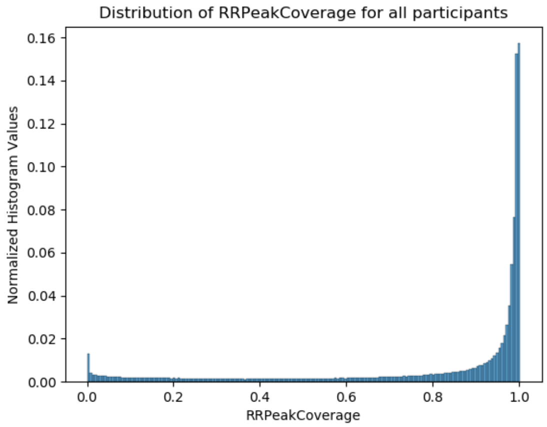


Figure 2.12 – Distribution of RRPeakCoverage

Table 2.2 – Summary of the datasets explored in the thesis

Dataset	No. Participants	Mental State	Duration/session	Stimulous	Signals Explored
DEAP	32	Valence/Arousal	1 min	Affective clips	EEG
WAUC	48	Mental Workload	10 min	MATB-II	EEG, ECG, Respiration, PPG, GSR, Skin Temparature
PASS	48	Stress	10 min	Video games	ECG
ENPQ	27	Stress/Mental Workload	3 h	Arrest/Shooting exercises	ECG, Breathing
TILES	200	Stress/Anxiety	8-12 h	Daily shift work	ECG

provides an estimate of the percentage of correct RR values recorded for each five minute interval. According to the manufacturer, *RRPeakCoverage* value of > 0.8 represents a good quality of the RR values recorded. The distribution of the quality metric is shown in Fig. 2.12. Almost 27% of the segments were below good *RRPeakCoverage* quality levels, hence showing the impact of noise on in-the-wild data collection settings.

The TILES data collection has been done in completely in-the-wild conditions with different types of hospital staff (nurses, lab technicians) over the duration of multiple full-day work shifts (approximately 8 hours each shift). As a result, this dataset allows for testing of features which could capture the long terms effects of mental states. Similar to other datasets, no baseline/rest data exists and the analysis is done between low/high states.

The summary of the different databases used and the signals explored in this thesis is shown in Table 2.2

2.9 Conclusion

This chapter presented the background of mental state monitoring using physiological signals. First, the various physiological modalities were discussed. Their acquisition methods, properties, pre-processing and benchmark features with their relation to mental states were presented. This was followed by a short description of different types of multimodal systems and their advantages. Next, the different components of the machine learning pipeline were discussed, ranging from feature selection and evaluation to performance metrics used. Following this, the available datasets were discussed along with their limitations. Finally, we described some of the new datasets collected and used for this thesis and their advantages. Overall, this chapter discusses the background of mental state monitoring literature while expanding upon the various tools required for moving research from in-the-lab to in-the-wild.

Chapter 3

Motif based analysis of EEG Signals

3.1 Preamble

This chapter is compiled from material extracted from the manuscript published in *Computational Intelligence and Neuroscience* [192].

3.2 Introduction

As described in Chapter 2, EEG signals are very sensitive to artefacts, such as eye blinks and muscle movement [116]. To overcome such issues, artefact removal algorithms can be used. Alternately, new noise-robust features can be developed and/or multimodal fusion strategies can be explored [79]. In this chapter, focus is placed on the latter and motif based features are proposed and tested alone or alongside alternate complementary features for affect recognition. Motif based analysis, as discussed in the previous chapter, has in the past been used to recognize sleep states [237], as well as the effects of anesthesia [235], to detect seizures [238, 117], and to measure alertness [69]. Motif-based methods are inherently robust to noise as they deal with the shape of the time series and are unaffected by the magnitude [262, 263, 117]. While PE is the most popular motif based metric, other approaches have also been developed. These include calculation of a similarity metric between two time series based on motif distributions [117], as well as a synchronization between two time series [118].

To the best of our knowledge, motif based metrics and their noise robust properties have yet to be explored for affective state monitoring, thus this chapter fills this gap. In particular, we compare the proposed features with spectral power and spectral asymmetry benchmark features. Notwithstanding, one main limitation of motif features concerns the loss of both amplitude and rate-of-change information when time series are converted into motif series [262, 264]. As such, we also explore three different fusion strategies to combine information from the proposed motif features and classical benchmark features. Experimental tests on a publicly available database [13] are performed and show the advantages of the proposed features over benchmark ones, as well as the benefits of fusion for affective state monitoring.

The remainder of this chapter is organized as follows. Section 3.3 introduces the motif based features. Next, Section 3.4 describes the experimental setup including the benchmark features, fusion methods tested, classification strategy and performance metrics used. Section 3.5 then presents and discusses the results obtained and conclusions are drawn in Section 3.6.

3.3 Motif Based Features

PE is the most commonly used motif based features and has been extracted here. Additionally, other features based on the statistics of recurring patterns within the motif series can be extracted. The features proposed herein are detailed in the subsections below and only consider motifs of degree $n = 3$ and lag value $\lambda = 1$. These parameters have been suggested in the past for related tasks [69, 118].

3.3.1 Ordinal distance dissimilarity

Ordinal distance based dissimilarity [117] is a metric with close parallel to the benchmark asymmetry index and measures the dissimilarity between two motif series for different electrode pairs using:

$$D_m(X,Y) = \sqrt{\frac{n!}{n!-1}} \sqrt{\sum_{i}^{n!} (p_x(i) - p_y(i))^2}, \tag{3.1}$$

where $p_x(i)$ and $p_y(i)$ are the relative frequencies of the motif pattern represented by i in electrode X and Y, respectively, and n is the degree of the motif. The ordinal dissimilarity is calculated on the electrode pairs described in Section 2.2.4.

3.3.2 Motif synchronization

Functional connectivity gives insight into the dynamic neural interaction of the different regions of the brain. Recently, motif synchronization has been proposed as a functional connectivity analysis tool [118] and measures the simultaneous appearance of motifs in two time series. For two motif series X_m and Y_m, $c(X_m;Y_m)$ is defined as the highest number of times in which the same motif can appear in Y_m shortly after it appeared in X_m for different delay times, i.e.:

$$c(X_m;Y_m) = c_{XY} = max(\sum_{i=1}^{l_m} J_i^{\tau_0}, \sum_{i=1}^{l_m} J_i^{\tau_1}, ..., \sum_{i=1}^{l_m} J_i^{\tau_n}) \tag{3.2}$$

with

$$J_i^{\tau}(i) = \begin{cases} 1 & \text{if } X_{M_i} = Y_{M_{i+\tau}} \\ 0 & \text{else} \end{cases}$$

The time delay τ ranges from $\tau_0 = 0$ to τ_n, where τ_n is the maximum value to be considered, and l_m is the size of the time varying window within the time series. Similarly, the opposite measure c_{YX} can be obtained by changing only the order of the time series to $Y_{M_i} = X_{M_{i+\tau}}$. Finally, the degree of synchronization Q_{XY} and the synchronization direction q_{XY} is given by:

$$Q_{xy} = \frac{max(c_{XY}, c_{YX})}{l_m}, \tag{3.3}$$

and

$$q_{XY} = \begin{cases} 0 & \text{if } c_{XY} = c_{YX} \\ sign(c_{XY} - c_{YX}) & \text{else.} \end{cases}$$

The degree of synchronization, Q_{XY}, is scaled between 0 and 1, with 0 representing no interaction and 1 suggesting very high interactions. Feature q_{XY}, in turn, gives the direction of information flow, with 0 indicating no preferred direction, 1 indicating direction from X to Y, and -1 indication direction from Y to X. For our calculation, τ_n has been chosen as 5 and the window size l_m is chosen as 256.

3.3.2.1 Graph features

The different functional connections obtained by motif synchronization analysis can be further extended by means of graph theoric analysis, where each electrode on the scalp represents a node on the brain network. Weighted graphs have weights that represent the level of interaction between the two nodes. Edges with smaller weights are believed to represent noisy/spurious connections [119], thus a thresholding is done to obtain an unweighted graph. Previously, graph theoretic features have been explored for affect recognition based on EEG spectral coherence measures [48]. Graph theoretic analysis based on motifs, however, has yet to be explored, thus both weighted and unweighted graphs (thresholded to the average value of the graph weighted) are tested herein. An advantage of motif synchronization over more popular connectivity approaches is the ability it provides to measure direction of information flow for the different nodes in the brain network. From the weighted and unweighted graphs, several features are extracted, namely: i) Degree of connectivity (k): Defined as k_i where i is a given node. For the unweighted network it is calculated as:

$$k_i = \sum_{j \epsilon N_e} a_{ij}, \tag{3.4}$$

where N_e represents the nodes in the network and $a_{ij}, i \neq j$, represents the value of the unweighted adjacency matrix. For the weighted network the formula is:

$$k_i^w = \sum_{j \epsilon N_e} w_{ij}, \tag{3.5}$$

where instead of a_{ij}, the weights w_{ij} assigned to each edge are used. The average degree of connectivity for the whole network is used as a feature in our analysis.

ii) Clustering coefficient (C): The mean clustering coefficient for an unweighted network is given by:

$$C = \frac{1}{N_e} \sum_{i}^{N_e} \frac{e_i}{k_i(k_i - 1)}, \tag{3.6}$$

where e_i is the number of existing edges between the neighbors of i and k_i is the degree of connectivity for the unweighted network. For a weighted network the clustering coefficient value is given by:

$$C^w = \frac{1}{N_e} \sum_{i}^{N_e} \frac{t_i^w}{k_i^w(k_i^w - 1)}, \tag{3.7}$$

where t_i is calculated as:

$$t_i = \frac{1}{2} \sum_{j,h \epsilon N_e} (w_{ij} w_{jh} w_{hi})^{\frac{1}{3}}, \tag{3.8}$$

and represents the geometric mean of the triangles constructed from the edges around a particular node i, and k_i^w represents the weighted degree of connectivity.

iii) Transitivity (Tr): Transitivity is defined as the ratio of "triangles to triplets" in a network and is defined as:

$$Tr = \frac{3\lambda}{k(k-1)}, \tag{3.9}$$

where λ represents the number of triangles in network, while k is the average degree of connectivity (weighted or unweighted) of a network. Transitivity is a global measure of clustering coefficient and is equal to it when the degree of connectivity of all nodes is equal to one another.

iv) Characteristic path length (L): For an unweighted network it is given by:

$$L = \frac{1}{N_e(N_e - 1)} \sum_{j=1 i \neq j}^{N_e} d_{ij}, \tag{3.10}$$

with d_{ij} being the minimum amount of edges required to connect node i and j and is replaced by the shortest weighted path length d_{ij}^w for the weighted characteristic path length L^w.

v) Global efficiency (G): This is calculated using the inverse of the shortest weighted or unweighted path for the network, i.e.:

$$G = \frac{1}{N_e(N_e - 1)} \sum_{j=1 i \neq j}^{N_e} d_{ij}^{-1}, \tag{3.11}$$

where $d_i j$ is replaced by the shortest weighted path length d_{ij}^w for weighted global efficiency measure.

vi) Small world features: The work in [265] has shown that human brain networks exhibit small world characteristics. A small world network is characterized by a high clustering coefficient and a small average path length from one node to another [266]. Here, three small world features are computed, namely a) the small world characteristics length:

$$L_s = \frac{L}{L_{rand}}, \tag{3.12}$$

b) the small world clustering coefficient:

$$C_s = \frac{C}{C_{rand}}, \tag{3.13}$$

and c) the small-worldness of a network [267]:

$$S = \frac{C_s}{L_s}, \tag{3.14}$$

where C_{rand} and L_{rand} are the corresponding clustering coefficient and characteristic path length values for a random network, respectively.

vii) Direction of flow (DoF): As motif synchronization also provides the direction of information flow in the brain network graph, a simple feature is explored here to represent the overall response of the brain network as either receiving or transmitting information, on average. DoF is defined as:

$$DoF = \sum_{ij} q_{ij}, \tag{3.15}$$

where q_{ij} is defined as the direction of information flow with 1 representing information flowing from i to j, -1 representing information flow in the opposite direction, and 0 being no preferred information flow direction.

3.4 Experimental Setup

3.4.1 Benchmark Features

As mentioned previously, spectral power features in different EEG subbands have been widely used for affective state monitoring, including for the DEAP database [13, 78]. Moreover, an inter-hemispheric asymmetry in spectral power has also been reported in the emotion state monitoring literature [28, 46], particularly in frontal brain regions [268]. EEG signals are band decomposed into different bands. For this analysis the theta, alpha, beta, and gamma bands were used. Following this, 48 asymmetry index (AI) features (12 inter-hemispheric electrode pairs x 4 subbands) were computed. In order to compare against the ordinal dissimilarity features, the same electrode pairs as used for features in Section 3.3.

Moreover, EEG subband ratios have also been explored for human mental state monitoring [10] and are included as benchmark features as well. The ratios computed include: $\frac{\gamma}{\beta}, \frac{\beta}{\theta}, \frac{\alpha}{\theta}, \frac{(\alpha+\beta)}{\gamma}$, and $\frac{(\gamma+\beta)}{\theta}$.

Lastly, the Shannon entropy [120] has been used as a feature to measure the complexity of the EEG time series. Shannon entropy can be calculated as follows:

$$SE = -\sum_{j} P_j \cdot \log(P_j), \tag{3.16}$$

where P_j is the power in sub-band j.

Table 3.1 provides a summary of the number of features extracted for each feature group and sub-group.

3.4.2 Feature selection

Previous work has shown that motif features convey complementary information to other amplitude and rate-of-change based features [262, 264]. As such, we explore the effects of combining the proposed motif based features with the benchmark ones on the DEAP database described in Chapter 2. Given the small dataset size, however, it is important to avoid issues with curse of

Table 3.1 – Summary and grouping of features extracted

Feature name	# features	Group
(Weighted) graph features	20	Motif based features
(Unweighted) graph features	20	
Direction of flow	4	
Small world features	12	
Permutation entropy	4	
Ordinal distance dissimilarity	48	
Spectral subband power ratio	5	Benchmark spectrum based features
Shannon entropy	1	
Spectral power	4	
Asymmetry index	48	

dimentionality and overfitting, thus feature selection is required. Here we explored three different feature selection approaches: ANOVA, RFE, mRMR.

For the experiments herein, 90% of the data (across all subjects) is set aside for feature selection and classifier training and the remaining 10% is left aside for testing. The split was performed with a random seed of 0 using the scikit-learn function in Python. The best feature selection algorithm and its corresponding optimal number of features is then selected by grid search. Classifier training and different fusion schemes are described next.

3.4.3 Classification

Here, SVM classifiers are trained on two different binary classification problems, namely, discriminating between low and high valence states, as well as low and high arousal states. An RBF kernel was used and implemented with the Scikit-learn library in Python [121]. As we are interested in exploring the benefits of the proposed motif features and compare them against benchmark features, we do not perform classifer hyper-parameter optimization and use default parameters instead, namely, $\lambda_{SVM} = 1$ and $\gamma_{RBF} = 0.01$.

Moreover, as the DEAP database relies on 9-point scale ratings, it has typically been the case where the mid-point is considered as a threshold, where ratings greater than the threshold are considered “high,” and those below are considered “low”. As was recently emphasized in [197], however, subjects have their own internal biases, thus leading to varying scales for grading and, consequently, different thresholds per participant. For example, as reported in [197], by using a mid-

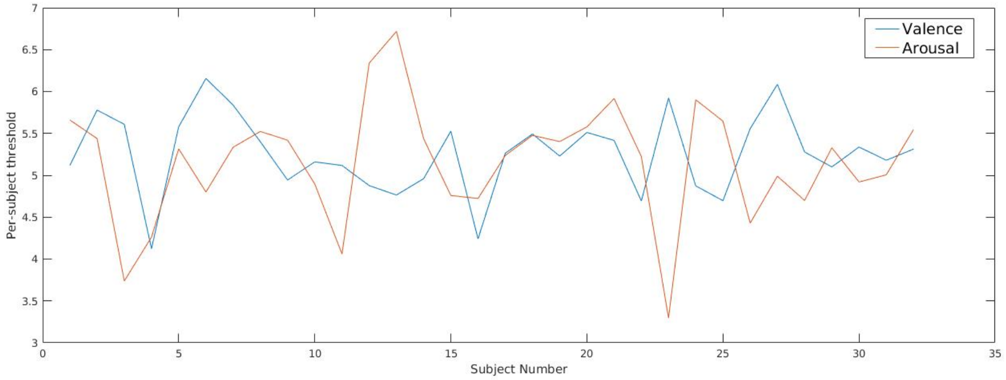


Figure 3.1 – Threshold variation for each subject for valence (blue) and arousal (orange) dimensions.

point threshold value of 5, a 60/40 ratio of high/low levels was obtained across all participants. In turn, if an individualized threshold was used corresponding to the value in which an almost balanced high/low ratio was achieved per participant, improved results were achieved [197]. Figure 3.1, for example, depicts the threshold found for each participant for arousal and valence in this latter scenario. As can be seen, on average a threshold of 5 was most often selected, though in a few cases, much higher or much lower values were found, thus exemplifying the need for the individualized approach used herein.

3.4.4 Fusion Strategies

As described in the previous chapter, fusion of different modalities has been commonly done to improve performance. However, those fusion strategies can also be used for different feature sets of the same modality. Here, we explore three different types of fusion strategies to combine motif and benchmark spectrum based features, namely: feature fusion, score level fusion, and output associative fusion.

3.4.4.1 Feature fusion

For this method the motif and benchmark feature vectors are combined prior to feature selection and corresponds to feature level fusion.

3.4.4.2 Score level fusion

Score level fusion is a decision level fusion method. The weighted decision fusion method proposed in [122] has been used. According to this technique, the fusion classification probability p_0^x for $x \epsilon [0, 1]$ for each class $x \epsilon 1, 2$ can be denoted by:

$$p_0^x = \sum_{i=1}^{N} \alpha_i p_i^x t_i, \tag{3.17}$$

where i is the index of a particular feature group, N is total number of groups used and α_i are the weights corresponding to each group $(\sum_i^N \alpha_i = 1)$. The parameter t_i is the training set performance of a particular feature group such that the fusion probabilities for all classes sum up to unity, and is given by:

$$t_i = \frac{F_i}{\sum_i^N \alpha_i F_i}, \tag{3.18}$$

where, F1 is the F1-score obtained on the training set using a particular feature group. The weight space was searched for best performance as this is indicative of the contribution to the outcome made by each of the feature groups.

3.4.4.3 Output associative fusion

Psychological evidence has suggested a strong inter-correlation between the valence and arousal dimensions [123, 124, 125, 126]. As such, the output associative fusion (OAF) method has been used to model the correlations for continuous prediction of valence and arousal scales [269]. The OAF framework has been explored here and is depicted by the block diagram in Fig. 3.2. As can be seen, first individual classifiers make the valence and arousal predictions for each individual feature group. This is then followed by a final prediction step which considers both the valence and arousal dimensions in order to better predict each of the two outputs.

3.4.5 Figure-of-merit

BACC has been used as the performance metric as it takes into account class unbalances. To test the significance of the attained performances against chance, an independent one-sample t-test against a random voting classifier was used ($p \leq 0.05$), as suggested in [13]. In order to

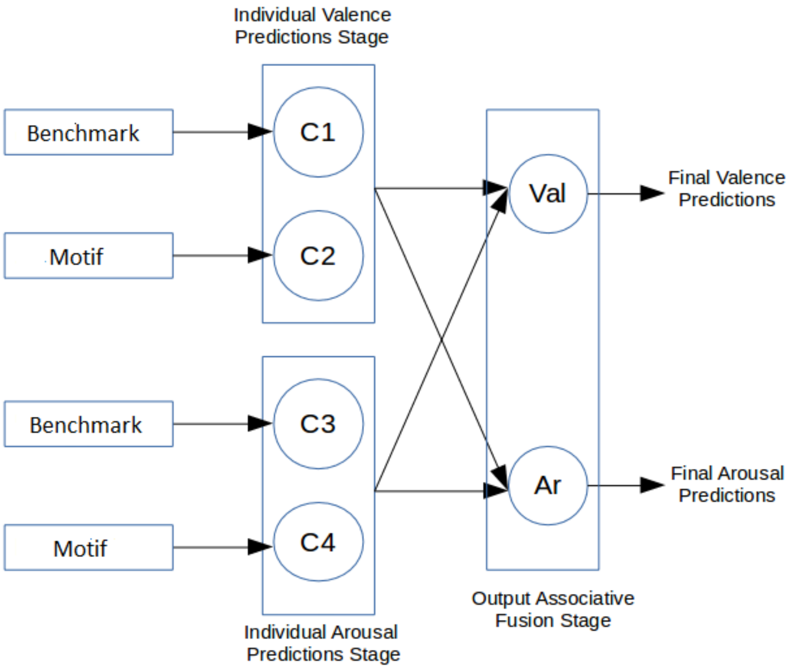

Individual Valence Predictions Stage
Benchmark
Motif
Benchmark
Motif
C1
C2
C3
C4
Val
Ar
Final Valence Predictions
Final Arousal Predictions
Output Associative Fusion Stage
Individual Arousal Predictions Stage

Figure 3.2 – Block diagram of OAF strategy for the two feature groups.

have a more generalized performance of the classifier, once the feature selection step is performed, classifier training and testing is performed 100 times with different train/test partitioning. This setup provides a more generalized performance of the features and their (in)variance to the training set used. The BACC values reported in the table correspond to the mean ± the standard deviation of all BACC values attained on the test set over all of the 100 iterations.

3.5 Experimental Results and Discussion

In this section, we show and discuss the obtained results in terms of impact of feature selection, feature group, and fusion strategy on overall performance.

3.5.1 Feature selection

As mentioned previously, three different feature selection schemes were explored and tested. Feature selection was implemented in the benchmark features alone, proposed motif feature alone, and in the combined benchmark-motif set. The optimal BACC values obtained are shown in Tables 3.2-3.4, respectively, along with the final number of features (nof) used in the models.

Table 3.2 – Comparison of different feature selection algorithms and number of features (nof) for benchmark feature set

Feature Groups	Valence		Arousal	
	BACC	nof	BACC	nof
ANOVA	0.5490	9	0.5316	8
mRMR	0.5404	3	0.5281	4
RFE	0.5531	3	0.5318	15

Table 3.3 – Comparison of different feature selection algorithms and number of features (nof) for motif-based feature set (Motif feature set includes permutation entropy, ordinal distance disimilarity and graph-theoritical features)

Feature Groups	Valence		Arousal	
	BACC	nof	BACC	nof
ANOVA	0.5818	40	0.5362	42
mRMR	0.5757	44	0.5385	20
RFE	0.5872	20	0.5500	16

Table 3.4 – Comparison of different feature selection algorithms and number of features (nof) for combined benchmark-motif feature set (Motif feature set includes permutation entropy, ordinal distance disimilarity and graph-theoritical features)

Feature Groups	Valence		Arousal	
	BACC	nof	BACC	nof
ANOVA	0.5930	40	0.5446	39
mRMR	0.5816	29	0.5645	17
RFE	0.6010	38	0.5598	20

As can be seen, for ANOVA-based feature selection, fewer than 10 features were used in the models for both valance and arousal dimensions with the benchmark features, thus representing roughly one-sixth of the total amount of available features. For the motif group, in turn, roughly 40 were shown to be useful, thus amounting to roughly one-third of the available feature pool. When combining both feature sets, the optimal model also relied on roughly 40 features, thus one quarter of the available feature pool.

The mRMR algorithm, in turn, generally resulted in fewer top features but with similar overall BACC, thus corroborating the results in [247, 248]. For the benchmark feature set, for example, $BACC \approx 0.54$ was achieved with just three features for valence, thus in line with the ≈ 0.55

achieved with ANOVA-selected features. For arousal and motif features, similar BACC was achieved but relying on roughly half the number of features relative to ANOVA-based selection. With the combined feature set, in fact, improved BACC was achieved for the arousal dimension but with fewer than half the number of features chosen by ANOVA.

Lastly, RFE selection typically resulted in the highest accuracy with the best $BACC$ vs. nof tradeoff. This is expected as RFE considers the interaction of features among themselves and the final outcome. Overall, the best accuracy was achieved with the combined set, followed closely by the models trained on the proposed motif features. These findings corroborate the complementarity of the two different feature types and show the importance of motif features for affective state recognition.

A one-way ANOVA was computed between the different pair of feature selection algorithms (ANOVA vs. mRMR, ANOVA vs, RFE, RFE vs. mRMR) for the benchmark, motif and combined feature sets to assess the algorithm performance. For the benchmark feature set, in the arousal dimension, the three algorithms perform similarly with no statistical differences observable. However, for the valence dimension, the RFE performs significantly better than the mRMR algorithm ($p_{val} < 0.05$). There were no significant differences observed between RFE and ANOVA performances, but the RFE obtained similar performance with fewer features. For the motif feature set, in the arousal dimension, we observe the RFE perform significantly better than both ANOVA ($p_{val} < 0.01$) and mRMR ($p_{val} \approx 0.01$). In the valence dimension, we observe a significant difference in algorithm performance between RFE and mRMR; however the performance of ANOVA is not significant compared to both the algorithms. However, we again observe that RFE gives similar performance to ANOVA with half the number of features, thus being more efficient. Finally, for the combined feature set, in the arousal dimension, both mRMR and RFE perform significantly better than ANOVA ($p_{val} < 0.01$) while there are no differences between mRMR and RFE performances with mRMR reaching equivalent performance with fewer features than RFE. In the valence dimension, we observe ANOVA ($p_{val} \approx 0.05$) and RFE ($p_{val} < 0.01$) perform significantly better than mRMR, while there is no performance difference between ANOVA and RFE in this case. It is interesting to note that the number of features for both ANOVA and RFE is almost the same. In general, we find the RFE gives significant or equal performance compared to ANOVA and mRMR with fewer number of features. For feature fusion, the algorithm giving the highest average performance has been considered the algorithm of choice.

Table 3.5 – Top-20 features used in the best valence models for the different feature groups

Benchmark (nof=3, FS= RFE)	Motif (nof=20, FS= RFE)	Combined (nof=38, FS= RFE)
$\frac{\gamma}{\beta}$	$Tr\ (\alpha)$	$C\ (\alpha)$
$\frac{\beta}{\theta}$	$PE\ (\gamma)$	$\frac{\gamma}{\beta}$
Spectral power (α)	$G^w\ (\theta)$	$\frac{(\gamma+\beta)}{\theta}$
	$k^w\ (\theta)$	$\frac{\alpha}{\theta}$
	$C^w\ (\theta)$	$G^w\ (\theta)$
	$C\ (\theta)$	$PE\ (\gamma)$
	$PE\ (\beta)$	$D_m(P3, P4)\ (\beta)$
	$S\ (\beta)$	$D_m(O1, O2)\ (\theta)$
	$S\ (\gamma)$	$k^w\ (\theta)$
	$L_s\ (\gamma)$	$Tr^w\ (\theta)$
	$D_m(T7, T8)\ (\gamma)$	$C\ (\theta)$
	$D_m(Fc5, Fc6)\ (\beta)$	Spectral Power (α)
	$DoF\ (\theta)$	$C\ (\beta)$
	$D_m(P3, P4)\ (\beta)$	$k_w\ (\beta)$
	$D_m(O1, O2)\ (\beta)$	$DoF\ (\theta)$
	$D_m(F3, F4)\ (\theta)$	$C^w\ (\theta)$
	$Tr^w\ (\theta)$	$AI(C3, C4)\ (\beta)$
	$L_s\ (\theta)$	$AI(P3, P4)\ (\gamma)$
	$D_m(O1, O2)\ (\theta)$	$D_m(F3, F4)\ (\theta)$
	$D_m(F3, F4)\ (\beta)$	$D_m(T7, T8)\ (\gamma)$

Tables 3.5 and 3.6, in turn, report the top-20 features used in the models that achieved the best $BACC$ for valence and arousal, respectively. As can be seen for valence (Table 3.5), the $\frac{\gamma}{\beta}$ and $\frac{\beta}{\theta}$ power ratios showed to be important, along with alpha band spectral power. This corroborates previous work which has linked $\frac{\gamma}{\beta}$ and $\frac{\beta}{\theta}$ to audio comprehension [270, 271] and, consequently, to perceived valence in low quality text-to-speech systems [272]. For the motif-based features, in turn, small worldness (γ and β band) and weighted graph features (θ band) showed to be important, alongside PE for γ and β bands. Previous studies have indicated to a time-locked theta band synchronization occurring during affective picture processing [273] related to the valence dimension. This synchronization seems to be captured by motif based graph theoretic and ordinal similarity features, as eight of the top 20 features come from the θ band.

Lastly, for the combined feature set, it can be seen that a mix of benchmark and motif features are selected, thus exemplifying the complementarity of the two feature sets. Over the entire $nof = 38$ features used in the model, 11 are benchmark features and 27 motif based. In particular, 17 of

Table 3.6 – Top-20 features used in the best arousal models for the different feature groups

Benchmark (nof=15, FS= RFE)	Motif (nof=16, FS= RFE)	Combined (nof=17, FS= mRMR)
$\frac{(\alpha+\beta)}{\gamma}$	PE (β)	$D_m(O1,O2)$ (θ)
$\frac{(\gamma+\beta)}{\theta}$	Tr (β)	DoF (γ)
AI(O1,O2) (β)	C_s (β)	k^w (θ)
$\frac{\gamma}{\beta}$	$D_m(T7,T8)$ (α)	DoF (α)
AI(P7,P8) (γ)	L^w (α)	$D_m(T7,T8)$ (β)
AI(F3,F4) (β)	$D_m(FC1,FC2)$ (α)	$D_m(P3,P4)$ (β)
$\frac{\beta}{\theta}$	PE (θ)	$D_m(T7,T8)$ (α)
Spectral Power (β)	$D_m(P7,P8)$ (γ)	PE (θ)
AI(Cp5,Cp6) (θ)	L^w (γ)	$D_m(F3,F4)$ (θ)
AI(FC1,FC2) (α)	C^w (β)	$D_m(C3,C4)$ (γ)
AI(P3,P4) (θ)	$D_m(C3,C4)$ (β)	C^w (θ)
AI(P3,P4) (β)	C (α)	DoF (θ)
AI(C3,C4) (θ)	$D_m(Cp1,Cp2)$ (β)	L_s (α)
AI(Cp1,Cp2) (α)	$D_m(P3,P4)$ (α)	$D_m(Fc5,Fc6)$ (θ)
AI(T7,T8) (γ)	$D_m(F7,F8)$ (α)	C^w (α)
	k (α)	$D_m(Fc5,Fc6)$ (α)
		$D_m(Fc5,Fc6)$ (γ)

the top motif features showed importance across the motif and combined sets, as well as all of the top benchmark features across benchmark and combined sets. Additionally, for the combined set, 6 asymmetry features are also in the top selected features, of these 3 are from the same electrode pairs as the top ordinal dissimilarity measures, thus showing the complementary nature of the two feature sets. The power ratios $\frac{\alpha}{\theta}$and $\frac{(\gamma+\beta)}{\theta}$ also appear in the combined feature sets. From the motif feature sets, apart from the overlapping features, additional D_m and clustering coefficient features appear in the combined feature set along with two DoF features from the θ and γ bands.

