

UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Facoltà di Ingegneria
Corso di Laurea Triennale in Ingegneria Biomedica

Valutazione del livello di stress da segnali PPG acquisiti da siti anatomici non convenzionali

Stress Level Assessment from PPG Signals Acquired at Non-Conventional Anatomical Sites

Candidato:
Lorenzo Zappacosta Gasparrini

Relatore:
Prof. Lorenzo Palma

Anno Accademico 2025/2026

Università Politecnica delle Marche
Facoltà di Ingegneria
Corso di Laurea Triennale in Ingegneria Biomedica
Via Brecce Bianche – 60131 Ancona (AN), Italy

Sommario

Il riconoscimento continuo dello stress con sensori indossabili richiede segnali acquisibili in modo poco invasivo e per lunghi periodi. In questo lavoro si valuta se il solo segnale fotopletismografico (PPG), acquisito sulla parte superiore del braccio e nella regione lombare, sedi già usate dai sensori per il monitoraggio continuo del glucosio, permetta di classificare tre livelli di stress (basso, medio e alto) su soggetti non usati per l'addestramento dei modelli.

Sono state analizzate le registrazioni di 25 soggetti per sito, acquisite con sensori Polar OH1+ a 135 Hz durante un protocollo composto da una fase di riposo e quattro compiti mentali. Il segnale è stato filtrato con un passa-banda Chebyshev di tipo II (0,5–8 Hz), corretto nella polarità e suddiviso in finestre di 60 s con sovrapposizione del 40%, da cui sono state estratte 55 feature temporali, spettrali, di variabilità delle pulsazioni e morfologiche; le brevi pause tra i compiti sono state escluse. Sono stati confrontati una Random Forest (RF), una rete neurale a strati densi (NN) e una rete convoluzionale monodimensionale (1D-CNN) applicata direttamente al segnale. Selezione delle feature, bilanciamento delle classi e iperparametri sono stati scelti con una validazione annidata sui soli soggetti di training, e la valutazione è stata condotta per soggetto a 5 fold, ripetuta 10 volte, con intervalli di confidenza al 95% stimati con un bootstrap sui soggetti.

Per singola finestra la balanced accuracy è risultata di 0,582 (RF) e 0,593 (NN) sul braccio e di 0,615 e 0,624 sulla schiena, a fronte di 0,33 per un classificatore casuale; aggregando le finestre di ciascun compito sale fino a 0,680 sul braccio e 0,713 sulla schiena. La 1D-CNN si è fermata a 0,44–0,54. Lo stress medio è la classe più difficile da riconoscere (recall 0,37–0,48) e la variabilità tra soggetti supera quella tra modelli e tra sedi: il confronto appaiato tra braccio e schiena non mostra differenze significative (test di Wilcoxon, $p \geq 0{,}58$). Nella classificazione binaria riposo/stress la balanced accuracy è di 0,71–0,72 sul braccio e di 0,75–0,76 sulla schiena. Il PPG di entrambe le sedi contiene quindi informazione sul livello di stress anche su soggetti nuovi, ma la classificazione della singola finestra non è ancora affidabile: per un impiego pratico sono necessari campioni più ampi, etichette individuali e una validazione su una popolazione esterna.

Parole chiave: fotopletismografia, stress, dispositivi indossabili, machine learning, braccio, regione lombare.

Abstract

Continuous stress recognition with wearable sensors requires signals that can be acquired unobtrusively over long periods. This work assesses whether the photoplethysmographic (PPG) signal alone, recorded on the upper arm and on the lower back, two sites already used by continuous glucose monitoring sensors, can classify three stress levels (low, medium and high) in subjects not used to train the models.

Recordings from 25 subjects per site were analysed, acquired with Polar OH1+ sensors at 135 Hz during a protocol including a rest phase and four mental tasks. The signal was band-pass filtered with a type II Chebyshev filter (0.5–8 Hz), polarity-corrected and segmented into 60-s windows with 40% overlap, from which 55 temporal, spectral, pulse-variability and morphological features were extracted; the short pauses between tasks were excluded. A Random Forest (RF), a dense neural network (NN) and a one-dimensional convolutional network (1D-CNN) applied to the raw signal were compared. Feature selection, class balancing and hyperparameters were chosen through nested validation on training subjects only, and performance was assessed with 5-fold subject-wise cross-validation repeated 10 times, with 95% confidence intervals from a subject-level bootstrap.

At window level, balanced accuracy was 0.582 (RF) and 0.593 (NN) on the arm and 0.615 and 0.624 on the back, against 0.33 for a random classifier; aggregating the windows of each task raised it up to 0.680 on the arm and 0.713 on the back. The 1D-CNN reached only 0.44–0.54. The medium stress class was the hardest to recognise (recall 0.37–0.48), and inter-subject variability exceeded the differences between models and sites: the paired comparison between arm and back showed no significant difference (Wilcoxon test, $p \geq 0.58$). In the binary rest/stress task, balanced accuracy was 0.71–0.72 on the arm and 0.75–0.76 on the back. PPG from both sites therefore carries information on stress level even for new subjects, but single-window classification is not yet reliable: practical use will require larger samples, individual labels and validation on an external population.

Keywords: photoplethysmography, stress, wearable devices, machine learning, upper arm, lower back.

Indice

Sommario iii

Abstract iv

Elenco degli acronimi x

Introduzione 1

1 Background 3
- 1.1 Basi fisiologiche del sistema cardiovascolare . . . 3
  - 1.1.1 Cuore e circolazione sanguigna . . . 3
  - 1.1.2 Ciclo cardiaco . . . 4
  - 1.1.3 Regolazione autonomica dell'attività cardiaca . . . 4
- 1.2 Stress, segnali fisiologici e dispositivi wearable . . . 5
  - 1.2.1 Stress e risposta fisiologica . . . 5
  - 1.2.2 Valutazione dello stress e segnali fisiologici . . . 8
  - 1.2.3 Principi della fotopletismografia . . . 9
  - 1.2.4 Componenti e morfologia del segnale PPG . . . 10
  - 1.2.5 Dispositivi indossabili, siti di acquisizione e artefatti . . . 12
- 1.3 Machine Learning . . . 13
  - 1.3.1 Apprendimento supervisionato e classificazione . . . 13
  - 1.3.2 Estrazione e selezione delle feature . . . 15
  - 1.3.3 Sbilanciamento delle classi . . . 15
  - 1.3.4 Validazione incrociata per soggetto . . . 16
  - 1.3.5 Classificatori considerati . . . 17
  - 1.3.6 Metriche di valutazione . . . 19
- 1.4 Stato dell'arte . . . 20
  - 1.4.1 Review sulla stress detection con dispositivi indossabili . . . 21
  - 1.4.2 Classificazione multiclasse con il solo PPG . . . 21
  - 1.4.3 PPG in sedi di misura alternative . . . 21
  - 1.4.4 Lavori precedenti del gruppo di ricerca . . . 22
- 1.5 Contributo della tesi . . . 23

2 Materiali e metodi 25
- 2.1 Dataset e protocollo di acquisizione . . . 26
- 2.2 Definizione delle classi . . . 29

2.3 Sviluppo della pipeline di elaborazione 31
2.4 Preprocessing del segnale 31
2.5 Segmentazione in finestre 32
2.6 Estrazione delle feature 32
2.7 Classificazione e validazione 36
2.7.1 Modelli 36
2.7.2 Selezione annidata di feature, bilanciamento e iperparametri . 36
2.7.3 Validazione 39
2.7.4 Metriche e livelli di valutazione 40
2.7.5 Strumenti 40

3 Risultati **41**
3.1 Analisi esplorativa della frequenza di pulsazione 41
3.2 Prestazioni dei classificatori 42
3.3 Valutazione cumulativa 43
3.4 Prestazioni per classe 43
3.5 Confronto tra i siti e variabilità inter-soggetto 46
3.6 Effetto della correzione di polarità e della selezione delle feature . . . 47
3.7 Effetto dell'esclusione delle pause 49
3.8 Importanza delle feature 49
3.9 Classificazione binaria riposo/stress 51

4 Discussione **53**
4.1 Interpretazione dei risultati 53
4.2 Confronto con la letteratura 53
4.3 Confronto tra i siti di misura 54
4.4 Riconoscimento della classe intermedia e qualità delle etichette . . . 55
4.5 Variabilità inter-soggetto 55
4.6 Limiti dello studio 56

Conclusioni **58**

Appendice A – Dettagli tecnici **60**
A.1 Elenco completo delle feature 60
A.2 Parametri dei modelli 62

Ringraziamenti **68**

Elenco delle figure

1.1 Struttura interna del cuore 3
1.2 Flusso del sangue in diastole e in sistole 5
1.3 Il ciclo cardiaco 6
1.4 Il ciclo cardiaco nel segnale PPG 6
1.5 Le due vie della risposta allo stress 7
1.6 Configurazioni di acquisizione del PPG 9
1.7 Componenti continua e pulsatile del segnale PPG 10
1.8 Punti caratteristici della pulsazione PPG 11
1.9 Esempi di dispositivi PPG indossabili 12
1.10 Esempio di segnale PPG pulito e con artefatti 14
1.11 Esempio schematico di SMOTE 16
1.12 Schema della Random Forest 18

2.1 Schema della configurazione finale 25
2.2 Posizionamento dei sensori 26
2.3 Protocollo di induzione dello stress e frequenza di pulsazione 28
2.4 Postazione di acquisizione 29
2.5 Risposte al questionario finale 30
2.6 Risposta in ampiezza del filtro passa-banda 32
2.7 Fasi del preprocessing del segnale 33
2.8 Verifica della polarità del segnale 33
2.9 Registrazione completa e finestre di analisi 34
2.10 Esempi di feature su una finestra reale 35
2.11 Schema della rete neurale densa 37
2.12 Schema della 1D-CNN 37
2.13 Schema della validazione annidata 38
2.14 Suddivisione dei soggetti nei fold 39

3.1 Frequenza di pulsazione nelle fasi del protocollo 41
3.2 Balanced accuracy dei modelli ai tre livelli 44
3.3 Matrici di confusione 45
3.4 Curve ROC per classe 46
3.5 Accuratezza per soggetto: braccio e schiena 47
3.6 Accuratezza per soggetto 48
3.7 Effetto della correzione della polarità 49
3.8 Predizioni per il riposo iniziale e per le pause 50

3.9 Feature più importanti nella Random Forest 51

Elenco delle tabelle

1.1 Studi citati nello stato dell'arte 23

2.1 Righe di un file di registrazione 27
2.2 Corrispondenza tra fasi del protocollo e classi di stress. 30

3.1 Prestazioni per finestra di tutti i modelli 43
3.2 Prestazioni ai tre livelli di valutazione 44
3.3 Risultati con e senza le pause tra i compiti 50

A.1 Le 55 feature estratte da ogni finestra, divise per gruppo. 60
A.2 Parametri della configurazione finale 62

Elenco degli acronimi

1D-CNN	rete neurale convoluzionale monodimensionale (*one-dimensional Convolutional Neural Network*)
ACTH	ormone adrenocorticotropo
AUC	area sotto la curva ROC (*area under the curve*)
bpm	battiti al minuto (*beats per minute*)
CRH	ormone di rilascio della corticotropina
ECG	elettrocardiogramma
EDA	attività elettrodermica (*electrodermal activity*)
HRV	variabilità della frequenza cardiaca (*heart rate variability*)
IC	intervallo di confidenza
LED	diodo a emissione di luce (*light-emitting diode*)
LOSO	*leave-one-subject-out*
NN	rete neurale (*Neural Network*) a strati densi
pNN50	percentuale di intervalli successivi che differiscono di più di 50 ms
PP	intervallo picco-picco
PPG	fotopletismografia (*photoplethysmography*)
PRV	variabilità della frequenza di pulsazione (*pulse rate variability*)
RF	Random Forest
ROC	*receiver operating characteristic*
RMSSD	radice della media dei quadrati delle differenze tra intervalli successivi
SD1, SD2	deviazioni standard del diagramma di Poincaré
SDNN	deviazione standard degli intervalli tra battiti
SMOTE	*Synthetic Minority Over-sampling Technique*
SVM	*Support Vector Machine*

Introduzione

Lo stress fa parte della vita di tutti i giorni, ma quando si prolunga o si ripete spesso ha effetti negativi sia sulla salute fisica sia su quella mentale [1]. Riconoscerlo in tempo, prima che diventi un problema, richiederebbe una misurazione continua e poco invasiva. È qui che entrano in gioco i dispositivi indossabili: sensori piccoli ed economici che registrano segnali fisiologici durante la giornata e che, insieme a tecniche di machine learning, possono stimare lo stato di chi li indossa [2, 3].

Tra i segnali disponibili, il fotopletismogramma (PPG) è uno dei più usati. Si acquisisce con un semplice sensore ottico, è già presente in molti dispositivi commerciali e contiene informazioni sull'attività cardiaca e sulla circolazione periferica, entrambe influenzate dalla risposta allo stress [4]. La maggior parte degli studi sulla stress detection, però, usa il PPG al polso o al dito, spesso insieme ad altri segnali come l'attività elettrodermica, e affronta un problema binario: stress o non stress [3].

Questa tesi fa parte di un progetto più ampio, dedicato al monitoraggio del diabete di tipo 1 con sensori indossabili [5]. L'idea è acquisire il PPG in due sedi meno convenzionali, la parte superiore del braccio e la regione lombare, le stesse usate dai sensori per il monitoraggio continuo del glucosio. Il legame con il diabete non riguarda solo la sede del sensore. Lo stress psicologico libera catecolamine e cortisolo, che aumentano la produzione di glucosio da parte del fegato e riducono la sensibilità all'insulina: in una persona sana l'aumento della glicemia viene compensato, mentre nel diabete di tipo 1, in cui l'insulina deve essere somministrata dall'esterno, lo stress può tradursi in iperglicemia e in un maggiore fabbisogno di insulina [6, 7]. Riconoscere lo stress con lo stesso sensore, o con un sensore nella stessa sede di quello per il glucosio, potrebbe, quindi, dare un'informazione in più ai sistemi di monitoraggio e di somministrazione dell'insulina. Prima di arrivarci occorre però capire se il PPG registrato in queste sedi contenga abbastanza informazione sullo stress, ed è questa la domanda da cui parte la tesi.

Obiettivi del lavoro

L'obiettivo della tesi è valutare se, usando soltanto il segnale PPG acquisito sul braccio e sulla schiena, sia possibile distinguere tre livelli di stress (basso, medio e alto) in persone che il modello non ha mai visto durante l'addestramento. Il lavoro cerca di rispondere a tre domande:

1. con quale accuratezza si riconoscono i tre livelli di stress dal solo PPG?

2. emergono differenze tra braccio e schiena, o nessuno dei due siti risulta superiore all'altro?

3. quali scelte di elaborazione del segnale e di classificazione influenzano davvero le prestazioni?

Per rispondere sono state usate le registrazioni di 25 soggetti per sito, acquisite con un protocollo di induzione dello stress composto da una fase di riposo e quattro compiti mentali. Dai segnali sono state estratte feature temporali, spettrali e morfologiche, poi usate per addestrare e confrontare diversi classificatori: Random Forest, reti neurali dense e una rete convoluzionale applicata direttamente al segnale. Ogni scelta è stata valutata con una validazione incrociata per soggetto, ripetuta più volte per tenere conto della variabilità dovuta al campione ridotto.

Struttura della tesi

- Il **Capitolo 1** introduce le basi del lavoro: la fisiologia del sistema cardiovascolare, la risposta allo stress, i principi del PPG e dei dispositivi indossabili, i concetti di machine learning usati nella tesi e lo stato dell'arte. Si chiude con il contributo della tesi.

- Il **Capitolo 2** descrive il dataset, la definizione delle classi e tutte le fasi dell'elaborazione, dal preprocessing alla validazione dei modelli.

- Il **Capitolo 3** riporta i risultati, per modello, sito, classe e soggetto.

- Il **Capitolo 4** interpreta i risultati, li confronta con la letteratura e ne discute i limiti.

- Le **Conclusioni** riassumono i risultati principali e indicano i possibili sviluppi futuri.

Capitolo 1

Background

1.1 Basi fisiologiche del sistema cardiovascolare

1.1.1 Cuore e circolazione sanguigna

Il sistema cardiovascolare è formato dal cuore e dai vasi sanguigni. Il cuore funziona come una pompa ed è diviso in quattro camere: i due atri ricevono il sangue, mentre i due ventricoli lo spingono nella circolazione polmonare e in quella sistemica [8, cap. 9]. La struttura interna del cuore e il percorso del sangue sono mostrati nella Figura 1.1.

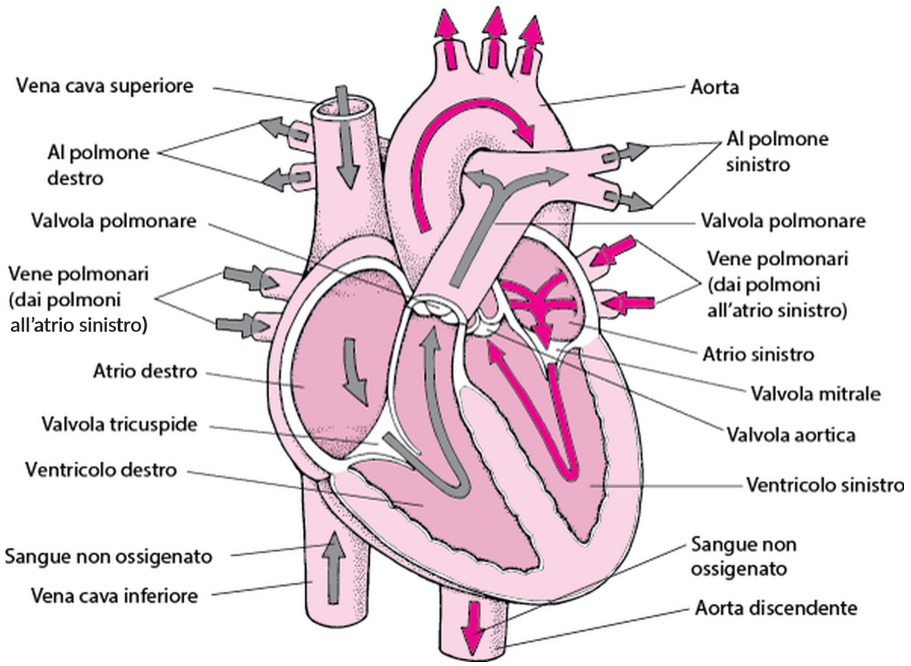


Figura 1.1: Struttura interna del cuore e direzione del flusso sanguigno: in grigio il sangue non ossigenato, in rosa quello ossigenato. Adattata da [9], con la correzione di un'etichetta (le vene polmonari portano il sangue all'atrio sinistro).

Il ventricolo destro invia il sangue ai polmoni, dove avvengono gli scambi gassosi. Il sangue ossigenato torna all'atrio sinistro, passa al ventricolo sinistro e da qui viene

spinto nell'aorta, che lo distribuisce ai tessuti. Attraverso le vene il sangue rientra poi nell'atrio destro e nel ventricolo destro, chiudendo il circuito [8, cap. 9].

Le arterie portano il sangue dal cuore agli organi, le vene lo riportano indietro. Tra le due reti si trovano i capillari, dove avvengono gli scambi di ossigeno, nutrienti e prodotti del metabolismo. Il passaggio del sangue attraverso questo letto vascolare è chiamato perfusione tissutale [8].

1.1.2 Ciclo cardiaco

Con ciclo cardiaco si intende la sequenza di eventi che si ripete a ogni battito. Per ogni camera si distinguono una fase di contrazione, la sistole, e una di rilasciamento, la diastole. Sono le differenze di pressione ai due lati delle valvole a farle aprire e chiudere, così che il sangue scorra in una sola direzione [8, cap. 9].

All'inizio della sistole ventricolare la pressione sale a valvole chiuse. Quando nel ventricolo sinistro supera quella dell'aorta, e nel destro quella del tronco polmonare, si aprono le valvole semilunari e comincia l'eiezione, cioè l'espulsione del sangue dai ventricoli [8, cap. 9].

Durante il rilasciamento la pressione ventricolare scende. Le valvole semilunari si chiudono non appena diventa inferiore a quella arteriosa; quando scende anche sotto la pressione degli atri, si aprono le valvole atrioventricolari e i ventricoli tornano a riempirsi. È questa alternanza tra riempimento ed eiezione a generare le pulsazioni arteriose [8, cap. 9]. La Figura 1.2 mostra il percorso del sangue nelle due fasi.

L'andamento di pressioni e volumi durante un battito è riassunto nella Figura 1.3. La sistole comprende la contrazione isovolumetrica, con tutte le valvole chiuse, e l'eiezione; la diastole comprende il rilasciamento isovolumetrico e il riempimento dei ventricoli, completato dalla contrazione degli atri. Il complesso QRS dell'ECG precede di poco la contrazione dei ventricoli, mentre la chiusura delle valvole produce i toni cardiaci [11].

Il ciclo cardiaco si ritrova, con un certo ritardo, nella pulsazione registrata in periferia dal PPG, descritto nella Sezione 1.2.3. La Figura 1.4 mostra alcuni battiti registrati sulla schiena di uno dei partecipanti allo studio. Nella fase di salita il sangue espulso dal cuore dilata i vasi e il segnale cresce fino al picco sistolico, la cui ampiezza dipende, tra l'altro, dal volume di sangue espulso. Nella discesa può comparire l'incisura dicrota, dovuta alle riflessioni dell'onda di polso lungo le arterie, e il segnale torna poi lentamente al livello di partenza mentre il sangue defluisce [12]. L'intervallo tra l'inizio del battito e l'incisura dicrota viene usato come stima del tempo di eiezione del ventricolo sinistro [4]: per questo, nella figura, l'incisura separa la fase sistolica da quella diastolica.

1.1.3 Regolazione autonomica dell'attività cardiaca

L'attività del cuore è modulata dal sistema nervoso autonomo. La componente simpatica tende ad aumentare la frequenza cardiaca e la forza di contrazione; quella

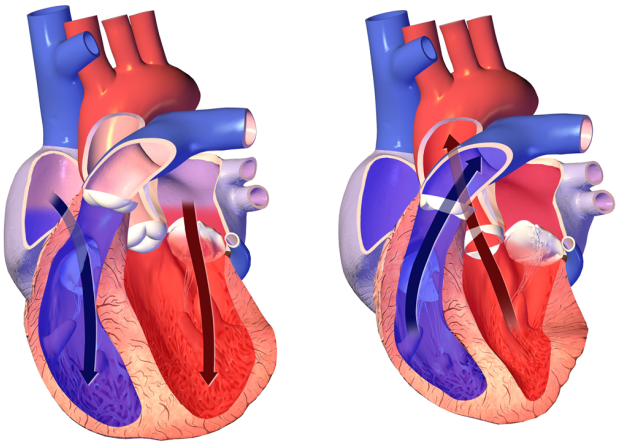


Figura 1.2: Flusso del sangue nel cuore durante la diastole (a sinistra), quando le valvole atrioventricolari sono aperte e i ventricoli si riempiono, e durante la sistole (a destra), quando si aprono le valvole semilunari e il sangue viene spinto nell'aorta e nell'arteria polmonare. In blu il sangue povero di ossigeno, in rosso quello ossigenato. Adattata da [10], immagine con licenza Creative Commons: scritte tradotte in italiano e ordine dei due pannelli invertito.

parasimpatica, che agisce soprattutto attraverso il nervo vago, la rallenta. In questo modo il cuore si adatta di continuo alle richieste dell'organismo [8, cap. 61]. È proprio su questo equilibrio che agisce la risposta allo stress, descritta nella prossima sezione.

1.2 Stress, segnali fisiologici e dispositivi wearable

1.2.1 Stress e risposta fisiologica

Lo stress si può descrivere come la risposta dell'organismo a richieste o stimoli che ne mettono alla prova i meccanismi di adattamento. Gli stimoli che la provocano sono chiamati *stressor* e possono essere di natura fisica o psicologica. Nella risposta entrano in gioco sia processi fisiologici sia il modo in cui la persona vive la situazione: per valutare lo stress bisogna quindi guardare a entrambi gli aspetti [2].

Dal punto di vista fisiologico, tutto parte dall'ipotalamo, che di fronte a uno stressor attiva due vie. La prima è rapida ed è la classica risposta di attacco o fuga (*fight or flight*): attraverso il sistema nervoso simpatico l'ipotalamo stimola la midollare del surrene, che libera adrenalina e noradrenalina. Questi ormoni aumentano la frequenza cardiaca, la forza di contrazione del cuore e la pressione arteriosa e rendono disponibile più glucosio, preparando l'organismo ad affrontare la minaccia o a sfuggirle [1, 8].