For arousal (Table 3.6) and benchmark feature set, almost all power ratios showed to be important alongside several asymmetry index features, particularly those in the frontal and parietal regions. Such findings corroborate previous literature showing the relationship between i) arousal and frontal asymmetry [268] in alpha band (e.g., [274]) and other bands (e.g., [275]), ii) an inherent asymmetry in the right parietal-temporal regions, responsible for modulating autonomic and behavioural arousal, and iii) a relationship between arousal and EEG band power ratios [276].

For motif-based features, in turn, roughly half the top features corresponded to ordinal distance dissimilarity measures, thus corroborating the literature on EEG asymmetry and arousal [277, 275].

Moreover, the majority of the top features are from the beta and alpha bands (13 of the top 16), which have been linked to attention based arousal changes [278] and to changes in visual selective attention [279] [280], which is very closely linked to arousal [281].

Interestingly, for the combined sets, none of the top features were from the benchmark feature set, thus suggesting that the proposed motif features conveyed improved arousal information relative to benchmark features. The majority of the features corresponded to ordinal distance dissimilarity across all EEG bands. Moreover, the best achieving model for motif only and combined feature sets were attained using different feature selection algorithms (RFE and mRMR, respectively). Notwithstanding, two features coincided as being important, namely $PE(\theta)$, $D_m(T7,T8)(\alpha)$, and a third showed similar behaviour ($C(\alpha)$ and $C^w(\alpha)$), thus suggesting their importance for arousal prediction. In the combined set, θ showed up in seven of the $nof = 17$, thus also corroborating previous findings [277, 275]. Lastly, most of ordinal dissimilarity features come from frontal, parietal or temporal regions, thus inline with previous research connecting parietal-temporal regions with autonomic and behavioural arousal, as well as frontal regions with arousal [282].

3.5.2 Individual feature groups

So far we have explored the performance achieved with benchmark, motif and combined feature sets. It is interesting, however, to gauge how each individual feature subgroup contributes toward affective state recognition. Table 3.7 reports the balanced accuracy for each individual feature subgroup for the best achieving model found after RFE feature selection.

As can be seen, for valence the weighted and unweighted graph features achieve similar performances, though the model based on the former feature subgroup relies on $nof = 2$, as opposed to $nof = 8$. In fact, all motif-based features achieved similar performance, with small worldness features being the only ones not significantly better than the benchmark (i.e., $p < 0.01$ and indicated by an asterisk in the table). For arousal, in turn, it is observed that graph and small world feature subgroups do not significantly improve over the benchmark, whereas other motif features, such as permutation entropy and ordinal distance dissimilarity, do. Overall, models relying on these two feature subgroups showed to provide the most discriminatory information for valence and arousal models.

Table 3.7 – Performance comparison of different individual feature groups and subgroups

Feature (sub)group	Valence		Arousal	
	BACC	nof	BACC	nof
Weighted graph	0.5662*	2	0.5066	6
Unweighted graph	0.5581*	8	0.5006	6
Small world	0.5533	6	0.5208	2
Other motif	0.5578*	9	0.5632*	12
Spectral power, AI	0.5400	15	0.5344	11
Power ratio	0.5467	3	0.5000	1

Additionally, among the EEG features, we observe that SE, θ and γ spectral power never appear as top-selected features. This could be due to the fact that power and entropy measures are averaged over all electrodes, thus removing any spatial information relevant for the features. Notwithstanding, averaging ensures that the proposed features are invariant and robust to the electrode set considered, as seen with the global graph theoretic features using motif synchronization. For valence, in turn, we observe that none of the AI features show up among the top in the EEG feature set alone scenario. When using only motif features, on the other hand, seven D_m features (out of $nof = 20$) are selected, thus suggesting that motif features may carry more relevant asymmetry signatures for the task at hand. With the combined feature set, it can be seen that proposed features from all groups appear in the top list for both valence and arousal.

3.5.3 Fusion strategies

As mentioned previously, three fusion schemes were explored: feature, score, and output associative fusion. Tables 3.2-3.4 show the effects of feature fusion and the gains attained with the combined set relative to using only a feature group individually. For the valence dimension, for example, gains of 8.6% and 2.4% were achieved with feature fusion relative to using benchmark and motif feature alone, respectively. As shown on Table 3.5, the model based on the combined set relied on features from both feature groups, thus emphasizing their complementarity for valence prediction.

Table 3.8 – Performance comparison of different fusion methods and a random voting classifier with chance levels

Fusion Methods	Valence BACC	Arousal BACC
Feature	0.6010	0.5645
Score	0.5875	0.5807
OAF	0.5873	0.5568
Random voting	0.5018	0.5028

For arousal, feature fusion resulted in more modest gains relative to the benchmark (i.e., 6.1%) and to motif features (2.6%). Interestingly, the best model relied on mRMR selected features which did not include benchmark ones. The second best model, on the other hand, was achieved with RFE feature selection and the top-20 features included seven benchmark ones (i.e., $\frac{(\alpha+\beta)}{\gamma}$, $\frac{\beta}{\theta}$, AI(01,02) (β), AI(Fc1,Fc2) (β), AI(C3,C4) (γ), AI(F3,F4) (β), and AI(Fc1,Fc2) (γ)), three of which overlap with the top features selected from the benchmark alone set. The remaining 13 features were from the motif group, nine of which showed to be top features selected in the motif alone set, namely: $PE(\beta)$, $PE(\theta)$, $L^w(\alpha)$, $L^w(\gamma)$, $C_s(\beta)$, $D_m(P7,P8)(\gamma)$, $D_m(Fc1,Fc2)(\alpha)$, $D_m(T7,T8)(\alpha)$, $D_m(C3,C4)(\beta)$. By comparing the feature sets selected by mRMR and RFE, it seems the former is capable of removing redundancies that may exist between D_m and AI asymmetry features, but favouring the motif ones as they provide maximum relevance. Four features overlap between the two feature selection algorithms, namely: $PE(\theta)$, $D_m(Fc1,Fc2)(\alpha)$, $D_m(T7,T8)(\alpha)$, and $D_m(C3,C4)(\beta)$, thus further suggesting their importance for the task at hand.

For decision fusion, in turn, the weight space was searched in steps of 0.1 and it was found that for valence, the benchmark feature set resulted in a weight of 0.2 (i.e., 0.8 for motif features), whereas a weight of 0.3 was found for arousal (i.e., 0.7 weight for motifs). Such findings highlight the importance of motif features over the benchmark ones for both valence and arousal prediction. The BACC results shown in Table 3.8 show the effect of score-level fusion over feature fusion. As can be seen, gains are attained only for the arousal dimension, thus further suggesting the complementarity of the two feature groups. For comparison purposes, a random voting classifier is also shown for comparison and all attained BACCs are shown to be significantly better than chance ($p \leq 0.01$).

Lastly, the output associative fusion method was outperformed by all other fusion methods, despite showing to be significantly better than chance. Notwithstanding, for the valence dimension it achieved results similar to score level fusion without the need for an exhaustive search of weights.

Here, only two feature groups were explored, thus such advantage may become more critical in more complex scenarios involving additional feature groups (e.g., amplitude modulation [197]). Overall, feature-level fusion showed to be the best strategy for valence and was observed to be significantly better than score-level ($p_{val} \approx 0.01$) and output associative fusion ($p_{val} \approx 0.01$), whereas score-level fusion for arousal being significantly better than both feature ($p_{val} < 0.01$) and output associative fusion ($p_{val} < 0.01$). In both cases, the proposed motif features showed to provide important discriminatory information and to be complementary to existing benchmark features.

3.5.4 Comparison with previous work

There is increased interest in affective state recognition from EEG and different methods have been recently proposed in the literature, many of which have also relied on the DEAP database. The work in [48], for example, explored graph theoretic features computed from magnitude square coherence values. Such features were shown to outperform several other spectral and wavelet-based methods and on the DEAP dataset achieved an F1 score of 0.63 for valence and 0.60 for arousal using an SVM classifier. For direct comparisons, the best models proposed herein achieved an F1 score of 0.5883 for valence and 0.6960 for arousal, thus representing a 16% increase in arousal, but a drop of 6.6% for valence. It is important to emphasize, however, that the results in [283] relied on leave-one-sample-out (LOSO) cross-validation using a global threshold (equal to 5) for binarization of valence and arousal labels, thus the reported results are likely higher than what is achieved with the method described herein.

More recently, in turn, the work in [197] proposed new amplitude modulation coupling features to gauge connectivity patterns as a function of valence and arousal. BACC values of 0.594 and 0.598 were reported for valence and arousal, respectively, using SVM classifiers and feature fusion, whereas somewhat lower values were attained with score-level fusion for arousal (no changes seen for valence). The values reported in [197] were obtained using a LOSO cross validation scheme. Under the same testing setup, our proposed schemes achieves a BACC of 0.614 and 0.581 for valence and arousal, thus representing a 3.3% increase and a 2.85% decrease in performance, respectively. It is important to point out that motif based methods did not rely on amplitude or rate of change information, therefore fusing them with amplitude modulation features might further improve performance.

3.5.5 Study limitations

This chapter has taken the first steps at gauging the advantages of motif based features over exiting spectrum-based benchmarks. To this end, no optimization was done on the classifiers per se in order to directly compare performances achieved with the same classifier setup but with varying feature inputs. As such, it is expected that further gains may be observed not only with classifier hyperparameter optimization, but also with more complex classification methods or alternate fusion schemes. The work in [283], for example, showed that relevance vector machines (RVM) and fusion of RVMs outperformed SVMs, especially for the arousal dimension. Recent work using deep neural networks has also shown to be a promising route [284]. Future work should explore these more complex machine learning principles combined with motif-based features.

3.6 Conclusion

Noise in physiological signals remains one of the biggest challenges in moving from the lab to in-the-wild conditions. Previously, motif based measures have been used in various clinical applications and have shown robustness to noise. Further, motifs are conceptually simple and computationally inexpensive. In this chapter, we propose the use of motif series and graph theoretic features for improved emotional state recognition. Experiments on the widely used DEAP database show the proposed motif features outperforming several spectrum-based benchmark features while simultaneously providing robustness to noise. Feature-level fusion showed to provide important accuracy gains for both emotional dimensions, thus highlighting the complementarity of the two feature groups for affective state recognition. Score-level fusion, in turn, provided further improvements for arousal prediction. Overall, gains of 8.6% for valence and 9.2% for arousal could be achieved with the proposed system relative to the benchmark and gains up to 16% could be achieved relative to prior art. Overall, this chapter establishes the importance of noise-robust motifs for EEG-based emotional state recognition.

Chapter 4

Non-linear, multi-scale ECG features for robust mental state assessment

4.1 Preamble

This chapter is compiled from material extracted from the manuscript published in *Entropy* [193].

4.2 Introduction

As described in the previous chapters, physiological signals exhibit non-linear properties which have been exploited previously across various clinical applications. In this chapter, we explore the non-linear properties of ECG using multi-scale entropy features and test their robustness for in-the-wild mental state assessment.

Multi-scale entropy (MSE) [127] has been proposed to characterize the complexity of physiologic time series at multiple scales. The algorithm is based on obtaining sample entropy at different time scales using a scaling algorithm. However, the originally proposed scaling algorithm (known as coarse graining) is sub-optimal and may lead to imprecise or undefined entropy values [128]. As a result, several variants of the multi-scale algorithm have been proposed in recent years [128], including PE [129, 132]. Several variants of the coarse-graining method have also been proposed, including replacing it with moving average for short time series [130], a composite procedure that reduces the variance of entropy at higher scales [131], and the recently-proposed generalized multiscale entropy measure [135], which quantifies the dynamics of the volatility (variance) of the time series over different scales [285, 135].

Entropy based measures have been used in the past to characterize aging and to diagnose different cardiac diseases [286, 287, 288]. Moreover, multi-scale analysis of the volatility series of the RR intervals [135] has also shown non-linear behavior and capable of successfully distinguishing between healthy subjects and those with congestive heart failure. In [135], the authors argue that the coarse grained volatility series encapsulates additional information about the time series missed by the normal coarse graining procedure. Additionally the magnitude of the difference intervals of the RR series (i.e., $dRR_i = abs(RR_{i+1} - RR_i)$ has exhibited similar long-range correlations [289]. This property was used in [290] and showed better performance in distinguishing patients with congestive heart failure at lower scales compared to the RR series. Further, several improvements to the PE measure have been proposed. The modified permutation entropy (mPE), for example, takes into account cases in which instantaneous heart rate measures remain the same for two consecutive beats [133], while the weighted permutation entropy [134] tries to incorporate the amplitude information of the time series being analyzed.

While such multi-scale measures have been used in cardiac disease monitoring, they have received little attention for mental state monitoring. In this chapter, we are particularly interested in assessing user mental states in an ambulatory setting, in which movement may not only introduce artefacts that play a detrimental role in signal quality, but also cause changes in cardiac dynamics that may alter HRV measurement. As such, we explore a number of existing multi-scale features and propose new ones for mental state monitoring. We hypothesize that the noise robustness provided by PE coupled with studying complexity at different scales would help better quantify heart rate changes due to mental workload at different levels of physical activity.

The remainder of this chapter is organized as follows. Section 4.3 introduces the multi-scale entropy features. Next, Section 4.4 describes the experimental setup including the pre-processing steps for the data, benchmark features, and performance metrics used. Section 4.5 then presents and

Table 4.1 – Different scaling and entropy algorithms used

Scaling Algorithms	Entropy Algorithms
Coarse graining (*cg*)	Sample Entropy
moving average (*mavg*)	Modified Permutation Entropy
2^{nd} moment *cg* (*mom*)	Weighted Modified Permutation Entropy
moving average *mom* (*mavg_mom*)	
composite coarse graining (*comp_cg*)	

discusses the results obtained. Following this, Section 4.6 evaluates the bias in feature performance due to window overlap size and finally conclusions are drawn in Section 4.7.

4.3 Multi-scale entropy features

The multi-scale entropy methods rely on two steps: i) scaling and ii) entropy calculation over the different scales. Here, we explored different algorithms for both steps, as summarized in Table 4.1.

4.3.1 Scaling algorithms

Several scaling algorithms have been proposed in the literature and attempt to convey fractal information at different scales. All of these methods take the original time series ($x(i)$) with an index i and produce the time series for a different scale. Details about the methods explored herein are given next.

- Coarse graining (cg): This is the original algorithm proposed to obtain different scales. A point j on the scaled series $y_s(j)$ for a scale s is given by:

$$y_s(j) = \frac{\sum_{i=(j-1)s+1}^{js} x(i)}{s}, \tag{4.1}$$

  where, $1 \leq j \leq N/s$. This method has been shown to be sub-optimal [128] and to lead to scaled series that decrease in size, which could increase the variance in the estimated entropy [131]. As a result, several variants have since been proposed, including the ones listed below.
- Moving average ($mavg$): With moving average, a point j on the scaled series $y_s(j)$ for a scale s is given by:

$$y_s(j) = \frac{\sum_{i=j}^{i=j+s-1} x(i)}{s}, \tag{4.2}$$

where $1 \leq j \leq N - s + 1$.

- Composite coarse graining (*comp_cg*) : The composite method generates s different (from $k = 1 \ldots s$) scaled series for a given scale s. The entropy estimates from the different series for the scale s are then averaged to get the entropy estimate. This helps reduce the error in the entropy estimation that occurs due to coarse graining produce. For a given scale s the point j of the k^{th} coarse grained series ($y_{k,s}(j)$) is given by:

$$y_{k,s}(j) = \frac{\sum_{i=(j-1)s+k}^{js+k-1} x(i)}{s}, \tag{4.3}$$

where $1 \leq j \leq N/s$, and $1 \leq k \leq s$ gives the next index of the scaled series. The composite multi-scale entropy ($CMSE$) for a given scale is then given by:

$$CMSE(s) = \frac{\sum_{k=1}^{s} Ent(y_{k,s})}{s}, \tag{4.4}$$

where Ent is the entropy calculation algorithm. In the original CMSE algorithm the sample entropy algorithm is used.

- 2^{nd} moment coarse graining (*mom*): This method quantifies the standard deviation of the scaled series (also called the volatility series) rather than its mean as done by the coarse graining procedure. A point i on the scaled series $y_s(i)$ for a scale s is given by: :

$$y_s(j) = \sigma|_{i=(j-1)s+1}^{js}(x(i)), \tag{4.5}$$

where $1 \leq j \leq N/s$ and σ represents the standard deviation.

- 2^{nd} moment moving average (*mavg_mom*): We propose the moving average procedure for calculation of the 2^{nd} moment to adapt to short time series. This replaces the non-overlapping windows used in coarse graining to a sliding window, as in the case of the moving average algorithm.

Figures 4.1 and 4.2 show the RR series and RR volatility series (using *mavg_mom*) for scales $s = 1$ to $s = 3$ and scales $s = 2$ to $s = 4$, respectively, for a five minute ECG segment. As can be seen, scaling removes some high frequency information from the series, commonly associated with artefacts.

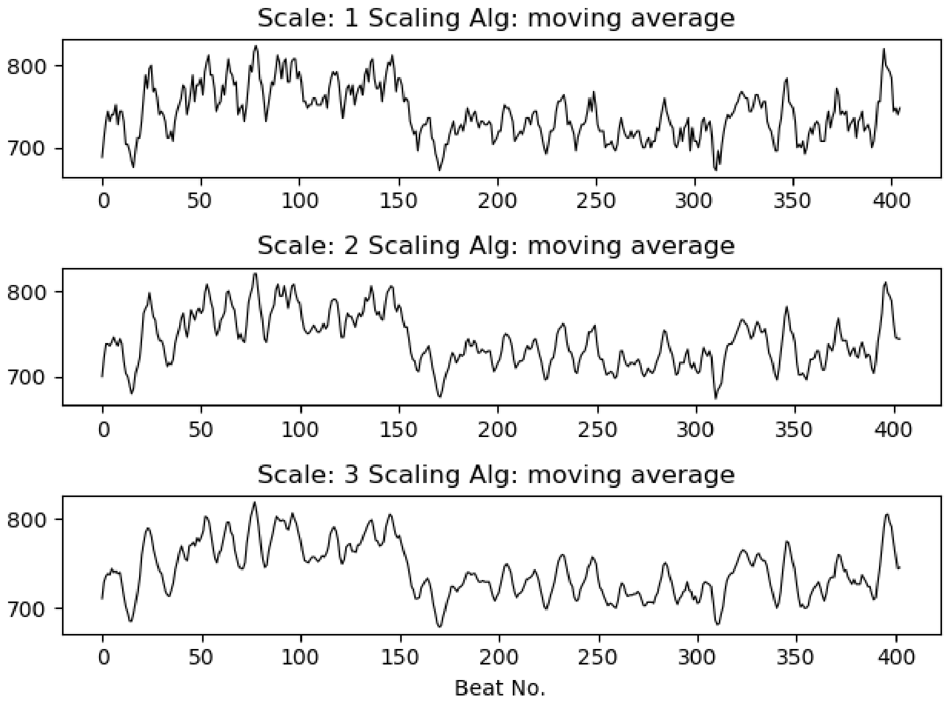


Figure 4.1 – Scaled RR time series with the moving average algorithm for scales s=1 (original series) to s=3

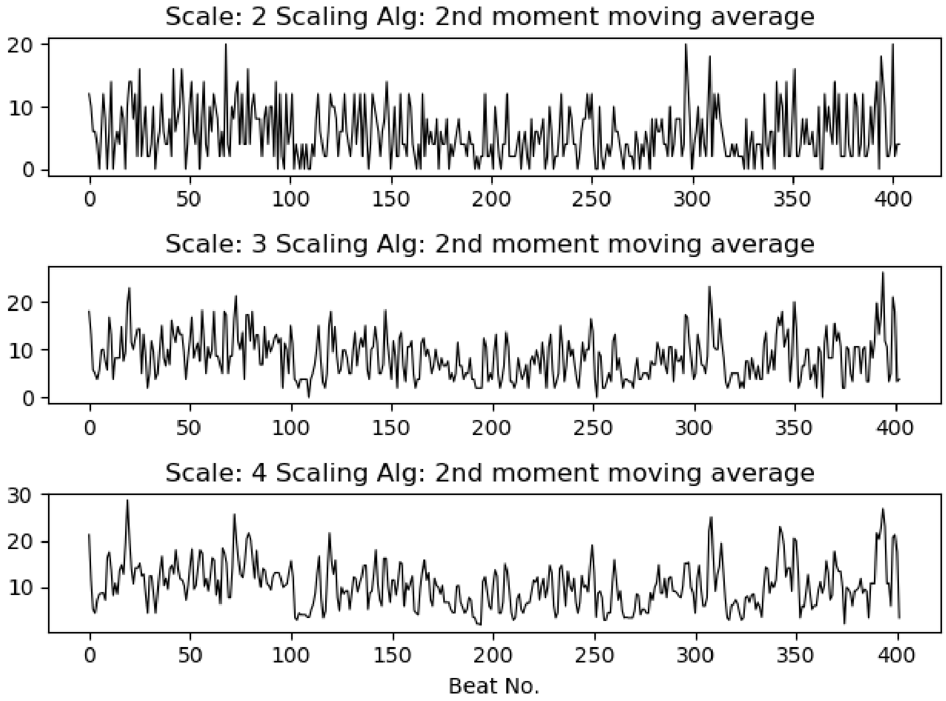


Figure 4.2 – Scaled RR time series with the 2^{nd} moment moving average algorithm for scales s=2 to s=4

Figure 4.3 – Original motifs of degree m=3 appearing in a time series

4.3.2 Entropy algorithms

- Sample Entropy: This has been the most commonly used entropy measures for HRV analysis and has been utilized here as a baseline to compare against permutation entropy based measures.
- Modified Permutation Entropy (mPE): A reduced ECG sampling frequency reduces the resolution at which R peaks are detected, for example, a ECG sampling frequency of 1000 Hz equates to R peak resolution of 1 ms while a lower 250 Hz sampling frequency leads to a peak resolution of 4 ms. Therefore, a reduced sampling can lead to detection of RR values which make seen stationary. Such occurrences require additional motif symbols. As a result, for modified permutation entropy four additional symbolic representations have been added (from 7 to 10 as shown in the equation below). The time series ($X(t)$) is first converted to the ordinal series ($X^{m,\lambda}(j)$), where $1 \geq j \leq N - m$ where N is the size of the time series using the following relations:

$$\mathrm{X}^{m,\lambda}(j) = \begin{cases} 1 & \text{if } X(i) < X(i+\lambda) \text{ \& } X(i+\lambda) < X(i+2\lambda) \text{ \& } X(i) < X(i+2\lambda) \\ 2 & \text{if } X(i) < X(i+\lambda) \text{ \& } X(i+\lambda) > X(i+2\lambda) \text{ \& } X(i) < X(i+2\lambda), \\ 3 & \text{if } X(i) > X(i+\lambda) \text{ \& } X(i+\lambda) < X(i+2\lambda) \text{ \& } X(i) < X(i+2\lambda), \\ 4 & \text{if } X(i) < X(i+\lambda) \text{ \& } X(i+\lambda) > X(i+2\lambda) \text{ \& } X(i) > X(i+2\lambda), \\ 5 & \text{if } X(i) > X(i+\lambda) \text{ \& } X(i+\lambda) > X(i+2\lambda) \text{ \& } X(i) > X(i+2\lambda), \\ 6 & \text{if } X(i) > X(i+\lambda) \text{ \& } X(i+\lambda) < X(i+2\lambda) \text{ \& } X(i) > X(i+2\lambda). \\ 7 & \text{if } X(i) == X(i+\lambda) \text{ \& } X(i+\lambda) < X(i+2\lambda). \\ 8 & \text{if } X(i) == X(i+\lambda) \text{ \& } X(i+\lambda) > X(i+2\lambda). \\ 9 & \text{if } X(i) > X(i+\lambda) \text{ \& } X(i+\lambda) == X(i+2\lambda). \\ 10 & \text{if } X(i) < X(i+\lambda) \text{ \& } X(i+\lambda) == X(i+2\lambda). \end{cases}$$

The modified permutation entropy (mPE) is then given by:

$$mPE = -\sum_{j}^{m!+n} p(\pi_j^{m,\lambda}) \cdot \log(p(\pi_j^{m,\lambda})), \tag{4.6}$$

where n is the number of additional motif patterns added for the modified permutation entropy and $p(\pi_j^{m,\lambda})$ is the relative frequency of the motif pattern represented by $\pi_j^{m,\lambda}$ and calculated as:

$$p(\pi_j^{m,\lambda}) = \frac{\sum_{j \leq m!+n} \nVdash_{u:type(u)=\pi_j}(X_j^{m,\lambda})}{\sum_{j \leq m!+n} \nVdash_{u:type(u)\in\Pi}(X_j^{m,\lambda})}, \tag{4.7}$$

where $\nVdash_A(u)$ denotes the indicator function of set A defined as $\nVdash_A(u) = 1$ if $u \in A$ and $\nVdash_A(u) = 0$ otherwise and $type(.)$ denotes the map from pattern space to symbol space.

- Weighted Modified Permutation Entropy (mPE_wt): Weighted permutation entropy was proposed to incorporate the amplitude information into the permutation entropy algorithm. This is done by calculating the variances (referred as weights w_i) corresponding to each motif pattern in the time series. The relative frequency $p(\pi_i^{m,\lambda})$ of a given pattern is then calculated as:

  $$p(\pi_j^{m,\lambda}) = \frac{\sum_{j \leq m!+n} \nVdash_{u:type(u)=\pi_i}(X_j^{m,\lambda})w_j}{\sum_{j \leq m!+n} \nVdash_{u:type(u)\in\Pi}(X_j^{m,\lambda})w_j}. \tag{4.8}$$

  The permutation entropy for this adjusted relative frequency is then calculated using the standard permutation entropy equation in (4.7).

The entropy values were calculated for both the RR series and the dRR series for a scale of $s = 1 \dots 10$. Moreover, the mean and standard deviation of the entropy measures across all scales were calculated, thus resulting in 24 total features for each type of entropy measure.

4.3.3 Ordinal distance dissimilarity

Ordinal distance based dissimilarity [117] can be used to calculate the difference between the two ordinal series. The distance between the motif distributions of two ordinal series X and Y is given by

$$D_m(X,Y) = \sqrt{\frac{m!}{m!-1}} \sqrt{\sum_{j}^{m!} (p_x(\pi_j^{m,\lambda}) - p_y(\pi_j^{m,\lambda}))^2}, \tag{4.9}$$

where $p_x(\pi^{m,\lambda})$ and $p_y(\pi^{m,\lambda})$ are the relative frequencies of the motif pattern represented by $\pi^{m,\lambda}$ in series X and Y, respectively, and m is the degree of the motif. As the RR series has statistical fractal properties (the statistical properties over different scales do not change), the ordinal distance has been calculated over the different scaled series (referred to as the inter-scale ordinal distance $(Isod_{s_x,s_y})$). To limit the feature space size, we have calculated the distances between scales $s_x =$

$1\ to\ s_x = 3$ and $s_y = 1\ to\ s_y = 10$ with $s_x \neq s_y$. This was done for both the RR and dRR series. Additionally, we calculate the statistics of $Isod$ for given $s_x = 1\ to\ s_x = 3$ relative to all s_y, we calculate the mean, standard deviation and first difference of $Isod_{s_x,\cdot}$, thus resulting in a total number of ordinal distance features of 66. For simplicity, only the original PE based motif structures are considered and the scaling is done by the moving average scaling algorithm.

4.4 Experimental Setup

4.4.1 Pre-processing

The proposed features has been tested for mental workload assessment using the WAUC dataset. In this study, we focus only on the ECG signal measured by the Bioharness Bh3 device, sampled at 250 Hz. This sampling rate allowed for continuous streaming throughout the experiment without the need to recharge the device. Such sampling rate has been successfully used in a number of different applications, including [291, 292, 293]. First, the ECG signal for all subjects was visually inspected and two subjects were removed as the data was corrupted due to sensor malfunction. For the remaining subjects the inter-beat interval series was extracted as follows. First, the ECG was filtered using a band-pass filter with a bandwidth 4-40 Hz to enhance the QRS complex. This was followed by an energy based QRS detection algorithm [136], which is an adaption of the popular Pan Tompkins algorithm [61]. The RR series was further filtered to remove outliers using range based detection ($\geq 280ms$ and $\leq 1500ms$), moving average outlier detection, and a filter based on percent change in consecutive RR values ($\leq 20\%$) as implemented in [146].

4.4.2 Benchmark features

Standard time- and frequency-domain HRV metrics were extracted and used as benchmark measures. A complete list of these conventional measures can be found in Table 6.6. The majority of these benchmark features have been shown in the literature to correlate with mental workload [32] and anxiety [38]. Complete details about these measures can be found in [55]. A total of 15 benchmark features were extracted over 5-minute segments of RR series with a 4-minute overlap, resulting in six RR series for each of the 10-minute experimental sessions. While 80% overlap

Table 4.2 – Benchmark HRV features extracted

Time domain features
mean, standard deviation, coefficient of variation, rmsdd, pNN50, mean of 1^{st} diff., standard deviation of absolute of 1^{st} diff., normalized mean of absolute 1^{st} diff
Frequency domain features
High frequency power (HF), normalized HF, Low frequency power (LF), normalized LF, very low frequency power, HF/LF, total power

between ECG epochs of a given session allows for increasing the number of samples as a form of data augmentation, the large overlap may cause optimistic bias in the results. The current analysis has been done with the 5 minute setting with the effects of optimistic bias discussed later in the chapter. The 5-minute windows for HRV analysis follows recommendations from [55]. Notwithstanding, shorter time series may cause problems with multi-scale entropy estimation. In order to overcome this limitation, focus has been placed on multi-scale entropy methods designed specifically for short term analysis of HRV recordings [131, 130].

4.4.3 Feature selection and classification

For evaluation, a five-fold cross validation setup was used on the full dataset across all subjects. Workload assessment is performed as a binary classification task, where the high and low mental workload ground truth labels are taken from the from the MATB-II task. A support vector machine (SVM) classifier with an RBF kernel is used. To explore the generalization performance, the above mentioned procedure is repeated 50 times with different random seeds. This leads to 250 (5-folds times 50 repetitions with different random seeds) training and test sets and classifications. To assess feature importance, we use feature selection and look at the frequency of features occurring in the top 20 sets for the 250 possible combinations.

To assess feature importance, recursive feature elimination was performed using the Extra Trees Classifier [294]. Given an external estimator that assigns weights to features (an Extra Trees Classifier in this case), the least important features are pruned from the current set of features. The procedure is recursively repeated on the pruned set until the desired number of features to be selected is reached. This technique considers the interaction of features with the learning algorithm to give the optimal subset of features. The feature selection is used to select the top 20 features for each fold of the cross-validation set. The implementation of the classifier and feature selection algorithms was done using sci-kit learn [121]. Accuracy (Acc) and F1-score (F1) have been used as figures-of-merit.

4.5 Experimental Results and Discussion

The performance of the different entropy and scaling methods, the inter-scale ordinal distance, and benchmark features were compared for mental workload assessment. Comparison was done for different levels of physical workload, thus allowing for the robustness of the features to be assessed relative to increases in movement artefacts and changing dynamics of the heart rate brought on by physical activity. We first compare the performance of the different combinations of entropy and scaling approaches for different activity levels. The best performing algorithms are then compared to the performance of ordinal distance scale similarity measure and benchmark features. Additionally, we perform feature fusion where all the different feature sets are combined to test for feature set complementarity.

4.5.1 Comparing different multi-scale entropy algorithms

We calculated the accuracy (Acc) for different activity levels with the combinations of the three entropy measures and five scaling algorithms. Figure 4.4, 4.5, and 4.6 show the performance of the algorithms for no, medium and high physical activity levels, respectively. As can be seen, generally across all physical activity cases and for all entropy algorithms, the short time moving average based scaling (mov_avg and $mavg\_mom$) and composite based scaling ($comp\_cg$) methods outperform coarse graining based approaches (cg, mom). It can also be seen that the modified permutation entropy based on second moment based scaling methods (mom and $mavg\_mom$) (referred to as generalized permutation entropy in [135]) typically achieve higher predictive power across all physical workload cases, hence indicating the importance of volatility series of the RR series, as well as dRR series. Lastly, the modified permutation entropy based algorithms (mPE and mPE_wt) performs better than sample entropy based methods.

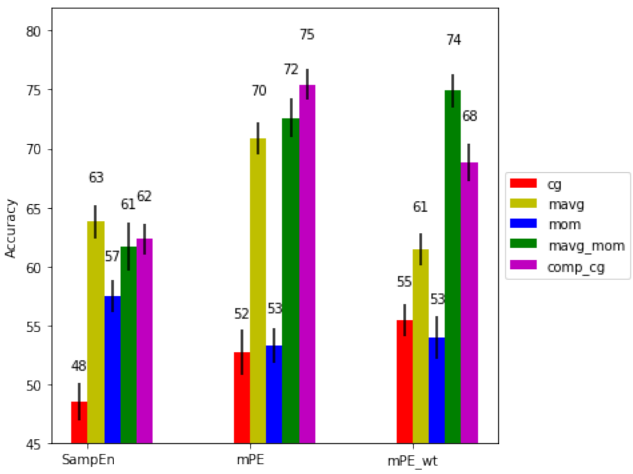


Figure 4.4 – Performance comparison for the no physical activity condition

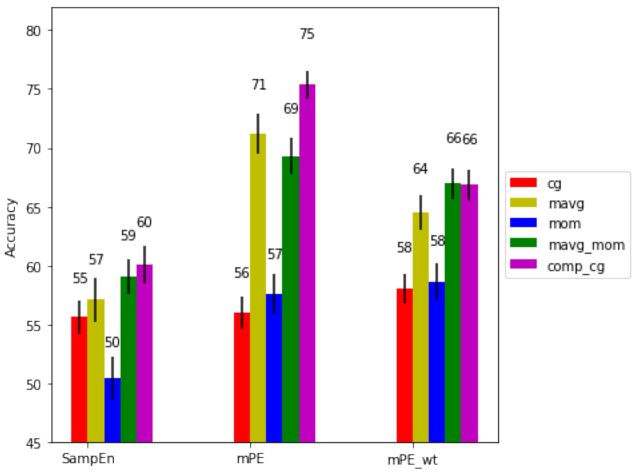


Figure 4.5 – Performance comparison for the medium physical activity condition

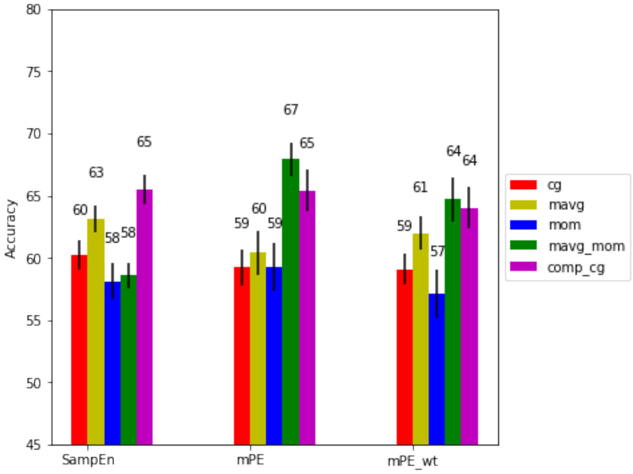


Figure 4.6 – Performance comparison for the high physical activity condition

Specifically, looking at the performance of the features across different physical activity level conditions, it can be seen for the no physical activity condition that mPE with (*comp_cg* and $mavg_m om$) achieved significantly higher performance ($p < 0.01$) than most of the other methods. Moreover, by including the amplitude information into the mPE via the mPE_{wt} measure, a drop in performance is seen for the moving average scaled RR series, but the best performance for the volatility scaled series is achieved with our proposed moving average scaling of the second moment (*mavg_mom*). These results suggest that amplitude information of the volatility series is important for mental workload assessment.

For the medium physical activity level condition, we observe similar performance trends to the no physical activity condition with mPE (*comp_cg* scaling) achieving significantly higher performance ($p < 0.01$) than the other methods. Interestingly, incorporating the amplitude information in this condition leads to a decrease in performance for both *comp_cg* and *mavg_mom*. This could be due to the changing cardiac dynamics during physical activity.

Finally, for the high physical activity level condition, we see an overall drop in performance compared to the other two physical activity levels. We observe that the *SampEn* performance is comparable to mPE and mPE_{wt} for certain scaling methods (*cg*, *mavg* and *comp_cg*). We also see a drop in performance for mPE when using *mavg* compared to *cg* scaling, though this is not observed for *SampEn* and *mPE_wt*, both of which show improvement on using the *mavg* scaling. Overall, we achieve the best performance using the modified permutation entropy for proposed short time 2^{nd} moment calculation and *comp_cg*, using the mPE i.e. excluding the amplitude information. Higher performance without incorporating amplitude information (not using mPE_{wt}) in both medium and high physical activity level conditions could be due to higher noise in the RR series arising from misdetections in the QRS complex caused by movement artefacts. As mPE_{wt} is sensitive to such artefacts, ignoring the amplitude information is better in such cases. Overall, mPE with *comp_cg* and *mavg_mom* achieved the best results, with *mavg_mom* based scaling giving significantly higher results than all the other methods tested ($p < 0.01$).

Moreover, as [135] emphasizes the complementary nature of the volatility series, we further investigate the fusion of *comp_cg* and *mavg_mom* base scaling methods using the mPE algorithm, as these resulted in consistently better performance across all three physical activity conditions. Table 4.3 shows the results of fusion for the different physical activity levels. As can be seen, for the

Table 4.3 – Performance of fused *comp_cg* and *mavg_mom* scaling with *mPE* algorithm for different physical workload levels (* represents cases which perform significantly better ($p < 0.01$) than chance)

Physical Activity Level	Acc	F1
No	0.7893 ±0.0122*	0.7886 ±0.0131*
Medium	0.7726 ±0.0114*	0.7701 ±0.0111*
High	0.6741 ±0.0150*	0.6698 ±0.0164*

Table 4.4 – Benchmark performance comparison for the no physical activity condition (* represents cases which perform significantly better ($p < 0.01$) than chance)

Feature (nof)	Acc	F1
benchmark (15)	0.5772 ±0.0192*	0.4991 ±0.0206
Isod (66)	0.7838 ±0.0137*	0.7882 ±0.0137*
multi-scale entropy (48)	0.7893 ±0.0122*	0.7886 ±0.0131*
fused (129)	0.8438 ±0.0126*	0.8428 ±0.0132*

Table 4.5 – Benchmark performance comparison for the medium physical activity condition (* represents cases which perform significantly better ($p < 0.01$) than chance)

Feature (nof)	Acc	F1
benchmark (15)	0.5318 ±0.0169*	0.6019 ±0.0231*
Isod (66)	0.8189 ±0.0133*	0.8203 ±0.0134*
multi-scale entropy (48)	0.7726 ±0.0114*	0.7701 ±0.0111*
fused (129)	0.8401 ±0.0128*	0.8410 ±0.0129*

no and medium physical activity levels, fusion gives a significant ($p < 0.01$) improvement of 3.53% in accuracy and 3.30% in F1-score and 1.90% accuracy and 1.63% F1-score, respectively, over the best performing *comp_cg* + *mPE* algorithm. However, no improvement is seen for high physical activity level. Such findings corroborate those of [135].

4.5.2 Gauging performance against the benchmark

Here, we compare the performance of the best performing algorithms from 4.5.1 with the benchmark features. We further perform feature fusion of the two sets to further explore their complementary. Tables 4.4, 4.5 and 4.6 shows the benchmark, inter-scale ordinal distance, and best performing multi-scale entropy methods and their fusion for no, medium and high physical activity levels, respectively. In the Tables, 'nof' indicates number of features used in each case.

As can be seen, for the no physical activity condition, both multi-scale entropy and the inter-scale ordinal distance features perform significantly better than the benchmark with improvements of 21.21% in accuracy and 28.25% in F1-score and 20.66% in accuracy and 28.91% in F1-score,

Table 4.6 – Benchmark performance comparison for the high physical activity condition (* represents cases which perform significantly better ($p < 0.01$) than chance)

Feature (nof)	Acc	F1
benchmark (15)	0.5751 ±0.0160*	0.5393 ±0.0245*
Isod (66)	0.7825 ±0.0128*	0.7818 ±0.0137*
multi-scale entropy (48)	0.6741 ±0.0150*	0.6698 ±0.0164*
fused (129)	0.8015 ±0.0152*	0.7987 ±0.0156*

respectively. Additionally, fusion provides significant ($p < 0.01$) improvements of 5.45% accuracy and 5.42% F1-score over the multi-scale features alone.

Similarly for medium physical activity levels, both multi-scale entropy and the inter-scale ordinal distance features perform significantly better than the benchmark with improvements of 24.08% in accuracy and 16.82% in F1-score and 27.20% in accuracy and 21.84% in F1-score, respectively. In this case, the inter-scale ordinal distance features perform significantly better ($p < 0.01$) than the multi-scale entropy features. Fusion also improves performance significantly ($p < 0.01$) and results in gains of 2.12% accuracy and 2.07% F1-score over the inter-scale ordinal distance features. Lastly, for the high physical activity level condition, both multi-scale entropy and the inter-scale ordinal distance features perform significantly better than the benchmark with improvements of 9.90% in accuracy and 13.05% in F1-score and 20.74% in accuracy and 24.25% in F1-score, respectively. Again, the inter-scale ordinal distance features perform significantly better ($p < 0.01$) than the multi-scale entropy features. Additionally, fusion gives a significant ($p < 0.05$) improvement of 1.90% accuracy and 1.69% F1-score over the inter-scale ordinal distance features. The improvement in performance achieved with fusion for all three activity levels further corroborates the results of [135].

4.5.3 Feature ranking

Feature importance was computed based on the outcomes of feature selection across the five cross validation steps, repeated 50 times. The top 20 features were selected for every fold. As such, the frequency of occurrence of a given feature in the top feature set was calculated over the 250 iterations. Features appearing more than 70% were further ranked according to their frequency of occurrence ($freq$) for no, medium and high physical activity levels. These values are reported in Tables 4.7, 4.8 and 4.9, respectively, along with the feature names.

Table 4.7 – Most frequently occurring features in the top-20 feature pool for the no physical activity condition

Feature Name	$freq$
mean of RR	99.2
$Isod_{s1,s4}$ dRR	98.8
$Isod_{s3,s9}$ RR	98.4
Coefficient of variation	93.2
lf/hf	82.8
std. absolute first difference RR	81.2
mean $Isod_{s3,:}$ RR	80
$(mPE + comp\_cg)_{s3}$ RR	70.4

Table 4.8 – Most frequently occurring features in the top-20 feature pool for the medium physical activity condition

Feature Name	$freq$
$Isod_{s1,s2}$ RR	99.2
$Isod_{s1,s7}$ RR	99.2
$Isod_{s2,s10}$ RR	98
mean RR	95.2
$(mPE + mavg\_mom)_{s1}$ RR	91.2
$(mPE + comp\_cg)_{s9}$ RR	89.6
$(mPE + comp\_cg)_{s1}$ RR	88.8
$Isod_{s1,s4}$ dRR	83.2
$(mPE + comp\_cg)_{s10}$ RR	81.6
$(mPE + mavg\_mom)_{s8}$ dRR	74.4
$Isod_{s2,s5}$ RR	74.4
$(mPE + mavg\_mom)_{s8}$ RR	70.4

As can be seen, for the no physical activity condition, we observe that 3 of the 8 top-ranked features are from the inter-scale ordinal distance feature set with interaction of different scales with $s = 3$ being the case for 2 of the 3 $Isod$ features. Additionally, one multi-scale mPE feature shows up in the most frequent set with composite scaling based entropy of $s = 3$. Additionally, we see 4 benchmark features in the top feature set, with three statistical features as well as the ratio of low to high frequency (LF/HF). A consistent decrease in the mean of RR has been reported in the literature with increased mental stress, a similar trend in the standard deviation of RR intervals could explain the overall importance of the coefficient of variation which is a ratio of the two [32]. Similarly, an increase in the LF/HF ratio is reported for increased mental workload across various studies [32]. We also observe that one of the proposed feature was calculated over the dRR series ($Isod$ feature), reflecting the presence of long-term correlations and complexity in the magnitude

Table 4.9 – Most frequently occurring features in the top-20 feature pool for the high physical activity condition

Feature Name	$freq$
mean abs. first difference RR	100
$(mPE + comp\_cg)_{s8}$ RR	96
lfnu	95.2
$(mPE_{wt} + mavg\_mom)_{s3}$ dRR	93.6
hfnu	90.8
$(mPE + comp\_cg)_{s4}$ RR	90.8
$Isod_{s3,s4}$ RR	84.8
$Isod_{s2,s9}$ RR	79.6
$(mPE + comp\_cg)_{s6}$ dRR	79.6
$(mPE + comp\_cg)_{s2}$ dRR	70.8

difference of RR series as noted in [289]. The presence of the different feature sets along with the benchmark features further corroborates the complementary nature of the features.

Similarly, looking at medium physical activity level conditions, we observe that of the 12 most frequent features, 5 are from inter-scale ordinal distance features with interactions between $s = 1$ and $s = 2$ with other scales. Further, additionally 6 multi-scale mPE features show up in the most frequent set with composite scaling based entropy of the original time series ($s = 1$) (same as for mov_mom with $s = 1$ as well), along with $s = 9$ and $s = 10$ as well as of 2^{nd} moment from $s = 5$ and $s = 9$. Additionally, scales $s = 8$ for the mPE of 2^{nd} moment also shows up in the top features for both the RR and dRR series. We only have one benchmark feature (mean RR) in this case among the top features, thus suggesting their sensitivity to movement artefacts. One of features was calculated over the dRR series.

Lastly, for the high physical activity level conditions, of the top 10 most occurring features, we observe that 2 features are from the inter-scale ordinal distance features, 5 features are from the multi-scale mPE entropy with composite scaling based entropy of the scales $s = 2$, $s = 4$, $s = 6$ and $s = 8$, as well as of 2^{nd} moment from $s = 3$. Interestingly, no entropy feature from the original time series ($s = 1$) is seen in the top features. We also observe 3 benchmark features in the top set, with both normalized low and high frequency along with mean of absolute first difference.

Mental workload has reported a drop in HRV features [295, 296] attributed to sympathetic activation and/or para-sympathetic withdrawal [297, 295, 296, 298]. Time and frequency domain HRV features are focused on characterizing the balance between these two systems. However, a lack of clear unbalance of the ANS due to mental workload has been reported in the literature [299]. This has shifted focus on the use of non-linear descriptors based on complex systems approach to better characterize the fractal RR time series [35]. These methods often characterize the complexity of the RR time series [34]. Recent studies indicate that this complexity is a result of both sympathetic and parasympathetic components of the ANS [139]. A recent study [299] has shown that the correlation dimension, which measures the fractal self-similarity of signal, decreases to comparable pathological values during mental workload inducing tasks, which indicates an suppression of the parasympathetic activity in the heart [300] and breakdown of long term correlations in the RR series [65] which can be quantified by complexity at higher scales [286].

A few studies have looked at the effects of exercise on HRV features. The work in [301] reported an increase in overall complexity due to walking (4km/hr) along with a significant increase in normalized low frequency power and a decrease in normalized high frequency power. A similar trend for low intensity exercise was reported in [302] with a contradictory increase in the high frequency component with increased exercise intensity on a bicycle. This increase has further been explained by the influence of breathing on heart rate (respiratory sinus arrhythmia, RSA) which has a strong high frequency component during high intensity exercise [303]. Interestingly, when looking at the non-linear properties of the heart rate for high intensity exercise, entropy ($scale = 1$) decreases while complexity is still retained at different scales [304], something that can be exploited by multi-scale entropy measure.

The scaling process for the multi-scale entropy algorithm is equivalent to low pass filtered frequency bands with decreasing bandwidth with increasing scales [305]. This scaling can be achieved by different types of scaling operations. For this study, we have focused on two methods, namely composite coarse graining and moving average scaling methods. Given the presence of two distinct frequency regions in the heart rate due to parasympathetic activity (corresponding to high frequency fluctuations in the RR series) and sympathetic activity (corresponding to lower frequency fluctuations) [55], the multiscale entropy algorithm represents the complexity of the overall series due to interaction of both sympathetic and parasympathetic systems at lower scales, while representing the complexity of lower frequency component (mostly due to sympathetic activity) at higher scales. Furthermore, the inter-scale ordinal distance feature tries to quantify the complex interaction between the different frequency regions.

In keeping with the above variations in ANS balance with mental workload and exercise, we observe a scale $s = 3$ showed importance for the 'no physical' workload case which captures more lower frequency information compared to original scale. With medium physical activity further contributing to the increase lower frequency components in mental workload, we observe entropy of higher scale of $s = 8$ to $s = 10$ (capturing low frequency information) show up in the top feature sets. Finally, for high physical activity where high frequency components show important contribution due to the influence of RSA to the heart rate we see both low ($s = 2$ and $s = 3$) and high ($s = 4$, $s = 6$ and $s = 8$) scales for entropy show significance in the top feature set. Additionally, we see the normalized high and low frequency components among the top features, which show significance during exercise [301]. We hypothesize that the RSA component which usually causes cardio-respiratory coherence is disrupted due to added mental workload [306], hence making these features important for distinguishing between the two states. The inter-scale ordinal distance feature also shows significance for all three physical activity levels, thus hinting at non-linear interaction between the different frequency regions. The presence of features from the dRR series show the importance of complementary non-linear information present in the series which should be further investigated. Finally, the importance of generalized entropy features calculated on the volatility series hints at the multifractal characteristics holding vital information regarding mental workload. The link between generalized entropy and multifractal heart rate characteristics has been hypothesized in [135].

4.6 Optimistic bias in prediction

Due to the limited duration of physiological signals collected in mental state monitoring experiments, data augmentation is required for improving the number of samples available for the use of machine learning approaches. One of the most commonly used data augmentation approaches involves epoching the available signals with a fixed window length (e.g., 5 minutes for short term HRV analysis) and window shifts with some overlap. The overlap used allows for generating more feature samples leading to better generalization of the machine learning classifiers used. However, a high overlap leads to high correlation between two or more consecutive epochs generated. This can lead to optimistic bias in the results due to the high similarity between features in the training and test sets. Here, we propose a method for evaluation of this bias and re-evaluate the results after removing the bias.

The bias calculation method for the current setup works as follows:

1. First a random time series with uniformly distributed samples is generated. The length of this series is same as the original RR series for a given epoch. The random values range between 400 to 1000 ms to keep them within the physiological range of the original RR series.
2. Next, benchmark HRV features were extracted from the random time series for all epochs.
3. Next, the features are input into the machine learning pipeline. Ideally, these random features should perform equal to chance. However, the performance seen with these features is due to the optimistic bias introduced due to the overlap between signal epochs.

For this analysis, the randomly generated HRV benchmark features were used with the same ML pipeline as used in Sub-section 4.4.3. Two different window size and overlap conditions are explored: i) window: 300 s and overlap: 240 s (i.e., 80% overlap) and ii) window: 240 s and overlap: 120 s (i.e., 50% overlap). While the 240 s window is smaller than the standard 5 min recommend window size for HRV, this window size still allows for clear estimation of LF component and can therefore be used for short term measurement [54]. The 240 s window allows for optimal number of epochs to be extracted for a 10 minute session of the experiment from the WAUC dataset. The 80% overlap leads to a total 1624 samples while the 50% overlap reduces the number of points to 1071 (a decrease of 34%). This decrease in sample size could also play a role in decreasing feature performance. However, using randomly generated features removes this improvement and gives the bias due to amount of overlap.

Tables 4.10 and 4.11 show the performance of randomly generated benchmark features for different physical activity levels for 80% and 50% overlaps, respectively. The performance was compared against a random voting classifier. We observe that the random benchmark features outperform chance for 80% overlap for all physical activity conditions. This improvement in performance is due to the optimistic bias introduced by large overlap between epochs. In contrast, for the 50% overlap the random features perform equal to chance, thus removing the bias. As a result, the analysis conducted henceforth will use a 240 s window size with a 120 s overlap.

Table 4.10 – Performance of randomly generated benchmark HRV features for 300 s window with 80% overlap for different physical activity levels (* performance significantly better than chance (p<0.01))

Physical Activity Level	Acc	F1
No	0.579* ± 0.017	0.544* ± 0.018
Medium	0.559* ± 0.017	0.578* ± 0.018
High	0.579* ± 0.014	0.623* ± 0.015

Table 4.11 – Performance of randomly generated benchmark HRV features for 240 s window with 50% overlap for different physical activity levels

Physical Activity Level	Acc	F1
No	0.523 ± 0.021	0.512 ± 0.025
Medium	0.482 ± 0.017	0.478 ± 0.018
High	0.497 ± 0.019	0.542 ± 0.02

Table 4.12 – Benchmark performance comparison for the no physical activity condition for both 80% and 50% epoch overlap (* represents cases which perform significantly better ($p < 0.01$) than chance)

	Overlap: 80%		Overlap: 50%	
Features	Acc	F1	Acc	F1
benchmark	0.577 ± 0.019*	0.499 ± 0.021	0.537 ± 0.054*	0.422 ± 0.083
Isod	0.784 ± 0.014*	0.788 ± 0.014*	0.625 ± 0.054*	0.626 ± 0.063*
multi-scale entropy	0.789 ± 0.012*	0.788 ± 0.013*	0.572 ± 0.053*	0.566 ± 0.063*
fuse	0.844 ± 0.012*	0.843 ± 0.013*	0.633 ± 0.055*	0.612 ± 0.064*

4.6.1 Re-evaluating multi-scale feature performance

We re-evaluated the performance of the proposed features with the new 50% overlap setting. The performance is evaluated for the benchmark HRV features, best entropy measured selected from section 4.5.1 (the fusion of ($comp\_cg$ + mPE) and ($mavg\_mom$ + mPE) for both RR and dRR series) and $Isod$ features. Mental workload prediction using the different features sets for no, medium, and high physical activity levels for both 80% and 50% window overlap settings are shown in Tables 4.12, 4.13, and 4.14, respectively. As we can observe, there is a performance drop for all feature sets after removing the optimistic bias. More specifically, for the 50% overlap setting, we observe the benchmark feature set performing equal to chance for medium and high physical activity conditions. Additionally, the $Isod$ feature set consistently performs better than the multi-scale entropy features. This trend is consistent for the 80% overlap performance.

Optimistic biases can often overestimate algorithm performance and make it harder to compare against state-of-the-art algorithms. As a result, quantifying this bias can help improve the robustness of results and the methods used. While data augmentation is a popular method for

Table 4.13 – Benchmark performance comparison for the medium physical activity condition for both 80% and 50% epoch overlap (* represents cases which perform significantly better ($p < 0.01$) than chance)

	Overlap: 80%		Overlap: 50%	
Features	Acc	F1	Acc	F1
benchmark	0.532 ± 0.017*	0.602 ± 0.023*	0.484 ± 0.053	0.496 ± 0.095
lsod	0.819 ± 0.013*	0.82 ± 0.013*	0.656 ± 0.05*	0.658 ± 0.058*
multi-scale entropy	0.773 ± 0.011*	0.77 ± 0.011*	0.529 ± 0.053*	0.516 ± 0.077*
fuse	0.84 ± 0.012*	0.841 ± 0.013*	0.639 ± 0.055*	0.637 ± 0.065*

Table 4.14 – Benchmark performance comparison for the high physical activity condition for both 80% and 50% epoch overlap (* represents cases which perform significantly better ($p < 0.01$) than chance)

	Overlap: 80%		Overlap: 50%	
Features	Acc	F1	Acc	F1
benchmark	0.575 ± 0.016*	0.539 ± 0.024*	0.5 ± 0.047	0.518 ± 0.1
lsod	0.783 ± 0.012*	0.783 ± 0.014*	0.611 ± 0.049*	0.626 ± 0.057*
multi-scale entropy	0.674 ± 0.015*	0.669 ± 0.016*	0.522 ± 0.049*	0.548 ± 0.068*
fuse	0.801 ± 0.015*	0.798 ± 0.015*	0.596 ± 0.054*	0.598 ± 0.065*

increasing number of samples, the trade-off between overlap based data augmentation and induced optimistic bias needs to be taken into account. For the purpose of this thesis, in cases where data augmentation is required for ECG signals a window size of 240 s and an overlap of 120 s have been used.

4.7 Conclusions

In the past decade, non-linear features, such as multi-scale entropy, have emerged as popular measures for clinical applications. such features, however, remain to be tested for mental state monitoring, especially under noisy ambulatory conditions. In this chapter, we used multi-scale entropy to quantify the non-linearity of the RR series from individuals performing mental workload tasks in ambulatory conditions. We further combined the noise-robustness provided by motif analysis. Motif based multi-scale metrics outperformed traditionally used sample entropy metrics further emphasizing its noise robustness properties focused on in Chapter 3. Further, different scaling algorithms were tested and the optimal methods for mental workload assessment were derived. Finally, we addressed the optimistic bias in various applications due to high overlap used for data augmentation, thus finding optimal window/overlap parameters for subsequent experiments.

Chapter 5

Separating Confounding Factors using HRV Subband Features for Improved "In-the-Wild" Mental State Monitoring

5.1 Preamble

The content in this chapter is extracted from the manuscript under review for the IEEE Journal of Biomedical and Health Informatics [194].

5.2 Introduction

As discussed in Chapter 4, non-linear features that quantify the RR series complexity have shown improvements over benchmark linear features for mental state monitoring applications in realistic conditions. This is because of their ability to not only deal with movement artefact and noise present in the signal (by using motifs), but also by extracting relevant information in the presence of confounding factors.

While the exact mechanisms behind the changes seen in the complexity of the RR series are still unknown, recent findings [137, 138, 139] have suggested an influence of the SNS and PNS systems under various different clinical conditions. For example, the chaotic behavior in the HF component was linked to circadian (sleep/wake cycle) variability that is independent of age related changes to HF band power [140]. Moreover, the synchronization of the HF band characteristics with respiration and blood pressure was observed when performing tasks related to higher mental workload [82]. In fact, the interaction between the SNS and PNS (i.e., LF and HF bands) has been shown to follow a non-linear coupling behavior, where an increase in SNS activity may not necessarily lead to PNS withdraw, or vice versa; in fact, it can often increase or cause no change in the PNS activity [54]. Metrics quantifying this non-linear interaction have been proposed to discriminate individuals with congestive heart failure [141] and obstructive sleep apnea [142] from healthy controls.

Most of these complexity based features have been calculated over the entire spectrum of the RR series. It is expected that subband based complexity measures may provide some additional insights [128]. For example, changes in HF band peak frequency have been linked to physical activity [143] and entropy of LF and HF bands individually have been shown useful for obstructive sleep apnea detection [144]. In fact, characterizing the properties of different subbands has shown useful in other biosignals. These signals consist of both periodic and aperiodic components, which can directly impact the signal spectrum [307]. For electroencephalograms, for example, these aperiodic aspects of the spectrum have been related to cognitive states and task demands [308]. In this chapter, we propose subband complexity measures and new spectral descriptor features in order to better characterize stress and anxiety under ambulatory and "in-the-wild" conditions. Experiments with two public datasets show that the proposed features not only perform as well as or better than benchmark HRV features, but that they exhibit complementary insights that further improve accuracy when fused together.

The remainder of this chapter is organized as follows: Section 5.3 introduces the proposed features. Next, Section 5.4 presents the experimental setup where the pre-processing, feature extraction, classification pipelines, and figures-of-merit used are described. Section 5.5 presents and discusses the obtained results. Finally, Section 5.6 presents the conclusions.

5.3 Proposed features

The proposed feature set can be divided into two sets: subband-complexity based and subband spectral descriptors based. The subband-complexity features require a separation of the LF and HF time series. This was done by creating the tachogram series (sampled at 4 Hz) from the non-uniformly sampled RR series. Next, two band-pass filters in the range 0.04-0.15 Hz and 0.15-0.4 Hz were used to separate the LF and HF components of the tachogram, hence generating to new time series, namely rr_{lf} and the rr_{hf}, respectively. A representative tachogram series (top) along with rr_{lf} (middle) and the rr_{hf} (bottom) series are shown in Fig. 5.1. Finally, non-linear features (as described in Section 5.4.2.2) are extracted from both rr_{lf} and rr_{hf}. Additionally, the non-linear interaction between the two series has been quantified using the transfer entropy metric [145]. Transfer entropy (TE) is a time-asymmetric measure of amount of directed transfer of information between two time series X and Y. TE quantifies the reduction in uncertainty about X_t from knowing Y_{tk} after considering the reduction in uncertainty about X_t from knowing X_{tk}. Where, k is a lag period and t is the current time period. Therefore, it can be expressed as the difference between two conditional mutual information computations:

$$T_{Y \to X} = I(X_t|X_{t-k}, Y_{t-k}) - I(X_t|X_{t-k}). \tag{5.1}$$

Here, $I(.|.)$ represents the lagged mutual information between probability distributions. We used the Kraskov [309] estimator which makes use of k-nearest neighbors for TE estimation. This metric has been previously shown to be useful in the prediction of congestive heart failure [141] and obstructive sleep apnea [142]. As it is an asymmetric measure, information transfer for both LF to HF (TE-LF-HF) and HF to LF (TE-HF-LF) were calculated using the PyIF toolbox [145].