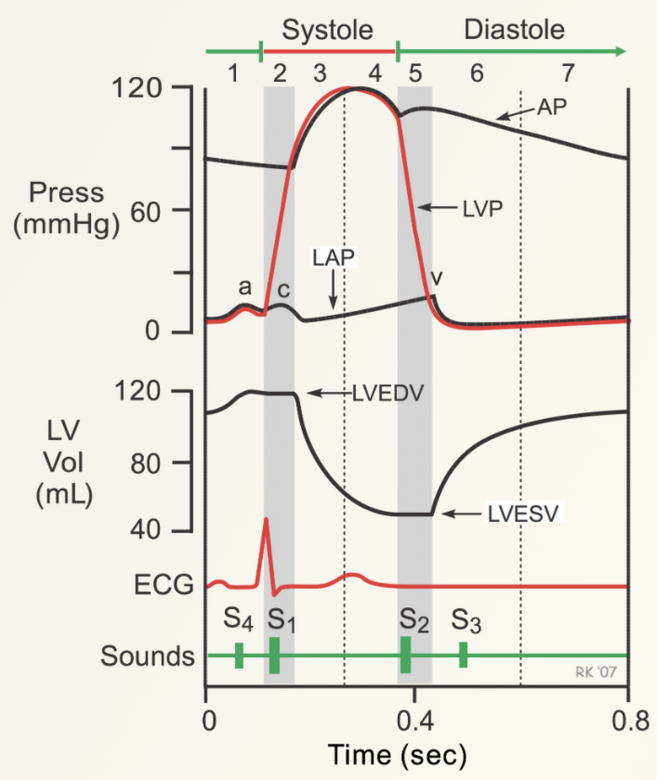


Figura 1.3: Il ciclo cardiaco: pressione aortica (AP), ventricolare sinistra (LVP) e atriale sinistra (LAP), volume del ventricolo sinistro, ECG e toni cardiaci in un battito. Le fasi numerate vanno dalla contrazione atriale (1) al riempimento ventricolare lento (7) [11].

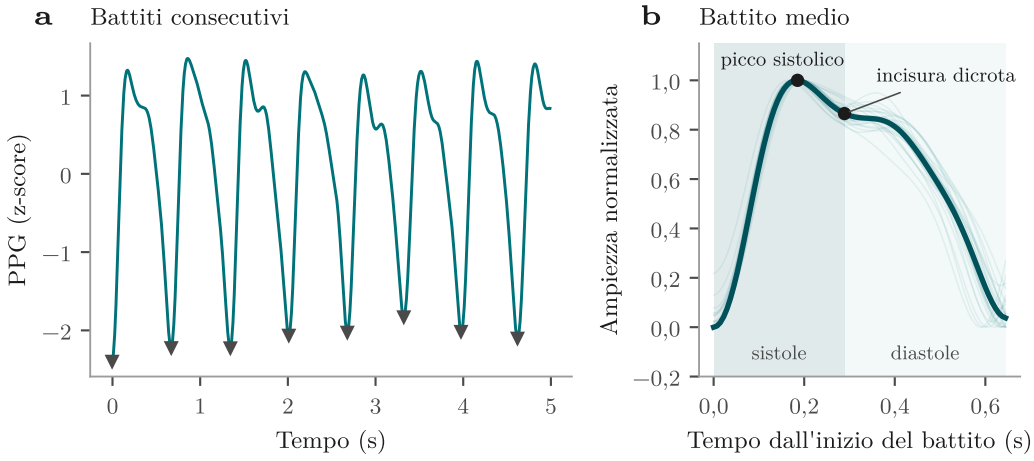


Figura 1.4: Il ciclo cardiaco nel segnale PPG (soggetto 2, schiena, riposo): cinque secondi di battiti consecutivi, con l'inizio di ciascun battito indicato dai triangoli (a), e battito medio di una finestra di 60 s, con le fasi sistolica e diastolica della pulsazione periferica separate dall'incisura dicrota (b). In trasparenza i singoli battiti.

In parallelo si attiva una seconda via, più lenta: l'asse ipotalamo-ipofisi-surrene. Risponde anche a stress acuti e brevi, come quelli indotti in laboratorio [38], ma i suoi effetti compaiono dopo alcuni minuti e diventano importanti soprattutto quando lo stress si prolunga. L'ipotalamo secerne l'ormone di rilascio della corticotropina (CRH), che fa produrre all'ipofisi anteriore l'ormone adrenocorticotropo (ACTH); questo stimola a sua volta la corticale del surrene a liberare glucocorticoidi, in particolare il cortisolo. Il cortisolo mobilita le riserve energetiche e permette all'organismo di sostenere la risposta nel tempo [8, cap. 78]. Una risposta allo stress adeguata è utile, ma se il sistema resta attivato troppo a lungo, o reagisce in modo sproporzionato, può contribuire a disturbi del metabolismo, del sonno e dell'umore [1].

Le due vie sono riassunte nella Figura 1.5. Per questa tesi la più interessante è quella simpatica, perché i suoi effetti su cuore e vasi si riflettono direttamente sul segnale cardiovascolare periferico.

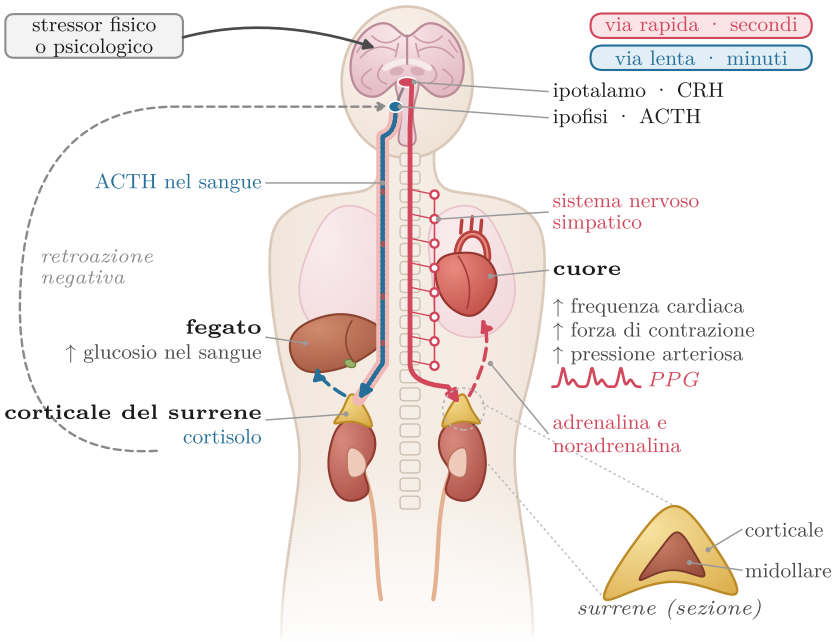


Figura 1.5: Le due vie della risposta allo stress. Nella via simpatico-midollare (in rosso) l'ipotalamo attiva, attraverso il midollo spinale e il sistema nervoso simpatico, la midollare del surrene, che rilascia adrenalina e noradrenalina: in pochi secondi aumentano frequenza cardiaca, forza di contrazione e pressione arteriosa, con effetti visibili nel segnale PPG. Nell'asse ipotalamo-ipofisi-surrene (in blu) il CRH ipotalamico stimola l'ipofisi a rilasciare ACTH nel sangue; l'ACTH raggiunge la corticale del surrene, che produce cortisolo. Il cortisolo agisce in minuti, stimola il fegato a liberare glucosio e a sua volta frena ipotalamo e ipofisi (retroazione negativa). Nel riquadro, sezione del surrene con la corticale esterna e la midollare interna. Schema elaborato sulla base di [1, 8].

L'effetto dello stress sul ritmo cardiaco è stato studiato soprattutto attraverso la

variabilità della frequenza cardiaca. Una meta-analisi di 37 studi ha trovato che la variazione riportata più spesso durante lo stress è una ridotta attività parasimpatica, che si manifesta con un calo della potenza in alta frequenza e un aumento di quella in bassa frequenza [13]. È il motivo per cui le misure di frequenza e di variabilità del battito sono tra le più usate per riconoscere lo stress.

I segnali fisiologici però non raccontano solo lo stress. Il segnale fotopletismografico (PPG), descritto nella Sezione 1.2.3, risente anche della respirazione, dei movimenti, della temperatura e di come il sensore aderisce alla pelle [4]. Una variazione del segnale non è quindi, di per sé, un indice di stress. Per questo motivo il lavoro distingue il segnale, che fornisce l'informazione fisiologica, dalle etichette, che indicano le condizioni sperimentali che i modelli devono riconoscere.

1.2.2 Valutazione dello stress e segnali fisiologici

Lo stress può essere valutato con questionari e con misure fisiologiche. I questionari raccolgono ciò che la persona riferisce, le misure fisiologiche descrivono come risponde l'organismo: sono due punti di vista sullo stesso fenomeno e possono essere usati insieme [2].

Il significato di un questionario dipende dalle domande e dal momento a cui si riferiscono. Un giudizio dato alla fine di un'attività riassume l'esperienza complessiva, ma non descrive per forza ogni istante dell'attività. Questa differenza va tenuta presente quando lo si confronta con l'andamento nel tempo dei segnali fisiologici.

Tra i segnali più usati nello studio dello stress ci sono il PPG e l'attività elettrodermica. Dal PPG si ricavano informazioni sulla frequenza e sulla forma delle pulsazioni periferiche; l'attività elettrodermica misura invece le variazioni di conduttanza della pelle legate alle ghiandole sudoripare, controllate dal sistema simpatico [2, 4]. In questa tesi l'attività elettrodermica compare solo come riferimento alla letteratura: l'analisi sperimentale riguarda esclusivamente il PPG.

Per impostare un problema di classificazione bisogna associare a ogni osservazione un'etichetta, cioè la categoria che il modello deve imparare a riconoscere. L'etichetta è una scelta operativa fatta sulla base delle informazioni disponibili, non una misura dello stress contenuta nel segnale. Assegnare la stessa classe a un intero intervallo sperimentale, ad esempio, non significa che la risposta fisiologica resti costante per tutta la sua durata.

In questo lavoro i questionari dicono come i partecipanti hanno percepito le attività, mentre le annotazioni temporali indicano quando ciascuna attività si colloca nelle registrazioni. Da queste due informazioni nascono le classi associate alle finestre di segnale PPG; il criterio adottato è descritto nella Sezione 2.2. Va ricordato anche quando si leggono i risultati: le prestazioni dei modelli misurano la capacità di riconoscere queste classi, non lo stress in senso assoluto.

1.2.3 Principi della fotopletismografia

La fotopletismografia (*photoplethysmography*, PPG) è una tecnica ottica non invasiva che rileva le variazioni di volume sanguigno nei tessuti. Nei sensori a contatto una sorgente luminosa, di solito uno o più LED, illumina la pelle, e un fotodetettore misura la luce che gli arriva dopo aver attraversato i tessuti [4].

Qui il segnale si collega alla fisiologia vista nella Sezione 1.1. La componente pulsatile del PPG segue soprattutto le variazioni di volume del sangue arterioso a ogni battito, e la forma dell'onda dipende dall'attività del cuore, dalle proprietà dei vasi e dal modo in cui le onde di polso, generate dall'eiezione, si propagano e si riflettono lungo l'albero arterioso. Il PPG dà quindi informazioni sulla circolazione periferica, ma non misura direttamente né l'attività elettrica del cuore né la pressione arteriosa [4].

Nel suo percorso la luce viene in parte assorbita e in parte diffusa. A ogni battito il volume di sangue arterioso cambia, e con esso la quantità di luce che raggiunge il fotodetettore: da qui nasce la componente pulsatile del segnale. Il fotodetettore trasforma la luce in un segnale elettrico che viene campionato nel tempo. Poiché la luce interagisce con più componenti del tessuto, il PPG non è una misura del volume di sangue in unità assolute [4].

A seconda di dove si trovano sorgente e fotodetettore si parla di trasmissione o di riflessione. In trasmissione i due elementi stanno ai lati opposti del tessuto e il ricevitore misura la luce che lo attraversa, come nei pulsossimetri da dito. In riflessione stanno dallo stesso lato e il ricevitore raccoglie la luce diffusa che torna verso la superficie: è la soluzione usata nella maggior parte dei dispositivi indossabili [4]. Le due configurazioni sono schematizzate nella Figura 1.6. Anche il sensore usato in questo lavoro, il Polar OH1+, misura in riflessione.

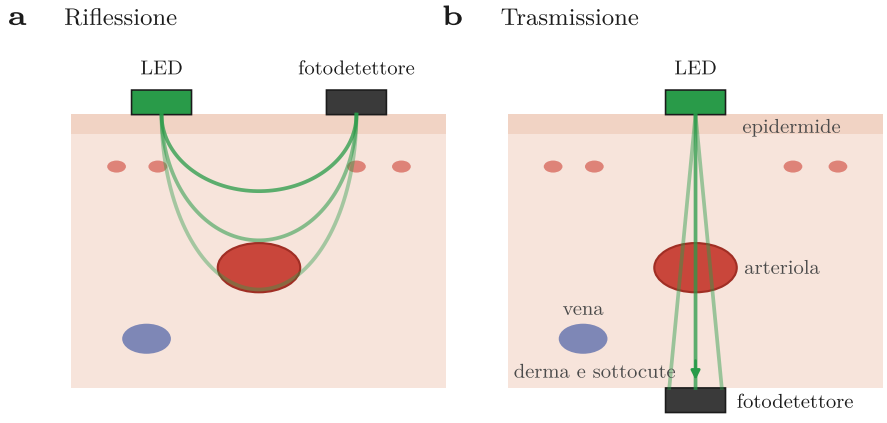


Figura 1.6: Configurazioni di acquisizione del PPG. In riflessione (a) LED e fotodetettore stanno dallo stesso lato e il fotodetettore raccoglie la luce diffusa dai tessuti, che segue un percorso curvo; in trasmissione (b) la luce attraversa il tessuto e viene raccolta dal lato opposto. Schema elaborato sulla base di [4].

Anche la lunghezza d'onda conta, perché determina quanto la luce penetra nei tessuti: la luce verde arriva in genere meno in profondità di quella rossa e infrarossa. Il segnale finale dipende quindi dalla configurazione ottica, dal sito anatomico e dal contatto con la pelle, e per lo stesso motivo il PPG viene di solito espresso in unità arbitrarie: la sua ampiezza dipende anche dal dispositivo e dalle condizioni di misura [4]. Per questa tesi è un aspetto centrale, perché i segnali registrati sul braccio e sulla schiena vengono analizzati separatamente.

1.2.4 Componenti e morfologia del segnale PPG

Nel segnale PPG si distinguono una componente di fondo, detta continua o DC, e una componente pulsatile, detta AC. La prima dipende dall'interazione della luce con i tessuti e con il sangue non pulsatile; la seconda segue le variazioni di volume del sangue arterioso a ogni battito [4].

Il nome "continua" non va preso alla lettera: il livello di fondo non resta fisso durante la registrazione. La respirazione e le condizioni di acquisizione possono introdurre variazioni lente e modificare anche l'ampiezza delle pulsazioni [4]. La Figura 1.7 mostra in modo schematico come sangue e tessuti contribuiscono al segnale ottico.

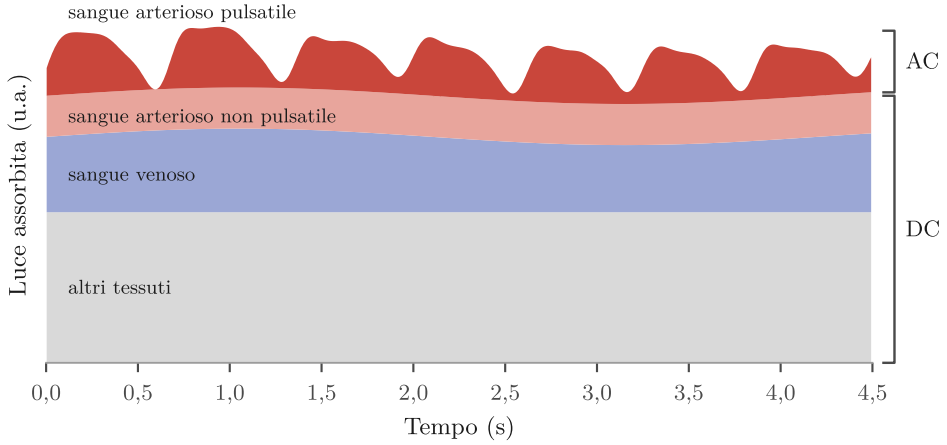


Figura 1.7: Contributi di tessuti e sangue alla luce assorbita. La componente continua (DC) dipende dagli altri tessuti, dal sangue venoso e dal sangue arterioso non pulsatile, e varia lentamente con la respirazione; la componente pulsatile (AC) segue il volume di sangue arterioso a ogni battito. Schema adattato da [4]; la forma delle pulsazioni è ricavata da un segnale del dataset (soggetto 2, schiena).

Se si rappresenta una pulsazione con il picco sistolico verso l'alto, si riconoscono un punto iniziale, detto piede dell'onda, una salita fino al picco e una fase di discesa. Il picco sistolico è il massimo principale. Nella discesa possono comparire l'incisura dicrota, una piccola depressione del profilo, e subito dopo il picco diastolico. Non sempre questi punti sono ben visibili: la forma dell'onda cambia con le caratteristiche

cardiovascolari della persona, con il sito di misura e con la qualità del segnale [4]. Un esempio è riportato nella Figura 1.8.

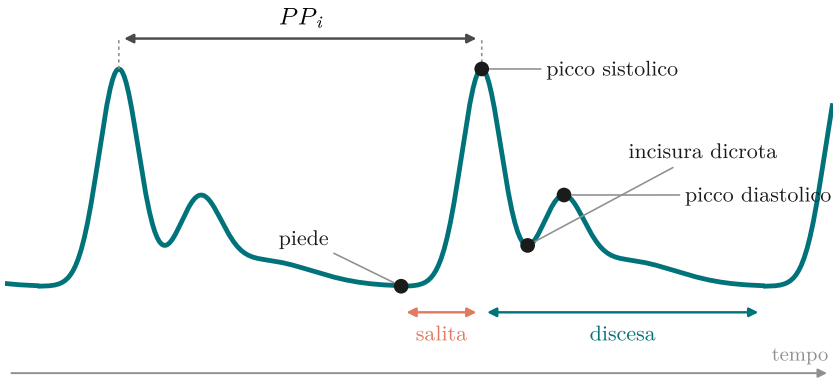


Figura 1.8: Punti caratteristici di una pulsazione PPG e intervallo picco-picco PP_i tra due picchi sistolici consecutivi. La forma è schematica: nei segnali reali l'incisura e il picco diastolico possono essere poco visibili o assenti. Schema elaborato sulla base di [4].

Dai punti caratteristici si ricavano misure temporali e di ampiezza. Se t_i e t_{i+1} sono gli istanti di due picchi sistolici consecutivi (Figura 1.8), l'intervallo picco-picco è

$$PP_i = t_{i+1} - t_i. \tag{1.1}$$

Con PP_i in secondi, la frequenza di pulsazione corrispondente, in pulsazioni al minuto, è

$$PR_i = \frac{60}{PP_i}. \tag{1.2}$$

Il calcolo funziona solo se i picchi individuati corrispondono davvero a pulsazioni consecutive: un picco perso o un falso rilevamento alterano sia gli intervalli sia la frequenza [4].

La variabilità degli intervalli ricavati dal PPG prende il nome di variabilità della frequenza di pulsazione (*pulse rate variability*, PRV). Non coincide con la variabilità della frequenza cardiaca (*heart rate variability*, HRV), che si misura sull'ECG. Il PPG registra l'arrivo dell'onda di polso in periferia, l'ECG l'attività elettrica del cuore, e il tempo che separa i due eventi non è costante. Soprattutto durante stress o attività fisica, le due misure non possono essere considerate equivalenti [4]. La PRV è stata proposta come sostituto della HRV quando non è disponibile un ECG, ma il rapporto tra le due non è semplice e servono ancora studi per capire quali processi influenzano la PRV oltre a quelli che agiscono sulla HRV [12]. Per questo, nel descrivere le feature temporali di questa tesi, si parlerà di intervalli tra pulsazioni periferiche.

1.2.5 Dispositivi indossabili, siti di acquisizione e artefatti

I dispositivi indossabili, o *wearable*, acquisiscono segnali fisiologici tramite sensori mantenuti a contatto con il corpo. Nel caso del PPG, i componenti ottici sono abbastanza piccoli da poter essere inseriti in bracciali e fasce. La qualità della registrazione dipende però molto da come il dispositivo è fissato e da come viene usato [4].

Il sito di misura cambia le caratteristiche del segnale. Differenze anatomiche e vascolari tra una zona e l'altra modificano ampiezza e forma delle pulsazioni, e conta anche la posizione del sensore rispetto al cuore. Il fissaggio deve garantire un contatto stabile, perché una pressione diversa sulla pelle può alterare la forma e i tempi dell'onda [4]. La maggior parte dei dispositivi commerciali misura al polso, mentre il braccio è usato raramente. Il braccio potrebbe però risentire meno degli artefatti da movimento rispetto a sedi più periferiche come il polso o il dito [14]. La Figura 1.9 mostra alcuni dispositivi commerciali che misurano il PPG. Tra questi c'è il Polar OH1, versione precedente del sensore usato per il dataset di questa tesi, indossato con la sua fascia sulla parte superiore del braccio; nella vista dal basso si vedono i sei LED verdi disposti intorno al fotodetettore.



Figura 1.9: Esempi di dispositivi indossabili che misurano il PPG, visti indossati (sopra) e dal lato del sensore (sotto). Immagine di V. Marozas e P. H. Charlton, licenza CC BY 4.0 [14].

Il problema principale dei wearable sono gli artefatti da movimento: quando la persona si muove, il segnale si deforma, le pulsazioni diventano meno riconoscibili e possono comparire picchi che non corrispondono a battiti reali. Anche la luce ambientale e un cattivo contatto tra sensore e pelle possono rovinare la registrazione [4].

Non tutto ciò che altera il PPG è però un errore. Respirazione, temperatura e variazioni della perfusione periferica sono fenomeni fisiologici veri; se disturbano o meno dipende da cosa si vuole misurare, perché una componente utile per stimare un parametro può diventare un'interferenza per un altro [4].

Per questi motivi il segnale va preparato prima dell'analisi, con operazioni di preprocessing. Il filtraggio attenua le componenti indesiderate ma può modificare la forma dell'onda, quindi va scelto in base alle feature che si vogliono estrarre; una valutazione della qualità aiuta invece a individuare i tratti in cui le stime potrebbero essere poco affidabili [4]. Sono aspetti che pesano direttamente sul confronto tra braccio e schiena, perché i due siti possono dare segnali di qualità diversa.

Un esempio di come gli artefatti alterano il segnale è riportato nella Figura 1.10, ottenuta da due registrazioni del dataset usato in questa tesi. Nel tratto pulito ogni battito ha la stessa forma, con picco sistolico e incisura dicrota ben riconoscibili. Nel tratto con artefatti, registrato durante lo speech test, compaiono in due punti oscillazioni ampie e irregolari, con ogni probabilità dovute a movimenti, che per alcuni secondi coprono del tutto le pulsazioni. Rispetto a queste oscillazioni, i battiti regolari del resto della finestra risultano schiacciati dalla normalizzazione.

1.3 Machine Learning

1.3.1 Apprendimento supervisionato e classificazione

Nell'apprendimento supervisionato il modello viene addestrato su esempi di cui si conoscono sia i dati di ingresso sia la risposta corretta, chiamata etichetta o *label*. Quando ogni esempio è descritto da valori numerici, detti *feature*, il modello impara a collegare queste feature all'etichetta [15].

Si parla di classificazione quando l'obiettivo è assegnare un'osservazione a una categoria tra quelle previste; nella regressione, invece, si stima un valore numerico. Codificare le classi con numeri interi non cambia la natura del problema, che resta una classificazione [15].

Durante l'addestramento, o *training*, il modello impara dagli esempi etichettati. In fase di predizione riceve dati nuovi, senza etichetta, e propone una classe. Per valutarlo si confrontano le sue predizioni con le etichette vere di un insieme di test che non ha mai visto [15].

Andare bene sui dati di training non basta: un modello utile deve generalizzare, cioè fare predizioni affidabili su dati che non ha usato per imparare [15]. Le scelte descritte nelle prossime sottosezioni servono a stimare questa capacità in modo corretto.

Il problema affrontato in questa tesi è di classificazione: a ogni finestra di segnale va assegnato un livello di stress. Da qui in avanti si farà quindi riferimento solo a questo tipo di problema, e quanto descritto su feature, bilanciamento, validazione e metriche va inteso per la classificazione.

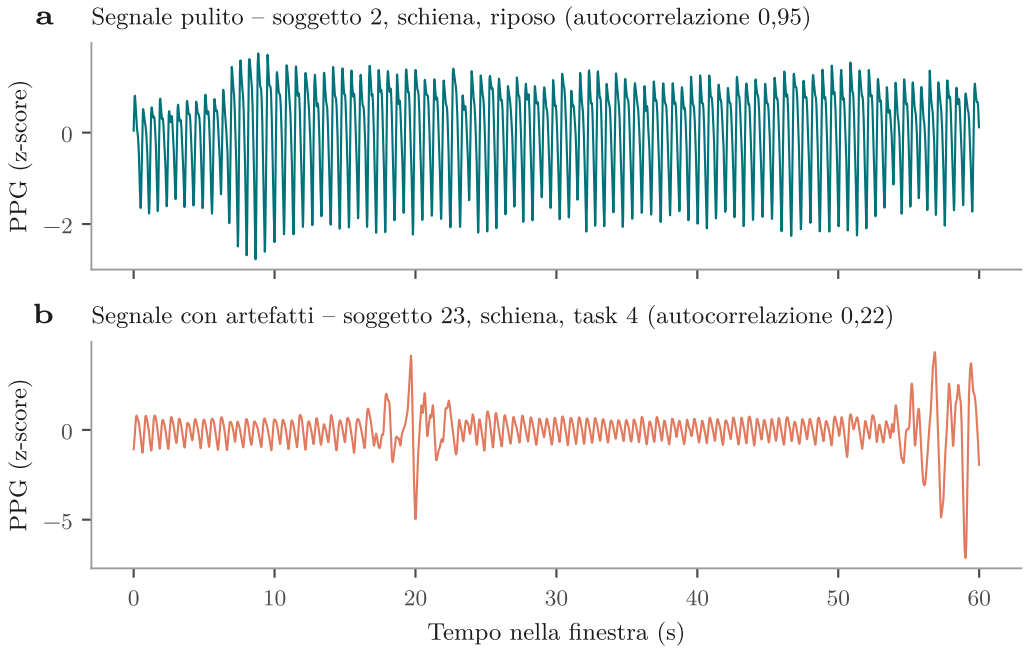


Figura 1.10: Finestre di 60 s di segnale PPG acquisito sulla schiena: segnale pulito (a) e segnale con artefatti (b). Nel titolo di ogni pannello è riportata l'autocorrelazione della finestra, usata come indice di qualità del segnale: è il massimo dell'autocorrelazione normalizzata per ritardi tra 0,33 e 1,5 s, cioè nell'intervallo della durata di un battito. Vale circa 1 quando i battiti si ripetono con forma e ritmo regolari e si abbassa quando gli artefatti rendono il segnale irregolare. Ciascuna finestra è normalizzata con il proprio z-score, quindi le ampiezze dei due pannelli non sono confrontabili.

1.3.2 Estrazione e selezione delle feature

Estrarre le feature significa ricavare dai dati un insieme di valori numerici che ne descrivano gli aspetti utili all'analisi. Per un segnale fisiologico possono essere proprietà statistiche, temporali o spettrali. Ogni osservazione diventa così un vettore di numeri che il classificatore può ricevere in ingresso [2].

La selezione lavora invece su feature già calcolate e ne sceglie un sottoinsieme. L'estrazione costruisce la rappresentazione dei dati, la selezione la riduce. Togliere variabili poco informative può semplificare il modello, ma non porta automaticamente a prestazioni migliori [16].

I criteri di selezione sono diversi. Alcuni misurano il legame statistico tra ogni feature e la classe, altri si basano sull'importanza attribuita da un modello, altri ancora confrontano più sottoinsiemi. Il criterio deve essere adatto al problema: una misura pensata per la regressione, ad esempio, non va bene necessariamente per la classificazione [16].

La selezione va appresa solo sui dati di training e ripetuta a ogni suddivisione della validazione; le feature scelte vengono poi applicate al test così come sono. Se il test partecipasse alla scelta, informazioni che servono per la valutazione finirebbero nello sviluppo del modello. Questo errore si chiama *data leakage* e porta a stime delle prestazioni troppo ottimistiche [17].

1.3.3 Sbilanciamento delle classi

Un dataset è sbilanciato quando le classi hanno numerosità diverse. Se una categoria è molto più frequente delle altre, il modello tende a favorirla e fa più fatica a riconoscere quelle rare. Lo sbilanciamento va quindi considerato sia nell'addestramento sia nella scelta delle metriche.

Una prima soluzione è dare pesi diversi alle classi: gli esempi delle classi meno rappresentate contano di più durante l'addestramento, senza bisogno di crearne di nuovi. Di solito i pesi sono inversamente proporzionali alla numerosità delle classi nel training [18].

L'alternativa è il sovracampionamento, che aumenta il numero di esempi delle classi minoritarie. La tecnica SMOTE (*Synthetic Minority Over-sampling Technique*) crea esempi sintetici interpolando tra osservazioni vicine della stessa classe nello spazio delle feature. Non duplica esempi esistenti: genera nuovi vettori di feature, che però non corrispondono a nuove acquisizioni reali [19, 20]. Per un'osservazione $\mathbf{x}_i$ della classe minoritaria si sceglie a caso uno dei suoi k vicini più prossimi della stessa classe, $\mathbf{x}_j$, e si genera

$$\mathbf{x}_{\text{nuovo}} = \mathbf{x}_i + \lambda\,(\mathbf{x}_j - \mathbf{x}_i), \qquad 0 \le \lambda \le 1, \tag{1.3}$$

con λ estratto a caso. Il nuovo esempio si trova quindi sul segmento che unisce le due osservazioni reali, come mostrato nella Figura 1.11 [19].

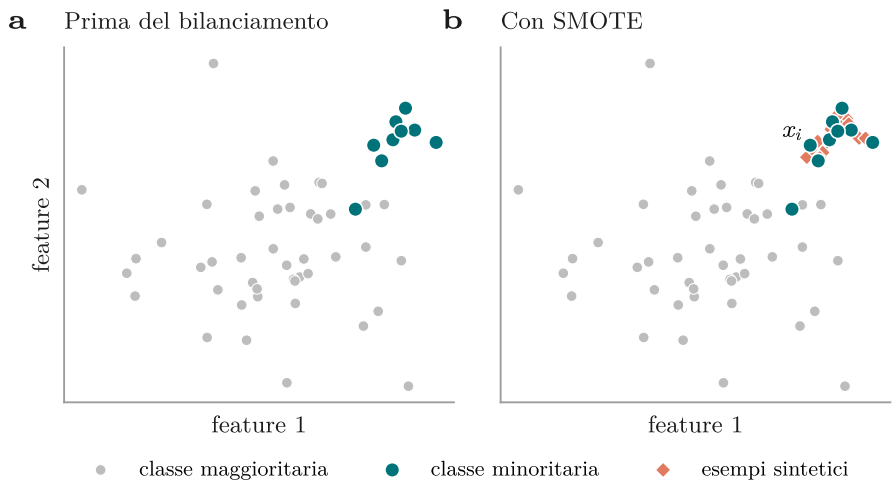


Figura 1.11: Esempio schematico di SMOTE in due dimensioni, con dati simulati. Gli esempi sintetici (b) si trovano sui segmenti che uniscono ogni osservazione della classe minoritaria ai suoi vicini della stessa classe; per $\mathbf{x}_i$ sono tratteggiati i collegamenti con i tre vicini più prossimi.

Nessuna delle due strategie garantisce da sola un miglioramento, e vanno entrambe verificate sui dati. SMOTE, in particolare, rende poco quando ci sono osservazioni anomale o classi che si sovrappongono molto [20]. In ogni caso il bilanciamento riguarda solo il training: i pesi si calcolano sui dati di addestramento, SMOTE genera esempi solo a partire da questi e il test non viene toccato, così da conservarne la distribuzione reale [21].

1.3.4 Validazione incrociata per soggetto

Quando un dataset contiene più osservazioni per ciascun partecipante, i dati della stessa persona possono condividere caratteristiche individuali. Una suddivisione casuale delle singole osservazioni tra training e test non è quindi adeguata a valutare la generalizzazione a persone nuove, poiché può collocare dati dello stesso individuo in entrambi gli insiemi [22].

La validazione incrociata per soggetto mantiene separati i partecipanti. I soggetti vengono suddivisi in k gruppi, detti *fold*: a ogni iterazione le osservazioni dei soggetti di un fold sono riservate al test e quelle degli altri formano il training. Si eseguono così k iterazioni, addestrando ogni volta un nuovo modello, e ogni partecipante finisce nel test una sola volta [22].

Nella versione stratificata, la suddivisione cerca anche di mantenere in ogni fold una distribuzione delle classi simile a quella complessiva, sempre senza dividere i dati di uno stesso soggetto. Il caso limite, in cui ogni fold contiene un solo partecipante, è la *leave-one-subject-out* (LOSO): con N soggetti si fanno N iterazioni. Usare meno fold riduce il numero di addestramenti, ma ogni modello viene addestrato su meno soggetti [22].

Questa separazione riguarda l'intera procedura di apprendimento, non soltanto il classificatore. Le trasformazioni che richiedono una stima dai dati, come la standardizzazione delle feature, devono essere adattate al training e poi applicate al test. Lo stesso vale per la selezione delle feature e per il bilanciamento [17, 21].

Se occorre scegliere gli iperparametri, ossia le impostazioni che regolano l'addestramento, il confronto tra configurazioni va fatto con una validazione interna ai soli dati di training. Si ottiene così una procedura annidata: il livello interno guida la scelta, mentre i soggetti esclusi nel livello esterno restano riservati alla valutazione. Anche la suddivisione interna deve tenere separati i soggetti, se l'obiettivo è valutare la generalizzazione da una persona all'altra [23, 22].

La validazione per soggetto permette quindi di stimare le prestazioni su persone non viste durante l'addestramento, all'interno del dataset disponibile. Non sostituisce però una verifica su una popolazione esterna o in condizioni di acquisizione diverse.

1.3.5 Classificatori considerati

I classificatori si differenziano sia per il modo in cui imparano sia per il tipo di dati che ricevono. In questa tesi ne vengono usati tre: la Random Forest (RF), la rete neurale a strati densi, indicata come NN, e la rete convoluzionale monodimensionale (1D-CNN). Qui se ne descrive il funzionamento generale; le configurazioni usate negli esperimenti sono riportate nella Sezione 2.7.

Random Forest La Random Forest combina le predizioni di molti alberi decisionali. Ogni albero applica una serie di regole sui valori delle feature e divide progressivamente le osservazioni fino ad arrivare a una predizione. Nella versione più comune, ogni albero viene addestrato su un campione estratto con reinserimento dal training e, a ogni divisione, considera solo un sottoinsieme casuale delle feature. In questo modo gli alberi risultano diversi tra loro, e la loro combinazione è meno sensibile alle variazioni dei dati rispetto a un albero singolo [24, 25]. Per classificare una nuova osservazione si combinano le risposte dei singoli alberi: nella formulazione originale con un voto di maggioranza, nell'implementazione di scikit-learn usata in questa tesi con la media delle probabilità stimate dagli alberi (Figura 1.12).

Rete neurale a strati densi (NN) Una rete neurale (*neural network*, NN) è formata da unità di calcolo collegate da pesi che vengono appresi durante l'addestramento. In una rete a strati densi ogni unità combina gli ingressi provenienti dallo strato precedente e applica una funzione di attivazione. Gli strati intermedi permettono di rappresentare relazioni non lineari tra feature e classi. L'addestramento modifica i pesi in modo da ridurre una funzione di perdita, che misura quanto le predizioni si discostano dalle etichette [26].

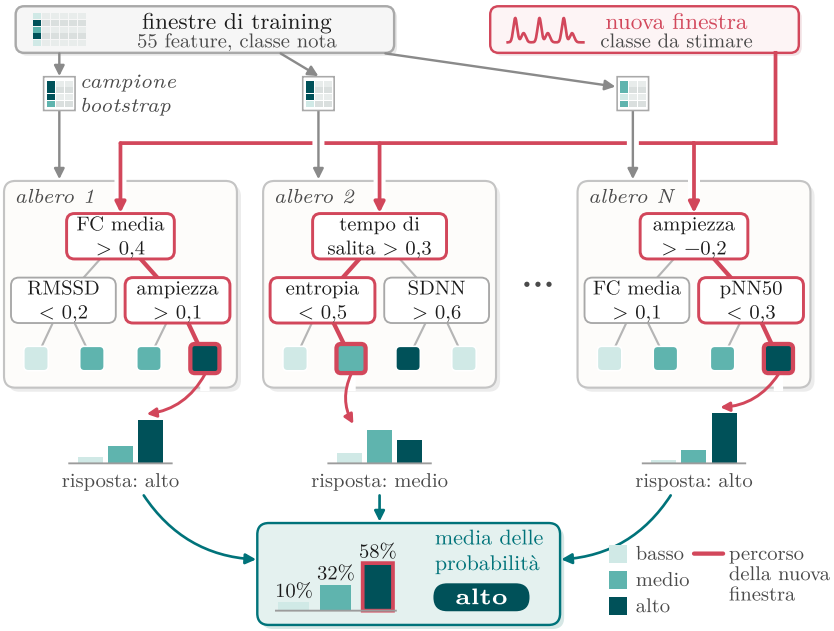


Figura 1.12: Schema della Random Forest. Ogni albero è addestrato su un campione bootstrap del training e a ogni divisione considera solo un sottoinsieme casuale delle feature; ogni nodo confronta una feature con una soglia. Una nuova finestra percorre ciascun albero fino a una foglia (percorso evidenziato), che fornisce le probabilità delle tre classi. La predizione finale è la classe con la probabilità media più alta su tutti gli alberi; feature e soglie sono puramente illustrative.

1D-CNN Anche la 1D-CNN è una rete neurale. Usa filtri appresi che scorrono lungo una dimensione dell'ingresso, ad esempio l'asse del tempo. Applicata a sequenze PPG, può imparare direttamente dalla forma d'onda gli andamenti locali utili alla classificazione, invece di ricevere feature calcolate in anticipo [27].

Quando in questa tesi si distinguono NN e 1D-CNN, quindi, si confrontano due architetture, non due famiglie di modelli separate. La differenza pratica è nell'ingresso: la NN e la Random Forest lavorano sulle feature, la 1D-CNN sul segnale. Anche se la 1D-CNN impara da sola la propria rappresentazione, il segnale va comunque preparato e training e test vanno tenuti separati. Le architetture usate in questa tesi sono descritte nella Sezione 2.7.

1.3.6 Metriche di valutazione

Per valutare un classificatore si confrontano le classi predette con le etichette vere del test. Con classi sbilanciate un solo numero può nascondere grosse differenze tra una categoria e l'altra, per questo conviene usare più metriche insieme alla matrice di confusione.

Matrice di confusione. La matrice di confusione indica, per ogni classe reale, quante osservazioni sono state assegnate a ciascuna classe predetta. Con le classi reali sulle righe e quelle predette sulle colonne, la diagonale contiene le classificazioni corrette e gli altri elementi mostrano quali classi il modello confonde.

Accuracy. L'accuracy è la frazione di osservazioni classificate correttamente:

$$\text{Accuracy} = \frac{N_{\text{corrette}}}{N_{\text{totali}}}.$$

Se le classi hanno numerosità diverse, l'accuracy può essere alta anche quando il modello riconosce bene solo la classe più frequente. Per questo va affiancata da metriche calcolate classe per classe.

Precision e recall. Precision e recall si calcolano separatamente per ciascuna classe, considerando le osservazioni come appartenenti a quella classe o a tutte le altre. I veri positivi (TP) sono i casi assegnati correttamente alla classe, i falsi positivi (FP) quelli assegnati per errore, i falsi negativi (FN) i casi della classe che il modello attribuisce ad altre categorie.

La precision dice quante delle osservazioni assegnate alla classe vi appartengono davvero:

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

Il recall dice quante delle osservazioni della classe vengono effettivamente riconosciute:

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

La prima misura riguarda l'affidabilità delle assegnazioni, la seconda la capacità di non perdere i casi della classe.

F1-score. L'F1-score è la media armonica di precision e recall della stessa classe:

$$F1 = 2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

e risulta alto solo se lo sono entrambe.

Balanced accuracy e macro-F1. Per riassumere le prestazioni su tutte le classi si usano medie che danno a ogni classe lo stesso peso. La balanced accuracy è la media dei recall delle K classi, il macro-F1 la media dei loro F1-score:

$$\text{Balanced accuracy} = \frac{1}{K}\sum_{\text{classi}} \text{Recall}, \qquad \text{Macro-F1} = \frac{1}{K}\sum_{\text{classi}} F1.$$

In questo modo una classe poco numerosa conta quanto una molto frequente. La media pesata, al contrario, tiene conto della numerosità delle classi e risente di più di quelle frequenti. Per questo le misure complessive vanno sempre lette insieme ai risultati per classe.

Curva ROC. Le metriche precedenti dipendono dalla regola di decisione, che assegna a ogni osservazione la classe con la probabilità stimata più alta. La curva ROC (*receiver operating characteristic*) valuta invece direttamente le probabilità: considerando una classe contro tutte le altre, riporta il tasso di veri positivi in funzione di quello di falsi positivi al variare della soglia. L'area sotto la curva (AUC) vale 1 per una separazione perfetta e 0,5 per un classificatore casuale, e non dipende dalla soglia scelta [28].

Interpretazione nella validazione per soggetto. Nella validazione per soggetto ogni partecipante viene valutato quando il fold che lo contiene è escluso dall'addestramento. Le metriche si possono calcolare riunendo le predizioni di test di tutti i fold, oppure separatamente per fold o per soggetto. Non sono equivalenti: il primo modo descrive le prestazioni sull'insieme delle osservazioni, il secondo mette in evidenza le differenze tra individui. Nel riportare i risultati va quindi sempre indicato quale dei due è stato usato.

1.4 Stato dell'arte

Questa sezione colloca la tesi rispetto agli studi esistenti. Il punto di partenza sono stati i due articoli del gruppo di ricerca, i capitoli sul PPG di Charlton e colleghi e alcune review sulla stress detection. A queste fonti si è aggiunta una ricerca bibliografica su stress detection, PPG e dispositivi indossabili, con l'obiettivo di

trovare studi che usassero il solo PPG, più di due livelli di stress o una validazione per soggetto, cioè le caratteristiche più vicine a questo lavoro.

1.4.1 Review sulla stress detection con dispositivi indossabili

Due review recenti danno un'idea di come si muove il settore. Bolpagni et al. [3] hanno analizzato 32 studi sulla stress detection personalizzata con dispositivi indossabili. I segnali più usati sono l'attività elettrodermica e il PPG, e 19 studi su 32 combinano più sensori. I dataset pubblici coinvolgono in media 42 persone, quasi sempre in laboratorio, con uno sbilanciamento evidente per età e sesso. Tra i dispositivi censiti, quelli da polso sono i più numerosi. Tra i modelli di machine learning sono molto usati quelli basati su alberi decisionali, scelti anche perché più interpretabili, mentre le reti convoluzionali monodimensionali sono la soluzione di deep learning più diffusa [3].

La review di Namvari et al. [29], dedicata ai dispositivi basati sul PPG, conferma il quadro. Il polso è la sede di misura più frequente, la variabilità del battito è il parametro più usato e i campioni sono piccoli: solo uno degli studi inclusi supera i 70 partecipanti. Gli autori indicano come limiti principali la scarsità di studi ampi e la debolezza delle etichette di riferimento, spesso ottenute da questionari o da protocolli di laboratorio con stress brevi e intensi [29].

1.4.2 Classificazione multiclasse con il solo PPG

Il dataset WESAD, uno dei più usati in letteratura, permette di vedere quanto vale il PPG preso da solo. Su 15 soggetti, con validazione *leave-one-subject-out*, il PPG acquisito al polso da un Empatica E4 distingue stress e non stress con un'accuracy dell'85,8% e un F1-score di 0,83. Quando le classi diventano tre (riposo, stress e divertimento), l'accuracy scende al 70,2% e l'F1-score a 0,55. Combinando tutti i segnali del sensore toracico, il problema a tre classi arriva all'80,3% di accuracy [30].

Lo stesso calo si osserva in altri lavori. Seo et al., per esempio, riconoscono due livelli di stress da lavoro con un'accuracy del 73,3%, combinando respirazione e immagini del volto, ma ne riconoscono tre solo nel 54,4% dei casi, combinando ECG, respirazione e immagini del volto con una rete neurale profonda [31]. I due risultati non usano gli stessi segnali, quindi il calo non dipende solo dal numero di classi. Aggiungere livelli intermedi rende il problema molto più difficile, perché le risposte fisiologiche a stimoli di intensità diversa tendono a sovrapporsi.

1.4.3 PPG in sedi di misura alternative

Pochi lavori hanno studiato il PPG lontano dal polso e dal dito, e quasi sempre con obiettivi diversi dal riconoscimento dello stress. Longmore et al. [32] hanno confrontato otto sedi (dito, fronte, tempia, collo, torace, polso, regione lombare e tibia) su quattro soggetti, con lo stesso sensore a riflessione a basso costo e finestre

di 60 s. A riposo l'errore mediano sulla frequenza cardiaca era dell'1,3–1,4% su dito e fronte, mentre superava il 50% su polso, collo, regione lombare e tibia. La sede, quindi, può cambiare molto la qualità del segnale, anche se con un sensore e un fissaggio diversi da quelli usati in questa tesi.

Hettiarachchi et al. [33] hanno valutato il Polar OH1, la versione precedente del sensore usato in questa tesi, su 24 soggetti durante esercizio su tapis roulant e cyclette, indossato su avambraccio, parte superiore del braccio e tempia. Rispetto all'ECG, la correlazione intraclasse della frequenza cardiaca era di 0,99 e la differenza media di circa 0,3 bpm in tutte e tre le sedi. Sulla parte superiore del braccio questo tipo di sensore misura quindi in modo affidabile la frequenza cardiaca, anche durante il movimento.

Per lo stress, Barki e Chung [34] hanno acquisito il PPG nell'orecchio, sul trago, su 14 soggetti sottoposti a Stroop test e a calcoli aritmetici. Dopo aver attenuato gli artefatti da movimento con un filtro adattivo che usa come riferimento un accelerometro, una rete convoluzionale applicata alle trasformate wavelet del PPG ha distinto stress e non stress con un'accuracy del 92%, usando 11 soggetti per l'addestramento e 3 per il test. Il risultato è promettente, ma riguarda due classi e un test su soli tre soggetti.

1.4.4 Lavori precedenti del gruppo di ricerca

Lo studio di Campanella et al. [2] è il riferimento di metodo più vicino a questa tesi. Su 29 soggetti, con un Empatica E4 al polso, sono state estratte 27 feature da PPG e attività elettrodermica e confrontati Random Forest, Support Vector Machine e regressione logistica, con e senza selezione delle feature. La Random Forest con tutte le feature ha raggiunto un'accuracy del 76,5% nel distinguere stress e non stress. La validazione era però a 10 fold sui segmenti, quindi segmenti della stessa persona potevano finire sia nel training sia nel test [2]. Il percorso preprocessing, feature, classificazione è lo stesso seguito qui. Cambiano però i segnali (qui solo PPG), il numero di classi e il modo di valutare.

Il secondo studio, di Campanella, Seepold e Palma, ha prodotto il dataset usato in questa tesi. Confronta il PPG acquisito sulla parte superiore del braccio sinistro e nella regione lombare, due sedi scelte perché già usate dai sensori per il monitoraggio continuo del glucosio nel diabete di tipo 1, anche se le acquisizioni riguardano soggetti sani. Il protocollo comprende 20 minuti di attività libera e una fase di induzione dello stress con compiti cognitivi e di esposizione sociale [5].

Nella parte di attività libera la regione lombare ha mostrato un segnale leggermente più pulito e stabile, il braccio una maggiore variabilità e più rumore. Nella classificazione binaria tra riposo e stress, con una Random Forest, il braccio ha raggiunto un'accuracy del 66–67% con e senza selezione delle feature, la schiena il 72% usando tutte le feature [5]. Anche qui la validazione stratificata a 5 fold divideva i segmenti e non i soggetti. Questi risultati giustificano l'analisi separata dei due siti,

ma non dimostrano che la schiena sia sempre la scelta migliore. Lo dicono gli stessi autori, che presentano lo studio come preliminare.

La Tabella 1.1 riassume gli studi citati in questa sezione, escluse le review, e li mette a confronto con questo lavoro.

Tabella 1.1: Studi citati nello stato dell'arte (escluse le review) e confronto con questa tesi. FC: frequenza cardiaca; ICC: correlazione intraclasse.

Studio	Segnali e sede	Classi	Validazione	Risultato
Schmidt et al. [30]	solo PPG, polso (15 soggetti)	2 / 3	LOSO	acc. 85,8% / 70,2%
Seo et al. [31]	respirazione e immagini del volto (2 classi); ECG, respirazione e immagini del volto (3 classi)	2 / 3	–	acc. 73,3% / 54,4%
Barki e Chung [34]	PPG all'orecchio come ingresso del classificatore, accelerometro per la rimozione degli artefatti (14 soggetti)	2	11 soggetti di training, 3 di test	acc. 92,0%
Campanella et al. [2]	PPG ed EDA, polso (29 soggetti)	2	10 fold su segmenti	acc. 76,5%
Campanella, Seepold e Palma [5]	solo PPG, braccio e schiena (27 soggetti)	2	5 fold su segmenti	acc. 66–67% / 72%
Longmore et al. [32]	PPG in 8 sedi, tra cui polso e regione lombare (4 soggetti)	–	confronto con un riferimento	errore FC a riposo 1,3–1,4% (dito, fronte), oltre 50% (polso, lombare)
Hettiarachchi et al. [33]	Polar OH1 su avambraccio, braccio e tempia (24 soggetti)	–	confronto con l'ECG	ICC 0,99 per la FC
Questa tesi	solo PPG, braccio e schiena (25 soggetti per sito)	3	5 fold per soggetto, ripetuta 10 volte	vedi Capitolo 3

Dal confronto emergono gli aspetti meno esplorati, sui quali si concentra questa tesi:

- il PPG acquisito in sedi diverse dal polso e dal dito, in particolare la regione lombare, per la quale non esiste ancora una letteratura sistematica [5];
- la classificazione in più di due livelli di stress con il solo PPG;
- una valutazione su soggetti mai visti dal modello. Molti studi dividono i segmenti in modo casuale e possono quindi sovrastimare le prestazioni attese su una persona nuova.