Next, to calculate the subband-spectral descriptor features, the FFT of the tachogram is first calculated and the power spectral density of the LF and HF frequency components were extracted. Several spectral descriptors were then calculated for each region. The spectral descriptors include: i)

1. Centroid – calculated as the frequency-weighted sum of power normalized by the unweighted sum of power, i.e.:

$$cent = \frac{\sum_{k=b_1}^{b_2} f_k \cdot p_k}{\sum_{k=b_1}^{b_2} p_k}. \tag{5.2}$$

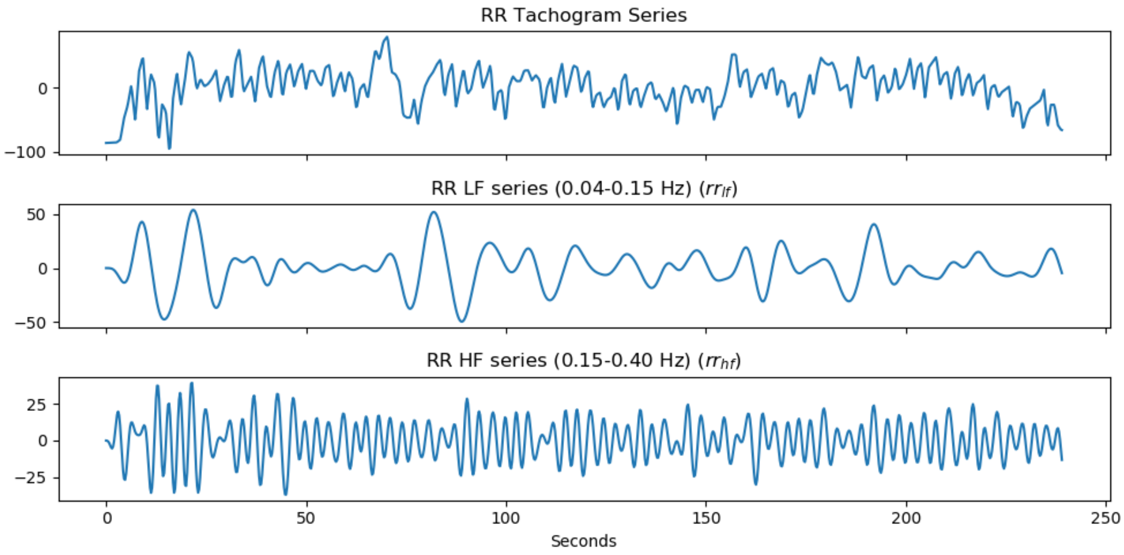


Figure 5.1 – RR tachogram series (top) and the band-filtered LF (middle) and HF (bottom) series.

The centroid represents the “center of gravity” of the spectrum and might be different from the maximum spectral peak.

2. Spread – the standard deviation around the spectral centroid, given by:

$$spread = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - cent)^2 p_k}{\sum_{k=b_1}^{b_2} p_k}}. \tag{5.3}$$

The spread represents the instantaneous bandwidth of the spectrum.

3. Skewness – calculated from the third order moment by:

$$skew = \frac{\sum_{k=b_1}^{b_2} (f_k - cent)^3 p_k}{(spread)^3 \sum_{k=b_1}^{b_2} p_k}. \tag{5.4}$$

The skewness is a measure of the symmetry around the centroid of the spectral band.

4. Kurtosis – computed from the fourth order movement by:

$$kurt = \frac{\sum_{k=b_1}^{b_2} (f_k - cent)^4 p_k}{(spread)^4 \sum_{k=b_1}^{b_2} p_k}. \tag{5.5}$$

The kurtosis measures the flatness, or non-Gaussianity, of the spectrum around its centroid.

5. Crest – measures the ratio of the maximum of the spectrum to the mean of the spectrum, i.e.:

$$crest = \frac{max(p_{k\epsilon(b_1:b_2)})}{\frac{1}{b_2-b_1}\sum_{k=b_1}^{b_2} p_k}. \tag{5.6}$$

   Crest is a measure of the peakedness of the spectrum.

6. Spectral entropy – calculated as:

$$ent = \frac{-\sum_{k=b_1}^{b_2} pk \log(pk)}{\log(b_2 - b_1)}, \tag{5.7}$$

   where b_1 and b_2 represent the range of the relevant frequency bands. f_k represents the frequency in Hz for bin k and p_k represents the spectral power for that bin. Spectral entropy is the measure of uniformity of the spectrum.

5.4 Experimental Setup

5.4.1 Pre-processing

The PASS and TILES datasets described in Chapter 2 have be used for this analysis. The PASS dataset measures stress elicited by video games on 48 participants under different physical activity conditions while the TILES dataset measures stress and anxiety in 200 hospital workers during their work shifts.

For the PASS dataset, the raw ECG signal collected from the BH3 and sampled at 250 Hz was used to assess the HRV. The signal was first cleaned using the 5-25 Hz bandpass filter. This was followed by QRS-complex detection which was done using an energy based QRS detector [146] to create the RR time series. As artefacts (e.g., muscle artefacts, electrode movement, ectopic beats) can cause errors in the RR series, an additional filter to remove RR outliers was used. This filter was based on the following criteria: i) psychological range based removal of RR segments with only RR values in the range of ($< 400ms$ and $< 1200ms$) are kept, ii) moving average based filtering, and iii) quotient based filtering, based on relative change in RR value from the next. In turn, for the TILES dataset, the smart-shirt directly provides RR values. A maximum of four RR intervals are detected by the smart-shirt per second. The RR series is reconstructed from the provided RR values and passed through the RR outlier filter described above.

Table 5.1 – Benchmark HRV features

Time domain features
mean, standard deviation, coefficient of variation, rmsdd, pNN50, pNN20
Frequency domain features
High frequency power (HF), normalized HF, Low frequency power (LF), normalized LF, very low frequency power, HF/LF, total power

5.4.2 Feature Extraction

After pre-processing, the following benchmark and non-linear features along with the proposed feature set are extracted:

5.4.2.1 Benchmark HRV features

Here, commonly used time- and frequency-domain features are used as benchmark [55] and are listed in Table 6.6. These features describe the SNS and PNS system response and have been widely used for stress and anxiety prediction in the past [38, 32]. More details about these features can be found in [55, 54].

5.4.2.2 Non-linear (RR Complexity) features

The RR series exhibits complex non-linear behavior. This behavior has been observed to change based on different physical and psychological conditions [35]. Non-linearity in the RR time series has been quantified using different measures, including entropy, fractal, and chaotic/dynamical system measures. The ANS adapts the heart rate based on the current demands of the body and this adaptation process occurs continuously, thus leading to irregularity in the RR series.

This irregularity can be quantified using sample (SE) and permutation entropy measures (PE) (described in Chapter 2). SE has shown to be an important predictor of various mental states such as stress [32] and anxiety [70], as well as mental fatigue [233]. PE of RR series has also been previously used for prediction of different emotional states [236]. Fractal measures used to quantify complexity include DFA, LE, and CorrDim. DFA has been widely used for stress prediction [32] while LE has been used to predict anxiety [70]. Additionally, CorrDim has been been a robust indicator of long-term mental workload [37] in the past. Here, these non-linear features have been extracted using the NonLinear measures for Dynamical Systems toolbox [310].

For the PASS database, all features are computed using 240-second windows with a 120 second overlap for each session. Overall, 42 HRV features (13 benchmark, 5 non-linear, 24 proposed (12 band spectral and 12 band complexity)) are available for analysis. In turn, for the TILES database, due to the presence of long duration physiological time series, the features were first extracted over non-overlapping 5-minute windows for each day to account for short-term HRV variability, as done in [147, 148]. As the data may be noisy in certain windows, features were only extracted for windows where the $RRPeakCoverage$ quality metric is > 0.3. The value was chosen empirically to ensure at least 1.5 minutes (0.3×5 minutes) of RR information is available to allow for ultra-short-term HRV measures to be extracted. Such measures have been shown to be reliable surrogates for short-term HRV measures for stress prediction [56].

Following this, features were aggregated over an entire day using the following 11 statistical functionals: mean, standard deviation, coefficient of variation, median, min, max, range, 1st and 3st quartile, skewness and kurtosis. Additionally, the $RRPeakCoverage$ was used to create three new quality-aware functionals, including quality weighted mean, standard deviation, and coefficient of variation. Overall, because 14 functionals were calculated over each feature, we have a total of 588 (182 benchmark, 70 non-linear, 336 proposed (168 band spectral and 168 band complexity)) features available for analysis.

5.4.3 Classification and Figures-of-Merit

Binary classification was performed for both stress and anxiety assessment. For the PASS database, the ground truth of the games was used as the stress label. For the TILES database, in turn, a global threshold was used to binarize the stress and anxiety ratings. As a result, a threshold of 1.8 and 1.5 were used for stress and anxiety ratings, respectively. For evaluation, a five-fold cross validation setup was used on the full dataset across all subjects. To explore the generalization performance, the above mentioned procedure is repeated 10 times with different random seeds. This leads to 50 (5-folds times 10 repetitions with different random seeds) training/test sets; classification results reported are the average and standard deviation over the 50 runs. To assess feature importance, we use feature selection and look at the features which rank in the top set more than 70% of the 50 trial runs.

As both dataset labels are imbalanced (54% high stress labels in the PASS dataset; 58% high stress labels and 43% high stress labels for anxiety in the TILES dataset), an SVM classifier with an RBF kernel and a 'balanced' class weight was used. The 'balanced' mode uses the values of the labels to automatically adjust weights inversely proportional to class frequencies in the input training data; such procedure is generally recommended for imbalanced datasets [121].

Additionally, BACC, F1, and MCC were used as figures-of-merit. These are metrics known to be robust to class imbalances. More specifically, MCC takes into account all four values in the confusion matrix and has shown to be more robust to data imbalances when compared to f1-score and accuracy [250]. The MCC value ranges between -1 to 1, with 1 representing perfect prediction, 0 representing random prediction, and -1 indicating inverse prediction. The performances were also compared against a random voting classifier by calculating the significance ($p < 0.01$).

To assess feature importance, RFE was performed using the Extra Trees Classifier [294]. The feature selection is used to select the top 13 features (equal to the number of benchmark HRV features) for the PASS database and the top 100 features for the TILES database for each fold of the cross-validation set. Due to aggregation, the number of features increases by a factor of 14 (number of functionals). The top 100 features have been used to guard against over-fitting [147]. The implementation of the classifier and feature selection algorithms was done using the sci-kit learn toolbox [121].

5.5 Experimental Results and Discussion

In this section, we describe the results obtained with the benchmark, proposed, and combined features sets for both datasets, as well as describe the top-features found. For the PASS database, stress was evaluated across all physical activity conditions.

5.5.1 Classification

Stress classification performance for the PASS and TILES datasets are shown in Tables 5.2 and 5.3, respectively. Anxiety classification performance for the TILES dataset, in turn, is available in Table 5.4. The first two rows present the results for the benchmark features, whereas the proposed features are given in rows 3 and 4. The next six rows represent the fusion of different features sets. In particular, 'Band-All' corresponds to the fusion of band -spectral and complexity features, 'Fuse-Complexity' to fusion of benchmark set with band-complexity features, 'Fuse spectral' to fusion of benchmark set with band-spectral features, 'Fuse RR-Complexity' to fusion of benchmark set with RR complexity features, 'Fuse-Band-All' to fusion of benchmark set with 'Band-All' features, and 'Fuse-All' to the fusion of all extracted features sets. Features highlighted in bold in each Table show the best performing feature set (based on MCC value).

As can be seen, for the stress prediction on the PASS dataset, the best performance is achieved by the fusion of the benchmark features with the proposed band-complexity and spectral descriptor features (Fuse-Band-All) with significant improvements ($p < 0.01$) of 4.64% in BACC, 14.7% in F1, and 24.2% in MCC over the benchmark feature set alone. This feature set combination also shows a significant improvement ($p < 0.01$) of 7.66% in F1 over the fusion of benchmark set with commonly used RR complexity features (Fuse-RR-Complexity). Without feature fusion with the benchmark set, the combination of band-complexity and band descriptor features performs similarly to the benchmark set in terms of BACC and MCC with a significant improvement of 13.5% in F1. We also observe that that these features do not perform as well as the benchmark individually, however fusion of the two (Band-All) feature sets leads to comparable performance with the benchmark feature set, thus showing complementarity between the two proposed feature sets. Finally, compared to the RR complexity features, the combination of the proposed features showed a significant improvement ($p < 0.01$) of 12.5% in BACC, 10.3% in F1 and 158% in MCC.

In turn, for the TILES dataset, for stress prediction, the best performance is achieved with the fusion of all the feature sets (Fuse-All) with a significant improvement of 6.13% in BACC, 5.5% in F1 and 31.6% in MCC over the benchmark feature set alone. This performance is comparable to the performance of fusion of benchmark with the proposed features, as well as fusion of the benchmark with only band spectral features. Individually, the benchmark, RR complexity, and

Table 5.2 – Performance comparison for stress (PASS)

Features	BACC	F1	MCC
Benchmark	0.603 ± 0.029	0.563 ± 0.047	0.211 ± 0.058
RR Complexity	0.542 ± 0.033	0.579 ± 0.040	0.085 ± 0.067
Band-Complexity	0.545 ± 0.036	0.577 ± 0.043	0.090 ± 0.073
Band-Spectral	0.580 ± 0.034	0.625 ± 0.040	0.161 ± 0.069
Band-All	0.610 ± 0.033	0.639 ± 0.045	0.220 ± 0.067
Fuse-Complexity	0.617 ± 0.036	0.624 ± 0.042	0.234 ± 0.073
Fuse-Spectral	0.618 ± 0.035	0.634 ± 0.039	0.237 ± 0.070
Fuse-RR-Complexity	0.611 ± 0.035	0.600 ± 0.042	0.223 ± 0.070
Fuse-Band-All	**0.631 ± 0.030**	**0.646 ± 0.039**	**0.262 ± 0.060**
Fuse-All	0.629 ± 0.033	0.648 ± 0.036	0.258 ± 0.066

proposed features perform comparable to one another, thus further suggesting the complementarity between them.

Lastly, for anxiety prediction, the best performance is again achieved by the combination of all feature sets with significant improvements of 6.45% in BACC, 9.89% in F1, and 36.4% in MCC over the benchmark feature set alone. Additionally, the fusion of benchmark with the proposed features shows a significant improvement of 3.56% in BACC, 6.02% in F1 and 17.9% in MCC compared to the combination of benchmark with RR complexity features. Individually, the proposed features perform comparably to the benchmark and RR complexity feature sets in terms of BACC and MCC with significant improvements of 4.76% and 5.34% in F1 over benchmark and RR complexity features, respectively.

Overall, across both the datasets and mental states, both band-complexity and spectral descriptor features show complementary behaviour not only with each other, but also with existing benchmark features. These findings suggest that future studies could rely on these feature set combinations for improved accuracy and robustness.

5.5.2 Feature Ranking

Feature importance was computed based on the outcomes of feature selection across the five cross validation steps, repeated 10 times. The top-13 and top-100 features were selected for every fold for the PASS and TILES databases, respectively. As such, the frequency of occurrence of a given

Table 5.3 – Performance comparison for stress (TILES)

Features	BACC	F1	MCC
Benchmark	0.620 ± 0.015	0.655 ± 0.013	0.237 ± 0.029
RR Complexity	0.621 ± 0.017	0.664 ± 0.019	0.239 ± 0.032
Band-Complexity	0.612 ± 0.017	0.660 ± 0.020	0.221 ± 0.033
Band-Spectral	0.617 ± 0.014	0.665 ± 0.017	0.232 ± 0.028
Band-All	0.626 ± 0.017	0.670 ± 0.020	0.250 ± 0.033
Fuse-Complexity	0.647 ± 0.014	0.681 ± 0.015	0.291 ± 0.028
Fuse-Spectral	0.655 ± 0.015	0.685 ± 0.016	0.305 ± 0.029
Fuse-RR-Complexity	0.644 ± 0.012	0.681 ± 0.014	0.291 ± 0.024
Fuse-Band-All	0.654 ± 0.014	0.690 ± 0.015	0.305 ± 0.028
Fuse-All	**0.658 ± 0.015**	**0.691 ± 0.015**	**0.312 ± 0.031**

Table 5.4 – Performance comparison for anxiety (TILES)

Features	BACC	F1	MCC
Benchmark	0.604 ± 0.016	0.546 ± 0.020	0.209 ± 0.031
RR Complexity	0.599 ± 0.014	0.543 ± 0.019	0.197 ± 0.028
Band-Complexity	0.605 ± 0.014	0.564 ± 0.019	0.208 ± 0.029
Band-Spectral	0.593 ± 0.014	0.549 ± 0.018	0.185 ± 0.028
Band-All	0.612 ± 0.014	0.572 ± 0.017	0.223 ± 0.028
Fuse-Complexity	0.630 ± 0.012	0.589 ± 0.016	0.258 ± 0.023
Fuse-Spectral	0.639 ± 0.013	0.592 ± 0.016	0.277 ± 0.026
Fuse-RR-Complexity	0.617 ± 0.015	0.565 ± 0.021	0.234 ± 0.030
Fuse-Band-All	0.639 ± 0.014	0.592 ± 0.017	0.277 ± 0.027
Fuse-All	**0.643 ± 0.014**	**0.600 ± 0.018**	**0.285 ± 0.027**

feature in the top feature set was calculated over the 50 iterations. Features appearing more than 70% were further ranked according to their frequency of occurrence ($freq$) for the best performing feature set for the PASS datasets. For the TILES dataset, due to the added functionals calculated on top of the daily feature series, the top 70% features were first separated from their corresponding functionals and feature frequencies for same features with different functionals were aggregated and renormalized. Additionally, the frequency of functionals across the top features were also noted. The top features for stress and their frequency of occurrence for the PASS dataset are shown in Table 5.5. Table 5.6, in turn, shows the top features for stress and anxiety for the TILES dataset.

For the PASS database, overall, 4 of the 11 features are from the proposed feature set with three features from the HF band (2 spectral and 1 complexity) while one from the LF band (spectral). From the commonly used complexity features, we observe DFA, PE and CorrDim in the top features. Of the top 5 features, 3 are from the proposed feature set. The HF spectral entropy (HF-ent) appears as a top ranked feature along with meanRR and coefficient of variation.

An in-depth analysis on the effects of physical activity on HF-ent showed an increase with increasing physical activity level. This can be explained by the effects of respiration on the HF band, termed respiratory sinus arrhythmia (RSA) [54]. Typically, RSA is one of the major factors influencing HF power, hence an increased respiration rate due to physical activity can move the RSA component to a higher frequency range even outside the HF band range for very high respiration rates [311]. This shift in RSA is reflected in the increased spectral entropy value with increasing physical activity levels. Similarly, mental stress can also increase respiration rate [73], thus leading to increase in spectral entropy of the HF band. Indeed, further investigation showed an increase in the spectral entropy of the HF band with increased mental stress.

This shift in RSA to a higher frequency range can further explain the observed increase in the skewness of the HF spectrum (HF-skew), which also appeared as a top feature. In turn, the spectral entropy of the LF band (LF-ent) also showed a similar increase with mental stress, likely due to sympathetic activation caused by stress leading to more overall spectral activity in this region [32]. We also observe a decrease in PE of the rr_{hf} (HF-PE) time series with increased mental stress. This might be caused by the combined effects of the RSA component shifting towards higher frequencies due to physical activity, as well as parasympathetic withdrawal caused by both physical activity and stress reducing complexity series. [301, 32, 70].

A decrease in meanRR and pNN50 have been shown to be indicators of increased SNS modulation [55], reduction of PNS activity [54], and have been reported with increased stress levels [32]. Additionally, changes in both DFA and CorrDim have been linked to stress [32, 312]. Moreover, a decline in HF power has generally been linked to negative emotions [54], which explains the occurrence of HF band-complexity and spectral descriptor features in the top feature set.

Compared to the PASS dataset, which looks at stress in the presence of physical activity for a short duration, the TILES dataset measured stress with long duration stimulus confounded by physical activity and social interaction. While physical activity shifts the RSA towards higher fre-

Table 5.5 – Top features for stress (PASS)

Feature Name	freq (%)
meanRR	100
Coef-Var	100
HF-ent	100
LF-ent	94
HF-PE	94
RR-CorrDim	90
HF-skew	86
sdRR	82
RR-DFA	78
pNN50	39
RR-PE	38

quencies [301], speaking has been shown to shift it towards lower frequencies by lowering respiration rates [312]. Additionally, onset of long-term mental stress (1 hour) causes an initial fluctuation for the HRV parameters followed by gradual return to baseline levels for most metrics [37]. Such fluctuations are captured by the various statistical functionals used to aggregate the feature series over a given day.

For the TILES dataset, for stress, nine of the top 15 features are from the proposed feature set. Of the top proposed features, six of the nine features are from the HF band (2 complexity, 4 spectral), with one from the LF band (spectral) along with both the LF-to-HF the HF-to-LF transfer entropy features. Similar to the PASS dataset, the HF band features are ranked higher than the LF band features. The non-linear interaction of the LF and HF series also appears as a top features. This quantifies the complex non-linear interaction between the SNS and PNS, which are not captured by LF/HF ratio [142, 141].

Among the benchmark feature set, meanRR, pNN20 and coefficient of variation are among the top features, similar to the top PASS dataset features. Additionally, pNN50 and rmssd features, which are highly correlated to each other [55], appear in PASS and TILES stress features, respectively. The PE of the RR series is also among the top features and has been previously shown to distinguish emotions [236]. HF-PE and HF-skew appear as a top proposed features for the stress prediction for both the PASS and TILES datasets showing robustness of these features over different datasets and experimental conditions.

The skewness of HF-PE series appears as a top feature. The distribution skews towards the right with increased stress. A higher value of skewness indicates a greater positive fluctuation of the HF-PE daily series. This type of fluctuation has been observed for long0term mental stress [37] measures using HRV. The minimum value of HF-skew series decreases with increased stress and is among the top features. Further, the HF-skew series distribution skews towards the left. Stressful days for nurses could be characterized by a large amount of talking and team coordination, thus moving the RSA component towards lower frequencies and skewing the spectral power distribution towards this region. Overall, median of HF-PE still shows a decrease with increasing stress consistent with the findings from the PASS dataset. However, the median was not among the top feature set showing higher importance of fluctuations in long term physiological series.

Moreover, functionals representing fluctuation (i.e.„ standard deviation, weighted standard deviation, and range) appear as top features for the LF-HF-TE and show an increase with stress. This could mean changing levels of LF to HF information transfer occurring on more stressful days, while the transfer of information is constant for low stress days. Stress events cause a parasympathetic withdrawal and such withdrawal could cause a change in information transfer between LF to HF region increasing the variability of the transfer entropy. The range and minimum value of the HF-CorrDim series also show up as top features. While the range of the HF-CorrDim series increases with stress, there is a decrease in the minimum value with stress. Together, these functionals indicate a lowered HF-CorrDim due to stress. A decreased CorrDim for RR series indicates lower complexity and has been observed with an increase in stress [32, 37]. This is in line with the finding of lowered HF-PE values for both the TILES and PASS datasets with increasing stress.

Analyzing the top anxiety features, it can be seen that four features appear uniquely among the top features for stress. These are: rmssd, DFA of the RR series, LF-spread and HF-cent. The skewness of the HF-cent appears as a top feature and shows a shift towards the higher values with stress. The HF centroid value may shift towards higher values due to influence of physical activity on the RSA component. A stressful day could lead to higher levels of physical activity for hospital workers, thus increasing the HF centroid value. The maximum value of LF spread was a top feature and increased with the increase in spread. Sympathetic activation caused due to stress could increase the activity in the LF band leading to a higher spread value.

For anxiety prediction, the top features show a large overlap with the top stress features with 10 of the top 15 features appearing in both feature sets. This could be due to the two mental states being highly correlated, as continuously-high stress can lead to anxiety [149]. This fact is further validated by the high correlation (0.69) between stress and anxiety ratings reported by the participants. While top features for anxiety may overlap with the top stress features, different statistical functionals may be relevant.

We observe various functionals indicating fluctuation (i.e., range, standard deviation, and weighted standard deviation) for HF-PE in the top feature set. These functionals show an increase during anxiety conditions indicating greater fluctuation of the HF-PE feature on higher anxiety days. Notably, for anxiety, the mean of the HF-PE series shows an increase in value but is not in the top feature set. Such a change is indicative of increased complexity of the rr_{hf} and is contradictory to decrease in HF-PE with stress as observed previously. In contrast to the PASS dataset, daily mean of HF-skew series shows an overall decrease with anxiety; this could be due to the more dominant effect of speaking compared to physical activity on the RSA for higher anxiety levels.

Similar to HF-PE, standard deviation and weighted standard deviation of HF-skew appear as top features increased with increased anxiety. The minimum value of HF-skew series decreases with increased anxiety showing a greater shift of the distribution towards the left; this behavior was captured by the HF-skew top feature, thus further confirming the dominating effect of speaking on RSA with higher anxiety. The median of HF-kurt is also among the top features and shows a decrease with increased anxiety. A decrease in its value with a corresponding increase in HF-skew could indicate the predominant effect of speaking on the RSA along with a parasympathetic withdrawal caused by anxiety.

Moreover, the maximum value of the HF-CorrDim appears as a top feature showing an increase with anxiety. An increase in CorrDim value is indicative of increasing complexity of the rr_{hf} time series. This increase is also observed as an increase in HF-PE value. Therefore, the overall indicators of increased complexity with anxiety are in contrast to decreasing complexity for stress, as suggested by the features. Similar to stress, both transfer entropy metrics appear in the top feature set for anxiety prediction. Additionally, both weighted and standard deviation show a similar increase in LF-HF-TE metric with anxiety, as observed with stress indication higher variability in transfer entropy on high anxiety days.

Table 5.6 – Top features for stress and anxiety (TILES)

Stress		Anxiety	
Feature Name	freq (%)	Feature Name	freq (%)
meanRR	100	meanRR	100
RR-PE	69	RR-PE	76
Coef-Var	53	pNN20	59
LF-HF-TE	39	HF-PE	45
pNN20	37	Coef-Var	45
HF-CorrDim	27	HF-skew	37
HF-skew	24	RR-SE	34
rmsdd	24	HF-kurt	25
HF-crest	23	HF-LF-TE	24
HF-PE	14	LF-HF-TE	23
HF-kurt	14	HF-crest	11
HF-LF-TE	13	RR-lyap	11
LF-spread	13	HF-spread	10
RR-DFA	11	HF-CorrDim	10
HF-cent	10	LF-skew	10

SE and LE of the RR series along with HF-spread and LF-skewness uniquely appear as top features in the anxiety prediction. Previously, both SE and LE have been linked to state anxiety during an academic examination setting [70]. LE has also been linked to high arousal states elicited using images [313]. The weighted mean of LF-skewness shows a decrease with increased anxiety indicating a shift towards the center of the distribution. This could be indicative of more slower LF components (around 0.015 Hz) becoming relevant with sympathetic activation caused by anxiety.

No frequency domain features appear in the top feature set for either the PASS or TILES datasets for stress or anxiety. This could be due to the power spectral features being less robust to missing RR intervals [314]. On the other hand, the proposed features derived from spectral descriptors and complexity were among the top features, showing both the usefulness and robustness of these features to noise and confounding factors. These findings further point to the importance of subband HRV information for stress and anxiety monitoring.

Lastly, among the functionals used for aggregation, quality-weighted mean and standard deviation were among the most commonly used functionals for both stress and anxiety in the top feature set. This shows the importance of signal quality in mental state prediction, thus corroborating the findings of [147]. We also observe the max and min values among the top functionals. This indicates the highest values of certain HRV metrics for a given shift is a good indicator of perceived stress for the entire shift.

5.6 Conclusions

In this chapter, we propose novel HRV features for in-the-wild mental state monitoring applications. The features are based on separately analyzing the complexity and spectral characteristics of the HF and LF bands of the HRV tachogram. We show that these features outperform benchmark HRV and commonly used non-linear inter-beat-interval complexity features for stress and anxiety monitoring. The separation of the signal into its subband components helps deal with confounding factors effecting the RR series due to socialization, physical activity, and circadian rhythm.

Chapter 6

Multi-modal systems for in-the-wild mental state monitoring

6.1 Preamble

This chapter is compiled from material extracted from the manuscripts published at the *IEEE Systems, Man and Cybernetics Society Annual Conference, 2019* [196] and the *IEEE Engineering in Medicine and Biology Society Annual Conference, 2020* [195].

6.2 Introduction

Multi-modal mental state monitoring systems are known to provide improved performance compared to systems relying on a single modality. However, most multi-modal system studies have been performed in controlled laboratory conditions. In this chapter, we explore the potential of multi-modal systems for mental state monitoring in in-the-wild conditions in the presence of confounders, such as physical activity and social interaction.

This chapter is divided in two parts. In the first part, we explore a multi-modal system for mental workload monitoring on data collected in ambulatory conditions. Physical activity can not only influence the quality of the measured physiological signals [315], but also requires mental resources leading to changes in mental workload [316]. These combined effects make it extremely challenging to measure mental workload in ambulatory settings. For example, intense physical activity effects physiological signals such as decreased HRV [317], decreased skin temperature [318] and a shift of electrodermal activity to higher frequency regions [319]. Lastly, physical activity results in artefacts on the signals collected via wearable devices, thus may lead to measurement errors [320]. While such errors may not be crucial for consumer applications, they cannot be disregarded in safety-critical applications.

In the second part, we explore multi-modal prediction using ultra-short term HRV and respiration. Though short-term analysis of cardiac and respiratory processes have shown useful for offline behavioural analyses, several applications exist in which faster time responses are needed [321], especially in life-saving situations that first responders face on a weekly basis. To this end, so-called ultra-short-term HRV analyses have been explored in which window durations smaller than 5 minutes are used. While some applications have been reported in the literature (e.g., [56, 71, 321]), these have relied in controlled laboratory environments, thus the transferability to highly ecological settings may be unwarranted. Moreover, ultra-short-term analysis for breathing has not been widely explored.