1.5 Contributo della tesi

Gli aspetti meno esplorati emersi dallo stato dell'arte portano a una domanda pratica: quanto si riesce a riconoscere il livello di stress usando solo il PPG, acquisito

in sedi poco convenzionali come il braccio e la schiena, quando il modello viene valutato su persone che non ha mai visto? È la domanda da cui parte questa tesi.

Il lavoro, svolto durante il tirocinio, usa registrazioni PPG acquisite sul braccio e sulla schiena durante un protocollo con fasi di riposo e quattro compiti. Le fasi del protocollo sono raggruppate in tre livelli di stress (basso, medio e alto) e i due siti sono analizzati sempre separatamente, così da poterli confrontare senza mescolarne i dati.

Il lavoro ha riguardato l'intera elaborazione dei dati, già acquisiti dal gruppo di ricerca:

- la lettura dei file grezzi, la sincronizzazione delle etichette con i segnali e la correzione degli orari anomali;
- il preprocessing del segnale, con la verifica e la correzione della polarità dell'onda sulla forma del battito;
- la segmentazione in finestre e l'estrazione di 55 feature temporali, spettrali e morfologiche;
- la classificazione con Random Forest, reti neurali dense e una rete convoluzionale, con la scelta di feature, bilanciamento e iperparametri all'interno del training tramite una procedura annidata;
- la valutazione per finestra e per compito, ripetuta più volte, con intervalli di confidenza calcolati sui soggetti;
- l'analisi delle brevi pause tra i compiti, che ha portato a escluderle dal dataset.

Braccio e schiena sono sempre analizzati con la stessa procedura, ma separatamente. L'obiettivo non è dimostrare che un sito sia migliore dell'altro in assoluto, ma capire quali scelte contano davvero per ciascuno dei due, e con quali limiti. I metodi sono descritti nel Capitolo 2, i risultati nel Capitolo 3.

Il tirocinio si colloca nell'ambito dell'elettronica per la salute. I sensori indossabili come il Polar OH1+ sono sistemi elettronici completi: una sorgente ottica a LED, un fotodetettore, l'elettronica di acquisizione e la trasmissione dei dati senza fili. Il lavoro si è concentrato sulla parte che segue l'acquisizione, cioè su come trasformare il segnale misurato da questi dispositivi in un'informazione utile.

Capitolo 2

Materiali e metodi

In questo capitolo sono descritti i dati e i passaggi dell'elaborazione, dalle registrazioni al protocollo di validazione dei modelli. La Figura 2.1 riassume la configurazione usata, i cui passaggi vengono poi ripresi uno per volta.

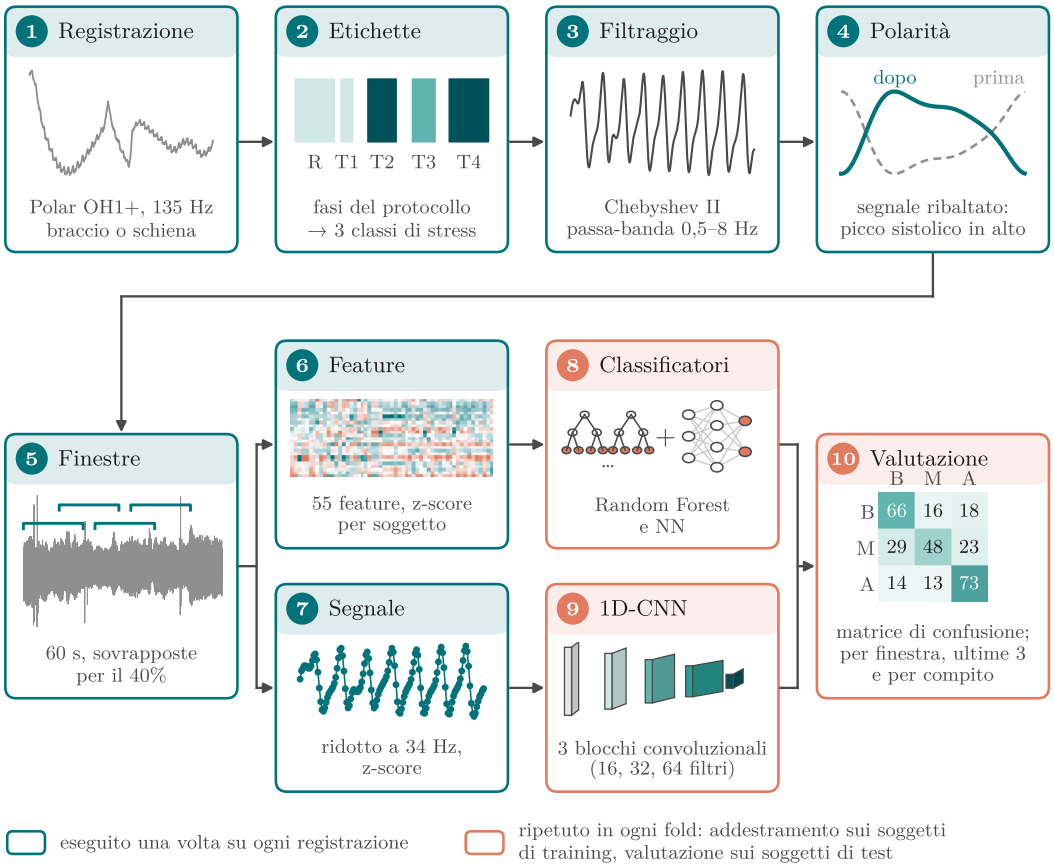


Figura 2.1: Schema della configurazione finale, dalle registrazioni alla valutazione dei modelli. Le miniature mostrano dati reali del soggetto 2 (schiena); la matrice di confusione è quella della NN sulla schiena, in percentuale per riga (B, M, A: stress basso, medio, alto). Braccio e schiena sono elaborati separatamente con la stessa pipeline; le pause tra i compiti sono escluse.

2.1 Dataset e protocollo di acquisizione

I dati provengono dallo studio di Campanella, Seepold e Palma descritto nella Sezione 1.4 [5]. Allo studio hanno partecipato 27 soggetti sani, reclutati tra l'Università Politecnica delle Marche e la HTWG di Costanza, 16 uomini e 11 donne, di età compresa tra i 18 e i 50 anni (la maggior parte tra i 21 e i 30). Il PPG è stato acquisito con due sensori Polar OH1+, che campionano il segnale ottico a 135 Hz: uno posto sulla parte superiore del braccio sinistro e fissato con la fascia elastica in dotazione, l'altro sulla regione lombare, applicato con un cerotto di 8×10 cm [5]. Il Polar OH1+ è stato scelto perché è pratico, poco ingombrante e di dimensioni simili a quelle dei sensori per il glucosio; il modulo ottico ha sei LED e il dispositivo trasmette i dati via Bluetooth. Le registrazioni sono state fatte con Polar Sensor Logger, un'applicazione Android che può collegarsi a più sensori contemporaneamente e salvare il segnale PPG insieme a data e ora, installata su un tablet [5]. Le due registrazioni sono sincrone, quindi per ogni soggetto si dispone dello stesso protocollo visto da due siti diversi. Il posizionamento dei due sensori è mostrato nella Figura 2.2.



Figura 2.2: Posizionamento dei sensori Polar OH1+ durante l'acquisizione: regione lombare, con medicazione adesiva (a sinistra), e parte superiore del braccio sinistro, con fascia elastica (a destra) [5].

Il protocollo di induzione dello stress comincia con 5 minuti di riposo con musica rilassante, che fanno da baseline, seguiti da quattro compiti, tutti svolti in inglese per avere le stesse condizioni tra partecipanti di lingua diversa [5]:

1. *Stroop test* [35] al computer, realizzato con PsyToolkit [36, 37]: compaiono nomi di colori scritti in un inchiostro di colore diverso, e il partecipante deve indicare prima il colore dell'inchiostro e poi, al contrario, la parola letta. Un rumore di cantiere in sottofondo aumenta il carico cognitivo;

2. *test aritmetico*: sottrarre 13 a voce alta partendo da 509, ricominciando da capo a ogni errore, per 5 minuti e con lo stesso rumore di fondo. È una variante del

compito aritmetico usato nel Trier Social Stress Test e nel Montreal Imaging Stress Task [38, 39];

3. *sing-a-song test*, nella versione breve proposta da van der Mee et al. [40]: lettura silenziosa di una frase neutra per 20 s, lettura ad alta voce per 30 s e infine canto di una canzone nota per 30 s, ciascuna fase preceduta da un conto alla rovescia di 30 s;

4. *speech test*: un minuto per preparare un breve discorso su un argomento non di competenza e due minuti per esporlo davanti allo sperimentatore, che mantiene un'espressione neutra, come nella prova di discorso del Trier Social Stress Test [38].

I compiti sono di tipo mentale e non richiedono attività fisica. In questo modo si limita la sovrapposizione tra le variazioni cardiovascolari dovute al movimento e quelle dovute allo stress [5], anche se parlare e cantare comportano comunque piccoli movimenti e modifiche della respirazione. La sequenza delle fasi e la loro durata tipica sono mostrate nella Figura 2.3a; la Figura 2.3b riporta, nelle stesse fasi, l'andamento medio della frequenza di pulsazione stimata dal PPG. Tra il riposo e lo Stroop test non è prevista una pausa etichettata: il compito inizia subito dopo la fine del riposo. Tutti i partecipanti hanno dato il consenso informato prima delle acquisizioni [5]. La postazione usata durante le prove è mostrata nella Figura 2.4.

Il materiale ricevuto comprende 52 file, uno per sito, relativi ai soggetti dal 2 al 27, un file con gli orari di inizio e fine di ogni fase del protocollo e i questionari compilati dai partecipanti. Da ogni file è stata usata la colonna del segnale PPG (`col_4`) insieme a data e ora di ogni campione. La struttura dei file è mostrata nella Tabella 2.1: ogni riga è un campione, con le colonne numeriche registrate dal sensore e la marca temporale. Tra un campione e il successivo passano circa 7,4 ms, cioè 1/135 di secondo; solo il primo pacchetto di dati di ogni file ha un passo leggermente più lungo, di circa 7,7 ms.

Tabella 2.1: Sei righe consecutive di un file di registrazione (soggetto 10, schiena, circa 12 minuti dopo l'inizio), così come esportato dall'applicazione di acquisizione. In grassetto la colonna del segnale PPG usata nell'analisi.

col_2	col_3	col_4	col_5	col_6	date	time
		...				
538155061452424064	554298	**370853**	309790	5744	2025-10-17	16:58:58.134
538155061459807744	554098	**370737**	309514	5780	2025-10-17	16:58:58.142
538155061467191488	554365	**370842**	309595	5760	2025-10-17	16:58:58.149
538155061474575232	554458	**370776**	309731	5791	2025-10-17	16:58:58.156
538155061481958912	554372	**370975**	309750	5718	2025-10-17	16:58:58.164
538155061489342656	554259	**370710**	309742	5810	2025-10-17	16:58:58.171
		...				

Nel file degli orari è stato corretto un refuso relativo al riposo del soggetto 5, segnato un'ora prima dell'inizio effettivo della registrazione. Il materiale ricevuto

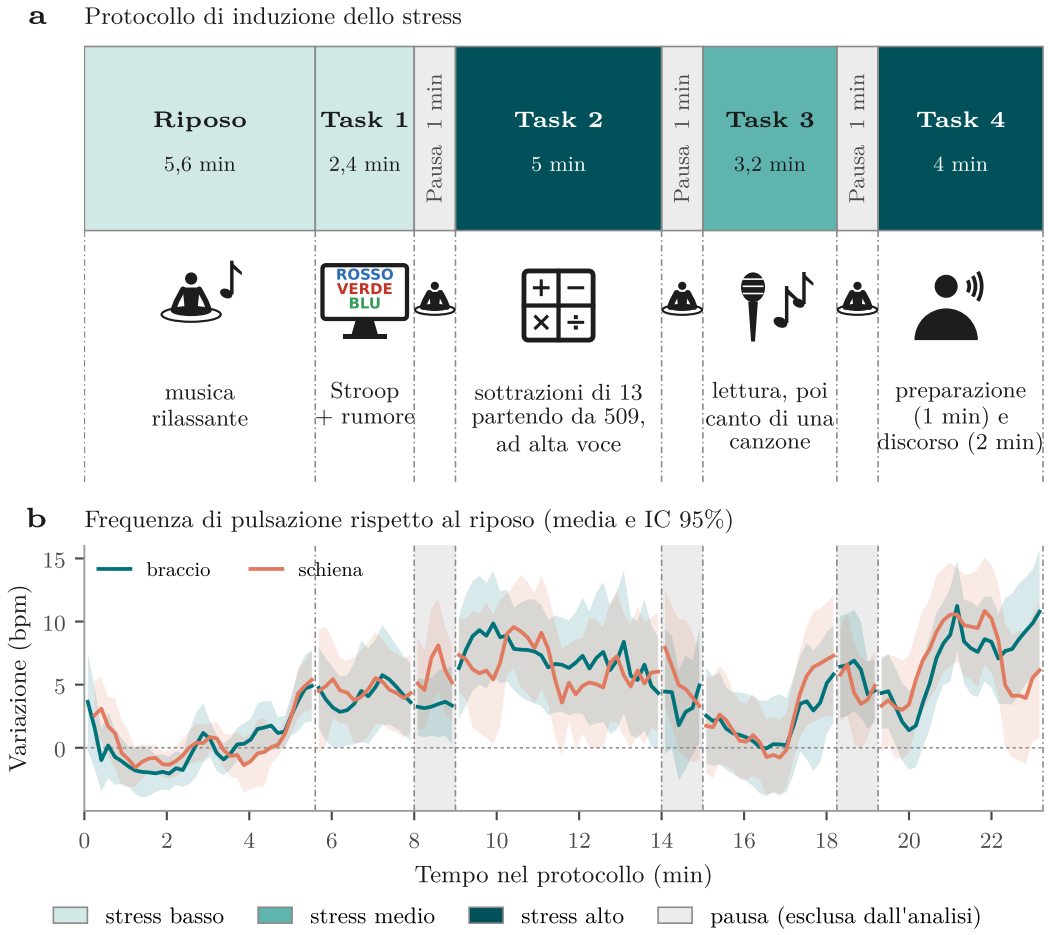


Figura 2.3: Protocollo di induzione dello stress (a) e andamento della frequenza di pulsazione stimata dal PPG nelle stesse fasi (b). Le durate sono le mediane calcolate sul file degli orari; i colori indicano la classe assegnata a ciascuna fase (Sezione 2.2). In (b), per ogni registrazione la frequenza è ricavata dagli intervalli tra picchi consecutivi (0,4–1,5 s), presa come mediana su 30 s e riferita alla mediana del riposo iniziale dello stesso soggetto; ogni fase è riportata alla sua durata mediana. Le linee sono la media tra i soggetti con riposo utilizzabile (22 sul braccio, 24 sulla schiena), le fasce l'intervallo di confidenza al 95%; le variazioni oltre ±30 bpm sono state scartate come artefatti. In grigio le pause tra i compiti, escluse dall'analisi.



Figura 2.4: Postazione di acquisizione durante il sing-a-song test. Sullo schermo la diapositiva che indica l'inizio del canto, preceduta dal conto alla rovescia; sul tavolo il tablet usato per la registrazione e le confezioni dei sensori Polar. Foto di S. Campanella.

riguarda i soggetti dal 2 al 27, cioè 26 registrazioni per sito; i dati demografici riportati sopra si riferiscono invece ai 27 partecipanti dello studio originale. Sono state escluse due registrazioni incomplete, braccio del soggetto 26 e schiena del soggetto 24, come previsto dalla pipeline. Restano quindi 25 soggetti per ciascun sito, 24 dei quali hanno entrambe le registrazioni e sono usati per il confronto diretto tra i due siti.

2.2 Definizione delle classi

Le etichette sono state costruite a partire dalle fasi del protocollo e dai questionari. Alla domanda su quale compito fosse il più impegnativo, 15 partecipanti hanno indicato il test aritmetico (task 2), 11 lo speech test (task 4) e 1 il sing-a-song test (task 3). Nell'ordinamento dei quattro compiti per stress percepito, il task 2 è in media il più vicino al primo posto, seguito dal task 4, dal task 3 e infine dallo Stroop test (task 1). Le risposte sono riassunte nella Figura 2.5, insieme al numero di partecipanti che hanno riferito battito accelerato o sudorazione durante ciascun compito.

Sulla base di questi risultati le fasi del protocollo sono state raggruppate in tre livelli, riportati nella Tabella 2.2. Lo Stroop test è stato unito al riposo nella classe di stress basso, perché è il compito percepito come meno stressante e perché, distinto dal riposo in una classe a sé, veniva riconosciuto in meno di un caso su cinque.

Nel file degli orari anche le brevi pause tra un compito e l'altro, di circa un minuto, sono segnate come riposo. Queste finestre sono state escluse dall'analisi. Nelle pause la frequenza di pulsazione resta in media sopra il valore del riposo iniziale (Figura 2.3b): è possibile che, subito dopo un compito impegnativo, l'organismo non sia ancora tornato alle condizioni di partenza, e che l'etichetta di riposo non rispecchi lo stato del partecipante. I dati sono compatibili con questa ipotesi ma

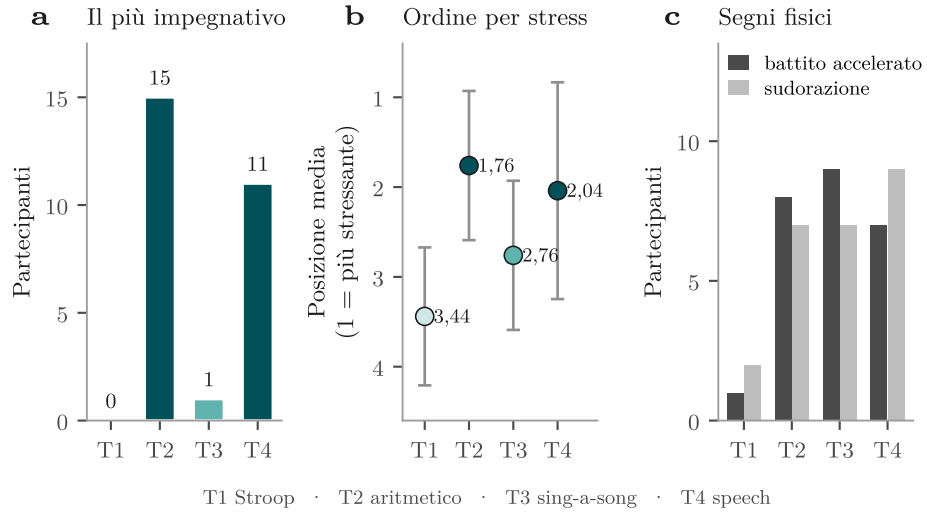


Figura 2.5: Risposte al questionario finale: compito indicato come più impegnativo (a), posizione media nell'ordinamento per stress percepito, calcolata sulle 25 risposte complete (b, media ± deviazione standard; l'asse è invertito, in alto i compiti più stressanti), e partecipanti che hanno riferito battito accelerato o sudorazione durante ciascun compito (c). I colori corrispondono alla classe assegnata a ogni compito.

Tabella 2.2: Corrispondenza tra fasi del protocollo e classi di stress.

Classe	Fasi del protocollo	Finestre (braccio / schiena)
Basso	riposo iniziale, task 1 (Stroop)	249 / 255
Medio	task 3 (sing-a-song)	110 / 113
Alto	task 2 (aritmetico), task 4 (speech)	285 / 286

non la dimostrano, e l'esclusione va quindi intesa come una scelta prudente sulla definizione delle classi. Si tratta di 47 finestre sul braccio e 48 sulla schiena; l'effetto della loro esclusione è riportato nella Sezione 3.7. La classe di stress basso comprende quindi il riposo iniziale e lo Stroop test.

La classe media contiene un solo compito ed è quindi meno numerosa delle altre due (Tabella 2.2). Di questo sbilanciamento si tiene conto sia nell'addestramento sia nella scelta delle metriche.

2.3 Sviluppo della pipeline di elaborazione

La configurazione descritta nelle sezioni seguenti è stata definita in modo progressivo. Una prima versione della pipeline, costruita a partire dai dati grezzi, comprendeva la lettura dei file, la sincronizzazione delle etichette con i segnali, il filtraggio, la segmentazione, l'estrazione delle feature e la classificazione in quattro classi (riposo e tre livelli di stress) con validazione *leave-one-subject-out*. In quella forma lo stress basso, cioè lo Stroop test, veniva confuso quasi sempre con il riposo, e l'uso di classificatori e feature alternativi non portava miglioramenti chiari. Si è quindi scelto di unire riposo e Stroop test in un'unica classe, di usare finestre sovrapposte e un filtro a banda più ampia, di selezionare feature, bilanciamento e iperparametri durante il training e di valutare anche l'aggregazione delle finestre per compito. Tutti i risultati del Capitolo 3 si riferiscono alla configurazione finale.

2.4 Preprocessing del segnale

La frequenza di campionamento effettiva di ogni registrazione è stata ricavata dalle marche temporali e vale circa 135 Hz, in accordo con le specifiche del sensore. Righe non numeriche presenti all'inizio di alcuni file sono state scartate.

Il segnale è stato filtrato con un passa-banda Chebyshev di tipo II del quarto ordine, con frequenze di taglio 0,5 e 8 Hz e attenuazione di 20 dB in banda oscura. Il filtro è applicato in avanti e all'indietro, così da non introdurre ritardi di fase che altererebbero la forma dell'onda. In un filtro di tipo II le frequenze di taglio indicano il punto in cui l'attenuazione raggiunge il valore di banda oscura, quindi la banda in cui il segnale passa quasi inalterato è un po' più stretta. Con la doppia applicazione, l'attenuazione resta sotto i 3 dB tra circa 0,66 e 6,1 Hz (Figura 2.6). L'intervallo comprende con margine la frequenza cardiaca, da 0,7 Hz (42 battiti al minuto) a 3 Hz (180 battiti al minuto), e le sue prime armoniche. Il limite inferiore elimina le variazioni lente del livello di fondo, quello superiore conserva le armoniche che descrivono la morfologia della pulsazione. Se nella registrazione c'è un'interruzione di oltre 0,1 s, i tratti continui vengono filtrati separatamente.

Osservando i segnali filtrati ci si accorge che sono capovolti: il picco sistolico, che in un PPG corretto ha una salita rapida e una discesa più lenta, compare come minimo (Figura 2.7). Per verificarlo in modo oggettivo, in ogni registrazione è stato

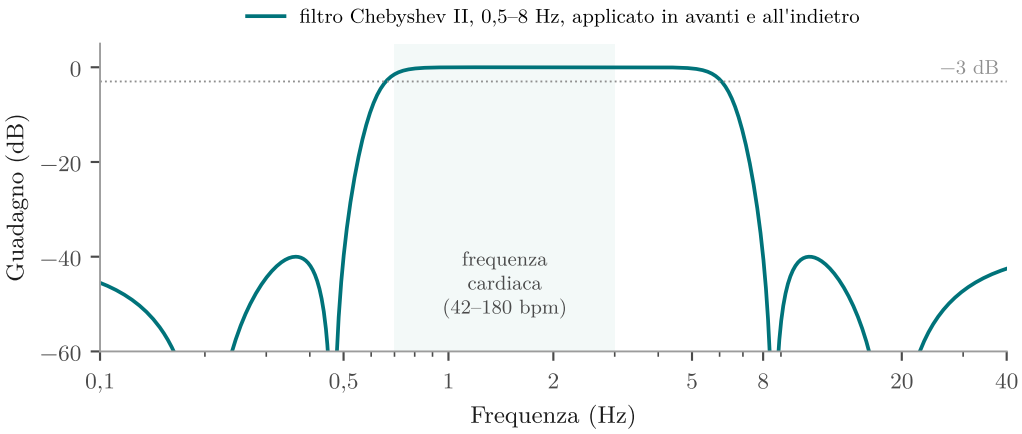


Figura 2.6: Risposta in ampiezza del filtro Chebyshev di tipo II applicato in avanti e all'indietro, con l'attenuazione di 3 dB indicata dalla linea punteggiata. La fascia colorata indica l'intervallo tipico della frequenza cardiaca.

calcolato il rapporto tra il tempo di salita e la durata del battito, che in un segnale orientato correttamente è inferiore a 0,5. Il segnale risulta capovolto in tutte le 25 registrazioni valutabili del braccio e in 22 su 26 della schiena; nelle quattro restanti il rapporto è vicino a 0,5 o opposto, e tra queste c'è la registrazione già esclusa, la schiena del soggetto 24 (Figura 2.8). Trattandosi di una caratteristica del sensore, tutti i segnali sono stati ribaltati prima dell'estrazione delle feature.

2.5 Segmentazione in finestre

Il segnale è stato diviso in finestre di 60 s con una sovrapposizione del 40%, cioè una nuova finestra ogni 36 s. Le finestre sono prese solo all'interno delle fasi del protocollo, senza mai attraversare il confine tra due fasi, ed ereditano l'etichetta della fase da cui provengono (Figura 2.9). La porzione finale di una fase più corta di 60 s viene scartata, così come le finestre che contengono un'interruzione del segnale; una fase più breve di 60 s non produce quindi nessuna finestra. Una finestra di un minuto contiene circa 60–100 battiti: abbastanza per stimare la frequenza media delle pulsazioni, mentre gli indici di variabilità calcolati su un intervallo così breve vanno intesi come indici a brevissimo termine [12]. Escluse le pause tra i compiti (Sezione 2.2), si ottengono 644 finestre per il braccio e 654 per la schiena.

2.6 Estrazione delle feature

Da ogni finestra sono state calcolate 55 feature, divise in sei gruppi:

- **dominio del tempo** (13): media, deviazione standard, minimo, massimo, escursione, mediana, scarto interquartile, asimmetria, curtosi, valore efficace, passaggi per lo zero e parametri di Hjorth [41];

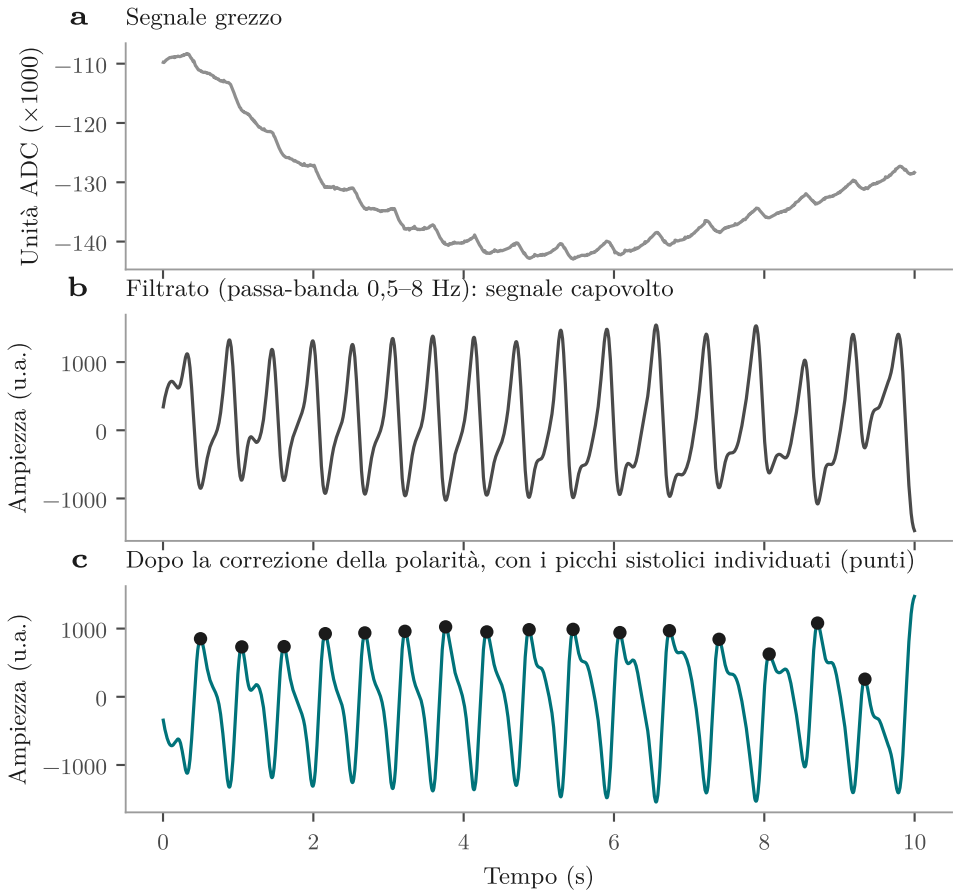


Figura 2.7: Fasi del preprocessing su 10 s di segnale (soggetto 2, schiena, riposo): segnale grezzo (a), filtrato (b) e dopo la correzione della polarità, con i picchi sistolici individuati (c).

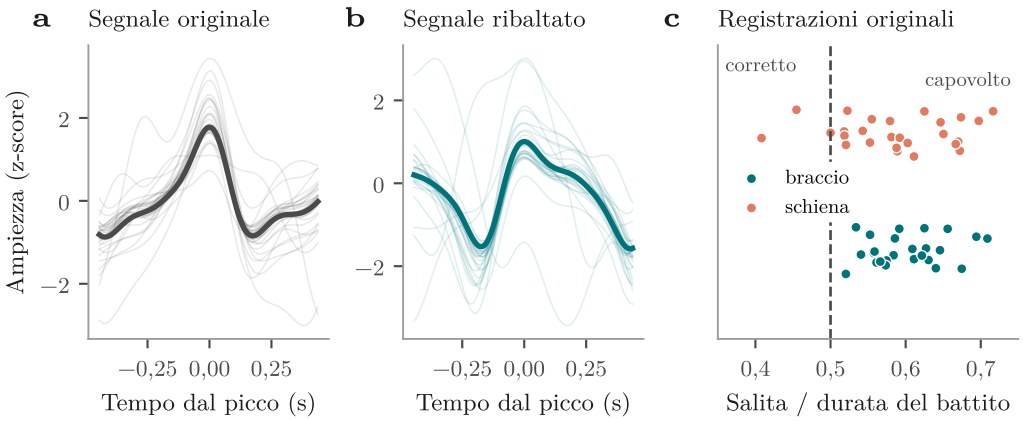


Figura 2.8: Verifica della polarità: battiti di una finestra e loro media prima (a) e dopo (b) il ribaltamento; rapporto tra salita e durata del battito nelle registrazioni originali (c).

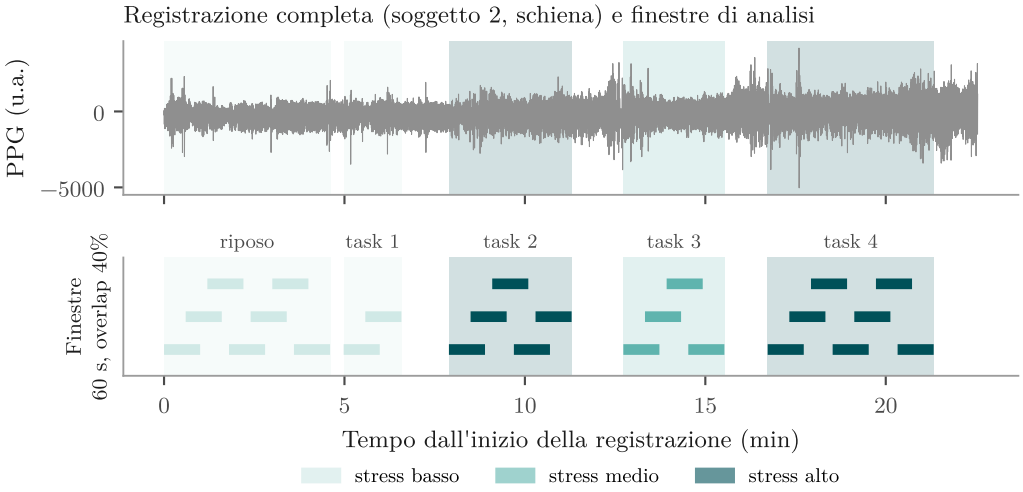


Figura 2.9: Registrazione completa del soggetto 2 (schiena), fasi del protocollo e finestre di analisi.

- **dominio della frequenza** (9): frequenza dominante, centroide, dispersione ed entropia dello spettro, potenza relativa nelle bande 0,5–1, 1–2, 2–4 e 4–8 Hz e potenza totale, stimate con il metodo di Welch;
- **frequenza e variabilità delle pulsazioni** (9): numero di picchi, frequenza media, intervallo PP medio e suo coefficiente di variazione, SDNN, RMSSD, pNN50 e indici SD1 e SD2 del diagramma di Poincaré, calcolati sugli intervalli tra picchi compresi tra 0,3 e 1,5 s;
- **ampiezza dei picchi** (5): media, deviazione standard, coefficiente di variazione, scarto interquartile e RMSSD dell'ampiezza dei picchi sistolici;
- **morfologia del battito** (16): larghezza a metà altezza, ampiezza piede-picco, tempi e pendenze di salita e di discesa, area del battito, ciascuno con le sue statistiche;
- **inviluppo** (3): media, deviazione standard e coefficiente di variazione dell'inviluppo ottenuto con la trasformata di Hilbert.

La Figura 2.10 mostra come sono calcolate alcune di queste grandezze su una finestra reale. Le misure morfologiche si ricavano battito per battito: il piede è il minimo che precede il picco sistolico, l'ampiezza piede-picco è la differenza tra i due, i tempi di salita e di discesa separano il picco dai minimi adiacenti, la larghezza è misurata a metà altezza e l'area è quella compresa tra il battito e la retta che unisce i due minimi. Per ogni finestra se ne calcolano poi media e dispersione. Nel grafico di Poincaré ogni intervallo picco-picco è messo in relazione con il successivo: SD1 descrive la dispersione perpendicolare alla bisettrice, legata alle variazioni rapide da un battito all'altro, mentre SD2 quella lungo la bisettrice, legata alle variazioni più lente. Lo

spettro, infine, è stimato con il metodo di Welch, che divide la finestra in segmenti di 8 s sovrapposti per metà e ne media i periodogrammi [42]. Il picco principale corrisponde alla frequenza di pulsazione, quelli successivi alle sue armoniche, che dipendono dalla forma dell'onda.

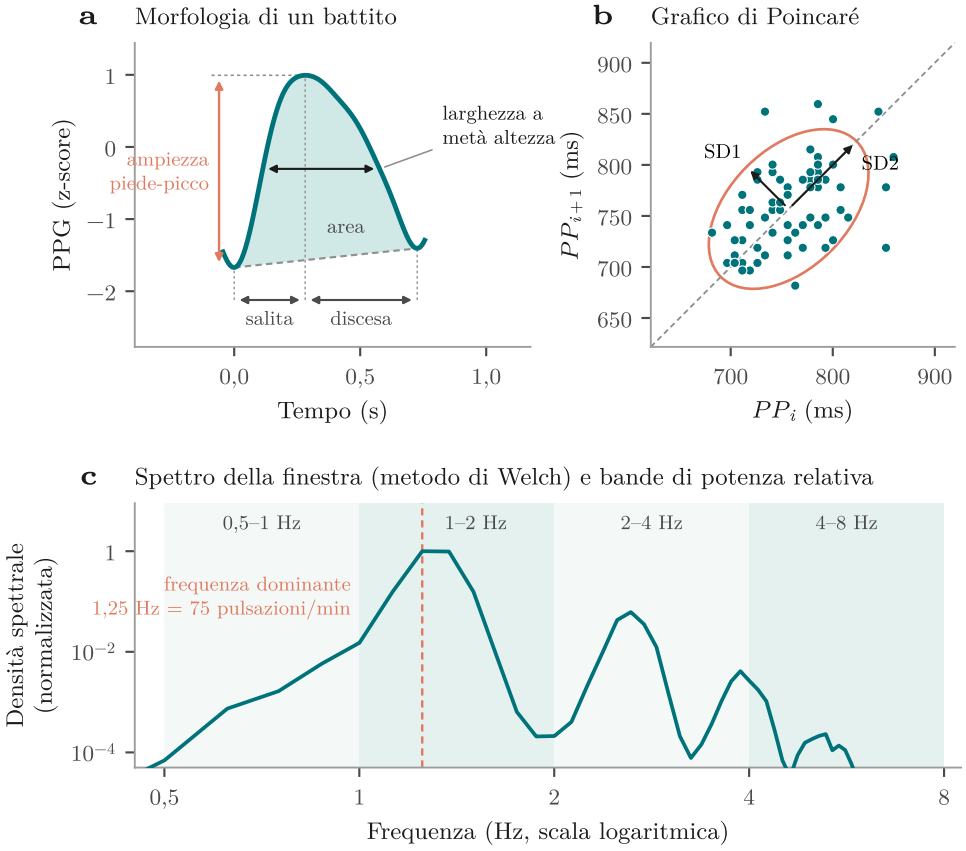


Figura 2.10: Esempi di feature calcolate su una finestra di riposo (soggetto 4, schiena). (a) Misure morfologiche su un battito: ampiezza piede-picco, tempi di salita e di discesa, larghezza a metà altezza e area sopra la retta che unisce i due minimi. (b) Grafico di Poincaré degli intervalli picco-picco della finestra, con l'ellisse a due deviazioni standard e le direzioni di SD1 e SD2. (c) Spettro di Welch con le quattro bande usate per la potenza relativa e la frequenza dominante.

L'elenco completo, con la definizione di ciascuna feature, è riportato nell'Appendice A (Tabella A.1). Gli indici di variabilità sono calcolati sul PPG e vanno quindi intesi come variabilità della frequenza di pulsazione, non come HRV in senso stretto (Sezione 1.2.4).

Le feature vengono poi normalizzate in due passaggi distinti. Il primo è uno z-score individuale: per ciascun soggetto e sito, ogni feature viene centrata sulla propria media e divisa per la propria deviazione standard, calcolate su tutte le finestre di quel soggetto usate nell'analisi. In questo modo il modello vede come cambiano le feature rispetto ai valori tipici di quella persona, invece dei valori assoluti, che

variano molto da un individuo all'altro. Questo passaggio non usa le etichette e non mescola dati di persone diverse, ma richiede di disporre dell'intera registrazione del soggetto, compreso quello di test. Il secondo passaggio è la standardizzazione appresa dal classificatore, descritta nella Sezione 2.7, che usa solo i dati di training. La valutazione descritta in questa tesi è quindi *offline*: corrisponde allo scenario in cui di una persona nuova è disponibile l'intera registrazione, non a quello in cui si classifica una singola finestra appena acquisita senza nessuna calibrazione.

2.7 Classificazione e validazione

2.7.1 Modelli

Sono stati confrontati tre classificatori:

- una **Random Forest**, che riceve le 55 feature;
- una **NN**, anch'essa basata sulle feature, in due versioni: quella con tuning annidato descritta più avanti, realizzata con `scikit-learn`, e una a configurazione fissa, realizzata con `Keras`, con due strati da 64 e 32 neuroni, normalizzazione del batch e dropout (Figura 2.11). Le versioni delle librerie sono indicate nella Sezione 2.7.5;
- una **1D-CNN**, che riceve direttamente il segnale della finestra, sottocampionato di un fattore 4 (circa 34 Hz) e normalizzato per soggetto e sito. L'architettura comprende tre blocchi convoluzionali con 16, 32 e 64 filtri, normalizzazione del batch e max pooling, seguiti da un pooling globale e da uno strato denso (Figura 2.12). La 1D-CNN usa una configurazione fissa, senza tuning: l'addestramento si interrompe quando la perdita su un 10% di validazione interna al training smette di diminuire. Il bilanciamento con SMOTE è applicato solo alla parte restante del training: ogni finestra di segnale viene trattata come un vettore di 2025 campioni, e le finestre sintetiche sono combinazioni lineari di due finestre reali vicine della stessa classe.

Le due architetture sono mostrate nelle Figure 2.11 e 2.12. Nella 1D-CNN ogni blocco convoluzionale riduce di quattro volte la lunghezza della sequenza e aumenta il numero di canali, finché un pooling globale riassume l'intera finestra in 64 valori.

Nel seguito NN indica la rete con tuning annidato realizzata con `scikit-learn`; la rete a configurazione fissa è indicata come NN (`Keras`).

2.7.2 Selezione annidata di feature, bilanciamento e iperparametri

Per la Random Forest e la NN di `scikit-learn` sono state confrontate alcune alternative:

- usare tutte le 55 feature oppure le 35 più informative secondo la mutua informazione con la classe;

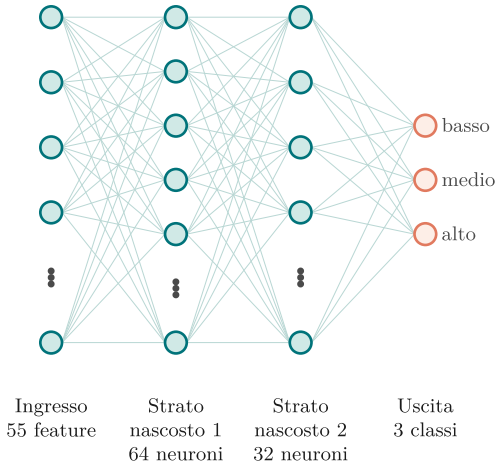


Figura 2.11: Schema della rete neurale densa usata in questa tesi, nella configurazione con due strati nascosti da 64 e 32 neuroni: le 55 feature di una finestra entrano a sinistra, e l'uscita fornisce la probabilità di ciascuna delle tre classi. Per chiarezza sono disegnati solo alcuni neuroni per strato.

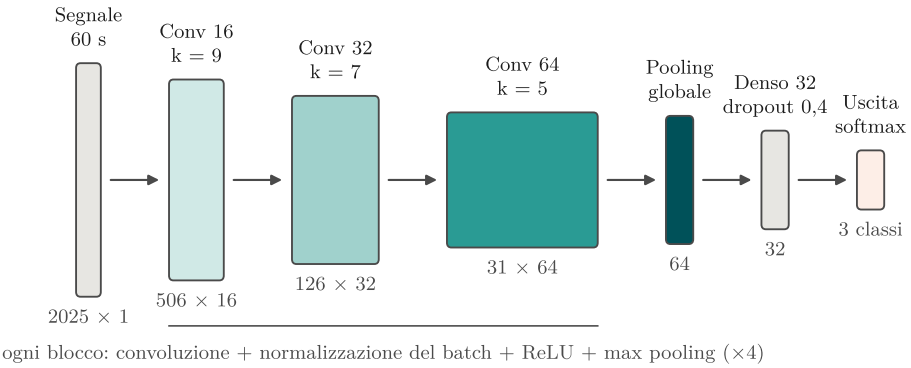


Figura 2.12: Schema della 1D-CNN usata in questa tesi. Sotto ogni blocco è indicata la dimensione dell'uscita (lunghezza × canali): la finestra di 60 s, sottocampionata a circa 34 Hz, entra come sequenza di 2025 campioni e viene ridotta a 64 valori prima degli strati finali.

- bilanciare le classi con SMOTE oppure con pesi inversamente proporzionali alla numerosità delle classi, assegnati a ogni esempio di training e passati ai modelli come `sample_weight`;
- due configurazioni degli iperparametri. Per la Random Forest: 300 alberi, con profondità libera oppure limitata a 12 livelli. Per la NN: due strati nascosti da 64 e 32 oppure da 128 e 64 neuroni, con regolarizzazione e passo di apprendimento diversi.

Le alternative danno otto combinazioni per modello. La scelta è fatta all'interno di ogni fold con una validazione interna a 2 fold, anch'essa per soggetto, che usa solo i soggetti di training: la combinazione con la balanced accuracy interna più alta viene poi riaddestrata su tutto il training e solo allora applicata ai soggetti di test. Imputazione dei valori mancanti, standardizzazione, selezione delle feature e bilanciamento sono appresi solo sul training, così i soggetti di test non intervengono nell'addestramento [23, 21]. La procedura all'interno di un fold esterno è schematizzata nella Figura 2.13.

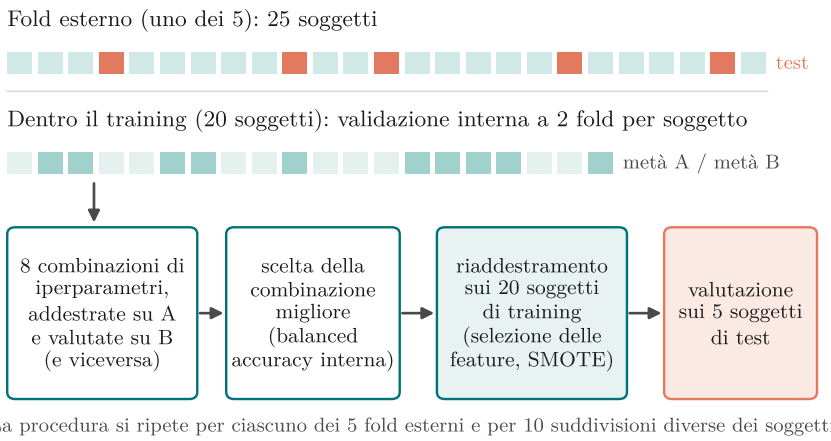


Figura 2.13: Schema della validazione annidata all'interno di un fold esterno. Ogni quadrato è un soggetto. I soggetti di training sono divisi in due metà per scegliere gli iperparametri; la combinazione migliore viene poi riaddestrata su tutto il training e valutata sui soggetti di test, che non partecipano a nessuna delle fasi precedenti.

Le reti neurali usano anche l'early stopping, cioè l'arresto anticipato dell'addestramento, controllato su un 10% delle finestre di training tenute da parte. Nelle reti in `Keras` la metrica controllata è la perdita di validazione (`val_loss`): l'addestramento si ferma dopo 15 epoche senza miglioramento e vengono ripristinati i pesi dell'epoca migliore. Nella NN di `scikit-learn` è invece l'accuratezza sulla stessa quota di validazione, con arresto dopo 10 epoche consecutive senza miglioramento. Questa suddivisione interna è casuale per finestre e non per soggetto: finestre dello stesso soggetto, anche parzialmente sovrapposte, possono finire sia nella parte usata per

addestrare sia in quella usata per decidere quando fermarsi. Nella NN in `Keras` e nella 1D-CNN la suddivisione avviene prima di SMOTE, quindi la validazione interna contiene solo dati reali; nella NN di `scikit-learn` avviene dopo, e può contenere anche esempi sintetici. Questo può rendere l'arresto anticipato meno severo, ma non riguarda i soggetti di test, che restano esclusi da ogni fase dell'addestramento. Tutti i parametri dei modelli sono riassunti nella Tabella A.2, in appendice.

2.7.3 Validazione

Braccio e schiena sono stati analizzati sempre separatamente. La validazione è incrociata per soggetto e stratificata, con 5 fold: a ogni iterazione circa 5 soggetti vengono tenuti fuori per il test [22]. Lo schema è mostrato nella Figura 2.14. Con soli 25 soggetti il risultato dipende sensibilmente da come vengono formati i fold. Per questo l'intera procedura è stata ripetuta 10 volte con suddivisioni diverse. In ogni ripetizione le metriche sono calcolate riunendo le predizioni di test dei 5 fold. Si riportano la media e la deviazione standard delle 10 ripetizioni, insieme a un intervallo di confidenza al 95% ottenuto con un bootstrap sui soggetti [43]: in ciascuno dei 2000 ricampionamenti si estraggono con reinserimento 25 soggetti, si sommano le matrici di confusione di tutte le loro finestre nelle 10 ripetizioni e si ricalcola la balanced accuracy; l'intervallo è dato dai percentili 2,5 e 97,5. La deviazione standard misura quindi quanto il risultato cambia con la suddivisione in fold, l'intervallo bootstrap quanto potrebbe cambiare con un altro gruppo di persone. Le 10 ripetizioni riusano gli stessi 25 soggetti e non equivalgono a un campione più grande. L'accuratezza è calcolata anche per singolo soggetto, per mostrare la variabilità tra persone. Per confrontare braccio e schiena si sono usati un bootstrap appaiato sulla balanced accuracy, ricampionando gli stessi soggetti per i due siti, e il test di Wilcoxon [44] sull'accuratezza di ciascun soggetto: sono due analisi diverse, la prima sulla metrica complessiva, la seconda sui singoli soggetti.

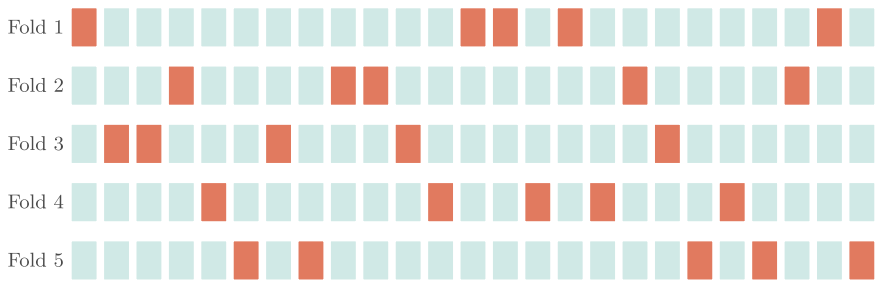


Figura 2.14: Esempio di suddivisione dei 25 soggetti nei 5 fold: in ogni iterazione i soggetti in corallo formano il test e non contribuiscono in alcun modo all'addestramento. La suddivisione cambia a ogni ripetizione.