The remainder of the chapter is organized as follows. First, Section Section 6.3 describes the analysis conducted for multi-modal mental workload assessment in ambulatory conditions including the dataset used, signals analysed and the results. Next, Section 6.4 covers multi-modal ultra-short term analysis for mental workload and stress assessment including the dataset, features, signal epochs used and the results. Finally, Section 6.5 provides a conclusion for the analysis.

6.3 Multi-modal mental workload assessment in ambulatory conditions

This section presents multi-modal assessment of mental workload from data collected in ambulatory conditions. The section is divided as follows: Sub-section 6.3.1 presents the experimental setup including the pre-processing required, the different features extracted and the analysis performed. Sub-section 6.3.2 then presents and discusses the results obtained.

6.3.1 Experimental Setup

6.3.1.1 Pre-processing

In order to assess mental workload in ambulatory conditions, the WAUC dataset (as described in Chapter 2 (Section 2.8.1)) was used for this analysis. Different physiological signals need different artefact removal and preprocessing techniques. Here, EEG signals were first filtered by a band-pass FIR filter with a bandwidth 1-45 Hz. Following this, ocular and face muscle movement artefacts were filtered using the widely-used used wavelet enhanced Independent Component Analysis (wICA) algorithm [150, 151]. The wICA makes use of wavelet thresholding on the independent components (IC) of the signal. The algorithm can be divided into five steps: i) First the EEG signal is decomposed into its ICs; ii) Next discrete wavelet transform is performed on the each IC; iii) This is followed by thresholding to attenuate the artefacts components from the IC; iv) An inverse discrete wavelet transform is then performed to recover the clean IC; and v) Finally, the enhanced IC are projected to obtain the enhanced EEG signal. The signal was then decomposed into the following conventional frequency bands: delta (δ, 1-4 Hz), theta (θ, 4-8 Hz), alpha (α, 8-12 Hz), beta (β, 12-30 Hz) and low-gamma (γ_1, 30-45 Hz).

Next, the inter-beat interval (RR) series was extracted from the ECG signal. First, the ECG was filtered using a 5^{th} order band-pass IIR filter with a bandwidth 4-40 Hz to enhance the QRS complex. This was followed by an energy based QRS detection algorithm [136], which is an adaptation of the Pan-Tompkins algorithm [61]. Visual inspection was performed on a sub-sample of the dataset to ensure beat detection was reliable. The RR series was further filtered to remove outliers using range-based detection ($\geq$ 280 ms and $\leq$ 1500 ms), moving average outlier detection, and a filter based on percent change in consecutive RR values ($\leq$ 20%) as implemented in [146]. For the breathing signal (BR), downsampling was performed from 18 to 6 Hz followed by filtering the signal using a low pass IIR filter with cutoff frequency of 2 Hz to remove noise.

For the skin temperature signal (TEMP), winsorization [322] (1% - 99% intervals) was first performed for the data to remove any existing outliers in the signal. Winsorization is a statistical method for removing the outlier values over the nth and (100 - n)th percentiles (n set to 5 for this analysis). This was followed by low pass filtering using a 40^{th} order FIR filter with cutoff frequency of 0.01 Hz to remove high frequency noise. The GSR signal was first down-sampled to

4 Hz. Following this, the phasic high frequency component, associated with sympathetic activity [319], was extracted using a 5^{th} order IIR filter with cutoff frequencies of 0.1-1 Hz. Blood volume pulse was band-pass filtered with a 5^{th} order IIR filter with cutoff frequencies of 8 to 30 Hz to reduce the effects of high frequency noise.

6.3.1.2 Feature extraction

As this chapter looks at the usefulness of multi-modal systems for in-the-wild conditions, focus has only been placed on widely used benchmark features for various physiological modalities to assess the improvements achieved solely by multi-modal techniques. As a result, benchmark features were extracted for each of the available signals [55, 207, 13]. For EEG, features were computed over 3-second windows (epochs) with 2-second overlap. Spectral sub-band power features were computed and normalized by the full band EEG power. A total of 40 (5 frequency bands × 8 electrodes) spectral power features were computed per epoch. These features have shown to correlate with mental workload [323].

Next, standard time- and frequency-domain HRV metrics were extracted and used as benchmark ECG measures. Time domain features include mean RR, standard deviation RR (SDRR), coefficient of variation, RMSSD, pNN50, mean of 1^{st} difference, mean of absolute 1^{st} difference and mean of absolute 1^{st} difference of normalized. These features have been adapted from [38] and [55]. Frequency domain features, in turn, included low frequency (LF), high frequency (HF), and very low frequency (VLF) powers, normalized LF and HF powers, LF/HF ratio, and total power. The majority of these benchmark features have been shown in the literature to correlate with mental workload [45]. Complete details about these measures can be found in [55]. For respiration signal analysis, statistical descriptors were calculated, namely mean, standard deviation, range, skewness, kurtosis, mean of 1^{st} difference. Additionally, spectral analysis was carried out for the breathing signal by calculating the energy of five equally-spaced bands between 0 to 1 Hz. The spectral energy ratio between 0.05-0.25 Hz and 0.25-0.50 Hz, breathing rate and spectral centroid were also calculated.

For BVP, in turn, the spectrum was divided into five equally-spaced bands between 0 and 2.5 Hz (where much of the signal was contained). Additionally, the spectral energy ratio between 0.04-0.15 Hz and 0.15-0.5 Hz. Skin temperature signal was quantified using statistical descriptors namely,

Table 6.1 – Distribution of features by device and modality.

Device	Modality	Feature type	Number
Enobio	EEG	Spectral	40
BioHarness 3	IBI	Time- HRV	8
		Frequency - HRV	7
	BR	Descriptive	6
		Spectral	8
E4	GSR	Descriptive	4
		Spectral	5
	BVP	Spectral	6
	TEMP	Descriptive	8
		Spectral	2

mean, standard deviation, range, mean of 1^{st} difference, min, max, skewness, kurtosis, along with spectral analysis where band energies between 0-0.1 Hz and 0.1-0.2 Hz were also calculated. Finally, for the phasic component of the GSR signal, descriptive statistics were calculated namely, mean, standard deviation, mean of 1^{st} difference, mean of negative 1^{st} difference (MNSSD). Additionally, band power features were calculated for five equally-spaced bands between 0-1 Hz. These bands are investigated as sympathetic activity has been reported to lie in this range with shifts towards higher frequency with physical activity [319]. The features extracted from the different sensors and modalities are summarized in Table 6.1. A total of 94 features were extracted. Unless stated otherwise above, the features were calculated for each session with a 1-minute window and 45-second overlap.

6.3.1.3 Classification, and figures-of-merit

Seven feature sets were explored for mental workload evaluation; these feature sets corresponded to the combination of features extracted from various sensors. These included BH3 only, E4 only, and Enobio only; BH3 and E4, BH3 and Enobio, E4 and Enobio combined; and then all devices combined. This was the case for all three physical conditions and explorations for bike and treadmill conditions were evaluated separately. In all cases, the mental workload level (controlled by MATB-II settings) was used as the target labels. A logistic regression model was used for classification. Linear models are simple to interpret and can provide information about feature importance. Additionally, ridge regression has the advantage of being tolerant to high-dimensional datasets. As such, the classification results reported here are to be considered as a lower bound on possible achievable performance with more complex models.

To quantify between-subject generalization, the models were evaluated in a leave-one-subject-out (LOSO) setting. Both epoch- and session-wise evaluations were performed. For epoch-based evaluation, the individual results for the given epochs for each session were considered as the final result. For session-wise evaluation, a majority voting of the epochs for a given session was made to get a session wise-result. Session-wise evaluation is done as some epochs might be corrupted by sensor noise, physiological artefacts or distractions of the user. The session-wise evaluation therefore exchanges temporal resolution for noise robustness against short term artefacts. Balanced accuracy (BACC) was used as classifier performance figure-of-merit. The implementation of the classifiers and testing algorithms relied on scikit-learn [121]. Significance of the per-epoch results are gauged by comparing against a random voting classifier.

6.3.2 Experimental results and discussion

6.3.2.1 Classification results

The epoch and session-wise performances for bike and treadmill conditions for different physical activity levels are given in Tables 6.2 and 6.3, respectively. For the treadmill condition, we see the best epoch-wise performance is achieved with the fusion of all devices for the ’no’ and ’high’ physical level conditions. The achieved BACC values of 0.704 and 0.645, respectively, show an improvement of 9.3% and 11.2% relative to the the highest performance achieved with an individual sensor (i.e., Enobio and E4). For medium physical level, in turn, the EEG (Enobio) modality showed the highest epoch-wise performance of 0.641, with other sensors showing lower accuracy. When looking at the session-wise majority voting performance, we observe the best no physical activity performance is achieved by two cases with the fusion of all feature sets, as well as by the individual E4 feature set of 0.735. Feature fusion is shown to be extremely important for the medium and high physical activity cases with 0.647 (Enobio + E4, and All) and 0.687 (all) achieved, respectively.

For the bike condition, we see the best epoch wise performance for no and high physical level conditions were from the EEG (Enobio) modality, reaching a BACC of 0.553 and 0.545, respectively. However, for the medium physical activity level, the best performance was achieved with BH3 + E4, with a BACC of 0.617. For the session-wise performance, in turn, we observe best no physical activity performance achieved by three feature sets: BH3, Enobio, and Enobio + E4, all achieving a BACC of 0.547. For medium physical activity, the best performance is achieved by four different feature sets of BH3, E4, (E4 + BH3), and Enobio + BH3 with a BACC of 0.575. Finally, for the high physical activity level, the best performance is achieved from E4 and E4 + BH3 feature sets with a BACC of 0.55.

Overall, the treadmill condition performance was greater than the bike subjects for all three physical activity levels, with differences of 27.3%, 3.9% and 18.3% between best performing models across the no, medium, and high physical activity levels, respectively. We observe that for most feature sets, session-wise performance was higher than epoch-wise performance for the treadmill condition; they were comparable to each other in case of the bike condition. Session-wise performance is more robust to short duration physiological artefacts as well as subject's distracted mental states which can corrupt epoch-wise results, hence a session-wise output is more useful in a realistic setting. The performance variability in the two conditions could be due to experimental differences between the bike and treadmill conditions. With treadmill activity involving head movement due to walking and running, this could have made focusing on a fixed screen harder compared to more stable upper body position for the bike. This can lead to additional visual processing and attention demands from the user [324], thus making the estimation of mental workload less complex. Overall, for all per-session conditions for both the treadmill and bike, sensor fusion showed to provide the best BACC, thus showing the importance of fusion for robust mental workload assessment for ambulant users.

6.3.2.2 Feature importance

Analysis of the weights for different features for a logistic regression classifier provides information on how much classifiers relied on individual features. As a result, Tables 6.4 and 6.5 show the top-three features from each sensor along with their relative rank in the fused feature set ranking for the no, medium, and high physical activity cases for the treadmill and bike conditions, respectively. Feature ranks are defined by the magnitude of the average feature weight across all subjects.

For the treadmill condition, we observe that Enobio (EEG) features are always the highest ranked ones for all physical activity levels. For no physical activity level, the frontal γ energy was shown to be a top feature. Frontal gamma oscillations have been previously linked to engagement in mental activity [325], along with visuo-spatial focused attention [326] and changes during mental arithmetic task [327]. Additionally, θ and δ energies from the right parietal lobe (P4) appeared as top

Table 6.2 – Mental workload prediction performance for different physical activity levels for the treadmill (* represents significance ($p < 0.05$) compared to random voting classifier)

Physical level	Features	BACC (epoch)	BACC (session)
No	BH3	0.576 ± 0.086*	0.647
	E4	0.616 ± 0.118*	**0.735**
	Enobio	0.644 ± 0.191*	0.706
	E4 + BH3	0.611 ± 0.126*	0.588
	Enobio + BH3	0.679 ± 0.186*	0.706
	Enobio + E4	0.662 ± 0.18*	0.647
	All	$\mathbf{0.704 \pm 0.188}^*$	**0.735**
Medium	BH3	0.509 ± 0.099	0.470
	E4	0.519 ± 0.11	0.558
	Enobio	$\mathbf{0.641 \pm 0.18}^*$	0.617
	E4 + BH3	0.492 ± 0.105	0.529
	Enobio + BH3	0.573 ± 0.17*	0.617
	Enobio + E4	0.604 ± 0.208*	**0.647**
	All	0.568 ± 0.169*	**0.647**
High	BH3	0.553 ± 0.125	0.531
	E4	0.580 ± 0.176	0.625
	Enobio	0.522 ± 0.17	0.531
	E4 + BH3	0.553 ± 0.073	0.562
	Enobio + BH3	0.611 ± 0.159	0.625
	Enobio + E4	0.608 ± 0.18	0.625
	All	$\mathbf{0.645 \pm 0.159}^*$	**0.687**

Table 6.3 – Mental workload prediction performance for different physical activity levels for the bike (* represents significance ($p < 0.05$) compared to random voting classifier)

Physical level	Features	BACC (epoch)	BACC (session)
No	BH3	0.529 ±0.164	**0.547**
	E4	0.486 ±0.158	0.524
	Enobio	**0.553 ± 0.118**	**0.547**
	E4 + BH3	0.486 ±0.186	0.452
	Enobio + BH3	0.523 ±0.142	0.452
	Enobio + E4	0.527 ±0.148	**0.547**
	All	0.500 ±0.185	0.524
Medium	BH3	0.564 ±0.096*	**0.575**
	E4	0.554 ±0.078*	**0.575**
	Enobio	0.471 ±0.151	0.450
	E4 + BH3	$\mathbf{0.617 \pm 0.109}^*$	**0.575**
	Enobio + BH3	0.560 ±0.117*	**0.575**
	Enobio + E4	0.486 ±0.149	0.500
	All	0.540 ±0.141	0.500
High	BH3	0.515 ±0.073	0.525
	E4	0.507 ±0.103	**0.550**
	Enobio	**0.545 ± 0.156**	0.500
	E4 + BH3	0.495 ±0.076	**0.550**
	Enobio + BH3	0.524 ±0.142	0.475
	Enobio + E4	0.538 ±0.128	0.525
	All	0.523 ±0.147	0.500

features. This region plays an important role in temporal attention [328]. Functional connectivity between the frontal and parietal regions is related to performance in visual discrimination tasks requiring attention shifts [329].

With increasing physical activity, we see alpha and theta frontal bands become more relevant. An increased theta and decreased alpha power has been reported in high mental workload conditions [323]. Additionally, the most important feature for high physical activity comes from the temporal lobe electrodes. The importance of the temporal lobe electrodes could be due to its role in global visual processing [330], visual discrimination and recognition [331] and also spatial motion and self-motion perception which becomes important with physical activity [332]. For the bike condition, EEG features are among the top features for low and high activity levels. They show similar behavior to the treadmill condition with features from the parietal and frontal electrodes being among the top features. For the medium physical activity case, the EEG features are among the top 6 features with BH3 features being the top features.

Within the BH3 sensor, in turn, top features for the treadmill conditions included SDRR, RMSSD, pNN50, mean of RR, and mean of Δ RR across all conditions. HRV is known to decrease with increasing mental workload, as a result of sympathetic activation and/or to parasympathetic withdrawal [299]. SDRR, RMSSD, and pNN50 are related to the high frequency component of the HRV and hence convey information about parasympathetic withdrawal [55]. With increased physical activity we observe the mean of Δ RR as the top feature. This gives us the mean rate-of-change of the time series and can quantify HF component. The work in [301] reported a significant increase in normalized low frequency power and a decrease in normalized high frequency power with physical activity. However, [302] reported a contradictory increase in the high frequency component with increased exercise intensity on a bicycle. This increase has further been explained by the influence of breathing on heart rate (respiratory sinus arrhythmia, RSA) which has a strong high frequency component during high intensity exercise [303]. Such changes coupled with mental workload could be reflected in the HF component of HRV.

Another top feature observed from BH3 was the standard deviation of BR, which measures the breathing rate variability. Increased sigh rate and respiratory variability have been linked to increased mental workload [333]. Increased sighs are known to induce feelings of relief in subjects and can therefore help perform tasks more efficiently [334]. For BH3 features in the bike condition,

Table 6.4 – Top 3 features for each sensor modality with their relative ranks for each physical activity level for treadmill

Sensor	No	Rank	Medium	Rank	High	Rank
Enobio	γ-FP1	1	α-FP2	1	θ-T10	1
	δ-P4	2	β-P4	2	δ-P4	2
	θ-P4	3	α-P4	3	δ-AF7	3
BH3	SDRR (HRV)	11	pNN50 (HRV)	14	std (BR)	9
	RMSSD (HRV)	12	mean RR (HRV)	20	pNN50 (HRV)	10
	std (BR)	13	SDRR (HRV)	28	mean Δ (HRV)	13
E4	std (GSR)	9	energy (1-1.5 Hz) (BVP)	22	MNSSD (GSR)	7
	MNSSD (GSR)	19	MNSSD (GSR)	24	energy (0-0.1 Hz) (TEMP)	15
	energy (1-1.5 Hz) (BVP)	25	mean Δ (TEMP)	35	energy (0.1-0.2 Hz) (TEMP)	16

Table 6.5 – Top 3 features for each sensor modality with their relative ranks for each physical activity level for bike

Sensor	No	Rank	Medium	Rank	High	Rank
Enobio	β-FP2	1	γ-P4	2	α-P4	1
	δ-T10	2	β-T9	5	δ-P4	2
	α-P3	3	θ-FP2	6	δ-FP1	3
BH3	pNN50 (HRV)	5	mean δ (HRV)	1	pNN50 (HRV)	12
	mean abs Δ norm (HRV)	20	std (BR)	3	std (BR)	14
	CoV (HRV)	21	pNN50 (HRV)	4	std abs Δ (HRV)	23
E4	energy (0.04-0.06 Hz) (GSR)	28	MNSSD (GSR)	18	std (TEMP)	29
	std (TEMP)	32	energy (0.5-1.0 Hz) (BVP)	38	range (TEMP)	33
	std (GSR)	36	energy (1.5-2.0 Hz) (BVP)	40	energy (1.5-2.0 Hz) (BVP)	37

in turn, for the no physical activity condition, the pNN50 and mean of normalized absolute Δ of RR series were among the top 20 features, with pNN50 being ranked highly among all physical activity levels. For the medium physical activity, BH3 features appear highly in the ranking, thus corroborating the high performance achieved by the BH3 sensor modality.

Finally, for the E4 sensor, we observe the mean of negative successive differences of GSR signal among the top features across various physical activity levels. This represents the average decrease rate during decay time for the GSR signal [13]. Mental workload levels have been distinguished [335] using GSR fluctuation duration, which is quantified by the decrease rate during decay time. For medium physical activity levels, the top E4 features ranked at the bottom of the list for both the treadmill and bike conditions. Finally, the temperature spectral features are also among the top ranked features for high physical activity levels. Skin temperature changes reflect the sympathetic nervous system activation [336]. Recently, skin temperature has been used to monitor fluctuations of attentions in an arithmetic task [337]. For the bike condition, we see the top three features are poorly ranked in the overall feature set for no and high physical activity levels. With medium physical activity levels, we see MNSSD (GSR) among the top 20 features which is similar to its performance for treadmill condition.

6.4 Ultra-short term HRV and breathing based mental state monitoring

This section explores the fusion of ultra short term HRV and breathing features for mental state monitoring in ecological conditions. The section is divided as follows: sub-section 6.4.1 presents the experimental setup including the pre-processing required, the different features extracted and the analysis performed; sub-section 6.4.2 then presents and discusses the results obtained.

6.4.1 Experimental Setup

6.4.1.1 Pre-processing

The ENPQ (as described in Chapter 2 (Section 2.8.3)) was used for this analysis. ECG data was visually inspected and subjects whose clear QRS was missing were removed along with the corresponding breathing data. Following this, a simple pre-processing step using a bandpass filter (5-25 Hz) was performed on the ECG signal. The RR series was then extracted using an energy-based QRS detection algorithm, which is an adaption of the popular Pan-Tompkins algorithm [61]. The RR series was further filtered to remove outliers using range-based detection ($\geq 280ms$ and $\leq 1500ms$), moving average outlier detection, and a filter based on percentage change in consecutive RR values ($\leq 20\%$). The ECG processing was done using the toolbox introduced in [146]. In turn, the breathing raw signal was first downsampled to 6 Hz, then low-pass filtered to remove noise using a Chebychev 8^{th} order filter with a 2 Hz cutoff frequency. Finally, several time- and frequency-domain features were extracted from the enhanced breathing curve.

6.4.1.2 Feature Extraction

As described in Section **??**, the purpose of this analysis is to show the advantages achieved by using multi-modal system. Hence, only benchmark feature sets from ECG and respiration amplitude have been used. As a result, from the enhanced RR series, standard time- and frequency-domain HRV features were extracted. These features, when computed from short-term durations less than 5 minutes, have been shown in the literature to correlate with mental workload [32] and stress [321]. Complete details about these measures can be found in [55].

Table 6.6 – Different groups of HRV and breathing features tested

HRV -Time domain
mean · standard deviation · rmssd · pNN50 · coefficient of variation · mean of 1^{st} diff. · std. dev. of abs. of 1^{st} diff. · normalized mean of abs. 1^{st} diff.
HRV -Frequency domain
High freq. power (HF) · normalized HF · low freq. power (LF) · normalized LF · very low freq. power (VLF) · HF/LF
Breathing -Time domain
mean · standard deviation · range · skewness · kurtosis · mean of 1^{st} diff.
Breathing -Frequency domain
Band power in 5 equal bands from 0 to 1 Hz · Band power ratio low freq (0.05-0.25 Hz) and high freq (0.25-0.5 Hz) · spectral centroid · breathing rate (spectral peak)

Several time- and frequency-domain features were extracted from the enhanced breathing curve. A complete list of breathing and HRV features is presented in Table 6.6. Overall, 15 HRV and 14 breathing features were computed for different window sizes (60, 90, 120, 180, 240, and 300 s), without overlap between consecutive windows.

6.4.1.3 Stress and mental workload classification

For evaluation, a 5-fold cross-validation setup was used. Here, stress and mental workload assessment was performed as a subject-wise binary classification task, where a classifier was trained to classify signal segments as high or low stress/mental workload. Binarization of the high/low labels was performed per subject and was based on the subject's reported NASA-TLX ratings. A support vector machine (SVM) classifier with a radial basis function (RBF) kernel was used. To explore the generalization performance of the classifier, the 5-fold test was repeated 50 times with different random seeds. To account for dataset class imbalance, we use balanced accuracy (BAC) as the performance figures-of-merit. Moreover, to assess feature importance recursive feature elimination was performed using the extra trees classifier (ETC) [121]. The sci-kit learn implementation of the SVM classifier and the ETC feature selection algorithm was used [121]. Overall, classifiers were trained with the top-15 features from the HRV, from the breathing, as well as from the fused feature sets.

Table 6.7 – Performance comparison for stress prediction for varying window durations.

	Window duration (s)					
Modality	60	90	120	180	240	300
HRV	0.557	0.558	0.567	0.570	0.576	0.594
Breathing	0.541	0.545	0.548	0.547	0.536	0.560
Fused	0.602	0.608	0.616	0.617	0.619	0.633

A recent focus has been to assess the stability of the ultra-short term HRV estimates with respect to the 5 minute estimate [321, 56]. In order to explore which features stood out for each ultra-short-term duration and classification task, we analyze the list of most frequently-selected features (features appearing at-least 80% of the time across the 250 trials). Moreover, to assess stability across window duration, we also analyze the features which showed to be important across all of the window durations.

6.4.2 Results and Discussion

Classification results for stress and mental workload are shown in Tables 6.7 and 6.8, respectively. For stress, we can observe an increase in performance as segment duration increases, going from BAC=0.557 for 60-second segments to BAC=0.594 to 5-minute segments, thus indicating a loss of roughly 7% in accuracy from an ultra-short-term analysis. For breathing, on the other hand, the changes are smaller and go from BAC=0.541 (60 s) to 0.560 (300 s), thus suggesting a loss of 3.5%. When both modalities are fused, substantial improvement is seen in terms of BAC, thus further corroborating the complementarity of the two modalities [205], even at ultra-short-term analyses. Overall, for when used the fused set, a BAC=0.602 is achieved for 60 s ultra-short-term segments, relative to BAC=0.633 for 5-minute segments (i.e., a 5.1% drop). Note that the results achieved with the fused set under ultra-short-term analysis slightly outperform the HRV and breathing results attained with short-term analysis, thus showing the importance of the multimodal approach for applications that rely on very fast decision making.

For mental workload assessment, the impact of ultra-short-term analysis seems to be even less pronounced than for stress. This may be due to the fact that stress is closely related to a sympathetic response [32] of the nervous system, which typically manifests itself in the lower frequency component of the heart rate [55, 32], which is harder to estimate using ultra-short-term HRV seg-

Table 6.8 – Performance comparison for mental workload prediction for varying window durations.

	Window duration (s)					
Modality	60	90	120	180	240	300
HRV	0.561	0.557	0.569	0.569	0.566	0.579
Breathing	0.536	0.545	0.538	0.542	0.555	0.546
Fused	0.597	0.599	0.599	0.605	0.607	0.610

ments [321]. As such, a BAC=0.561 is achieved for HRV segments of 60 s, whereas BAC=0.579 for 300 s (3.2% drop). Similar findings are seen for the breathing features, where a drop of 1.8% is seen between 1- and 5-minute segments. Again, as was the case with stress, fusion of cardiac and pulmonary information showed useful and substantial improvements were seen with the fused feature set. Overall, a BAC=0.597 could be achieved with a 1-minute segment, thus comparing favourably to BAC=0.610 achieved with 5 minutes (i.e., a drop of 2.2%). It is important to note that this ultra-short-term analysis resulted in accuracy that outperformed HRV and breathing classifiers using 5-minute durations, thus further corroborating the usefulness of the proposed method in adaptive operational settings.

Table 6.9 lists the top consistent features for stress and mental workload. As can be seen the 'meanRR' and 'rmssd' features showed up in both cases. These features have been reported in the stress measurement literature in controlled environments [321, 71]. Stress is associated with a parasympathetic withdrawal along with a sympathetic activation, which can observed by changes in the high frequency component of HRV [32], which is correlated to rmssd [55]. In turn, the top breathing features have been shown previously to correlate to different emotions [13]. Breathing power spectrum shifts towards larger frequencies during stress along with changing breathing patterns have been reported in [338, 207].

For mental workload, recent reports have shown a lack of sympathetic activation with HRV for long term continuous mental workload task while observing a parasympathetic withdrawal with rmsdd feature appearing as relevant over the full length of task [299]. The parasympathetic withdrawal also explains the occurrence of the high frequency components as a consistent feature. Various breathing band powers have shown to be consistent over all segment durations. This can explained by the fact that overall respiratory variability is reported to decrease along with increase in both respiration rate and sigh rate during mental tasks [333, 339]. These factors could shift the spectrum towards higher frequencies, along with a decrease in the spread of spectral peak captured

Table 6.9 – Consistent features for Stress and Mental Workload

Stress	Mental Workload
meanRR	meanRR
rmssd	rmssd
mean of 1^{st} diff	hf
normalized mean of abs 1^{st} diff	normalized mean of abs 1^{st} diff
breathing rate	br skewness
br skewness	br power (0-0.2Hz)
br power ratio	br power (0.2-0.4Hz)
br power (0.2-0.4Hz)	br power (0.4-0.6Hz)

by various power bands. Finally, breathing skewness appears to be a consistent feature for both constructs. A possible explanation could be related to increased sigh rates reported for both stress and mental workload [334, 333]. A sigh is characterized by a deep inhalation followed by slow exhalation, which could skew the amplitude distributions.

6.5 Conclusions

Physiological models for mental state assessment have typically relied on single modalities and with stationary participants. Such models, however, do not translate well into real-life settings where individuals are often subject to various confounding factors such as physical activity, social interaction or are corrupted by noise. Here, we show the importance of the fusion of multiple signal modalities to not only improve the performance of mental state assessment models of ambulant users, but to also provide robustness against movement artefacts while trading off some temporal resolution. Further, some applications, such as monitoring of first responders, may require quick temporal resolution. However, features extracted from shorter signal epochs may be more sensitive to artefacts and confounding factors. Using multi-modal fusion of ultra-short term HRV and breathing features, we show that the performance lost due to the use ultra-short term HRV features can be compensated. Finally, an in-depth feature ranking analysis shows the importance of different signal modalities for the task at hand and may provide future research insights on what sensing modalities to place focus on for mobile workload and stress assessment in highly ecological “in the wild” settings.

Chapter 7

Summary, Future Research Directions, and Conclusions

7.1 Summary

In this doctoral thesis we have investigated the challenges that arise with mental state monitoring in highly ecological settings, commonly termed "in-the-wild". These challenges include:

- The signals being measured are prone to noise and artefacts when collected in real life environments, thus severely hampering the performance of automated monitoring systems, and
- The signals being measured can be impacted by various confounding factors, such as speech, circadian rhythm, and physical activity, hence making it challenging to monitor mental states accurately.

To address these challenges, we have proposed: (*i*) the use of noise robust motif based features for emotion recognition for EEG applications, (*ii*) non-linear and noise robust multi-scale HRV features for prediction of mental workload in ambulatory conditions, (*iii*) features for individual HRV bands representing different nervous system components to better manage the effect of confounding factors on the RR series for prediction of mental states, and (*iv*) multi-modal systems with sensor and signal fusion strategies to improve performance and generalizability in ambulatory and real-life settings. In the following subsections, a summary and discussion is presented on the contributions of this thesis towards the development of noise robust and non-linear systems for in-the-wild mental state monitoring.

7.1.1 Noise-robust EEG motif features

Physiological signals are highly susceptible to noisy artefacts. Usually, artefacts manifest as short-term high amplitude fluctuations in the signals of interest and can severely distort the features extracted. A solution to this lies in the use of artefact filtering methods prior to the feature extraction step. However, these methods usually demand high computational resources and typically run in offline mode with some human intervention (e.g., independent component analysis for EEG artefact removal). To overcome this limitation, we proposed the use of noise robust motif representations for feature extraction. Motif based features build in robustness directly into the features by only considering the shape of the physiological time series for feature extraction. We investigated the use of various motif based features for EEG based emotion recognition application. Experimental results show that these methods outperform the benchmark EEG features for valence and arousal recognition in the binary classification setting. Further, an improvement in performance is observed by fusion of these proposed features with the benchmark, thus showing complementarity to existing feature sets. Ultimately, it is hoped that these low computational complexity noise robust features can be used for EEG based emotion recognition in noisy in-the-wild environments.