2.7.4 Metriche e livelli di valutazione

Come metrica principale è stata usata la balanced accuracy, affiancata da macro-F1, accuracy, matrici di confusione e metriche per classe (Sezione 1.3.6).

Le prestazioni sono state valutate a tre livelli:

1. **per finestra**, il livello principale: ogni finestra di 60 s è classificata da sola;
2. **media delle ultime tre finestre**: la probabilità di ogni finestra è mediata con quelle delle due precedenti della stessa fase. L'aggregazione usa solo finestre passate e potrebbe quindi funzionare in tempo reale; la pipeline valutata nel suo complesso resta però offline, perché il filtro è applicato in avanti e all'indietro e la normalizzazione individuale usa l'intera registrazione;
3. **per task**: le probabilità di tutte le finestre di una fase sono mediate e la fase riceve un'unica etichetta. Questo livello risponde a una domanda diversa, cioè riconoscere il livello di stress di un intero compito, e non va confrontato direttamente con il primo.

2.7.5 Strumenti

L'elaborazione è stata svolta in Python: versione 3.13 per preprocessing, feature, Random Forest, NN di `scikit-learn`, statistiche e grafici, versione 3.12 per le reti in `Keras`. Le librerie usate, con le rispettive versioni, sono:

- `NumPy` 2.5, `pandas` 3.0 e `SciPy` 1.18 per la lettura dei file, il preprocessing del segnale e il calcolo delle feature;
- `scikit-learn` 1.9.1 per Random Forest, NN con tuning annidato, selezione delle feature, validazione e metriche;
- `imbalanced-learn` 0.14.2 per SMOTE nelle reti in `Keras`. Per Random Forest e NN di `scikit-learn`, SMOTE è stato implementato direttamente con `NumPy`, cercando i vicini con la classe `NearestNeighbors` di `scikit-learn` e generando gli esempi sintetici con l'Equazione (1.3);
- `TensorFlow` 2.21.0 con `Keras` 3.15.1 per la 1D-CNN e per la NN a configurazione fissa;
- `Matplotlib` 3.11.2 per tutti i grafici e gli schemi originali della tesi, prodotti a partire dai dati e dai risultati di questo lavoro.

Le figure tratte da altre fonti, anche se adattate, sono indicate in didascalia. Il codice sviluppato è disponibile su richiesta.

Capitolo 3

Risultati

Il capitolo presenta i risultati della configurazione descritta nel Capitolo 2. Tutti i valori sono medie sulle 10 ripetizioni della validazione per soggetto e, se non è indicato diversamente, si riferiscono alla classificazione delle singole finestre di 60 s.

3.1 Analisi esplorativa della frequenza di pulsazione

Prima dei classificatori conviene guardare come cambia una sola grandezza, la frequenza di pulsazione stimata dal PPG, nelle diverse fasi (Figura 3.1). Per ogni soggetto è stata presa la mediana della frequenza nelle finestre di ciascuna fase, dopo la normalizzazione per soggetto. Nel test aritmetico la frequenza è più alta che a riposo in 19 soggetti su 23 sul braccio e in 19 su 24 sulla schiena; nello speech test in 21 su 23 e in 18 su 24. Lo Stroop test si colloca poco sopra il riposo. Il sing-a-song test, invece, ha valori mediani più bassi del riposo su entrambi i siti, e l'aumento compare solo in 13 soggetti su 23 sul braccio e in 12 su 24 sulla schiena. I soggetti considerati sono quelli con finestre valide sia nel riposo iniziale sia nel compito.

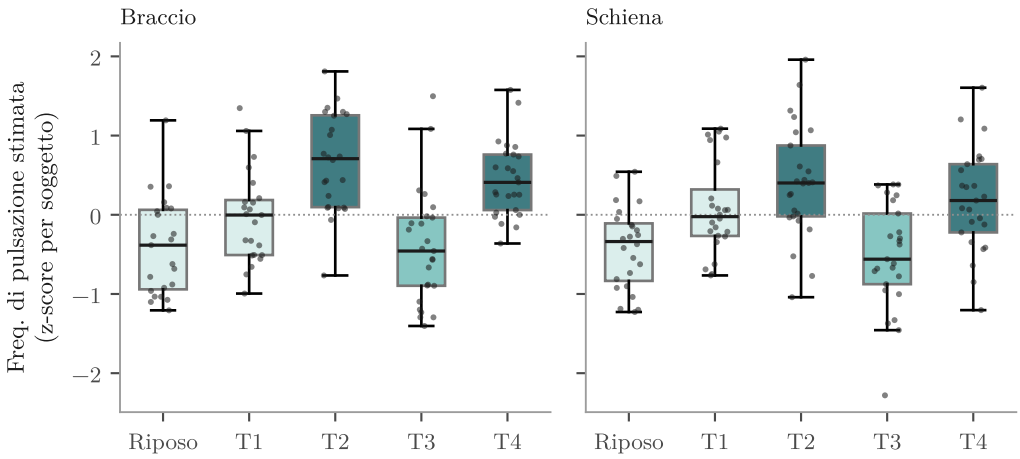


Figura 3.1: Frequenza di pulsazione stimata dal PPG in ciascuna fase (z-score per soggetto e sito). Ogni punto è la mediana di un soggetto; i colori indicano la classe assegnata alla fase.

La figura anticipa due aspetti che torneranno nei risultati. Le fasi della classe di stress alto si separano dal riposo in modo abbastanza netto, mentre il task 3 si

comporta diversamente da quanto ci si aspetterebbe da una classe intermedia. In più, la dispersione dei punti mostra che la risposta cambia molto da una persona all'altra.

L'andamento nel tempo della frequenza, riportato nella Figura 2.3b, aggiunge alcuni dettagli che le mediane per fase non mostrano. Durante il riposo la frequenza scende leggermente nei primi minuti e risale di circa 3 bpm nell'ultimo minuto. Una possibile spiegazione è l'attesa del primo compito; può contribuire anche il fatto che gli orari delle fasi sono annotati al minuto. Lo Stroop test produce un aumento modesto, in media di 4–5 bpm. Il test aritmetico dà la risposta più netta: la frequenza sale subito di circa 7 bpm, con punte di 9, e resta alta per tutti i cinque minuti. Nello speech test l'aumento è contenuto durante il minuto di preparazione e cresce durante il discorso, fino a circa 10 bpm.

Il sing-a-song test si divide in due parti. Durante la lettura, in silenzio e ad alta voce, la frequenza resta vicina al riposo; durante il canto sale di circa 6 bpm. Le finestre di questo compito contengono quindi momenti molto diversi tra loro, e questo può contribuire alla difficoltà di riconoscere la classe intermedia. Nelle pause, infine, la frequenza non torna al livello del riposo iniziale: un minuto potrebbe non bastare per un recupero completo. È il motivo per cui queste finestre sono state escluse dall'analisi (Sezione 2.2).

Questi valori vanno letti con cautela. Sono medie tra soggetti con intervalli di confidenza ampi, e la media risente di chi reagisce di più: considerando la mediana di ciascun soggetto, l'aumento nel sing-a-song compare solo in circa metà dei casi, in accordo con la Figura 3.1. Parlare e cantare, inoltre, modificano la respirazione, che da sola influenza la frequenza cardiaca, e nel canto i piccoli movimenti possono introdurre errori nel riconoscimento dei picchi. Le curve del braccio e della schiena sono quasi sovrapposte: le due sedi registrano la stessa dinamica.

3.2 Prestazioni dei classificatori

La Tabella 3.1 riporta le prestazioni per finestra di tutti i modelli. I modelli basati sulle feature si comportano in modo simile: la balanced accuracy va da 0,58 a 0,59 sul braccio e da 0,62 a 0,63 sulla schiena, ben sopra il valore di 0,33 di un classificatore casuale. La Random Forest ottiene 0,582 sul braccio e 0,615 sulla schiena, la NN 0,593 e 0,624; la NN in `Keras`, a configurazione fissa, arriva a valori quasi identici (0,592 e 0,628). Le differenze tra questi modelli sono di uno o due centesimi, inferiori alla variabilità dovuta alla scelta dei fold, e non permettono di indicare un modello migliore in modo netto.

La 1D-CNN, che lavora direttamente sul segnale, resta invece molto indietro: 0,444 sul braccio e 0,535 sulla schiena. È anche il modello più instabile tra una ripetizione e l'altra, soprattutto sul braccio.

Nel seguito le analisi di dettaglio sono riportate per i due modelli basati sulle feature con tuning annidato, la Random Forest e la NN, commentati separatamente.

Tabella 3.1: Prestazioni per finestra: balanced accuracy come media ± deviazione standard sulle 10 ripetizioni, macro-F1 come media. In grassetto il valore più alto per sito.

Sito	Modello	Balanced accuracy	Macro-F1
Braccio	Random Forest	0,582 ± 0,012	0,582
	NN (`scikit-learn`)	**0,593 ± 0,020**	**0,588**
	NN (`Keras`)	0,592 ± 0,022	0,587
	1D-CNN	0,444 ± 0,031	0,439
Schiena	Random Forest	0,615 ± 0,011	0,616
	NN (`scikit-learn`)	0,624 ± 0,015	0,620
	NN (`Keras`)	**0,628 ± 0,020**	**0,624**
	1D-CNN	0,535 ± 0,021	0,533

3.3 Valutazione cumulativa

Una singola finestra dura un minuto, mentre ogni compito del protocollo ne contiene diverse. Oltre alla classificazione delle singole finestre, quindi, si è valutato cosa succede quando si tiene conto anche delle finestre vicine dello stesso compito (Sezione 2.7). Il modello resta lo stesso: cambia solo il modo di leggerne le uscite. Nella media delle ultime tre finestre, le probabilità di ogni finestra sono mediate con quelle delle due precedenti; nella valutazione per task, le probabilità di tutte le finestre di un compito sono mediate e il compito riceve un'unica classe. Per esempio, se le finestre del test aritmetico di un soggetto sono classificate per lo più come stress alto, l'intero compito viene classificato come stress alto anche se qualche finestra isolata era stata assegnata a un'altra classe.

La Tabella 3.2 riporta i risultati per la Random Forest e la NN, la Figura 3.2 quelli di tutti i modelli. Per entrambi i modelli la media delle ultime tre finestre aggiunge 1–2 centesimi, mentre la valutazione per task aggiunge 7–9 centesimi: la Random Forest arriva a 0,653 sul braccio e 0,696 sulla schiena, la NN a 0,680 e 0,713, il valore più alto ottenuto. La NN in `Keras` segue lo stesso andamento, con 0,678 sul braccio e 0,712 sulla schiena per task, mentre per la 1D-CNN mediare più finestre serve a poco: per task sale a 0,473 sul braccio e scende a 0,520 sulla schiena, perché le sue previsioni restano deboli anche quando vengono combinate.

Gli intervalli di confidenza sono ampi, circa ±0,07 per finestra e ±0,07–0,08 per task. Con 25 soggetti per sito, il risultato dipende ancora molto da quali persone compongono il campione.

3.4 Prestazioni per classe

Le matrici di confusione della Figura 3.3 mostrano che non tutte le classi sono riconosciute allo stesso modo, e che i due modelli si comportano in modo simile. Lo

Tabella 3.2: Prestazioni della Random Forest e della NN ai tre livelli di valutazione: balanced accuracy come media ± deviazione standard sulle 10 ripetizioni, intervallo di confidenza al 95% (IC) e macro-F1.

Sito	Modello	Livello	Balanced accuracy	IC 95%	Macro-F1
Braccio	Random Forest	Per finestra	0,582 ± 0,012	0,513–0,649	0,582
		Ultime 3 finestre	0,597 ± 0,015	0,523–0,672	0,598
		Per task	0,653 ± 0,008	0,580–0,726	0,653
	NN	Per finestra	0,593 ± 0,020	0,521–0,661	0,588
		Ultime 3 finestre	0,615 ± 0,023	0,535–0,691	0,612
		Per task	0,680 ± 0,022	0,599–0,758	0,682
Schiena	Random Forest	Per finestra	0,615 ± 0,011	0,556–0,671	0,616
		Ultime 3 finestre	0,629 ± 0,013	0,562–0,695	0,635
		Per task	0,696 ± 0,014	0,613–0,779	0,701
	NN	Per finestra	0,624 ± 0,015	0,565–0,680	0,620
		Ultime 3 finestre	0,639 ± 0,019	0,576–0,703	0,639
		Per task	0,713 ± 0,024	0,643–0,785	0,716

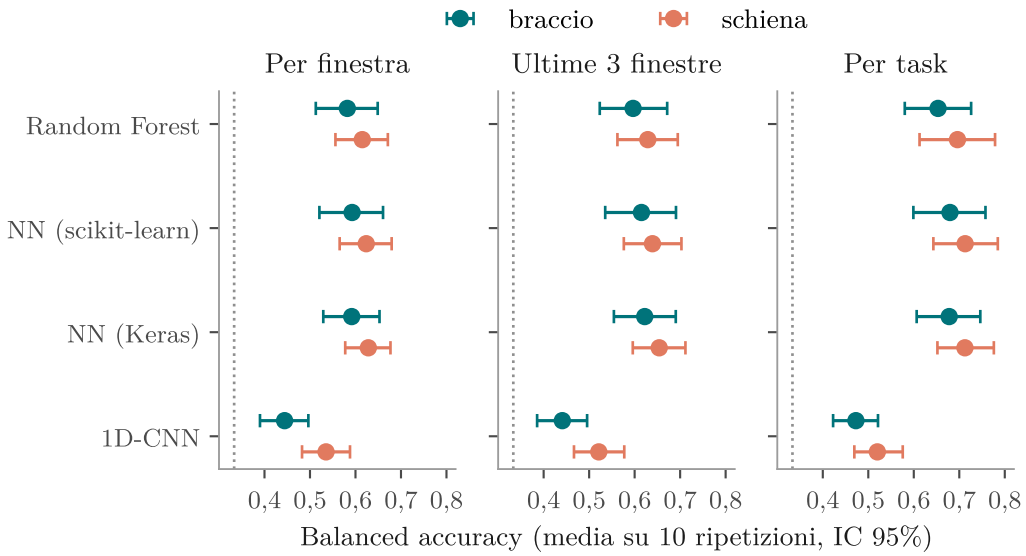


Figura 3.2: Balanced accuracy di tutti i modelli ai tre livelli di valutazione, con intervallo di confidenza al 95%. La linea tratteggiata indica il valore di un classificatore casuale.

stress alto è la classe identificata meglio: il recall è di 0,73 sul braccio e 0,79 sulla schiena per la Random Forest, di 0,69 e 0,73 per la NN. Seguono lo stress basso, con un recall di 0,64–0,66 per entrambi i modelli, e infine la classe media, la più difficile, confusa soprattutto con lo stress basso. La NN la riconosce un po' più spesso della Random Forest: il recall dello stress medio è di 0,37 sul braccio e 0,40 sulla schiena per la Random Forest, di 0,44 e 0,48 per la NN, che però riconosce leggermente meno lo stress alto. L'F1-score della classe media resta il più basso, tra 0,36 e 0,44, contro valori tra 0,66 e 0,76 per le altre due classi.

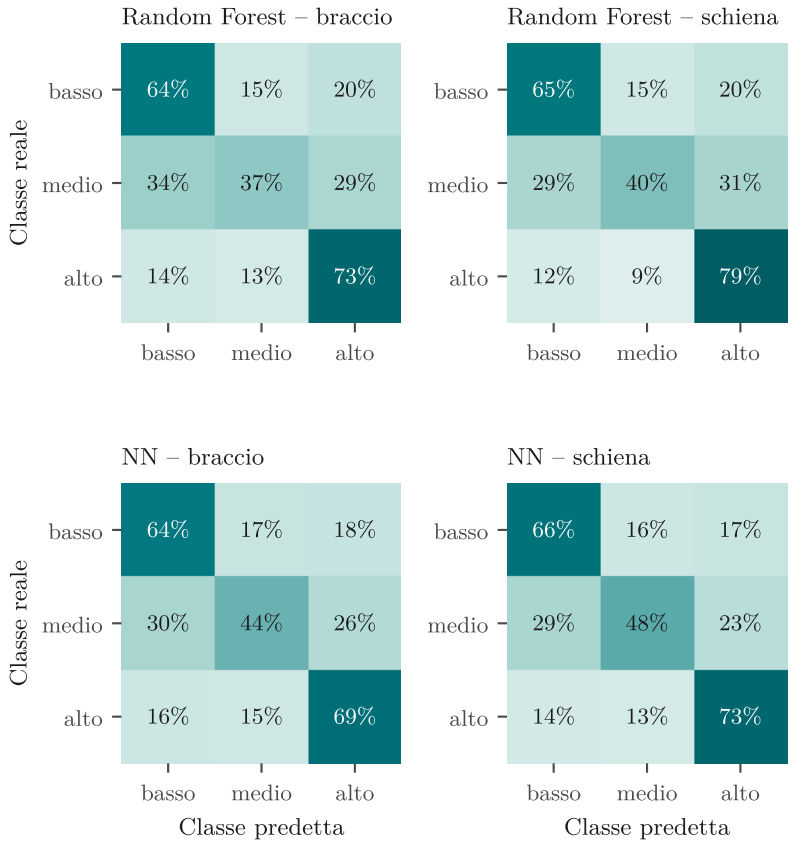


Figura 3.3: Matrici di confusione normalizzate per riga della Random Forest (sopra) e della NN (sotto), per finestra, somma delle 10 ripetizioni.

Una parte degli errori riguarda anche le due classi estreme: circa un quinto delle finestre di stress basso viene classificato come stress alto (il 20–21% con la Random Forest, il 17–18% con la NN), e tra il 12% e il 16% delle finestre di stress alto come stress basso.

Le curve ROC della Figura 3.4 completano il quadro, perché valutano le probabilità stimate indipendentemente dalla regola di decisione. Per la Random Forest l'AUC è di 0,84 e 0,87 per lo stress alto, di 0,80 e 0,85 per lo stress basso e di 0,71 e 0,76 per lo stress medio, rispettivamente sul braccio e sulla schiena; per la NN i valori corrispondenti sono 0,82 e 0,85, 0,79 e 0,82, 0,70 e 0,75 (medie delle 10 ripetizioni, con deviazioni standard tra 0,005 e 0,03). L'ordine delle classi è lo stesso delle matrici

di confusione, ma anche per la classe media i valori sono lontani da 0,5: le probabilità stimate contengono informazione, che però si traduce in un recall basso quando ogni finestra viene assegnata alla sola classe più probabile.

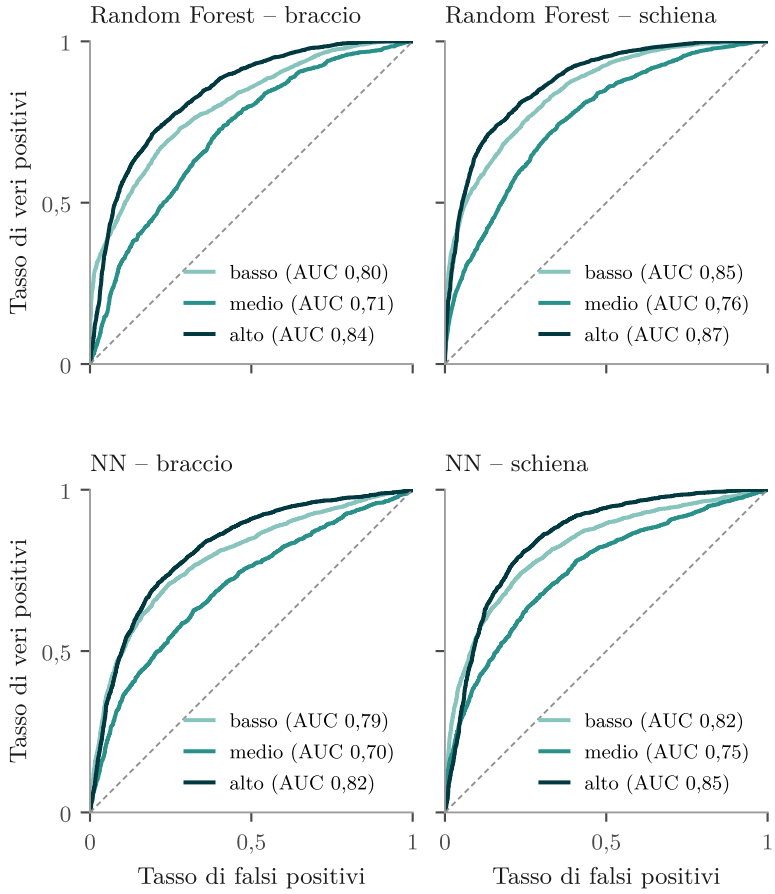


Figura 3.4: Curve ROC di ciascuna classe contro le altre due per la Random Forest (sopra) e la NN (sotto), per finestra, con le predizioni delle 10 ripetizioni riunite. Tra parentesi l'AUC media delle 10 ripetizioni; la diagonale tratteggiata corrisponde a un classificatore casuale.

3.5 Confronto tra i siti e variabilità inter-soggetto

Per finestra la schiena ottiene risultati leggermente migliori del braccio con entrambi i modelli, con circa 3 centesimi di balanced accuracy in più e una sensibilità maggiore per lo stress alto. Gli intervalli di confidenza dei due siti si sovrappongono però in gran parte. Per verificarlo in modo diretto, i due siti sono stati confrontati sugli stessi 24 soggetti, quelli disponibili per entrambi, con un bootstrap appaiato. La differenza a favore della schiena è di 0,031 per la Random Forest, con un intervallo di confidenza al 95% da $-0{,}055$ a $+0{,}124$, e di 0,036 per la NN, da $-0{,}047$ a $+0{,}123$: in entrambi i casi l'intervallo comprende lo zero. La schiena è migliore in 12 soggetti su 24 con la Random Forest e in 13 con la NN, e il test di Wilcoxon sulle accuratezze

per soggetto non indica una differenza ($p = 0{,}70$ e $p = 0{,}58$). I dati, quindi, non permettono di affermare che un sito sia superiore all'altro.

La differenza più evidente non è tra i siti ma tra le persone (Figura 3.6). Con la Random Forest l'accuratezza per soggetto va da 0,34 a 0,94 sul braccio e da 0,33 a 0,90 sulla schiena; con la NN da 0,30 a 0,89 e da 0,40 a 0,90. La mediana è di circa 0,67–0,72 per entrambi i modelli. Per alcuni soggetti i modelli funzionano molto bene, per altri, come i soggetti 10 e 15 sul braccio, restano vicini al livello casuale. Il buon risultato su un sito, inoltre, dice poco su quello nell'altro: la correlazione tra le accuratezze per soggetto dei due siti è di 0,14 per la Random Forest e di 0,28 per la NN. La Figura 3.5 mette i due siti direttamente a confronto: i punti si distribuiscono da entrambi i lati della diagonale, e alcuni soggetti, come il 12 e il 15, sono riconosciuti molto meglio sulla schiena con entrambi i modelli.

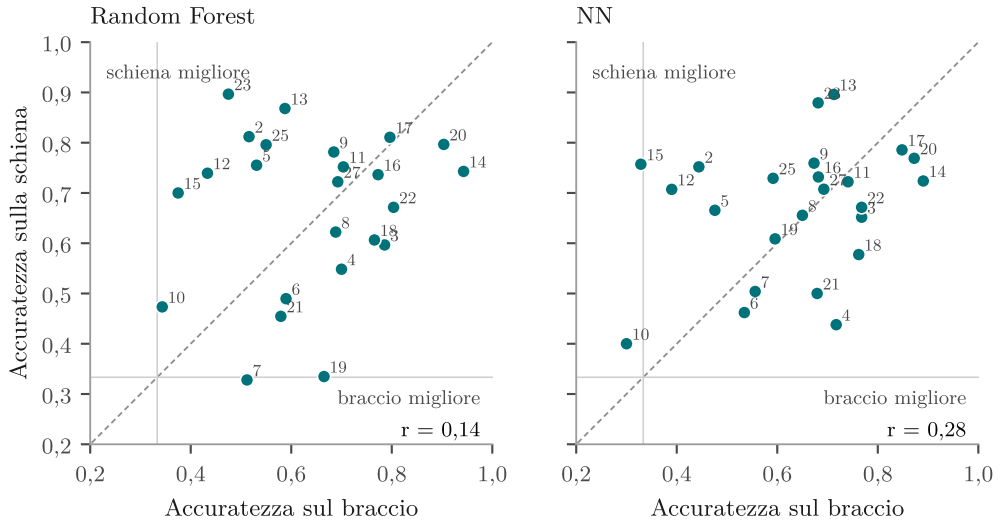


Figura 3.5: Accuratezza per soggetto sul braccio e sulla schiena per i 24 soggetti con entrambe le registrazioni, per la Random Forest (a sinistra) e la NN (a destra), per finestra. Sopra la diagonale la schiena è migliore, sotto il braccio; le linee grigie indicano il livello casuale (1/3). Accanto a ogni punto il numero del soggetto.

3.6 Effetto della correzione di polarità e della selezione delle feature

La correzione della polarità è stata valutata confrontando gli stessi modelli, con configurazione fissa, prima e dopo il ribaltamento del segnale (Figura 3.7). Solo per la Random Forest sul braccio l'intervallo di confidenza della differenza non comprende lo zero: il guadagno è di circa 2 centesimi di balanced accuracy, con un intervallo da +0,003 a +0,046. Considerato il numero di confronti eseguiti, va letto come un'indicazione e non come una prova. Negli altri casi le differenze sono piccole e

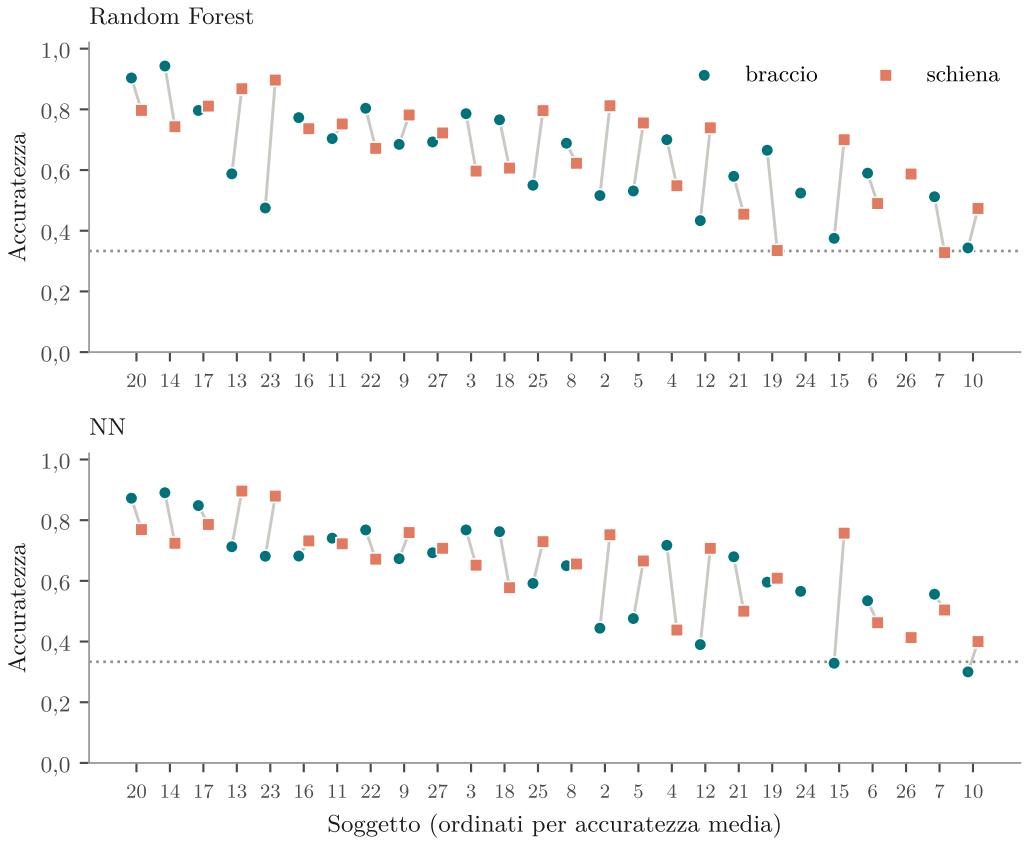


Figura 3.6: Accuratezza per soggetto della Random Forest (sopra) e della NN (sotto), per finestra, con i soggetti ordinati per valore medio. Le linee grigie collegano i due siti dello stesso soggetto; la linea tratteggiata indica 1/3, il valore atteso per una scelta casuale tra tre classi equiprobabili.

compatibili con zero. La correzione non è stata decisa in base a questi risultati, ma sulla forma del battito (Sezione 2.4), ed è stata mantenuta perché riporta l'onda all'orientamento convenzionale del PPG.

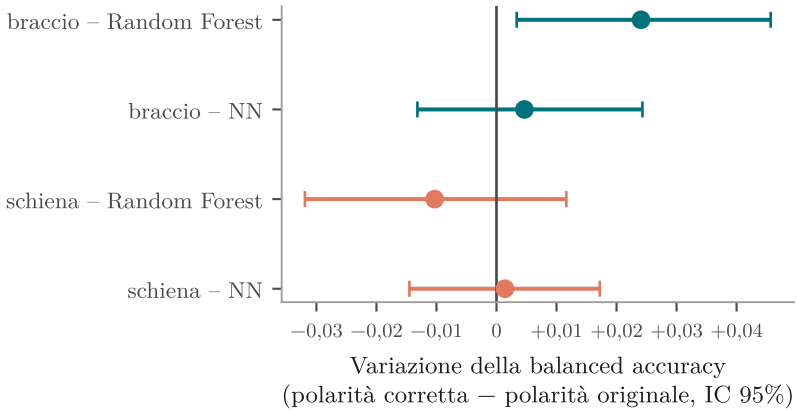


Figura 3.7: Variazione della balanced accuracy dovuta alla correzione della polarità, con intervallo di confidenza al 95%.

Il tuning annidato, infine, ha preferito più spesso tenere tutte le 55 feature sulla schiena: le 35 più informative sono state scelte solo nel 28–38% dei fold. Sul braccio sono state scelte nel 52–64% dei fold.

3.7 Effetto dell'esclusione delle pause

Per valutare l'effetto dell'esclusione delle pause, l'intera procedura, compreso il tuning annidato, è stata ripetuta anche tenendo le pause nella classe di riposo, come indicato nel file degli orari. In quel caso le finestre delle pause vengono classificate come stress alto nel 57% dei casi sul braccio e nel 65% sulla schiena con la Random Forest, nel 55% e nel 62% con la NN, contro il 10–15% delle finestre del riposo iniziale (Figura 3.8). Le pause, che sono circa un quinto delle finestre di riposo, rappresentano così tra il 37% e il 42% degli errori da stress basso a stress alto.

Il confronto tra le due versioni è riportato nella Tabella 3.3. Escludendo le pause la balanced accuracy per finestra sale di 2–4 centesimi e quella per task di 4–9 centesimi, e il recall dello stress basso passa da 0,56–0,62 a 0,64–0,66. Il confronto non isola un'unica causa, perché togliere le pause cambia anche la composizione della classe di riposo, ma è compatibile con l'ipotesi che una parte degli errori della versione con le pause dipendesse dall'etichetta di queste finestre più che dai modelli.

3.8 Importanza delle feature

La Figura 3.9 riporta le feature con l'importanza media più alta nella Random Forest, calcolata su tutti i fold e le ripetizioni. Su entrambi i siti le feature più usate descrivono la forma e l'ampiezza del battito: larghezza, pendenze di salita e

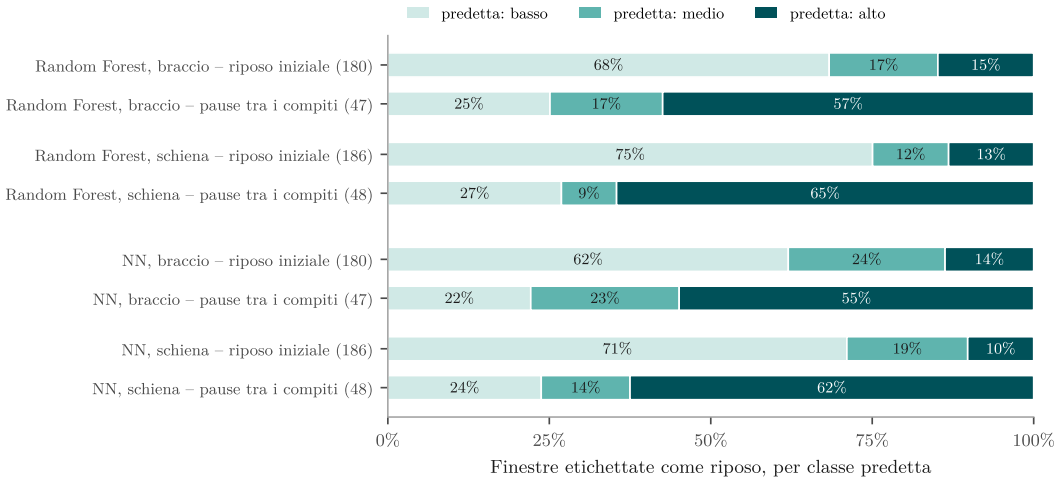


Figura 3.8: Classe predetta per le finestre etichettate come riposo nella versione con le pause, separando il riposo iniziale dalle pause tra i compiti (Random Forest e NN, per finestra, 10 ripetizioni). Tra parentesi il numero di finestre per ripetizione.

Tabella 3.3: Balanced accuracy della Random Forest e della NN senza le pause (configurazione finale) e con le pause nella classe di riposo (media su 10 ripetizioni e IC 95%).

Sito	Modello	Livello	Senza le pause	Con le pause
Braccio	Random Forest	Per finestra	0,582 (0,513–0,649)	0,566 (0,501–0,628)
		Ultime 3 finestre	0,597 (0,523–0,672)	0,581 (0,512–0,649)
		Per task	0,653 (0,580–0,726)	0,614 (0,532–0,693)
	NN	Per finestra	0,593 (0,521–0,661)	0,560 (0,498–0,623)
		Ultime 3 finestre	0,615 (0,535–0,691)	0,578 (0,508–0,648)
		Per task	0,680 (0,599–0,758)	0,620 (0,539–0,697)
Schiena	Random Forest	Per finestra	0,615 (0,556–0,671)	0,588 (0,528–0,649)
		Ultime 3 finestre	0,629 (0,562–0,695)	0,594 (0,523–0,663)
		Per task	0,696 (0,613–0,779)	0,609 (0,529–0,694)
	NN	Per finestra	0,624 (0,565–0,680)	0,589 (0,537–0,641)
		Ultime 3 finestre	0,639 (0,576–0,703)	0,603 (0,542–0,661)
		Per task	0,713 (0,643–0,785)	0,627 (0,563–0,693)

di discesa, ampiezza dei picchi. Sommando l'importanza per gruppo, la morfologia del battito pesa circa il 29% sul braccio e il 33% sulla schiena, mentre frequenza e variabilità delle pulsazioni pesano il 14% e l'11%. Il confronto tra gruppi va letto tenendo conto che contengono un numero diverso di feature: 16 la morfologia, 9 le pulsazioni. Sul braccio compaiono tra le prime anche frequenza media e intervallo PP medio, sulla schiena le statistiche di ampiezza dei picchi e l'inviluppo.

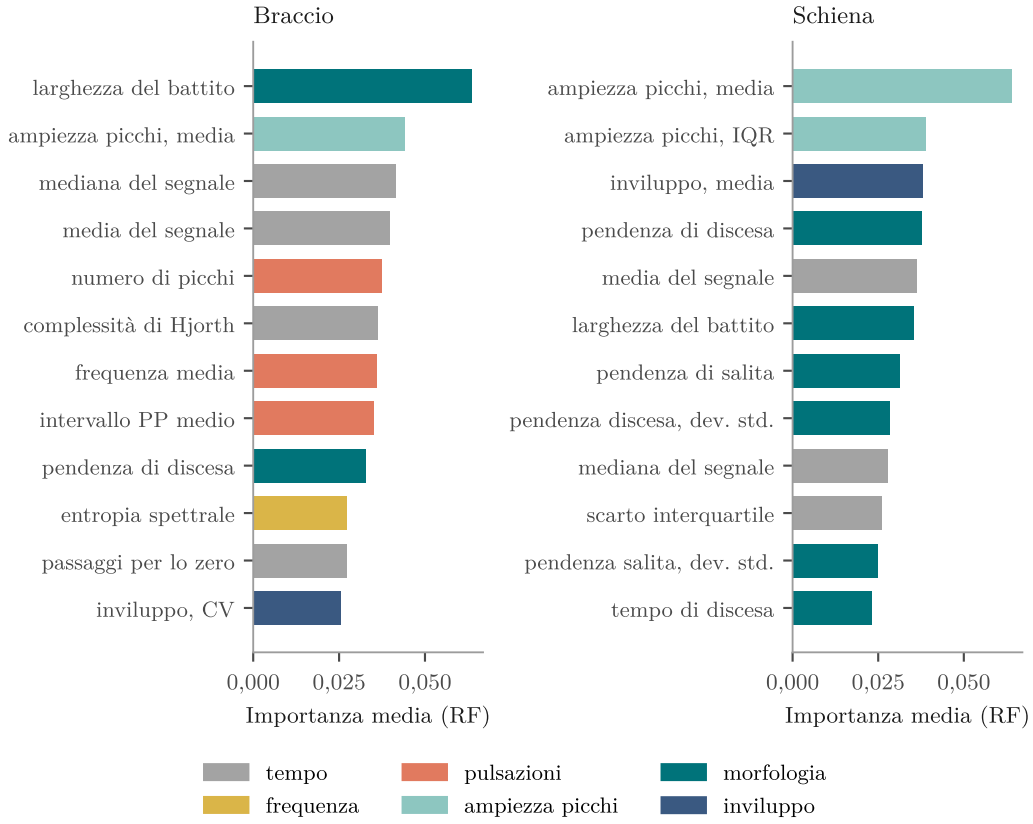


Figura 3.9: Le dodici feature con importanza media più alta nella Random Forest, per sito. Il colore indica il gruppo di appartenenza.

L'importanza della Random Forest va letta con cautela: tende a dividersi tra feature correlate, e molte delle 55 feature lo sono. Indica comunque che il modello non si limita alla frequenza cardiaca, ma usa anche la forma dell'onda.

3.9 Classificazione binaria riposo/stress

Per avere un termine di confronto con la letteratura, che affronta quasi sempre un problema binario, gli stessi dati sono stati classificati anche in due classi: riposo contro stress, unendo i quattro compiti. La procedura è la stessa, con Random Forest e NN a configurazione fissa e 10 ripetizioni. La balanced accuracy media è 0,714 per la Random Forest e 0,719 per la NN sul braccio, 0,751 e 0,759 sulla schiena, con una deviazione standard tra le ripetizioni tra 0,01 e 0,02; in questo caso un classificatore

casuale avrebbe 0,5. L'accuracy corrispondente è del 76–79% sul braccio e del 79–82% sulla schiena. Va però letta tenendo conto dello sbilanciamento: le finestre di stress sono circa il 72% del totale (464 su 644 sul braccio, 468 su 654 sulla schiena), e un classificatore che indicasse sempre stress raggiungerebbe già questa accuracy, con una balanced accuracy di 0,5. Anche in questo caso la schiena ottiene valori leggermente più alti.

Capitolo 4

Discussione

4.1 Interpretazione dei risultati

Il primo dato è che il PPG da solo, acquisito sul braccio o sulla schiena, contiene informazione sul livello di stress anche quando il modello viene valutato su persone che non ha mai visto. Con tre classi la balanced accuracy per finestra è di 0,58–0,59 sul braccio e di 0,62–0,63 sulla schiena. Sono valori ben sopra lo 0,33 del caso, ma lontani da una classificazione affidabile della singola finestra di un minuto. Le prestazioni salgono quando si combinano più finestre dello stesso compito, fino a 0,68 sul braccio e 0,71 sulla schiena con la NN. Questo suggerisce che una parte degli errori sia dovuta alla variabilità del segnale da un minuto all'altro, più che a un'assenza di informazione.

Il secondo dato riguarda i modelli. Random Forest e reti neurali dense danno risultati molto simili tra loro: le differenze sono di uno o due centesimi, entro la variabilità dovuta alla scelta dei fold. Con questo dataset, quindi, cambiare classificatore tra quelli provati sposta poco i risultati: il margine di miglioramento sembra legato soprattutto all'informazione contenuta nelle feature e alla definizione delle classi. La 1D-CNN applicata direttamente al segnale è invece nettamente peggiore, intorno a 0,44–0,54. Una spiegazione possibile è il numero ridotto di esempi: con 25 soggetti per sito e circa 650 finestre, una rete che deve imparare da sola cosa osservare nella forma d'onda ha pochi dati a disposizione, mentre le feature progettate a mano riassumono già frequenza, variabilità e morfologia del battito. Va però ricordato che la 1D-CNN è stata usata in una configurazione fissa, senza tuning: il risultato riguarda questa configurazione, non le reti convoluzionali in generale.

4.2 Confronto con la letteratura

Un confronto diretto con altri studi è difficile, perché cambiano i segnali, i protocolli, il numero di classi e soprattutto il modo di validare i modelli. Alcuni riferimenti permettono comunque di collocare i risultati.

Per il problema a tre classi il termine di paragone più vicino è il dataset WESAD, in cui il solo PPG del polso, valutato con la LOSO, raggiunge un F1-score di 0,55 nel distinguere riposo, stress e divertimento [30]. In questa tesi il macro-F1 per finestra

è di 0,58–0,59 sul braccio e di 0,62 sulla schiena. Le classi non sono le stesse, ma l'ordine di grandezza coincide: con il solo PPG, tre stati vengono riconosciuti in modo parziale anche nei dataset più usati. Anche Seo et al., con segnali multimodali, passano dal 73,3% di accuracy con due classi al 54,4% con tre, sebbene i due risultati siano ottenuti con combinazioni di segnali diverse [31].

Il confronto più interessante riguarda la classificazione binaria. Sugli stessi dati e con le stesse sedi, lo studio di Campanella, Seepold e Palma riporta un'accuracy del 66–67% sul braccio e del 72% sulla schiena [5]. In questa tesi, distinguendo riposo e stress, la balanced accuracy è di 0,71–0,72 sul braccio e di 0,75–0,76 sulla schiena, con un'accuracy tra il 76% e l'82%, a fronte del 72% circa che si otterrebbe indicando sempre stress. I due lavori arrivano a valori vicini, ma con una differenza importante: lo studio originale divideva i segmenti in modo casuale tra training e test, mentre qui i soggetti di test non sono mai stati visti durante l'addestramento. Una validazione più severa, che di solito abbassa le prestazioni, porta qui a valori simili o leggermente superiori. Il confronto è incoraggiante, ma le differenze tra le due procedure non consentono di attribuire il risultato a una specifica scelta di elaborazione.

I valori restano comunque inferiori a quelli riportati per WESAD in binario, dove il PPG del polso arriva all'85,8% [30]. I due dataset, però, differiscono per protocollo e durata delle fasi di stress. In questa tesi, inoltre, i quattro compiti sono molto diversi tra loro e uno di essi, lo Stroop test, è percepito come poco stressante, eppure fa parte della classe di stress.

4.3 Confronto tra i siti di misura

Su quasi tutte le misure la schiena ottiene risultati leggermente migliori del braccio: circa tre centesimi per finestra con entrambi i modelli. Il risultato è in linea con l'articolo di riferimento, che aveva trovato un segnale lombare un po' più pulito e stabile durante l'attività libera [5]. Il confronto appaiato sugli stessi soggetti, però, mostra che la differenza rientra nell'incertezza delle stime, e la schiena è migliore solo in circa metà dei soggetti. Con 25 soggetti non è possibile affermare che una sede sia davvero migliore dell'altra. Si può dire che entrambe forniscono un'informazione utilizzabile e confrontabile. La bassa correlazione tra le accuratezze dei due siti sullo stesso soggetto suggerisce anche che braccio e schiena non contengano esattamente la stessa informazione: è un motivo in più per valutare in futuro l'uso combinato dei due sensori.

Due osservazioni vanno nella stessa direzione dell'articolo di riferimento. Sulla schiena il tuning annidato ha preferito più spesso tenere tutte le feature, come nello studio originale, dove la regione lombare dava i risultati migliori con l'insieme completo. Sul braccio, invece, la correzione della polarità ha migliorato la Random Forest. Una spiegazione possibile è che sul braccio le feature morfologiche, calcolate dai picchi del segnale, risentano di più dell'orientamento dell'onda. Il dato però non basta per esserne sicuri.

4.4 Riconoscimento della classe intermedia e qualità delle etichette

La classe di stress medio è la più difficile da riconoscere: viene individuata nel 37–48% dei casi, a seconda del modello e della sede, e confusa soprattutto con lo stress basso. Ci sono almeno tre motivi.

Il primo è il numero di esempi: la classe media corrisponde a un solo compito, il sing-a-song test, e ha meno della metà delle finestre di ciascuna delle altre due classi. Il secondo è la natura del compito. Leggere ad alta voce e cantare modificano la respirazione e richiedono qualche movimento, e il PPG risente di entrambi. Nei questionari 9 partecipanti hanno riferito di aver avvertito il battito accelerato proprio durante il task 3, tanti quanti nel test aritmetico e nello speech test, anche se il compito è stato giudicato complessivamente meno stressante. Eppure, nell'insieme del task 3, la frequenza stimata dal PPG resta vicina a quella del riposo: leggermente più bassa considerando le finestre normalizzate (Figura 3.1), leggermente più alta nell'andamento continuo, dove l'aumento compare solo durante il canto (Figura 2.3b). Una spiegazione possibile è che la respirazione lenta e profonda del canto e i piccoli movimenti disturbino il riconoscimento dei picchi, ma con questi dati non è possibile verificarlo. Il segnale misurato e la percezione dei partecipanti, quindi, non vanno sempre nella stessa direzione.

Il terzo motivo riguarda le etichette in generale. Ogni fase del protocollo riceve la stessa classe per tutti i partecipanti, ma la risposta allo stress è individuale: lo Stroop test con rumore di fondo può essere quasi rilassante per una persona e impegnativo per un'altra. È uno dei limiti indicati più spesso dalle review sulla stress detection [29, 3]. In questa tesi se ne vede un esempio nelle pause tra i compiti. Nel file degli orari sono etichettate come riposo, ma quando vengono tenute nel dataset il modello le assegna più spesso allo stress alto che a quello basso (Sezione 3.7), e la frequenza di pulsazione nelle pause resta sopra il riposo iniziale (Figura 2.3b). Questi dati sono compatibili con un recupero ancora incompleto a un minuto dalla fine di un compito impegnativo, e quindi con un'etichetta di riposo poco adatta a queste finestre: per questo si è scelto di escluderle. Non dimostrano però che l'etichetta sia sbagliata, e per una conferma servirebbe un riferimento indipendente dello stato del partecipante durante le pause.

4.5 Variabilità inter-soggetto

La differenza più marcata non è tra modelli o tra sedi, ma tra persone. L'accuratezza per soggetto va da circa 0,3 a 0,9, con entrambi i modelli. Per alcuni partecipanti il modello riconosce bene quasi tutte le fasi, per altri resta vicino al caso. La normalizzazione delle feature rispetto ai valori tipici di ciascun soggetto riduce solo in parte questa variabilità. Ognuno reagisce allo stress con un'intensità e una

combinazione di risposte diverse, e un modello addestrato su altre persone non può conoscerle in anticipo.

Questo spiega anche perché gli intervalli di confidenza restano larghi e perché, con un singolo run della validazione, i risultati possono cambiare di qualche centesimo solo cambiando il modo in cui i soggetti vengono divisi nei fold. Ripetere la validazione e riportare media e intervallo di confidenza dà un'immagine più realistica di ciò che ci si può aspettare su un nuovo gruppo di persone.

4.6 Limiti dello studio

I risultati vanno letti tenendo conto di alcuni limiti.

- **Campione piccolo e omogeneo.** I soggetti sono 25 per sito, sani e in maggioranza tra i 21 e i 30 anni. Non si sa se i risultati valgano per altre fasce di età o per persone con diabete, che sono il contesto applicativo delle due sedi.
- **Protocollo di laboratorio.** Lo stress è indotto da compiti brevi e in condizioni controllate. Nella vita quotidiana movimento, postura e attività fisica renderebbero il segnale più rumoroso.
- **Classi legate alle attività.** I compiti si svolgono nello stesso ordine per tutti i partecipanti e sono diversi tra loro: alcuni richiedono di parlare o cantare, che modificano la respirazione, e comportano piccoli movimenti. Il modello potrebbe quindi riconoscere anche differenze tra attività, o tra momenti diversi del protocollo, oltre che tra livelli di stress. Riconoscere queste classi non equivale a misurare lo stress in qualunque situazione.
- **Etichette legate al compito.** La classe dipende dalla fase del protocollo e non da come si sentiva davvero ciascun partecipante. Le brevi pause tra i compiti sono state escluse proprio per questo motivo, ma anche all'interno di un compito la risposta può cambiare nel tempo, come mostra il sing-a-song test.
- **Early stopping per finestre.** La validazione interna usata per l'arresto anticipato delle reti è estratta per finestre e non per soggetto: finestre dello stesso soggetto possono finire sia nella parte usata per addestrare sia in quella usata per decidere quando fermarsi. I soggetti di test restano esclusi, ma l'arresto può risultare meno severo (Sezione 2.7).
- **Normalizzazione per soggetto.** Lo z-score è calcolato sull'intera registrazione del soggetto, compresa quella di test. Non usa le etichette, ma in un'applicazione reale servirebbe prima un periodo di registrazione della persona.
- **Finestre brevi.** Per l'analisi della variabilità delle pulsazioni le linee guida raccomandano registrazioni di almeno 5 minuti [12]. Le feature di variabilità calcolate su 60 s vanno quindi considerate indici a brevissimo termine.