7.1.2 Non-linear ECG Features

State-of-the-art features used for various mental state monitoring applications are unable to capture the non-linear complex behavior observed in physiological time series. The past decade has seen various non-linear measures outperform standard time- and frequency- domain methods, in particular for clinical applications. Here, we proposed the use of non-linear multi-scale HRV features for mental workload assessment in ambulatory applications. We have shown these features to outperform the benchmark feature set in mental workload prediction for different physical activity levels. Further, we show that the use of multi-scale permutation entropy features outperforms the standard multi-scale entropy metrics, thus further showing the noise robustness properties of motif based methods. Multi-scale permutation entropy is presented as a noise-robust, computationally simple non-linear feature for in-the-wild mental workload assessment.

7.1.3 Band complexity and spectral descriptor features for tachograms

The PNS and SNS impact the HRV differently. This knowledge has been used to derive frequency domain HRV features for various mental state monitoring applications. More specifically, the LF and HF band energies of the HRV tachogram are primarily influenced by the sympathetic and parasympathetic nervous systems, respectively. These bands, however, are affected differently by the various confounders, such as physical activity, speaking, and circadian rhythm. Additionally, little is known about the spectral characteristics of the HRV power spectrum for the individual bands. To overcome this limitation, we proposed complexity and spectral descriptor based analysis of LF and HF band time series for mental state monitoring. The proposed features show improved performance and complementarity to benchmark features for stress and anxiety prediction over various datasets. The results obtained suggest that complexity analysis of individual HRV bands, as well as characterization of the HRV spectrum components, helps deal with confounding factors otherwise not possible with traditional benchmark HRV features.

7.1.4 Multi-modal systems

In lab-controlled mental state monitoring research, the complementary properties of various physiological signals have been well documented, thus resulting in various multi-modal systems and applications that outperform single modality ones. Multi-modal systems may also become a useful tool for in-the-wild experiments to provide additional robustness to noise, where the sensitivity to noise of different sensors and modalities may be explored to our advantage. Here, we assess this hypothesis for two applications, namely mental workload assessment in ambulatory conditions, as well as tracking of stress/anxiety of police academy students during their field exercises. We show that sensor fusion along with session wise prediction can improve system performance in a LOSO setting. We also show that multi-modal setups are essential for ultra-short-term based analyses required for time-sensitive applications, such as monitoring first responders. The findings herein demonstrate the usefulness of multi-modal systems in tackling noisy data and confounding factors that are often found in highly ecological conditions.

7.2 Comparing proposed HRV features

In this thesis, we have proposed two separate sets of HRV features, namely multi-scale permutation entropy based features, as well as band-complexity and spectral descriptor features from tachograms. Here, we evaluate all of the proposed features and their combinations on the WAUC, PASS and TILES datasets in order to obtain a final ranking of the importance of each feature for monitoring different mental states in realistic settings.

For this analysis, the multi-scale HRV features (described in Chapter 4) and the band-complexity and spectral descriptor features (described in Chapter 5) were used. Additionally, time- and frequency- features and RR complexity features from Chapter 5 have also been extracted and used as benchmarks. For the WAUC and PASS datasets, the features were extracted on the filtered RR series for a window size of 240s and an overlap of 120s to minimize optimistic bias. In turn, for the TILES dataset, the features were extracted using 5 minute non-overlapping windows and aggregated for a given day using statistical functionals. The QRS-detection, RR filtering and functionals used have been described in Chapter 5.

For classification and figures-of-merit, steps described in Chapter 5 have been used. In particular, 5-fold CV is repeated 10 times (resulting in 50 unique train and test set iterations) with different random seeds and an SVM (RBF kernel) classifier is used. For feature selection, RFE is used to select the top 13 (equal to the number of benchmark features) features for WAUC and PASS datasets, while the top 100 features are used for the TILES dataset. The ground truth of the simulation and video game tasks were used as labels for WAUC and PASS datasets, whereas the binarized stress and anxiety ratings were used for the TILES dataset based on the global thresholding method. BACC, F1, and MCC have been used as figures-of-merit. The performances were also compared against a random voting classifier (for 50 iterations) by calculating the significance ($p < 0.01$).

Stress classification performance for the WAUC, PASS, and TILES datasets are shown in Tables 7.1, 7.2 and 7.3, respectively. Anxiety classification performance for the TILES dataset, in turn, is available in Table 7.4. The first two rows present the results for the benchmark and RR complexity features, whereas the proposed features and their combinations are given in rows 3 to 8. In particular, 'Band-All' corresponds to the fusion of band -spectral and complexity features, 'Multi-All' to the fusion of $Isod$ and multi-scale entropy (Multi-Scale) features, and 'Proposed-All' to the fusion of 'Band-All' and 'Multi-All' features. The next 9 rows represent the fusion of different features sets to the benchmark feature set. In particular, 'Fuse-RR-Complexity' to fusion of benchmark set with RR complexity features, 'Fuse-Complexity' to fusion of benchmark set with band-complexity features, 'Fuse-Spectral' to fusion of benchmark set with band-spectral features, 'Fuse-Band-All' to fusion of benchmark set with 'Band-All' features, 'Fuse-Isod' to fusion of benchmark with *Isod* features, 'Fuse-Multi-Scale' to the fusion of benchmark with multi-scale entropy features, 'Fuse-Multi-All' to fusion of benchmark with 'Multi-All' features, 'Fuse-Proposed-All' to the fusion of benchmark with all proposed features and 'Fuse-All' to the fusion of all extracted features sets. Features highlighted in bold in each Table show the best performing feature set (based on MCC value).

As can be seen, for the mental workload prediction on the WAUC dataset, the Isod feature set achieves the highest performance. The feature set shows a significant improvement ($p < 0.01$) of 14.31% in BACC, 28.72% in F1, and 111.2% in MCC over the benchmark feature set. The features also show significant improvements over the RR complexity features, which do not perform better than chance level. The band-complexity and spectral descriptor features, as well as the multi-scale entropy features, do not perform much better than chance levels. Fusion of features with the benchmark does not lead to any improvements, this could be because of poor performance of the benchmark set. Overall, for mental workload predictions in ambulatory environments (running and cycling), the Isod features on their own give the best performance.

For stress prediction on the PASS dataset, the best performance is achieved by the fusion of the benchmark features with both proposed feature sets (Fuse-All-Proposed) with significant improvements ($p < 0.01$) of 4.97% in BACC, 15.45% in F1, and 26.06% in MCC over the benchmark feature set alone. This shows the complementarity of the proposed and benchmark feature sets. On their own, the combination of the proposed features (Proposed-All) significantly outperforms the benchmark feature set with improvements of 4.14% in BACC, 14.92% in F1, and 21.33% in MCC. These results also show the complementarity of the multi-scale features with the band-complexity and spectral descriptor features for ambulatory stress detection.

In turn, for stress prediction on the TILES dataset, the best performance is achieved with two different feature sets. These are the Fuse-Spectral and Fuse-Band-All features. They show significant improvements of 5.48% in BACC for both, 4.58% and 5.34% in F1, and 31.6% in MCC

Table 7.1 – Performance comparison between all features for mental workload (WAUC)

Features	BACC	F1	MCC
Benchmark	0.503 ± 0.034	0.456 ± 0.079	0.004 ± 0.071
RR Complexity	0.497 ± 0.027	0.504 ± 0.038	-0.008 ± 0.055
Band-Complexity	0.519 ± 0.031	0.523 ± 0.041	0.037 ± 0.069
Band-Spectral	0.523 ± 0.023	0.545 ± 0.036	0.045 ± 0.045
Band-All	0.518 ± 0.028	0.534 ± 0.039	0.035 ± 0.058
Isod	**0.575 ± 0.041**	**0.587 ± 0.051**	**0.150 ± 0.082**
Multi-Scale	0.521 ± 0.032	0.526 ± 0.039	0.043 ± 0.067
Multi-All	0.566 ± 0.034	0.571 ± 0.044	0.132 ± 0.068
Proposed-All	0.563 ± 0.035	0.570 ± 0.045	0.124 ± 0.071
Fuse-RR-Complexity	0.517 ± 0.031	0.490 ± 0.051	0.035 ± 0.063
Fuse-Complexity	0.521 ± 0.030	0.519 ± 0.045	0.042 ± 0.061
Fuse-Spectral	0.520 ± 0.035	0.545 ± 0.049	0.038 ± 0.070
Fuse-Band-All	0.541 ± 0.024	0.546 ± 0.036	0.082 ± 0.048
Fuse-Isod	0.573 ± 0.038	0.574 ± 0.053	0.144 ± 0.076
Fuse-Multi-Scale	0.523 ± 0.032	0.534 ± 0.042	0.048 ± 0.064
Fuse-Multi-All	0.568 ± 0.034	0.571 ± 0.048	0.137 ± 0.069
Fuse-Proposed-All	0.564 ± 0.036	0.567 ± 0.042	0.128 ± 0.072
Fuse-All	0.556 ± 0.034	0.561 ± 0.047	0.111 ± 0.069

for both over the benchmark feature set alone. Individually, the Multi-All feature set performs best with only slight improvements over the benchmark and other feature sets. The difference in performance of the feature sets compared to the PASS dataset could also be due to the per-day aggregation of short-term HRV used for the TILES dataset in contrast to only short-term HRV evaluation done on the PASS dataset The increase in performance by fusing with benchmark features further corroborate their complementarity. Lastly, for anxiety prediction, the best performance is again achieved by the Fuse-Spectral and Fuse-Band-All feature sets with significant improvements of 5.47% in BACC, 8.42% and 9.71% in F1, and 32.5% in MCC for both over the benchmark feature set alone. The best performance on the TILES dataset achieved with fusion of benchmark with band-complexity and -spectral features for both stress *and* anxiety prediction highlights the importance of dealing with confounders in long term stress and anxiety detection in realistic settings.

Overall, across the different datasets and mental states, we observe the importance of the features proposed in the thesis for ambulatory and in-the-wild data. For the ambulatory datasets (i.e.,

Table 7.2 – Performance comparison between all features for stress (PASS)

Features	BACC	F1	MCC
Benchmark	0.603 ± 0.029	0.563 ± 0.047	0.211 ± 0.058
RR Complexity	0.542 ± 0.033	0.579 ± 0.040	0.085 ± 0.067
Band-Complexity	0.545 ± 0.036	0.577 ± 0.043	0.090 ± 0.073
Band-Spectral	0.580 ± 0.034	0.625 ± 0.040	0.161 ± 0.069
Band-All	0.610 ± 0.033	0.639 ± 0.045	0.220 ± 0.067
Isod	0.568 ± 0.036	0.585 ± 0.040	0.136 ± 0.072
Multi-Scale	0.582 ± 0.037	0.580 ± 0.042	0.165 ± 0.074
Multi-All	0.596 ± 0.032	0.597 ± 0.035	0.191 ± 0.064
Proposed-All	0.628 ± 0.036	0.647 ± 0.040	0.256 ± 0.072
Fuse-RR-Complexity	0.611 ± 0.035	0.600 ± 0.042	0.223 ± 0.070
Fuse-Complexity	0.617 ± 0.036	0.624 ± 0.042	0.234 ± 0.073
Fuse-Spectral	0.618 ± 0.035	0.634 ± 0.039	0.237 ± 0.070
Fuse-Band-All	0.631 ± 0.030	0.646 ± 0.039	0.262 ± 0.060
Fuse-Isod	0.600 ± 0.031	0.607 ± 0.039	0.201 ± 0.062
Fuse-Multi-Scale	0.626 ± 0.032	0.624 ± 0.043	0.253 ± 0.065
Fuse-Multi-All	0.614 ± 0.030	0.619 ± 0.037	0.228 ± 0.061
Fuse-Proposed-All	**0.633 ± 0.029**	**0.650 ± 0.035**	**0.266 ± 0.059**
Fuse-All	0.632 ± 0.032	0.648 ± 0.039	0.263 ± 0.063

WAUC and PASS), we observe either multi-scale and band-complexity and spectral descriptor features or their combination give the best performance. While for the TILES dataset, we see the band-complexity and spectral descriptor features give the best performance for both stress and anxiety prediction. This shows the importance of these features in dealing with confounding factors such as fatigue, socialization, and physical activity.

7.3 Future Research Directions

Below, several suggestions for future research directions are given:

1. **Using the developed methods for remote patient monitoring**: Recently, focus has been placed on using wearable devices for remote patient monitoring. Several such methods, however, involve measurement of physiological or other signal modalities from individuals in controlled settings or while performing specific tasks [41, 340]; some may even require the

Table 7.3 – Performance comparison between all features for stress (TILES)

Features	BACC	F1	MCC
Benchmark	0.620 ± 0.015	0.655 ± 0.013	0.237 ± 0.029
RR Complexity	0.621 ± 0.017	0.664 ± 0.019	0.239 ± 0.032
Band-Complexity	0.612 ± 0.017	0.660 ± 0.020	0.221 ± 0.033
Band-Spectral	0.617 ± 0.014	0.665 ± 0.017	0.232 ± 0.028
Band-All	0.626 ± 0.017	0.670 ± 0.020	0.250 ± 0.033
Isod	0.621 ± 0.017	0.649 ± 0.021	0.239 ± 0.033
Multi-Scale	0.618 ± 0.017	0.655 ± 0.021	0.233 ± 0.033
Multi-All	0.632 ± 0.016	0.661 ± 0.018	0.261 ± 0.031
Proposed-All	0.630 ± 0.017	0.663 ± 0.021	0.257 ± 0.034
Fuse-RR-Complexity	0.644 ± 0.012	0.681 ± 0.014	0.285 ± 0.024
Fuse-Complexity	0.647 ± 0.014	0.681 ± 0.015	0.291 ± 0.028
Fuse-Spectral	**0.655 ± 0.015**	**0.685 ± 0.016**	**0.305 ± 0.029**
Fuse-Band-All	**0.654 ± 0.014**	**0.690 ± 0.015**	**0.305 ± 0.028**
Fuse-Isod	0.636 ± 0.015	0.663 ± 0.018	0.269 ± 0.030
Fuse-Multi-Scale	0.64 ± 0.015	0.672 ± 0.017	0.276 ± 0.030
Fuse-Multi-All	0.639 ± 0.015	0.666 ± 0.018	0.275 ± 0.030
Fuse-Proposed-All	0.637 ± 0.015	0.668 ± 0.017	0.271 ± 0.030
Fuse-All	0.639 ± 0.015	0.674 ± 0.019	0.275 ± 0.031

presence of a physician [341]. A large number of diseases today are chronic and long-term “uncontrolled” monitoring is needed to track and prevent exacerbation of conditions, thus reducing the burden on the healthcare system. Unfortunately, as the COVID-19 pandemic has shown us, evaluation of disease severity using unobtrusive long-term monitoring of physiological signals is extremely challenging [342]. As a first step in this direction, we used the methods developed in this thesis for remote monitoring of chronic obstructive pulmonary disease (COPD) severity levels and exacerbation [211]. Heart rate and activity information was collected using smartwatches from individuals suffering from COPD. The optimal performance for severity and exacerbation detection was achieved with a multimodal system relying on HRV and activity features. Overall, the performance was above chance level and required only the first few days of each subject data for training. Future research should focus on testing the usefulness of proposed features for monitoring COPD and progression of other diseases (e.g., COVID-19) without placing any constraints on the user.

Table 7.4 – Performance comparison between all features for anxiety (TILES)

Features	BACC	F1	MCC
Benchmark	0.604 ± 0.016	0.546 ± 0.020	0.209 ± 0.031
RR Complexity	0.599 ± 0.014	0.543 ± 0.019	0.197 ± 0.028
Band-Complexity	0.605 ± 0.014	0.564 ± 0.019	0.208 ± 0.029
Band-Spectral	0.593 ± 0.014	0.549 ± 0.018	0.185 ± 0.028
Band-All	0.612 ± 0.014	0.572 ± 0.017	0.223 ± 0.028
Isod	0.607 ± 0.016	0.550 ± 0.019	0.216 ± 0.033
Multi-Scale	0.611 ± 0.015	0.567 ± 0.019	0.220 ± 0.027
Multi-All	0.619 ± 0.014	0.568 ± 0.020	0.238 ± 0.029
Proposed-All	0.619 ± 0.017	0.569 ± 0.021	0.239 ± 0.033
Fuse-RR-Complexity	0.617 ± 0.015	0.565 ± 0.021	0.234 ± 0.030
Fuse-Complexity	0.630 ± 0.012	0.589 ± 0.016	0.258 ± 0.023
Fuse-Spectral	**0.639 ± 0.013**	**0.592 ± 0.016**	**0.277 ± 0.026**
Fuse-Band-All	**0.639 ± 0.014**	**0.599 ± 0.017**	**0.277 ± 0.027**
Fuse-Isod	0.625 ± 0.016	0.571 ± 0.020	0.250 ± 0.032
Fuse-Multi-Scale	0.626 ± 0.013	0.583 ± 0.016	0.250 ± 0.026
Fuse-Multi-All	0.623 ± 0.014	0.572 ± 0.018	0.245 ± 0.029
Fuse-Proposed-All	0.623 ± 0.015	0.574 ± 0.019	0.245 ± 0.030
Fuse-All	0.624 ± 0.015	0.575 ± 0.019	0.247 ± 0.031

2. **Combining feature engineering with deep learning models**: Deep learning (DL) algorithms have recently emerged as a popular pattern recognition methodology. Deep neural networks (DNN) have shown great improvements in a large number of audio and image applications [343]. These algorithms typically rely on black box networks with a large number of trainable parameters. These parameters are trained using very large datasets and might require a large amount of computational resources and training time (up to weeks for bigger networks). With their popularity, several limitations of DL based approaches have also come to the forefront. These include their black box nature, susceptibility to adversarial attacks [344], and growing carbon footprint due to the computational resources used [345]. These issues have lead to a renewed push for use of more interpretable models to assure accountability and transparency [346].

   Recently, DNNs have also been used for mental state monitoring applications for data collected in controlled laboratory conditions [347, 348, 349]. However, DNNs have yet to be evaluated for data collected in-the-wild. Such data often requires added pre-processing and feature extraction before being input into DNNs. Recently, we compared the performance of DNN architectures of varying depths and relying on benchmark HRV features with the performance achieved with a traditional SVM with the proposed multi-scale entropy HRV features (see Chapter 4) [210]. Stress and anxiety monitoring on the TILES dataset showed that classical machine learning with feature engineering significantly outperformed the DNN-benchmark-HRV combination for both stress and anxiety classification. Additionally, the hyperparameter tuning times for DNNs **relying on multiple GPUs** were an order of magnitude higher than the tuning of the SVM with a single CPU. These results show the usefulness of feature engineering for in-the-wild experimentation, whilst ensuring interpretable results. Future work should focus on combining DL approaches with feature engineering to take advantage of the gains obtainable with both approaches.

3. **Integrating context with physiological feature engineering**: As discussed previously, physiological data collected in-the-wild is contaminated by a large number of confounding factors. These range from physiological (e.g., circadian rhythm, sleep quality) to social (e.g., current location, environment, amount of conversations) to physical activity. Recently, behavioral sensors that can capture location, environmental factors, physical activity, and sleep cycles have received attention. These sensors can capture important contextual information which could be leveraged to improve mental state monitoring in uncontrolled settings. The multi-modal systems presented in Chapter 6, for example, made use of various physiological modalities to minimize the impact of confounding factor on system performance. Future work should explore the inclusion of context-awareness to multi-modal systems to further boost performance. To the best of our knowledge, no work has looked into integrating context into mental state monitoring models. As a first step, we evaluated the effect of adding circadian rhythm information and location within the hospital to measure stress and anxiety based on the audio features available with the TILES dataset [350]. Our initial analysis has shown that adding such contextual information can help improve performance [350]. Future work should explore the use of other contextual cues directly into the feature engineering step (e.g., [147]), into the ML pipeline (e.g., [85]), or at both stages and measure the impact it can have on in-the-wild mental state assessment.

7.4 Conclusions

Various technological advances of the last decade have allowed for the collection of unobtrusive, long-term physiological data in highly ecological settings, thus enabling remote assessment of different mental states. Current systems, however, have numerous challenges which have limited their widespread use. For example, most models have been built using data collected in controlled laboratory environments, thus achieve poor results when taken outside of the lab. Most models have also been developed using simulated tasks, thus minimizing the impact of confounding factors, which are typically present in highly ecological settings. Lastly, avai;lable models are "data hungry," but large in-the-wild datasets are nonexistent. To address these challenges, this thesis has proposed several features and multimodal systems. The results obtained show the importance of *(i)* building noise robustness directly into physiological features, *(ii)* dealing with confounding factors via feature engineering, and *(iii)* utilizing multi-modal systems to overcome noise robustness and increase generalizability. Ultimately, it is hoped that the innovations presented herein will help advace the research and development of mental state monitoring models for highly ecological conditions and open doors for applications in other domains, including remote patient monitoring.

Bibliography

[1] M. A. Cohn, B. L. Fredrickson, S. L. Brown, J. A. Mikels, and A. M. Conway, "Happiness unpacked: positive emotions increase life satisfaction by building resilience." *Emotion*, vol. 9, no. 3, p. 361, 2009.

[2] J. Park, *Work stress and job performance.* Statistics Canada Ottawa, Canada, 2007.

[3] B. E. Cohen, D. Edmondson, and I. M. Kronish, "State of the art review: depression, stress, anxiety, and cardiovascular disease," *American journal of hypertension*, vol. 28, no. 11, pp. 1295–1302, 2015.

[4] H. Van Praag, "Can stress cause depression?" *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 28, no. 5, pp. 891–907, 2004.

[5] M.-F. Marin, C. Lord, J. Andrews, R.-P. Juster, S. Sindi, G. Arsenault-Lapierre, A. J. Fiocco, and S. J. Lupien, "Chronic stress, cognitive functioning and mental health," *Neurobiology of learning and memory*, vol. 96, no. 4, pp. 583–595, 2011.

[6] C. M. Doran and I. Kinchin, "A review of the economic impact of mental illness," *Australian Health Review*, vol. 43, no. 1, pp. 43–48, 2019.

[7] K. Patrick and J. F. Lavery, "Burnout in nursing," *Australian Journal of Advanced Nursing*, vol. 24, no. 3, p. 43, 2007.

[8] F. M. Dimou, D. Eckelbarger, and T. S. Riall, "Surgeon burnout: a systematic review," *Journal of the American College of Surgeons*, vol. 222, no. 6, p. 1230, 2016.

[9] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology.* Elsevier, 1988, vol. 52, pp. 139–183.

[10] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, 2014.

[11] A. Gaillard, "Comparing the concepts of mental load and stress," *Ergonomics*, vol. 36, no. 9, pp. 991–1005, 1993.

[12] N. H. Alsuraykh, M. L. Wilson, P. Tennent, and S. Sharples, "How stress and mental workload are connected," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2019, pp. 371–376.

[13] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[14] N. Thakur and C. Y. Han, "Framework for an intelligent affect aware smart home environment for elderly people," *Int. J. Recent Trends Hum. Comput. Interact.(IJHCI)*, vol. 9, no. 1, pp. 23–43, 2019.

[15] B. Kerous, F. Skola, and F. Liarokapis, "Eeg-based bci and video games: a progress report," *Virtual Reality*, vol. 22, no. 2, pp. 119–135, 2018.

[16] K.-J. Oh, D. Lee, B. Ko, and H.-J. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation," in *2017 18th IEEE international conference on mobile data management (MDM)*. IEEE, 2017, pp. 371–375.

[17] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[18] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[19] A. Luximon and R. S. Goonetilleke, "Simplified subjective workload assessment technique," *Ergonomics*, vol. 44, no. 3, pp. 229–243, 2001.

[20] T. M. Marteau and H. Bekker, "The development of a six-item short-form of the state scale of the spielberger state—trait anxiety inventory (stai)," *British journal of clinical Psychology*, vol. 31, no. 3, pp. 301–306, 1992.

[21] G. Borg, *Borg's perceived exertion and pain scales.* Human kinetics, 1998.

[22] F.-X. Lesage, S. Berjot, and F. Deschamps, "Clinical stress assessment using a visual analogue scale," *Occupational medicine*, vol. 62, no. 8, pp. 600–605, 2012.

[23] E. Facco, G. Zanette, L. Favero, C. Bacci, S. Sivolella, F. Cavallin, and G. Manani, "Toward the validation of visual analogue scale for anxiety," *Anesthesia progress*, vol. 58, no. 1, pp. 8–13, 2011.

[24] S. S. Sabet, C. Griwodz, and S. Möller, "Influence of primacy, recency and peak effects on the game experience questionnaire," in *Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems*, 2019, pp. 22–27.

[25] U. W. Müller, C. L. Witteman, J. Spijker, and G. W. Alpers, "All's bad that ends bad: there is a peak-end memory bias in anxiety," *Frontiers in psychology*, vol. 10, p. 1272, 2019.

[26] G. Eisele, H. Vachon, G. Lafit, P. Kuppens, M. Houben, I. Myin-Germeys, and W. Viechtbauer, "The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population," *Assessment*, p. 1073191120957102, 2020.

[27] C. Niemic, "Studies of emotion: A theoretical and empirical review of psychophysiological studies of emotion." 2004.

[28] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2017.

[29] J. Xu and B. Zhong, "Review on portable eeg technology in educational research," *Computers in Human Behavior*, vol. 81, pp. 340–349, 2018.

[30] P. Aspinall, P. Mavros, R. Coyne, and J. Roe, "The urban brain: analysing outdoor physical activity with mobile eeg," *British journal of sports medicine*, vol. 49, no. 4, pp. 272–276, 2015.

[31] R. Zink, B. Hunyadi, S. Van Huffel, and M. De Vos, "Mobile eeg on the bike: disentangling attentional and physical contributions to auditory attention tasks," *Journal of neural engineering*, vol. 13, no. 4, p. 046017, 2016.

[32] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, “Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis,” *Biomedical Signal Processing and Control*, vol. 18, pp. 370–377, 2015.

[33] R. Cassani, T. H. Falk, F. J. Fraga, P. A. Kanda, and R. Anghinah, “The effects of automated artifact removal algorithms on electroencephalography-based alzheimer’s disease diagnosis,” *Frontiers in aging neuroscience*, vol. 6, p. 55, 2014.

[34] P. Melillo, M. Bracale, and L. Pecchia, “Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination,” *Biomedical engineering online*, vol. 10, no. 1, p. 96, 2011.

[35] R. Sassi, S. Cerutti, F. Lombardi, M. Malik, H. V. Huikuri, C.-K. Peng, G. Schmidt, Y. Yamamoto, D. Reviewers:, B. Gorenek *et al.*, “Advances in heart rate variability signal analysis: joint position statement by the e-cardiology esc working group and the european heart rhythm association co-endorsed by the asia pacific heart rhythm society,” *Ep Europace*, vol. 17, no. 9, pp. 1341–1353, 2015.

[36] L. G. S. França, J. G. V. Miranda, M. Leite, N. K. Sharma, M. C. Walker, L. Lemieux, and Y. Wang, “Fractal and multifractal properties of electrographic recordings of human brain activity: toward its use as a signal feature for machine learning in clinical applications,” *Frontiers in physiology*, vol. 9, p. 1767, 2018.

[37] S. Delliaux, A. Delaforge, J. Deharo, and G. Chaumet, “Mental workload alters heart rate variability, lowering non-linear dynamics,” *Frontiers in physiology*, vol. 10, p. 565, 2019.

[38] W.-H. Wen *et al.*, “Toward constructing a real-time social anxiety evaluation system: Exploring effective heart rate features,” *IEEE Transactions on Affective Computing*, 2018.

[39] M. Peltola, “Role of editing of RR intervals in the analysis of heart rate variability,” *Frontiers in physiology*, vol. 3, p. 148, 2012.

[40] R. W. Homan, “The 10-20 electrode system and cerebral location,” *American Journal of EEG Technology*, vol. 28, no. 4, pp. 269–279, 1988.

[41] R. Cassani, T. H. Falk, F. J. Fraga, M. Cecchi, D. K. Moore, and R. Anghinah, “Towards automated electroencephalography-based alzheimer’s disease diagnosis using portable low-density devices,” *Biomedical Signal Processing and Control*, vol. 33, pp. 261–271, 2017.

[42] J. J. Bird, L. J. Manso, E. P. Ribeiro, A. Ekart, and D. R. Faria, “A study on mental state classification using eeg-based brain-machine interface,” in *2018 International Conference on Intelligent Systems (IS).* IEEE, 2018, pp. 795–800.

[43] J. G. Nicholls, A. R. Martin, B. G. Wallace, and P. A. Fuchs, *From neuron to brain.* Sinauer Associates Sunderland, MA, 2001, vol. 271.

[44] P. L. Nunez, R. Srinivasan *et al.*, *Electric fields of the brain: the neurophysics of EEG.* Oxford University Press, USA, 2006.

[45] R. Charles and J. Nixon, “Measuring mental workload using physiological measures: A systematic review,” *Applied ergonomics*, vol. 74, pp. 221–232, 2019.

[46] M. Palmiero and L. Piccardi, “Frontal eeg asymmetry of mood: A mini-review,” *Frontiers in Behavioral Neuroscience*, vol. 11, p. 224, 2017.

[47] D. Huang, C. Guan, K. K. Ang, H. Zhang, and Y. Pan, “Asymmetric spatial pattern for eeg-based emotion detection,” in *The 2012 International Joint Conference on Neural Networks (IJCNN).* IEEE, 2012, pp. 1–7.

[48] R. Gupta, T. H. Falk *et al.*, “Relevance vector classifier decision fusion and eeg graph-theoretic features for automatic affective state characterization,” *Neurocomputing*, vol. 174, pp. 875–884, 2016.

[49] A. Martínez-Rodrigo, B. García-Martínez, R. Alcaraz, P. González, and A. Fernández-Caballero, "Multiscale entropy analysis for recognition of visually elicited negative stress from eeg recordings," *International journal of neural systems*, vol. 29, no. 02, p. 1850038, 2019.

[50] H. T. Haverkamp, S. O. Fosse, and P. Schuster, "Accuracy and usability of single-lead ecg from smartphones-a clinical study," *Indian pacing and electrophysiology journal*, vol. 19, no. 4, pp. 145–149, 2019.

[51] J. C. Himmelreich, E. P. Karregat, W. A. Lucassen, H. C. van Weert, J. R. de Groot, M. L. Handoko, R. Nijveldt, and R. E. Harskamp, "Diagnostic accuracy of a smartphone-operated, single-lead electrocardiography device for detection of rhythm and conduction abnormalities in primary care," *The Annals of Family Medicine*, vol. 17, no. 5, pp. 403–411, 2019.

[52] D. Nepi, A. Sbrollini, A. Agostinelli, E. Maranesi, M. Morettini, F. Di Nardo, S. Fioretti, P. Pierleoni, L. Pernini, S. Valenti *et al.*, "Validation of the heart-rate signal provided by the zephyr bioharness 3.0," in *2016 Computing in Cardiology Conference (CinC)*. IEEE, 2016, pp. 361–364.

[53] B. Singh, D. Singh, A. Jaryal, and K. Deepak, "Ectopic beats in approximate entropy and sample entropy-based hrv assessment," *International Journal of Systems Science*, vol. 43, no. 5, pp. 884–893, 2012.

[54] F. Shaffer and J. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, p. 258, 2017.

[55] A. Camm *et al.*, "Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[56] L. Pecchia, R. Castaldo, L. Montesinos, and P. Melillo, "Are ultra-short heart rate variability features good surrogates of short-term ones? state-of-the-art review and recommendations," *Healthcare technology letters*, vol. 5, no. 3, pp. 94–100, 2018.

[57] P. Karthikeyan, M. Murugappan, and S. Yaacob, "Ecg signal denoising using wavelet thresholding techniques in human stress assessment," *International Journal on Electrical Engineering and Informatics*, vol. 4, no. 2, p. 306, 2012.

[58] S. Cuomo, R. Farina, and F. Piccialli, "An inverse bayesian scheme for the denoising of ecg signals," *Journal of Network and Computer Applications*, vol. 115, pp. 48–58, 2018.

[59] D. P. Tobón and T. H. Falk, "Adaptive spectro-temporal filtering for electrocardiogram signal enhancement," *IEEE J of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 421–428, 2018.

[60] L. Howell and B. Porr, "High precision ecg database with annotated r peaks, recorded and filmed under realistic conditions," 2018.

[61] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Trans. Biomed. Eng*, vol. 32, no. 3, pp. 230–236, 1985.

[62] L. Sathyapriya, L. Murali, and T. Manigandan, "Analysis and detection r-peak detection using modified pan-tompkins algorithm," in *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*. IEEE, 2014, pp. 483–487.

[63] M. Waser and H. Garn, "Removing cardiac interference from the electroencephalogram using a modified pan-tompkins algorithm and linear regression," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 2028–2031.

[64] F. Liu, S. Wei, Y. Li, X. Jiang, Z. Zhang, L. Zhang, and C. Liu, "The accuracy on the common pan-tompkins based qrs detection methods through low-quality electrocardiogram database," *Journal of Medical Imaging and Health Informatics*, vol. 7, no. 5, pp. 1039–1043, 2017.

[65] A. L. Goldberger, L. A. Amaral, J. M. Hausdorff, P. C. Ivanov, C.-K. Peng, and H. E. Stanley, "Fractal dynamics in physiology: alterations with disease and aging," *Proceedings of the national academy of sciences*, vol. 99, no. suppl 1, pp. 2466–2472, 2002.

[66] I. Papousek, K. Nauschnegg, M. Paechter, H. K. Lackner, N. Goswami, and G. Schulter, "Trait and state positive affect and cardiovascular recovery from experimental academic stress," *Biological Psychology*, vol. 83, no. 2, pp. 108–115, 2010.

[67] E. M. Medica-Torino, "Effects of anxiety due to mental stress on heart rate variability in healthy subjects," *Minerva Psichiatr*, vol. 52, pp. 227–231, 2011.

[68] Z. Visnovcova, M. Mestanik, M. Javorka, D. Mokra, M. Gala, A. Jurko, A. Calkovska, and I. Tonhajzerova, "Complexity and time asymmetry of heart rate variability are altered in acute mental stress," *Physiological measurement*, vol. 35, no. 7, p. 1319, 2014.

[69] A. Sengupta, A. Tiwari, A. Chaudhuri, and A. Routray, "Analysis of loss of alertness due to cognitive fatigue using motif synchronization of eeg records," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 1652–1655.

[70] D. Dimitriev, E. Saperova, and A. Dimitriev, "State anxiety and nonlinear dynamics of heart rate variability in students," *PloS one*, vol. 11, no. 1, p. e0146131, 2016.

[71] M. Zubair and C. Yoon, "Multilevel mental stress detection using ultra-short pulse rate variability series," *Biomedical Signal Processing and Control*, vol. 57, p. 101736, 2020.

[72] I. Homma and Y. Masaoka, "Breathing rhythms and emotions," *Experimental physiology*, vol. 93, no. 9, pp. 1011–1021, 2008.

[73] Y. Masaoka and I. Homma, "Anxiety and respiratory patterns: their relationship during mental stress and physical load," *International Journal of Psychophysiology*, vol. 27, no. 2, pp. 153–159, 1997.

[74] Y. Masaoka and I. Homma, "The effect of anticipatory anxiety on breathing and metabolism in humans," *Respiration physiology*, vol. 128, no. 2, pp. 171–177, 2001.

[75] E. Vlemincx, I. Van Diest, and O. Van den Bergh, "A sigh following sustained attention and mental stress: effects on respiratory variability," *Physiology & behavior*, vol. 107, no. 1, pp. 1–6, 2012.

[76] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *IEEE Transactions on information technology in biomedicine*, vol. 14, no. 2, pp. 410–417, 2009.

[77] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, "Galvanic skin response (gsr) as an index of cognitive load," in *CHI'07 extended abstracts on Human factors in computing systems*, 2007, pp. 2651–2656.

[78] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.

[79] J.-P. Thiran, F. Marques, and H. Bourlard, *Multimodal Signal Processing: Theory and applications for human-computer interaction.* Academic Press, 2009.

[80] G. Valenza, A. Lanata, and E. P. Scilingo, "Oscillations of heart rate and respiration synchronize during affective visual stimulation," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 683–690, 2012.

[81] G. Valenza, A. Lanatá, and E. P. Scilingo, "Improving emotion recognition systems by embedding cardiorespiratory coupling," *Physiological measurement*, vol. 34, no. 4, p. 449, 2013.

[82] H. Lackner, I. Papousek, J. Batzel, A. Roessler, H. Scharfetter, and H. Szalkay, "Phase synchronization of hemodynamic variables and respiration during mental challenge," *International journal of psychophysiology*, vol. 79, no. 3, pp. 401–409, 2011.

[83] E. Kroupi, J.-M. Vesin, and T. Ebrahimi, "Implicit affective profiling of subjects based on physiological data coupling," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 85–98, 2014.

[84] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using eeg and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* IEEE, 2014, pp. 5040–5043.

[85] R. Gupta, M. Khomami Abadi, J. A. Cárdenes Cabré, F. Morreale, T. H. Falk, and N. Sebe, "A quality adaptive multimodal affect recognition system for user-centric multimedia indexing," in *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, 2016, pp. 317–320.

[86] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1–45, 2017.

[87] F. Al-Shargie, T. B. Tang, N. Badruddin, and M. Kiguchi, "Towards multilevel mental stress assessment using svm with ecoc: an eeg approach," *Medical & biological engineering & computing*, vol. 56, no. 1, pp. 125–136, 2018.

[88] B. Cinaz, B. Arnrich, R. La Marca, and G. Tröster, "Monitoring of mental workload levels during an everyday life office-work scenario," *Personal and ubiquitous computing*, vol. 17, no. 2, pp. 229–239, 2013.

[89] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using matthews correlation coefficient metric," *PloS one*, vol. 12, no. 6, p. e0177678, 2017.

[90] Y.-H. Yang and H. H. Chen, *Music emotion recognition.* CRC Press, 2011.

[91] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[92] M. Ilves and V. Surakka, "Heart rate responses to synthesized affective spoken words," *Advances in Human-Computer Interaction*, vol. 2012, 2012.

[93] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart rate variability," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 385–394, 2015.

[94] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.

[95] A. T. Sohaib, S. Qureshi, J. Hagelbäck, O. Hilborn, and P. Jerčić, "Evaluating classifiers for emotion recognition using eeg," in *International conference on augmented cognition.* Springer, 2013, pp. 492–501.

[96] A. Yazdani, J.-S. Lee, J.-M. Vesin, and T. Ebrahimi, "Affect recognition based on physiological changes during the watching of music videos," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 1, pp. 1–26, 2012.

[97] P. J. Lang, M. M. Bradley, B. N. Cuthbert *et al.*, "International affective picture system (iaps): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, pp. 39–58, 1997.

[98] M. M. Bradley and P. J. Lang, "The international affective digitized sounds (; iads-2): Affective ratings of sounds and instruction manual," *University of Florida, Gainesville, FL, Tech. Rep. B-3*, 2007.

[99] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[100] K. Dedovic, R. Renwick, N. K. Mahani, V. Engert, S. J. Lupien, and J. C. Pruessner, "The montreal imaging stress task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain," *Journal of Psychiatry and Neuroscience*, vol. 30, no. 5, p. 319, 2005.

[101] A. R. Jensen and W. D. Rohwer Jr, "The stroop color-word test: a review," *Acta psychologica*, vol. 25, pp. 36–93, 1966.

[102] R. N. Shepard and J. Metzler, "Mental rotation of three-dimensional objects," *Science*, vol. 171, no. 3972, pp. 701–703, 1971.

[103] Y. Noto, T. Sato, M. Kudo, K. Kurata, and K. Hirota, "The relationship between salivary biomarkers and state-trait anxiety inventory score under mental arithmetic stress: a pilot study." *Anesthesia & Analgesia*, vol. 101, no. 6, pp. 1873–1876, 2005.

[104] H. Mansikka, K. Virtanen, and D. Harris, "Comparison of nasa-tlx scale, modified cooper–harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks," *Ergonomics*, vol. 62, no. 2, pp. 246–254, 2019.

[105] X.-Q. Huo, W.-L. Zheng, and B.-L. Lu, "Driving fatigue detection with fusion of eeg and forehead eog," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 897–904.

[106] Z. Cao, C.-H. Chuang, J.-K. King, and C.-T. Lin, "Multi-channel eeg recordings during a sustained-attention driving task," *Scientific data*, vol. 6, no. 1, pp. 1–8, 2019.

[107] Z. Li, H. Snieder, S. Su, X. Ding, J. F. Thayer, F. A. Treiber, and X. Wang, "A longitudinal study in youth of heart rate variability at rest and in response to stress," *International Journal of Psychophysiology*, vol. 73, no. 3, pp. 212–217, 2009.

[108] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, and G. Valenza, "Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors," *Scientific reports*, vol. 8, no. 1, pp. 1–15, 2018.

[109] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr, "The multi-attribute task battery II (MATB-II) software for human performance and workload research: A user's guide," 2011.

[110] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[111] J. A. Healey, "Wearable and automotive systems for affect recognition from physiology," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[112] R. Cassani, H. Banville, and T. Falk, "MuLES: An open source EEG acquisition and streaming server for quick and simple prototyping and recording," in *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, 2015, pp. 9–12.

[113] M. Parent, I. Albuquerque, A. Tiwari, R. Cassani, J. Gagnon, D. Lafond, S. Tremblay, and T. Falk, "PASS: A multimodal database of physical activity and stress for mobile passive body/brain-computer interface research," *Frontiers in Neuroscience*, vol. 14, p. 1274, 2020.

[114] K. Mundnich, B. Booth, M. l'Hommedieu, T. Feng, B. Girault, J. L'Hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. Villatte *et al.*, "TILES-2018: A longitudinal physiologic and behavioral data set of hospital workers," *arXiv preprint arXiv:2003.08474*, 2020.

[115] T. Feng, A. Nadarajan, C. Vaz, B. Booth, and S. Narayanan, "Tiles audio recorder: an unobtrusive wearable solution to track audio activity," in *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, 2018, pp. 33–38.

[116] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in eeg-bci for daily-life: Requirements for artifact removal," *Biomedical Signal Processing and Control*, vol. 31, pp. 407–418, 2017.

[117] G. Ouyang, C. Dang, D. A. Richards, and X. Li, "Ordinal pattern based similarity analysis for eeg recordings," *Clinical Neurophysiology*, vol. 121, no. 5, pp. 694–703, 2010.

[118] R. Rosário, P. Cardoso, M. Muñoz, P. Montoya, and J. Miranda, "Motif-synchronization: A new method for analysis of dynamic brain networks with eeg," *Physica A: Statistical Mechanics and its Applications*, vol. 439, pp. 7–19, 2015.

[119] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[120] N. Kannathal, M. L. Choo, U. R. Acharya, and P. Sadasivan, "Entropies for detection of epilepsy in eeg," *Computer methods and programs in biomedicine*, vol. 80, no. 3, pp. 187–194, 2005.

[121] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[122] S. Koelstra and I. Patras, "Fusion of facial expressions and eeg for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, pp. 164–174, 2013.

[123] A. M. Oliveira, M. P. Teixeira, I. B. Fonseca, and M. Oliveira, "Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity." *Proceedings of Fechner Day*, vol. 22, no. 1, pp. 245–250, 2006.

[124] N. Alvarado, "Arousal and valence in the direct scaling of emotional response to film clips," *Motivation and Emotion*, vol. 21, no. 4, pp. 323–348, 1997.

[125] R. D. Lane and L. Nadel, *Cognitive neuroscience of emotion.* Oxford University Press, 1999.

[126] P. A. Lewis, H. Critchley, P. Rotshtein, and R. Dolan, "Neural correlates of processing valence and arousal in affective words," *Cerebral cortex*, vol. 17, no. 3, pp. 742–748, 2006.

[127] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical review letters*, vol. 89, no. 6, p. 068102, 2002.

[128] A. Humeau-Heurtier, "The multiscale entropy algorithm and its variants: A review," *Entropy*, vol. 17, no. 5, pp. 3110–3123, 2015.

[129] S.-D. Wu, P.-H. Wu, C.-W. Wu, J.-J. Ding, and C.-C. Wang, "Bearing fault diagnosis based on multiscale permutation entropy and support vector machine," *Entropy*, vol. 14, no. 8, pp. 1343–1356, 2012.

[130] S.-D. Wu, C.-W. Wu, K.-Y. Lee, and S.-G. Lin, "Modified multiscale entropy for short-term time series analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 23, pp. 5865–5873, 2013.

[131] S.-D. Wu, C.-W. Wu, S.-G. Lin, C.-C. Wang, and K.-Y. Lee, "Time series analysis using composite multiscale entropy," *Entropy*, vol. 15, no. 3, pp. 1069–1084, 2013.

[132] Q. Li and F. Zuntao, "Permutation entropy and statistical complexity quantifier of nonstationarity effect in the vertical velocity records," *Physical Review E*, vol. 89, no. 1, p. 012905, 2014.

[133] C. Bian, C. Qin, Q. D. Ma, and Q. Shen, "Modified permutation-entropy analysis of heartbeat dynamics," *Physical Review E*, vol. 85, no. 2, p. 021906, 2012.

[134] B. Fadlallah, B. Chen, A. Keil, and J. Príncipe, "Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information," *Physical Review E*, vol. 87, no. 2, p. 022911, 2013.

[135] M. Costa and A. Goldberger, "Generalized multiscale entropy analysis: application to quantifying the complex volatility of human heartbeat time series," *Entropy*, vol. 17, no. 3, pp. 1197–1203, 2015.

[136] J. Behar, A. Johnson, G. D. Clifford, and J. Oster, "A comparison of single channel fetal ecg extraction methods," *Annals of biomedical engineering*, vol. 42, no. 6, pp. 1340–1353, 2014.

[137] A. Porta, T. Gnecchi-Ruscone, E. Tobaldini, S. Guzzetti, R. Furlan, and N. Montano, "Progressive decrease of heart period variability entropy-based complexity during graded head-up tilt," *Journal of applied physiology*, vol. 103, no. 4, pp. 1143–1149, 2007.

[138] A. Porta, P. Castiglioni, V. Bari, T. Bassani, A. Marchi, A. Cividjian, L. Quintin, and M. Di Rienzo, "K-nearest-neighbor conditional entropy approach for the assessment of the short-term complexity of cardiovascular control," *Physiological measurement*, vol. 34, no. 1, p. 17, 2012.

[139] M. Weippert, M. Behrens, A. Rieger, and K. Behrens, "Sample entropy and traditional measures of heart rate dynamics reveal different modes of cardiovascular control during low intensity exercise," *Entropy*, vol. 16, no. 11, pp. 5698–5711, 2014.

[140] G. Wu, N. Arzeno, L. Shen, D. Tang, D. Zheng, N. Zhao, D. Eckberg, and C. Poon, "Chaotic signatures of heart rate variability and its power spectrum in health, aging and heart failure," *PloS one*, vol. 4, no. 2, p. e4323, 2009.

[141] D. Luo, W. Pan, Y. Li, K. Feng, and G. Liu, "The interaction analysis between the sympathetic and parasympathetic systems in CHF by using transfer entropy method," *Entropy*, vol. 20, no. 10, p. 795, 2018.

[142] L. Zheng, W. Pan, Y. Li, D. Luo, Q. Wang, and G. Liu, "Use of mutual information and transfer entropy to assess interaction between parasympathetic and sympathetic activities of nervous system from HRV," *Entropy*, vol. 19, no. 9, p. 489, 2017.

[143] A. Pichon, C. Bisschop, M. Roulaud, A. Denjean, and Y. Papelier, "Spectral analysis of heart rate variability during exercise in trained subjects," *Medicine and science in sports and exercise*, vol. 36, pp. 1702–1708, 2004.

[144] S. Shao, T. Wang, C. Song, X. Chen, E. Cui, and H. Zhao, "Obstructive sleep apnea recognition based on multi-bands spectral entropy analysis of short-time heart rate variability," *Entropy*, vol. 21, no. 8, p. 812, 2019.

[145] K. Ikegwu, J. Trauger, J. McMullin, and R. Brunner, "PyIF: A fast and light weight implementation to estimate bivariate transfer entropy for big data," in *2020 SoutheastCon.* IEEE, 2020, pp. 1–6.

[146] J. Behar, A. Rosenberg, I. Weiser-Bitoun, O. Shemla, A. Alexandrovich, E. Konyukhov, and Y. Yaniv, "PhysioZoo: A novel open access platform for heart rate variability analysis of mammalian electrocardiographic data," *Frontiers in physiology*, vol. 9, p. 1390, 2018.

[147] A. Tiwari, S. Narayanan, and T. Falk, "Stress and anxiety measurement" in-the-wild" using quality-aware multi-scale HRV features," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* IEEE, 2019, pp. 7056–7059.

[148] A. tiwari, S. Narayanan, and T. Falk, "Breathing rate complexity features for "in-the-wild" stress and anxiety measurement," in *2019 27th European Signal Processing Conference (EUSIPCO).* IEEE, 2019, pp. 1–5.

[149] C. Grillon, R. Duncko, M. Covington, L. Kopperman, and M. Kling, "Acute stress potentiates anxiety in humans," *Biological psychiatry*, vol. 62, no. 10, pp. 1183–1186, 2007.

[150] N. P. Castellanos and V. A. Makarov, "Recovering EEG brain signals: Artifact suppression with wavelet enhanced independent component analysis," *Journal of Neuroscience Methods*, vol. 158, no. 2, pp. 300–312, Dec. 2006.

[151] O. Rosanne, I. Albuquerque, J.-F. Gagnon, S. Tremblay, and T. H. Falk, "Performance comparison of automated eeg enhancement algorithms for mental workload assessment of ambulant users," in *IEEE/EMBS Conf Neural Engineering*, 2019, pp. 61–64.

[152] D. Carroll, A. T. Ginty, G. Der, K. Hunt, M. Benzeval, and A. C. Phillips, "Increased blood pressure reactions to acute mental stress are associated with 16-year cardiovascular disease mortality," *Psychophysiology*, vol. 49, no. 10, pp. 1444–1448, 2012.

[153] T. Chandola, A. Heraclides, and M. Kumari, "Psychophysiological biomarkers of workplace stressors," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 1, pp. 51–57, 2010.

[154] J. Xiong, O. Lipsitz, F. Nasri, L. M. Lui, H. Gill, L. Phan, D. Chen-Li, M. Iacobucci, R. Ho, A. Majeed *et al.*, "Impact of covid-19 pandemic on mental health in the general population: A systematic review," *Journal of affective disorders*, 2020.

[155] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, C. T. Pike, and R. C. Kessler, "The economic burden of adults with major depressive disorder in the united states (2005 and 2010)," *The Journal of clinical psychiatry*, vol. 76, no. 2, pp. 155–162, 2015.

[156] D. A. Revicki, K. Travers, K. W. Wyrwich, H. Svedsäter, J. Locklear, M. S. Mattera, D. V. Sheehan, and S. Montgomery, "Humanistic and economic burden of generalized anxiety disorder in north america and europe," *Journal of affective disorders*, vol. 140, no. 2, pp. 103–112, 2012.

[157] P. D. McGorry, A. Ratheesh, and B. O'Donoghue, "Early intervention—an implementation challenge for 21st century mental health care," *JAMA psychiatry*, vol. 75, no. 6, pp. 545–546, 2018.

[158] J. Nash-Wright, "Dealing with anxiety disorders in the workplace: importance of early intervention when anxiety leads to absence from work," *Professional case management*, vol. 16, no. 2, pp. 55–59, 2011.

[159] T. Matsuo, D. Kobayashi, F. Taki, F. Sakamoto, Y. Uehara, N. Mori, and T. Fukui, "Prevalence of health care worker burnout during the coronavirus disease 2019 (covid-19) pandemic in japan," *JAMA network open*, vol. 3, no. 8, pp. e2 017 271–e2 017 271, 2020.

[160] S. Barello, L. Palamenghi, and G. Graffigna, "Burnout and somatic symptoms among frontline healthcare professionals at the peak of the italian covid-19 pandemic." *Psychiatry research*, vol. 290, p. 113129, 2020.

[161] I. Duarte, A. Teixeira, L. Castro, S. Marina, C. Ribeiro, C. Jácome, V. Martins, I. Ribeiro-Vaz, H. C. Pinheiro, A. R. Silva *et al.*, "Burnout among portuguese healthcare workers during the covid-19 pandemic," *BMC public health*, vol. 20, no. 1, pp. 1–10, 2020.

[162] G. F. Wilson and C. A. Russell, "Operator functional state classification using multiple psychophysiological features in an air traffic control task," *Human Factors*, vol. 45, no. 3, pp. 381–389, 2003.

[163] K. Laghari, R. Gupta, S. Arndt, J. Antons, R. Schleicher, S. Moller, and T. Falk, "Neurophysiological experimental facility for quality of experience (qoe) assessment," in *Proceedings of International Conference on Quality of Experience Centric Management (QCMan)*, 2013, pp. 1300–1305.

[164] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[165] M. Yadava, P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra, "Analysis of eeg signals and its application to neuromarketing," *Multimedia Tools and Applications*, vol. 76, no. 18, pp. 19 087–19 111, 2017.

[166] C. M. Whissell, "The dictionary of affect in language," in *The measurement of emotions*. Elsevier, 1989, pp. 113–131.

[167] H. Schlosberg, "Three dimensions of emotion." *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[168] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[169] G. Matthews and S. E. Campbell, "Dynamic relationships between stress states and working memory," *Cognition and emotion*, vol. 24, no. 2, pp. 357–373, 2010.

[170] J. W. Hinton, E. Rotheiler, and A. Howard, "Confusion between stress and state anxiety in a much used self-report 'stress' inventory," *Personality and individual differences*, vol. 12, no. 1, pp. 91–94, 1991.

[171] J. Torous, P. Staples, M. Shanahan, C. Lin, P. Peck, M. Keshavan, and J.-P. Onnela, "Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (phq-9) depressive symptoms in patients with major depressive disorder," *JMIR mental health*, vol. 2, no. 1, p. e8, 2015.

[172] K. Fujibayashi, H. Takahashi, M. Tanei, Y. Uehara, H. Yokokawa, and T. Naito, "A new influenza-tracking smartphone app (flu-report) based on a self-administered questionnaire: cross-sectional study," *JMIR mHealth and uHealth*, vol. 6, no. 6, p. e136, 2018.

[173] H. Sato and J.-i. Kawahara, "Selective bias in retrospective self-reports of negative mood states," *Anxiety, Stress & Coping*, vol. 24, no. 4, pp. 359–367, 2011.

[174] N. Mor and J. Winquist, "Self-focused attention and negative affect: A meta-analysis." *Psychological bulletin*, vol. 128, no. 4, p. 638, 2002.

[175] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*. IEEE, 2005, pp. 5–pp.

[176] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *18th International conference on pattern recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 1148–1153.

[177] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, 2010.

[178] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, 2018, pp. 3–13.

[179] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats." in *Interspeech*, 2018, pp. 122–126.

[180] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[181] R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, and R. Caldara, "Cultural confusions show that facial expressions are not universal," *Current biology*, vol. 19, no. 18, pp. 1543–1548, 2009.

[182] O. Rosanne, I. Albuquerque, R. Cassani, J.-F. Gagnon, S. Tremblay, and T. H. Falk, "Adaptive filtering for improved eeg-based mental workload assessment of ambulant users," *Frontiers in Neuroscience*, vol. 15, p. 341, 2021.

[183] T. H. Falk, M. Maier *et al.*, "Ms-qi: A modulation spectrum-based ecg quality index for telehealth applications," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1613–1622, 2014.

[184] R. Simpson, J. Langtree, and A. Mitchell, "Ectopic beats: How many count?" *EMJ Cardiology*, 2017.

[185] A. Eke, P. Herman, L. Kocsis, and L. Kozak, "Fractal characterization of complexity in temporal physiological signals," *Physiological measurement*, vol. 23, no. 1, p. R1, 2002.

[186] A. Bilan, A. Witczak, R. Palusiński, W. Myśliński, and J. Hanzlik, "Circadian rhythm of spectral indices of heart rate variability in healthy subjects," *Journal of electrocardiology*, vol. 38, no. 3, pp. 239–243, 2005.

[187] R. Perini and A. Veicsteinas, "Heart rate variability and autonomic activity at rest and during exercise in various physiological conditions," *European journal of applied physiology*, vol. 90, no. 3, pp. 317–325, 2003.

[188] M. Muehlhan, M. Marxen, J. Landsiedel, H. Malberg, and S. Zaunseder, "The effect of body posture on cognitive performance: a question of sleep quality," *Frontiers in human neuroscience*, vol. 8, p. 171, 2014.

[189] M. Causse, Z. Chua, V. Peysakhovich, N. Del Campo, and N. Matton, "Mental workload and neural efficiency quantified in the prefrontal cortex using fnirs," *Scientific reports*, vol. 7, no. 1, pp. 1–15, 2017.

[190] Z. M. Aghajan, P. Schuette, T. A. Fields, M. E. Tran, S. M. Siddiqui, N. R. Hasulak, T. K. Tcheng, D. Eliashiv, E. A. Mankin, J. Stern *et al.*, "Theta oscillations in the human medial temporal lobe during real-world ambulatory movement," *Current Biology*, vol. 27, no. 24, pp. 3743–3751, 2017.

[191] S. Fuchs, U. D. Reichel, and A. Rochet-Capellan, "Changes in speech and breathing rate while speaking and biking," 2015.

[192] A. Tiwari and T. Falk, "Fusion of motif-and spectrum-related features for improved EEG-based emotion recognition," *Computational intelligence and neuroscience*, vol. 2019, 2019.

[193] A. Tiwari, I. Albuquerque, M. Parent, J. Gagnon, D. Lafond, S. Tremblay, and T. Falk, "Multi-scale heart beat entropy measures for mental workload assessment of ambulant users," *Entropy*, vol. 21, no. 8, p. 783, 2019.

[194] A. Tiwari and T. H. Falk, "New measures of heart rate variability based on subband tachogram complexity and spectral characteristics for improved stress and anxiety monitoring in highly ecological settings," 2021, in preparation, to be submitted to IEEE Journal of Biomedical And Health Informatics.

[195] A. Tiwari, R. Cassani, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "Prediction of stress and mental workload during police academy training using ultra-short-term heart rate variability and breathing analysis," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).* IEEE, 2020, pp. 4530–4533.

[196] A. Tiwari, R. Cassani, J. Gagnon, D. Lafond, S. Tremblay, and T. Falk, "Movement artifact-robust mental workload assessment during physical activity using multi-sensor fusion," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC).* IEEE, 2020, pp. 3471–3477.

[197] A. Clerico, A. Tiwari, R. Gupta, S. Jayaraman, and T. H. Falk, "Electroencephalography amplitude modulation analysis for automated affective tagging of music video clips," *Frontiers in computational neuroscience*, vol. 11, p. 115, 2018.

[198] A. Tiwari and T. H. Falk, "Lossless electrocardiogram signal compression: a review of existing methods," *Biomedical Signal Processing and Control*, vol. 51, pp. 338–346, 2019.

[199] I. Albuquerque, A. Tiwari, M. Parent, R. Cassani, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "Wauc: a multi-modal database for mental workload assessment under physical activity," *Frontiers in Neuroscience*, vol. 14, 2020.

[200] I. Albuquerque, A. Tiwari, J.-F. Gagnon, D. Lafond, M. Parent, S. Tremblay, and T. Falk, "On the analysis of eeg features for mental workload assessment during physical activity," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC).* IEEE, 2018, pp. 538–543.

[201] A. Tiwari, I. Albuquerque, M. Parent, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "A comparison of two ecg inter-beat interval measurement methods for hrv-based mental-workload prediction of ambulant users," *CMBES Proceedings*, vol. 42, 2019.

[202] A. Tiwari, R. Cassani, S. Narayanan, and T. H. Falk, "A comparative study of stress and anxiety estimation in ecological settings using a smart-shirt and a smart-bracelet," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* IEEE, 2019, pp. 2213–2216.

[203] A. Tiwari, I. Albuquerque, J.-F. Gagnon, D. Lafond, M. Parent, S. Tremblay, and T. H. Falk, "Mental workload assessment during physical activity using non-linear movement artefact robust electroencephalography features," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC).* IEEE, 2019, pp. 4149–4154.

[204] A. R. Avila, S. R. Kshirsagar, A. Tiwari, D. Lafond, D. O'Shaughnessy, and T. H. Falk, "Speech-based stress classification based on modulation spectral features and convolutional neural networks," in *2019 27th European Signal Processing Conference (EUSIPCO).* IEEE, 2019, pp. 1–5.

[205] A. Tiwari, J. L. Villatte, S. Narayanan, and T. H. Falk, "Prediction of psychological flexibility with multi-scale heart rate variability and breathing features in an "in-the-wild" setting," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW).* IEEE, 2019, pp. 297–303.

[206] I. Albuquerque, J. Monteiro, O. Rosanne, A. Tiwari, J.-F. Gagnon, and T. H. Falk, "Cross-subject statistical shift estimation for generalized electroencephalography-based mental workload assessment," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC).* IEEE, 2019, pp. 3647–3653.

[207] M. Parent, A. Tiwari, I. Albuquerque, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "A multimodal approach to improve the robustness of physiological stress prediction during physical activity," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC).* IEEE, 2019, pp. 4131–4136.

[208] R. Cassani, A. Tiwari, and T. H. Falk, “Optimal filter characterization for photoplethysmography-based pulse rate and pulse power spectrum estimation,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).* IEEE, 2020, pp. 914–917.

[209] R. Cassani, A. Tiwari, I. Posner, B. Afonso, and T. H. Falk, “Initial investigation into neurophysiological correlates of argentine tango flow states: a case study,” in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC).* IEEE, 2020, pp. 3478–3483.

[210] A. Pimentel, A. Tiwari, and T. H. Falk, “Human mental state monitoring in the wild: Are we better off with deeper neural networks or improved input features?” *CMBES Proceedings*, vol. 44, 2021.

[211] A. Tiwari, S. Liaqat, D. Liaqat, M. Gabel, E. de Lara, and T. Falk, “Remote copd severity and exacerbation detection using heart rate and activity data measured from a wearable device,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* IEEE, 2021.

[212] G. Buzsáki, C. A. Anastassiou, and C. Koch, “The origin of extracellular fields and currents—eeg, ecog, lfp and spikes,” *Nature reviews neuroscience*, vol. 13, no. 6, pp. 407–420, 2012.

[213] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun, “Short-term emotion assessment in a recall paradigm,” *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, 2009.

[214] J. A. Urigüen and B. Garcia-Zapirain, “Eeg artifact removal—state-of-the-art and guidelines,” *Journal of neural engineering*, vol. 12, no. 3, p. 031001, 2015.

[215] S. Sanei and J. A. Chambers, *EEG signal processing.* John Wiley & Sons, 2013.

[216] F. Al-Shargie, M. Kiguchi, N. Badruddin, S. C. Dass, A. F. M. Hani, and T. B. Tang, “Mental stress assessment using simultaneous measurement of eeg and fnirs,” *Biomedical optics express*, vol. 7, no. 10, pp. 3882–3898, 2016.

[217] W. Chen, “Electrocardiogram,” in *Seamless Healthcare Monitoring.* Springer, 2018, pp. 3–44.

[218] L. Sörnmo and P. Laguna, *Bioelectrical signal processing in cardiac and neurological applications.* Academic Press, 2005, vol. 8.

[219] C. E. Kossmann, D. A. Brody, G. E. Burch, H. H. Hecht, F. D. Johnston, C. Kay, E. Lepeschkin, H. V. Pipberger, G. Baule, A. S. Berson *et al.*, “Recommendations for standardization of leads and of specifications for instruments in electrocardiography and vectorcardiography,” *Circulation*, vol. 35, no. 3, pp. 583–602, 1967.

[220] R. McFee and G. M. Baule, “Research in electrocardiography and magnetocardiography,” *Proceedings of the IEEE*, vol. 60, no. 3, pp. 290–321, March 1972.

[221] D. Kilpatrick and P. R. Johnston, “Origin of the electrocardiogram,” *IEEE Engineering in Medicine and Biology Magazine*, vol. 13, no. 4, pp. 479–486, Aug 1994.

[222] A. Oswald, “At the heart of the invention: The development of the holter monitor,” *National Museum of American History*, vol. 8, no. 01, 2014.

[223] J. Peake, G. Kerr, and J. Sullivan, “A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations,” *Frontiers in physiology*, vol. 9, p. 743, 2018.

[224] J. T. Bigger Jr, P. Albrecht, R. C. Steinman, L. M. Rolnitzky, J. L. Fleiss, and R. J. Cohen, “Comparison of time-and frequency domain-based measures of cardiac parasympathetic activity in holter recordings after myocardial infarction,” *The American journal of cardiology*, vol. 64, no. 8, pp. 536–538, 1989.

[225] L. Fei, X. Copie, M. Malik, and A. J. Camm, "Short-and long-term assessment of heart rate variability for risk stratification after acute myocardial infarction," *The American journal of cardiology*, vol. 77, no. 9, pp. 681–684, 1996.

[226] M. V. Kamath, M. Watanabe, and A. Upton, "Heart rate variability (hrv) signal analysis: clinical applications," 2012.

[227] G. Han, B. Lin, and Z. Xu, "Electrocardiogram signal denoising based on empirical mode decomposition technique: an overview," *Journal of Instrumentation*, vol. 12, no. 03, p. P03010, 2017.

[228] V. Marked, "Correction of the heart rate variability signal for ectopics and missing beats," *Heart rate variability*, 1995.

[229] C.-C. Chang, T.-C. Hsiao, Y. Chiang, and H. Hsu, "The usefulness of the coefficient of variation of electrocar-diographic rr interval as an index of cardiovascular function and its correlation with age and stroke," *Tungs' Med. J.*, vol. 6, pp. 41–48, 2012.

[230] S. Iwasaki, J. Kozawa, K. Fukui, H. Iwahashi, A. Imagawa, and I. Shimomura, "Coefficient of variation of rr interval closely correlates with glycemic variability assessed by continuous glucose monitoring in insulin-depleted patients with type 1 diabetes," *Diabetes research and clinical practice*, vol. 109, no. 2, pp. 397–403, 2015.

[231] G. Baselli, S. Cerutti, S. Civardi, F. Lombardi, A. Malliani, M. Merri, M. Pagani, and G. Rizzo, "Heart rate variability signal processing: a quantitative approach as an aid to diagnosis in cardiovascular pathologies," *International journal of bio-medical computing*, vol. 20, no. 1-2, pp. 51–70, 1987.

[232] G. E. Billman, "The lf/hf ratio does not accurately measure cardiac sympatho-vagal balance," *Frontiers in physiology*, vol. 4, p. 26, 2013.

[233] F. Wang, H. Wang, and R. Fu, "Real-time ECG-based detection of fatigue driving using sample entropy," *Entropy*, vol. 20, no. 3, p. 196, 2018.

[234] M. Zanin, L. Zunino, O. A. Rosso, and D. Papo, "Permutation entropy and its main biomedical and econophysics applications: a review," *Entropy*, vol. 14, no. 8, pp. 1553–1577, 2012.

[235] E. Olofsen, J. Sleigh, and A. Dahan, "Permutation entropy of the electroencephalogram: a measure of anaesthetic drug effect," *British journal of anaesthesia*, vol. 101, no. 6, pp. 810–821, 2008.

[236] Y. Xia, L. Yang, L. Zunino, H. Shi, Y. Zhuang, and C. Liu, "Application of permutation entropy and permutation min-entropy in multiple emotional states analysis of rri time series," *Entropy*, vol. 20, no. 3, p. 148, 2018.

[237] N. Nicolaou and J. Georgiou, "The use of permutation entropy to characterize sleep electroencephalograms," *Clinical EEG and Neuroscience*, vol. 42, no. 1, pp. 24–28, 2011.

[238] X. Li, G. Ouyang, and D. A. Richards, "Predictability analysis of absence seizures with permutation entropy," *Epilepsy research*, vol. 77, no. 1, pp. 70–74, 2007.

[239] S. Nayak, A. Bit, A. Dey, B. Mohapatra, and K. Pal, "A review on the nonlinear dynamical system analysis of electrocardiogram signal," *Journal of healthcare engineering*, vol. 2018, 2018.

[240] M. Rosenstein, J. Collins, and C. Luca, "A practical method for calculating largest lyapunov exponents from small data sets," *Physica D: Nonlinear Phenomena*, vol. 65, no. 1-2, pp. 117–134, 1993.

[241] P. Grassberger and I. Procaccia, "Characterization of strange attractors," *Physical review letters*, vol. 50, no. 5, p. 346, 1983.

[242] D. P. Tobon, S. Jayaraman, and T. H. Falk, “Spectro-temporal electrocardiogram analysis for noise-robust heart rate and heart rate variability measurement,” *IEEE journal of translational engineering in health and medicine*, vol. 5, pp. 1–11, 2017.

[243] S. D. Kreibig, “Autonomic nervous system activity in emotion: A review,” *Biological psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[244] C.-K. Peng, J. E. Mietus, Y. Liu, C. Lee, J. M. Hausdorff, H. E. Stanley, A. L. Goldberger, and L. A. Lipsitz, “Quantifying fractal dynamics of human respiration: age and gender effects,” *Annals of biomedical engineering*, vol. 30, no. 5, pp. 683–692, 2002.

[245] W. Boucsein, *Electrodermal activity.* Springer Science & Business Media, 2012.

[246] R. Hefron, B. Borghetti, C. Schubert Kabban, J. Christensen, and J. Estepp, “Cross-participant eeg-based assessment of cognitive workload using multi-path convolutional recurrent neural networks,” *Sensors*, vol. 18, no. 5, p. 1339, 2018.

[247] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[248] X.-W. Wang, D. Nie, and B.-L. Lu, “Eeg-based emotion recognition using frequency domain features and support vector machines,” in *International Conference on Neural Information Processing.* Springer, 2011, pp. 734–743.

[249] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, “Deep learning-based electroencephalography analysis: a systematic review,” *Journal of neural engineering*, vol. 16, no. 5, p. 051001, 2019.

[250] D. Chicco and G. Jurman, “The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC genomics*, vol. 21, no. 1, p. 6, 2020.

[251] B. M. Kudielka, D. H. Hellhammer, and C. Kirschbaum, “Ten years of research with the trier social stress test–revisited.” 2007.

[252] A. M. Johnson, “Speed of mental rotation as a function of problem-solving strategies,” *Perceptual and motor skills*, vol. 71, no. 3, pp. 803–806, 1990.

[253] K. Miller, C. Price, M. Okun, H. Montijo, and D. Bowers, “Is the n-back task a valid neuropsychological measure for assessing working memory?” *Archives of Clinical Neuropsychology*, vol. 24, no. 7, pp. 711–717, 2009.

[254] A. Drouin-Picaro, I. Albuquerque, J.-F. Gagnon, D. Lafond, and T. H. Falk, “Eeg coupling features: Towards mental workload measurement based on wearables,” in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC).* IEEE, 2017, pp. 28–33.

[255] D. N. Cassenti, T. D. Kelley, and R. A. Carlson, “Modeling the workload-performance relationship,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54, no. 19. SAGE Publications Sage CA: Los Angeles, CA, 2010, pp. 1684–1688.

[256] N. Zhuang, Y. Zeng, L. Tong, C. Zhang, H. Zhang, and B. Yan, “Emotion recognition from eeg signals using multidimensional information in emd domain,” *BioMed research international*, vol. 2017, 2017.

[257] W. Liu, W.-L. Zheng, and B.-L. Lu, “Emotion recognition using multimodal deep learning,” in *International conference on neural information processing.* Springer, 2016, pp. 521–529.

[258] G. K. Verma and U. S. Tiwary, “Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals,” *NeuroImage*, vol. 102, pp. 162–172, 2014.

[259] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 291–298.

[260] N. J. Christensen and H. Galbo, "Sympathetic nervous activity during exercise," *Annual review of physiology*, vol. 45, no. 1, pp. 139–153, 1983.

[261] G. F. Wilson, "Air-to-ground training missions: a psychophysiological workload analysis," *Ergonomics*, vol. 36, no. 9, pp. 1071–1087, 1993.

[262] U. Parlitz, S. Berg, S. Luther, A. Schirdewan, J. Kurths, and N. Wessel, "Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics," *Computers in biology and medicine*, vol. 42, no. 3, pp. 319–327, 2012.

[263] C. Bandt and B. Pompe, "Permutation entropy: a natural complexity measure for time series," *Physical review letters*, vol. 88, no. 17, p. 174102, 2002.

[264] K. Keller, A. M. Unakafov, and V. A. Unakafova, "Ordinal patterns, entropy, and eeg," *Entropy*, vol. 16, no. 12, pp. 6212–6239, 2014.

[265] D. S. Bassett and E. Bullmore, "Small-world brain networks," *The neuroscientist*, vol. 12, no. 6, pp. 512–523, 2006.

[266] C. J. Stam, B. Jones, G. Nolte, M. Breakspear, and P. Scheltens, "Small-world networks and functional connectivity in alzheimer's disease," *Cerebral cortex*, vol. 17, no. 1, pp. 92–99, 2007.

[267] C. Lithari, M. A. Klados, C. Papadelis, C. Pappas, M. Albani, and P. D. Bamidis, "How does the metric choice affect brain functional connectivity networks?" *Biomedical Signal Processing and Control*, vol. 7, no. 3, pp. 228–236, 2012.

[268] J. A. Coan and J. J. Allen, "Frontal eeg asymmetry as a moderator and mediator of emotion," *Biological psychology*, vol. 67, no. 1-2, pp. 7–50, 2004.

[269] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[270] O. Ghitza, "Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm," *Frontiers in psychology*, vol. 2, p. 130, 2011.

[271] O. Ghitza, "On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum," *Frontiers in psychology*, vol. 3, p. 238, 2012.

[272] R. Gupta, H. J. Banville, and T. H. Falk, "Multimodal physiological quality-of-experience assessment of text-to-speech systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 22–36, 2017.

[273] L. Aftanas, A. Varlamov, S. Pavlov, V. Makhnev, and N. Reva, "Affective picture processing: event-related synchronization within individually defined human theta band is modulated by valence dimension." *Neuroscience Letters*, vol. 303, no. 2, pp. 115 – 118, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304394001017037

[274] S. Arndt, J.-N. Antons, R. Gupta, R. Schleicher, S. Möller, T. H. Falk *et al.*, "The effects of text-to-speech system quality on emotional states and frontal alpha band power," in *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on.* IEEE, 2013, pp. 489–492.

[275] L. I. Aftanas, A. A. Varlamov, S. V. Pavlov, V. P. Makhnev, and N. V. Reva, "Time-dependent cortical asymmetries induced by emotional arousal: Eeg analysis of event-related synchronization and desynchronization in individually defined frequency bands," *International Journal of Psychophysiology*, vol. 44, no. 1, pp. 67–82, 2002.

[276] R. J. Barry, A. R. Clarke, S. J. Johnstone, R. McCarthy, and M. Selikowitz, "Electroencephalogram θ/β ratio and arousal in attention-deficit/hyperactivity disorder: Evidence of independent processes," *Biological psychiatry*, vol. 66, no. 4, pp. 398–401, 2009.

[277] G. G. Knyazev, "Motivation, emotion, and their inhibitory control mirrored in brain oscillations," *Neuroscience & Biobehavioral Reviews*, vol. 31, no. 3, pp. 377–395, 2007.

[278] F. M. Howells, D. J. Stein, and V. A. Russell, "Perceived mental effort correlates with changes in tonic arousal during attentional tasks," *Behavioral and Brain Functions*, vol. 6, no. 1, p. 39, Jul 2010. [Online]. Available: https://doi.org/10.1186/1744-9081-6-39

[279] J. J. Foxe and A. C. Snyder, "The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention," *Frontiers in psychology*, vol. 2, p. 154, 2011.

[280] A. Wróbel *et al.*, "Beta activity: a carrier for visual attention," *Acta neurobiologiae experimentalis*, vol. 60, no. 2, pp. 247–260, 2000.

[281] J. Coull, "Neural correlates of attention and arousal: insights from electrophysiology, functional neuroimaging and psychopharmacology," *Progress in Neurobiology*, vol. 55, no. 4, pp. 343 – 361, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301008298000112

[282] L. J. Metzger, S. R. Paige, M. A. Carson, N. B. Lasko, L. A. Paulus, R. K. Pitman, and S. P. Orr, "Ptsd arousal and depression symptoms associated with increased right-sided parietal eeg asymmetry." *Journal of abnormal psychology*, vol. 113, no. 2, p. 324, 2004.

[283] R. Gupta, K. Laghari, H. Banville, and T. H. Falk, "Using affective brain-computer interfaces to characterize human influential factors for speech quality-of-experience perception modelling," *Human-centric Computing and Information Sciences*, vol. 6, no. 1, p. 5, 2016.

[284] F. Movahedi, J. L. Coyle, and E. Sejdić, "Deep belief networks for electroencephalography: A review of recent contributions and future outlooks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 3, pp. 642–652, 2018.

[285] T. Kalisky, Y. Ashkenazy, and S. Havlin, "Volatility of linear and nonlinear time series," *Physical Review E*, vol. 72, no. 1, p. 011913, 2005.

[286] M. D. Costa, C.-K. Peng, and A. L. Goldberger, "Multiscale analysis of heart rate dynamics: entropy and time irreversibility measures," *Cardiovascular Engineering*, vol. 8, no. 2, pp. 88–93, 2008.

[287] M. Costa and J. Healey, "Multiscale entropy analysis of complex heart rate dynamics: discrimination of age and heart failure effects," in *Computers in Cardiology, 2003.* IEEE, 2003, pp. 705–708.

[288] Y.-L. Ho, C. Lin, Y.-H. Lin, and M.-T. Lo, "The prognostic value of non-linear analysis of heart rate variability in patients with congestive heart failure—a pilot study of multiscale entropy," *PloS one*, vol. 6, no. 4, p. e18699, 2011.

[289] Y. Ashkenazy, P. C. Ivanov, S. Havlin, C.-K. Peng, A. L. Goldberger, and H. E. Stanley, "Magnitude and sign correlations in heartbeat fluctuations," *Physical Review Letters*, vol. 86, no. 9, p. 1900, 2001.

[290] C. Liu and R. Gao, "Multiscale entropy analysis of the differential rr interval time series signal and its application in detecting congestive heart failure," *Entropy*, vol. 19, no. 6, p. 251, 2017.

[291] S. Mahdiani, V. Jeyhani, M. Peltokangas, and A. Vehkaoja, "Is 50 hz high enough ecg sampling frequency for accurate hrv analysis?" in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* IEEE, 2015, pp. 5948–5951.

[292] M. Merri, D. C. Farden, J. G. Mottley, and E. L. Titlebaum, “Sampling frequency of the electrocardiogram for spectral analysis of the heart rate variability,” *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 1, pp. 99–106, 1990.

[293] M. El-Yaagoubi, R. Goya-Esteban, Y. Jabrane, S. Muñoz-Romero, A. García-Alberola, and J. L. Rojo-Álvarez, “On the robustness of multiscale indices for long-term monitoring in cardiac signals,” *Entropy*, vol. 21, no. 6, p. 594, 2019.

[294] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[295] Z. Wang, L. Yang, and J. Ding, “Application of heart rate variability in evaluation of mental workload,” *Zhonghua lao dong wei sheng zhi ye bing za zhi= Zhonghua laodong weisheng zhiyebing zazhi= Chinese journal of industrial hygiene and occupational diseases*, vol. 23, no. 3, pp. 182–184, 2005.

[296] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard, “The effect of mental stress on heart rate variability and blood pressure during computer work,” *European journal of applied physiology*, vol. 92, no. 1-2, pp. 84–89, 2004.

[297] J. Taelman, S. Vandeput, E. Vlemincx, A. Spaepen, and S. Van Huffel, “Instantaneous changes in heart rate regulation due to mental load in simulated office work,” *European journal of applied physiology*, vol. 111, no. 7, pp. 1497–1505, 2011.

[298] S. M. Collins, R. A. Karasek, and K. Costas, “Job strain and autonomic indices of cardiovascular disease risk,” *American journal of industrial medicine*, vol. 48, no. 3, pp. 182–193, 2005.

[299] G. Chaumet, A. Delaforge, and S. Delliaux, “Mental workload alters heart rate variability lowering non-linear dynamics,” *Frontiers in Physiology*, vol. 10, p. 565, 2019.

[300] M. Osaka, H. Saitoh, H. Atarashi, and H. Hayakawa, “Correlation dimension of heart rate variability: a new index of human autonomic function.” *Frontiers of medical and biological engineering: the international journal of the Japan Society of Medical Electronics and Biological Engineering*, vol. 5, no. 4, pp. 289–300, 1993.

[301] M. Tulppo, R. Hughson, T. Mäkikallio, K. Airaksinen, T. Seppänen, and H. Huikuri, “Effects of exercise and passive head-up tilt on fractal and complexity properties of heart rate dynamics,” *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 280, no. 3, pp. H1081–H1087, 2001.

[302] F. Cottin, C. Médigue, P.-M. Leprêtre, Y. Papelier, J.-P. Koralsztein, and V. Billat, “Heart rate variability during exercise performed below and above ventilatory threshold,” *Medicine & Science in Sports & Exercise*, vol. 36, no. 4, pp. 594–600, 2004.

[303] G. Blain, O. Meste, and S. Bermon, “Influences of breathing patterns on respiratory sinus arrhythmia in humans during exercise,” *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 288, no. 2, pp. H887–H895, 2005.

[304] R. Goya-Esteban, O. Barquero-Pérez, E. Sarabia-Cachadina, B. de la Cruz-Torres, J. Naranjo-Orellana, and J. L. Rojo-Alvarez, “Heart rate variability non linear dynamics in intense exercise,” in *2012 Computing in Cardiology.* IEEE, 2012, pp. 177–180.

[305] J. F. Valencia, A. Porta, M. Vallverdu, F. Claria, R. Baranowski, E. Orlowska-Baranowska, and P. Caminal, “Refined multiscale entropy: Application to 24-h holter recordings of heart period variability in healthy and aortic stenosis subjects,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 9, pp. 2202–2213, 2009.

[306] R. Jerath and M. W. Crawford, “How does the body affect the mind? role of cardiorespiratory coherence in spectrum of emotions,” *Adv. Mind Body Med*, vol. 29, pp. 4–16, 2015.

[307] T. Donoghue, M. Haller, E. Peterson, P. Varma, P. Sebastian, R. Gao, T. Noto, A. Lara, J. Wallis, R. Knight *et al.*, "Parameterizing neural power spectra into periodic and aperiodic components," *Nature neuroscience*, vol. 23, no. 12, pp. 1655–1665, 2020.

[308] M. Immink, Z. Cross, A. Chatburn, J. Baumeister, M. Schlesewsky, and I. Bornkessel-Schlesewsky, "Resting-state aperiodic neural dynamics predict individual differences in visuomotor performance and learning," *bioRxiv*, 2021.

[309] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[310] C. Schölzel, "Nonlinear measures for dynamical systems," Jun. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3814723

[311] J. Schipke, G. Arnold, and M. Pelzer, "Effect of respiration rate on short-term heart rate variability," *Journal of Clinical and Basic Cardiology*, vol. 2, no. 1, pp. 92–95, 1999.

[312] A. Brugnera, C. Zarbo, M. Tarvainen, P. Marchettini, R. Adorni, and A. Compare, "Heart rate variability during acute psychosocial stress: A randomized cross-over trial of verbal and non-verbal laboratory stressors," *International journal of psychophysiology*, vol. 127, pp. 17–25, 2018.

[313] G. Valenza, P. Allegrini, A. Lanatà, and E. Scilingo, "Dominant lyapunov exponent and approximate entropy in heart rate variability during emotional visual elicitation," *Frontiers in neuroengineering*, vol. 5, p. 3, 2012.

[314] K. Kim, J. Kim, Y. Lim, and K. Park, "The effect of missing RR-interval data on heart rate variability analysis in the frequency domain," *Physiological measurement*, vol. 30, no. 10, p. 1039, 2009.

[315] S. Boettger, C. Puta, V. K. Yeragani, L. Donath, H.-J. Mueller, H. H. Gabriel, and K.-J. Baer, "Heart rate variability, qt variability, and electrodermal activity during exercise," *Medicine & science in sports & exercise*, vol. 42, no. 3, pp. 443–448, 2010.

[316] S. Sharples and T. Megaw, "The definition and measurement of human workload," *Evaluation of human work. Boca*, 2015.

[317] S. Michael, K. S. Graham, and G. M. Davis, "Cardiac autonomic responses during exercise and post-exercise recovery using heart rate variability and systolic time intervals—a review," *Frontiers in physiology*, vol. 8, p. 301, 2017.

[318] G. Tanda, "Skin temperature measurements by infrared thermography during running exercise," *Experimental Thermal and Fluid Science*, vol. 71, pp. 103–113, 2016.

[319] H. F. Posada-Quintero, N. Reljin, C. Mills, I. Mills, J. P. Florian, J. L. VanHeest, and K. H. Chon, "Time-varying analysis of electrodermal activity during exercise," *PloS one*, vol. 13, no. 6, 2018.

[320] D. Tobon, M. Maier, and T. H. Falk, "Ms-qi: A modulation spectrum-based ecg quality index for telehealth applications," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1613–1622, 2016.

[321] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term hrv features as surrogates of short term hrv: a case study on mental stress detection in real life," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–13, 2019.

[322] M. Wu, *Trimmed and winsorized estimators.* Michigan State University, 2006.

[323] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience and biobehavioral reviews*, vol. 44, p. 58—75, July 2014. [Online]. Available: https://doi.org/10.1016/j.neubiorev.2012.10.003

[324] S. Ladouce, D. I. Donaldson, P. A. Dudchenko, and M. Ietswaart, "Mobile eeg identifies the re-allocation of attention during real-world activity," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[325] S. Micheloyannis, E. Papanikolaou, E. Bizas, C. J. Stam, and P. G. Simos, "Ongoing electroencephalographic signal study of simple arithmetic using linear and non-linear measures," *International journal of psychophysiology*, vol. 44, no. 3, pp. 231–238, 2002.

[326] J. Kaiser and W. Lutzenberger, "Human gamma-band activity: a window to cognitive processing," *Neuroreport*, vol. 16, no. 3, pp. 207–211, 2005.

[327] R. Ishii, L. Canuet, T. Ishihara, Y. Aoki, S. Ikeda, M. Hata, T. Katsimichas, A. Gunji, H. Takahashi, T. Nakahachi *et al.*, "Frontal midline theta rhythm and gamma power changes during focused attention on mental calculation: an meg beamformer analysis," *Frontiers in human neuroscience*, vol. 8, p. 406, 2014.

[328] S. Agosta, D. Magnago, S. Tyler, E. Grossman, E. Galante, F. Ferraro, N. Mazzini, G. Miceli, and L. Battelli, "The pivotal role of the right parietal lobe in temporal attention," *Journal of cognitive neuroscience*, vol. 29, no. 5, pp. 805–815, 2017.

[329] K. Heinen, E. Feredoes, C. C. Ruff, and J. Driver, "Functional connectivity between prefrontal and parietal cortex drives visuo-spatial attention shifts," *Neuropsychologia*, vol. 99, pp. 81–91, 2017.

[330] J. Doyon and B. Milner, "Right temporal-lobe contribution to global visual processing," *Neuropsychologia*, vol. 29, no. 5, pp. 343–360, 1991.

[331] R. K. Lech and B. Suchan, "Involvement of the human medial temporal lobe in a visual discrimination task," *Behavioural brain research*, vol. 268, pp. 22–30, 2014.

[332] T. Brandt and M. Dieterich, "The vestibular cortex: its locations, functions, and disorders," *Annals of the New York Academy of Sciences*, vol. 871, no. 1, pp. 293–312, 1999.

[333] E. Vlemincx, J. Taelman, S. De Peuter, I. Van Diest, and O. Van Den Bergh, "Sigh rate and respiratory variability during mental load and sustained attention," *Psychophysiology*, vol. 48, no. 1, pp. 117–120, 2011.

[334] E. Vlemincx, I. Van Diest, and O. Van den Bergh, "A sigh of relief or a sigh to relieve: The psychological and physiological relief effect of deep breaths," *Physiology & behavior*, vol. 165, pp. 127–135, 2016.

[335] C. Collet, E. Salvia, and C. Petit-Boulanger, "Measuring workload with electrodermal activity during common braking actions," *Ergonomics*, vol. 57, no. 6, pp. 886–896, 2014.

[336] A. Merla and G. L. Romani, "Thermal signatures of emotional arousal: a functional infrared imaging study," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 247–249.

[337] T. Lara, E. Molina, J. A. Madrid, and Á. Correa, "Electroencephalographic and skin temperature indices of vigilance and inhibitory control," *Psicológica Journal*, vol. 39, no. 2, pp. 223–260, 2018.

[338] S. Gandhi, M. S. Baghini, and S. Mukherji, "Mental stress assessment-a comparison between hrv based and respiration based techniques," in *2015 Computing in Cardiology Conference (CinC)*. IEEE, 2015, pp. 1029–1032.

[339] M. Grassmann, E. Vlemincx, A. von Leupoldt, and O. Van den Bergh, "The role of respiratory measures to assess mental load in pilot selection," *Ergonomics*, vol. 59, no. 6, pp. 745–753, 2016.

[340] M. Rahman, E. Nemati, M. Rahman, V. Nathan, K. Vatanparvar, and J. Kuang, “Automated assessment of pulmonary patients using heart rate variability from everyday wearables,” *Smart Health*, vol. 15, p. 100081, 2020.

[341] C. Bellos, A. Papadopoulos, R. Rosso, and D. Fotiadis, “Identification of copd patients’ health status using an intelligent system in the chronious wearable platform,” *IEEE journal of biomedical and health informatics*, vol. 18, no. 3, pp. 731–738, 2013.

[342] I. Tomasic, N. Tomasic, R. Trobec, M. Krpan, and T. Kelava, “Continuous remote monitoring of copd patients—justification and explanation of the requirements and a survey of the available technologies,” *Medical & biological engineering & computing*, vol. 56, no. 4, pp. 547–569, 2018.

[343] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning.* MIT press Cambridge, 2016, vol. 1, no. 2.

[344] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath, “Deep learning models for electrocardiograms are susceptible to adversarial attack,” *Nature medicine*, vol. 26, no. 3, pp. 360–363, 2020.

[345] L. F. W. Anthony, B. Kanding, and R. Selvan, “Carbontracker: Tracking and predicting the carbon footprint of training deep learning models,” *arXiv preprint arXiv:2007.03051*, 2020.

[346] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[347] M. N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy, and V. Chandran, “Automatic driver stress level classification using multimodal deep learning,” *Expert Systems with Applications*, vol. 138, p. 112793, 2019.

[348] J. He, K. Li, X. Liao, P. Zhang, and N. Jiang, “Real-time detection of acute cognitive stress using a convolutional neural network from electrocardiographic signal,” *IEEE Access*, vol. 7, pp. 42 710–42 717, 2019.

[349] G. Giannakakis, E. Trivizakis, M. Tsiknakis, and K. Marias, “A novel multi-kernel 1d convolutional neural network for stress recognition from ecg,” in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW).* IEEE, 2019, pp. 1–4.

[350] A. Gaballah, A. Tiwari, S. Narayanan, and T. Falk, “Context-aware speech stress detection in hospital workers using Bi-LSTM classifiers,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.