- **Frequenza non verificata con l'ECG.** La frequenza di pulsazione e gli intervalli tra battiti sono stimati dai picchi del PPG, senza un ECG registrato in parallelo. Picchi persi o falsi rilevamenti, più probabili durante il parlato, il canto o piccoli movimenti, non possono quindi essere individuati con certezza (Sezione 1.2.4).
- **Nessuna validazione esterna.** Tutte le prove sono state svolte sullo stesso dataset, e anche le scelte che hanno portato alla configurazione finale, come l'esclusione delle pause, sono state fatte osservando questi dati: i risultati vanno confermati su dati nuovi.

Conclusioni

Il lavoro ha valutato la classificazione di tre livelli di stress a partire dal solo segnale PPG acquisito sulla parte superiore del braccio e nella regione lombare, con una pipeline sviluppata dai dati grezzi e una validazione per soggetto ripetuta. Con le feature estratte da finestre di 60 s, la Random Forest e la NN raggiungono una balanced accuracy di 0,58–0,59 sul braccio e di 0,62–0,63 sulla schiena, contro lo 0,33 di un classificatore casuale; aggregando le finestre di ciascun compito si arriva a 0,65–0,68 e a 0,70–0,71. Lo stress alto è riconosciuto in sette-otto casi su dieci, mentre lo stress medio, rappresentato dal solo sing-a-song test, resta la classe più critica. Nella classificazione binaria riposo/stress la balanced accuracy è di 0,71–0,76 e l'accuracy del 76–82%, a fronte del 66–72% riportato sugli stessi dati con una validazione per segmenti, meno severa; il confronto va letto tenendo conto che, indicando sempre stress, si otterrebbe un'accuracy del 72% circa.

Il confronto tra le sedi non mostra una differenza significativa. La regione lombare ottiene in media circa tre centesimi in più, in accordo con la maggiore stabilità del segnale osservata nello studio di riferimento, ma l'intervallo di confidenza della differenza appaiata comprende lo zero e la sede migliore cambia da un soggetto all'altro. Anche tra i classificatori basati sulle feature le differenze restano entro la variabilità dovuta alla suddivisione in fold, mentre la 1D-CNN applicata al segnale, addestrata su circa 650 finestre per sito e senza tuning, ottiene prestazioni nettamente inferiori. In questo lavoro, e con la configurazione di 1D-CNN provata, le feature progettate sulla base della fisiologia del segnale sono risultate più efficaci dell'apprendimento diretto dalla forma d'onda.

Più della scelta del classificatore hanno pesato le scelte metodologiche: la definizione delle etichette, come suggerisce l'effetto dell'esclusione delle pause tra i compiti, l'aggregazione delle finestre di uno stesso compito e una validazione che tenga separati i soggetti e venga ripetuta, così da non sovrastimare le prestazioni. La correzione della polarità ha un ruolo diverso: riporta la forma d'onda all'orientamento corretto, necessario per descriverne la morfologia, ma il suo effetto sulle prestazioni è piccolo e non uniforme tra modelli e sedi.

Nel complesso, i risultati indicano che un sensore PPG posto nelle stesse sedi dei dispositivi per il monitoraggio continuo del glucosio fornisce un'informazione sul livello di stress utilizzabile anche su soggetti nuovi. È un primo passo verso l'integrazione di questa informazione nei sistemi di monitoraggio per il diabete di tipo 1, che richiede però una verifica su campioni più ampi e in condizioni meno controllate.

Sviluppi futuri

Il primo sviluppo riguarda i dati. Un campione più ampio e una validazione su una popolazione esterna permetterebbero di stimare le prestazioni con maggiore precisione e di confermare le scelte fatte su questo dataset. Acquisizioni su persone con diabete di tipo 1, il contesto applicativo per cui le due sedi sono state scelte, sono necessarie per verificare che il segnale conservi le stesse caratteristiche. Anche il protocollo può essere migliorato: valutazioni dello stress raccolte subito dopo ogni fase consentirebbero di costruire etichette individuali, e pause più lunghe tra i compiti renderebbero possibile registrare il recupero invece di escluderlo dall'analisi.

Sul piano metodologico, la forte variabilità tra soggetti suggerisce di valutare modelli personalizzati, adattati a ciascuna persona con una breve registrazione di calibrazione. L'acquisizione sincrona nelle due sedi permette inoltre di studiare l'uso combinato dei due sensori, mentre finestre più lunghe renderebbero più affidabili gli indici di variabilità delle pulsazioni. Infine, prove in condizioni di vita quotidiana, con movimento e attività libere, sono indispensabili per verificare la robustezza del segnale e dei modelli al di fuori del laboratorio.

Appendice A – Dettagli tecnici

Questa appendice raccoglie i dettagli necessari a riprodurre la configurazione finale: l'elenco completo delle feature (Tabella A.1) e i parametri dei modelli (Tabella A.2).

A.1 Elenco completo delle feature

Tutte le feature sono calcolate su ciascuna finestra di 60 s del segnale filtrato e con polarità corretta, poi normalizzate con uno z-score per soggetto e sito (Sezione 2.6).

Tabella A.1: Le 55 feature estratte da ogni finestra, divise per gruppo.

Feature	Descrizione
Dominio del tempo (13)	
media, dev. std.	media e deviazione standard dei campioni della finestra
minimo, massimo	valore minimo e massimo
escursione	differenza tra massimo e minimo
mediana, IQR	mediana e scarto interquartile
asimmetria, curtosi	momenti di terzo e quarto ordine della distribuzione dei campioni
valore efficace	radice della media dei quadrati (RMS)
passaggi per lo zero	numero di cambi di segno del segnale filtrato al secondo
mobilità e complessità di Hjorth	rapporti tra le varianze del segnale e delle sue derivate, legati alla frequenza media e alla regolarità della forma d'onda
Dominio della frequenza (9), densità spettrale stimata con il metodo di Welch	
frequenza dominante	frequenza del massimo della densità spettrale tra 0,5 e 4 Hz
centroide, dispersione	baricentro dello spettro e sua deviazione standard
entropia spettrale	entropia dello spettro normalizzato: è bassa se la potenza è concentrata in poche frequenze
potenze relative	frazione della potenza nelle bande 0,5–1, 1–2, 2–4 e 4–8 Hz
potenza totale	potenza nella banda 0,5–8 Hz
Frequenza e variabilità delle pulsazioni (9), su intervalli PP tra 0,3 e 1,5 s	
numero di picchi	picchi sistolici individuati nella finestra
frequenza media	frequenza di pulsazione media, in battiti al minuto
PP medio, CV degli PP	media e coefficiente di variazione degli intervalli picco-picco
SDNN	deviazione standard degli intervalli
RMSSD	radice della media dei quadrati delle differenze tra intervalli successivi
pNN50	percentuale di differenze successive superiori a 50 ms

Feature	Descrizione
SD1, SD2	dispersioni del diagramma di Poincaré, legate alla variabilità a breve e a lungo termine
Ampiezza dei picchi (5)	
media, dev. std., CV, IQR, RMSSD	statistiche dell'ampiezza dei picchi sistolici
Morfologia del battito (16)	
larghezza (media, dev. std.)	durata del battito a metà altezza
ampiezza piede-picco (media, dev. std., CV, IQR, RMSSD)	differenza tra picco sistolico e minimo che lo precede
tempo di salita, tempo di discesa	durata media della fase di salita e di quella di discesa
pendenza di salita (media, dev. std.)	velocità di salita del battito
pendenza di discesa (media, dev. std.)	velocità di discesa del battito
area (media, dev. std., CV)	area del battito sopra la retta che unisce i due minimi adiacenti
Inviluppo (3)	
media, dev. std., CV	statistiche dell'inviluppo ottenuto con la trasformata di Hilbert

A.2 Parametri dei modelli

Tabella A.2: Parametri della configurazione finale. Dove sono indicate due alternative, la scelta è fatta dal tuning annidato all'interno di ogni fold (Sezione 2.7).

Elemento	Impostazione
Dati	finestre di 60 s con sovrapposizione del 40%, pause tra i compiti escluse; z-score per soggetto e sito calcolato sulle finestre usate
Selezione delle feature	tutte le 55 oppure le 35 con mutua informazione più alta
Bilanciamento	SMOTE sul training oppure pesi di classe inversamente proporzionali alla numerosità
Random Forest	300 alberi, $\sqrt{p}$ feature per divisione, con p numero di feature mantenute (55 o 35); profondità libera e almeno 1 osservazione per foglia, oppure profondità massima 12 e almeno 2 osservazioni per foglia
NN (`scikit-learn`)	strati nascosti 64–32 con $\alpha = 10^{-4}$ e passo 10^{-3}, oppure 128–64 con $\alpha = 10^{-3}$ e passo $3 \cdot 10^{-4}$; ottimizzatore Adam, al massimo 300 epoche con arresto anticipato
NN (`Keras`)	64 neuroni con normalizzazione del batch e dropout 0,3, poi 32 neuroni con dropout 0,3, uscita softmax; Adam con passo 10^{-3}, lotti da 32, al massimo 200 epoche, arresto dopo 15 epoche senza miglioramento della perdita su un 10% di validazione interna
1D-CNN	tre blocchi con 16, 32 e 64 filtri (dimensioni 9, 7 e 5), normalizzazione del batch, ReLU e max pooling di 4; pooling globale, strato denso da 32 con dropout 0,4, uscita softmax; stesso addestramento della rete in Keras, segnale sottocampionato di 4
Validazione esterna	5 fold per soggetto stratificati, ripetuti con 10 semi diversi
Validazione interna	2 fold per soggetto, solo sui soggetti di training
Intervalli di confidenza	bootstrap sui soggetti, 2000 ricampionamenti

Bibliografia

[1] G. P. Chrousos. "Stress and Disorders of the Stress System". *Nature Reviews Endocrinology*, vol. 5, n. 7, pp. 374–381, 2009. https://doi.org/10.1038/nrendo.2009.106

[2] S. Campanella, A. Altaleb, A. Belli, P. Pierleoni e L. Palma. "A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques". *Sensors*, vol. 23, n. 7, articolo 3565, 2023. https://doi.org/10.3390/s23073565

[3] M. Bolpagni, S. Pardini, M. Dianti e S. Gabrielli. "Personalized Stress Detection Using Biosignals from Wearables: A Scoping Review". *Sensors*, vol. 24, n. 10, articolo 3221, 2024. https://doi.org/10.3390/s24103221

[4] P. H. Charlton, P. A. Kyriacou, J. Mant, V. Marozas, P. Chowienczyk e J. Alastruey. "Wearable Photoplethysmography for Cardiovascular Monitoring". *Proceedings of the IEEE*, vol. 110, n. 3, pp. 355–381, 2022. https://doi.org/10.1109/JPROC.2022.3149785

[5] S. Campanella, R. Seepold e L. Palma. "Optimising Wearable Sensors Placement for Type 1 Diabetes Monitoring: A Preliminary Study". In *2026 IEEE Sensors Applications Symposium (SAS)*, IEEE, 2026, pp. 1–6. Articolo accettato, in corso di pubblicazione.

[6] R. S. Surwit, M. S. Schneider e M. N. Feinglos. "Stress and diabetes mellitus". *Diabetes Care*, vol. 15, n. 10, pp. 1413–1422, 1992. https://doi.org/10.2337/diacare.15.10.1413

[7] D. M. F. Ingrosso, M. Primavera, S. Samvelyan, V. M. Tagi e F. Chiarelli. "Stress and Diabetes Mellitus: Pathogenetic Mechanisms and Clinical Outcome". *Hormone Research in Paediatrics*, vol. 96, n. 1, pp. 34–43, 2023. https://doi.org/10.1159/000522431

[8] J. E. Hall e M. E. Hall. *Guyton and Hall Textbook of Medical Physiology*, 14ª ed. Elsevier, 2021. Capitoli 9 (*Cardiac Muscle; The Heart as a Pump and Function of the Heart Valves*), 61 (*The Autonomic Nervous System and the Adrenal Medulla*) e 78 (*Adrenocortical Hormones*).

[9] J. I. Gupta e M. J. Shea. "Biologia del cuore". *Manuale MSD, versione per i pazienti*, revisione di aprile 2025. https://www.

msdmanuals.com/it/casa/disturbi-cardiaci-e-dei-vasi-sanguigni/biologia-del-cuore-e-dei-vasi-sanguigni/biologia-del-cuore. Consultato il 3 ottobre 2026.

[10] Blausen.com staff. "Medical gallery of Blausen Medical 2014". *WikiJournal of Medicine*, vol. 1, n. 2, 2014. https://doi.org/10.15347/wjm/2014.010

[11] R. E. Klabunde. "Cardiac Cycle". *Cardiovascular Physiology Concepts*. https://cvphysiology.com/Heart%20Disease/HD002. Consultato il 3 ottobre 2026.

[12] E. Mejía-Mejía, J. Allen, K. Budidha, C. El-Hajj, P. A. Kyriacou e P. H. Charlton. "Photoplethysmography Signal Processing and Synthesis". In P. A. Kyriacou e J. Allen (a cura di), *Photoplethysmography: Technology, Signal Analysis and Applications*, Academic Press (Elsevier), 2022, pp. 69–146. https://doi.org/10.1016/B978-0-12-823374-0.00015-3

[13] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee e B.-H. Koo. "Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature". *Psychiatry Investigation*, vol. 15, n. 3, pp. 235–245, 2018. https://doi.org/10.30773/pi.2017.08.17

[14] P. H. Charlton e V. Marozas. "Wearable Photoplethysmography Devices". In P. A. Kyriacou e J. Allen (a cura di), *Photoplethysmography: Technology, Signal Analysis and Applications*, Academic Press (Elsevier), 2022, cap. 12, pp. 401–439. https://doi.org/10.1016/B978-0-12-823374-0.00011-6

[15] Scikit-learn developers. *Getting Started*. https://scikit-learn.org/stable/getting_started.html. Consultato il 30 settembre 2026.

[16] Scikit-learn developers. *Feature selection*. https://scikit-learn.org/stable/modules/feature_selection.html. Consultato il 30 settembre 2026.

[17] Scikit-learn developers. *Common pitfalls and recommended practices*. https://scikit-learn.org/stable/common_pitfalls.html. Consultato il 30 settembre 2026.

[18] Scikit-learn developers. *compute_class_weight*. https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html. Consultato il 30 settembre 2026.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall e W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. https://doi.org/10.1613/jair.953

[20] Imbalanced-learn developers. *Over-sampling*. https://imbalanced-learn.org/stable/over_sampling.html. Consultato il 30 settembre 2026.

[21] Imbalanced-learn developers. *Common pitfalls and recommended practices*. `https://imbalanced-learn.org/stable/common_pitfalls.html`. Consultato il 30 settembre 2026.

[22] Scikit-learn developers. *Cross-validation: evaluating estimator performance*. `https://scikit-learn.org/stable/modules/cross_validation.html`. Consultato il 30 settembre 2026.

[23] Scikit-learn developers. *Nested versus non-nested cross-validation*. `https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html`. Consultato il 30 settembre 2026.

[24] L. Breiman. "Random Forests". *Machine Learning*, vol. 45, n. 1, pp. 5–32, 2001. `https://doi.org/10.1023/A:1010933404324`

[25] Scikit-learn developers. *Ensembles: Random forests and other randomized tree ensembles*. `https://scikit-learn.org/stable/modules/ensemble.html`. Consultato il 30 settembre 2026.

[26] Scikit-learn developers. *Neural network models (supervised)*. `https://scikit-learn.org/stable/modules/neural_networks_supervised.html`. Consultato il 30 settembre 2026.

[27] Keras developers. *Conv1D layer*. `https://keras.io/api/layers/convolution_layers/convolution1d/`. Consultato il 30 settembre 2026.

[28] T. Fawcett. "An introduction to ROC analysis". *Pattern Recognition Letters*, vol. 27, n. 8, pp. 861–874, 2006. `https://doi.org/10.1016/j.patrec.2005.10.010`

[29] M. Namvari, J. Lipoth, S. Knight, A. A. Jamali, M. Hedayati, R. J. Spiteri e S. Syed-Abdul. "Photoplethysmography Enabled Wearable Devices and Stress Detection: A Scoping Review". *Journal of Personalized Medicine*, vol. 12, n. 11, articolo 1792, 2022. `https://doi.org/10.3390/jpm12111792`

[30] P. Schmidt, A. Reiss, R. Dürichen, C. Marberger e K. Van Laerhoven. "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection". In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*, pp. 400–408, 2018. `https://doi.org/10.1145/3242969.3242985`

[31] W. Seo, N. Kim, C. Park e S.-M. Park. "Deep Learning Approach for Detecting Work-Related Stress Using Multimodal Signals". *IEEE Sensors Journal*, vol. 22, n. 12, pp. 11892–11902, 2022. `https://doi.org/10.1109/JSEN.2022.3170915`

[32] S. K. Longmore, G. Y. Lui, G. Naik, P. P. Breen, B. Jalaludin e G. D. Gargiulo. "A Comparison of Reflective Photoplethysmography for Detection of Heart

Rate, Blood Oxygen Saturation, and Respiration Rate at Various Anatomical Locations". *Sensors*, vol. 19, n. 8, articolo 1874, 2019. https://doi.org/10.3390/s19081874

[33] I. T. Hettiarachchi, S. Hanoun, D. Nahavandi e S. Nahavandi. "Validation of Polar OH1 optical heart rate sensor for moderate and high intensity physical activities". *PLOS ONE*, vol. 14, n. 5, articolo e0217288, 2019. https://doi.org/10.1371/journal.pone.0217288

[34] H. Barki e W.-Y. Chung. "Mental Stress Detection Using a Wearable In-Ear Plethysmography". *Biosensors*, vol. 13, n. 3, articolo 397, 2023. https://doi.org/10.3390/bios13030397

[35] J. R. Stroop. "Studies of interference in serial verbal reactions". *Journal of Experimental Psychology*, vol. 18, n. 6, pp. 643–662, 1935. https://doi.org/10.1037/h0054651

[36] G. Stoet. "PsyToolkit: A software package for programming psychological experiments using Linux". *Behavior Research Methods*, vol. 42, n. 4, pp. 1096–1104, 2010. https://doi.org/10.3758/BRM.42.4.1096

[37] G. Stoet. "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments". *Teaching of Psychology*, vol. 44, n. 1, pp. 24–31, 2017. https://doi.org/10.1177/0098628316677643

[38] C. Kirschbaum, K.-M. Pirke e D. H. Hellhammer. "The 'Trier Social Stress Test' – a tool for investigating psychobiological stress responses in a laboratory setting". *Neuropsychobiology*, vol. 28, n. 1–2, pp. 76–81, 1993. https://doi.org/10.1159/000119004

[39] K. Dedovic, R. Renwick, N. K. Mahani, V. Engert, S. J. Lupien e 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, n. 5, pp. 319–325, 2005.

[40] D. J. van der Mee, Q. Duivestein, M. J. Gevonden, J. H. D. M. Westerink e E. J. C. de Geus. "The Short Sing-a-Song Stress Test: A Practical and Valid Test of Autonomic Responses Induced by Social-Evaluative Stress". *Autonomic Neuroscience*, vol. 224, articolo 102612, 2020. https://doi.org/10.1016/j.autneu.2019.102612

[41] B. Hjorth. "EEG analysis based on time domain properties". *Electroencephalography and Clinical Neurophysiology*, vol. 29, n. 3, pp. 306–310, 1970. https://doi.org/10.1016/0013-4694(70)90143-4

[42] P. D. Welch. "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms". *IEEE Transactions on Audio and Electroacoustics*, vol. 15, n. 2, pp. 70–73, 1967. https://doi.org/10.1109/TAU.1967.1161901

[43] B. Efron. "Bootstrap methods: another look at the jackknife". *The Annals of Statistics*, vol. 7, n. 1, pp. 1–26, 1979. https://doi.org/10.1214/aos/1176344552

[44] F. Wilcoxon. "Individual comparisons by ranking methods". *Biometrics Bulletin*, vol. 1, n. 6, pp. 80–83, 1945. https://doi.org/10.2307/3001968

Ringraziamenti

Innanzitutto vorrei ringraziare il Prof. Lorenzo Palma e la dottoranda Sara Campanella per l'infinita disponibilità durante il periodo di tirocinio e nella stesura dell'elaborato.

Essendo una persona molto pensierosa, inizialmente mi sono chiesto molte volte che cosa significasse "fare l'università" e perché dovesse essere così importante, come dicono le persone, tanto da dedicarle un periodo che, a quest'età e di questi tempi, è determinante per il resto della vita. Ebbene, non ho trovato una risposta in poco tempo. Ho dedicato questi tre anni e poco più a rispondere a questo dubbio attraverso due parti di me differenti: una vedeva questo periodo come studio e progresso nella carriera universitaria, concentrata completamente sugli obiettivi; l'altra, nettamente separata dalla prima, era l'autoanalisi di me stesso. Due parti che spesso entravano in conflitto, si alternavano continuamente e intensamente, ma non riuscivano mai a conciliarsi, e per molto tempo sono rimaste due estremi. Dopo una grande lotta, in un percorso entusiasmante, ho trovato le mie risposte, ma non ero solo: con me c'eravate voi, e vi ringrazio.

Ringrazio in primis me stesso, per essermi messo in gioco, per aver resistito e poi per aver preso la palla al balzo in un periodo in cui le difficoltà sono state tutt'altro che accademiche, varcando sempre la soglia del limite, approfondendo e dedicandomi all'autoconoscenza e alla crescita personale, oltre che allo sviluppo di competenze in ambito universitario.

Ho ringraziato e ringrazierò ancora per molto tempo mio fratello Federico, senza ombra di dubbio la persona più importante per me, partecipe di ogni minimo passo da sempre, ma ancora di più da quando eravamo ormai separati, a distanza, e abbiamo reso il nostro legame ancora più profondo. Lontani, ma sempre insieme; occupati, ma sempre disponibili l'uno per l'altro. Sei un esempio assoluto, per me il più significativo, e qui non parlo di periodi ma di vita, e anche il più forte, oggettivamente. Sono grato e mi reputo fortunato ad averti; è scontato dire che senza di te non ce l'avrei fatta, perché senza di te non sarei me stesso, ed è molto peggio di non farcela. Sono sicuro, inoltre, che il nostro rapporto abbia ancora molto margine per crescere.

E grazie al resto della mia famiglia, per tutto quello che ha potuto fare per me: mio padre Fabrizio, mia nonna Rita e mio nonno Antonio; mia zia Cinzia, di grande aiuto e spinta all'inizio del percorso; mio zio Joe, che ha temprato la mia personalità, accrescendo immensamente la mia motivazione nella carriera e nei miei obiettivi, e ricordandomi il valore e la gratitudine per le cose comuni.

Ringrazio Diego e Manuel: pur avendo preso strade diverse, con vite diverse, siamo rimasti quelli di sempre, anzi abbiamo rafforzato la nostra amicizia, trovando con il tempo sempre più punti in comune, obiettivi e aspirazioni simili, che ci hanno resi un gruppo inscindibile. Sono anni che ci conosciamo, e non ho mai dubitato di voi. Non vi scambierei per nessuno, indubbiamente i vicini migliori che possano capitarti. Sono sicuro che tutto continuerà così.

Ringrazio i miei amici Nazareno e Salvatore, conosciuti fin da subito, con cui ho condiviso momenti centrali, esperienze e grandi riflessioni sulla nostra personalità. Siamo stati benissimo e malissimo insieme. Oggi, guardando indietro, sono contento delle persone che siete diventati: siete due persone di grandissima curiosità, e sembra che non sappiate come si sbaglia con un amico. Salvatore, grazie per essere qui oggi, e per avermi sollevato in un periodo che è stato il più vorticoso e doloroso di sempre.

Grazie, Marco, perché è grazie a te che conosco la sensazione di avere un amico che, anche se non lo vedi per anni, resta sempre lo stesso nei tuoi pensieri, e quando vi incontrate dopo tanto tempo è come se foste stati insieme il giorno prima: sei sempre presente, anche a distanza. Ludovico, una persona a cui, a dir poco, si può attribuire l'aggettivo "generosa": in poco tempo siamo diventati tanto, e ci aspetta ancora molto altro, siamo solo alle fondamenta. Marco, i nostri destini si sono incrociati; con te, Ludovico, mi sono impegnato per renderli realtà. Quindi, ci vediamo fra poco a Torino.

Mattia, è grazie anche a te se oggi tutto questo è possibile. Ospitali l'uno con l'altro, dandoci tutto senza mai chiedere niente, abbiamo indossato gli stessi panni e mangiato dallo stesso piatto. C'eravamo quando non c'era nessuno e, facendoci forza e aiutandoci a vicenda, siamo andati avanti, fino a oggi. Siamo un ottimo incastro: ti aspetto.

Grazie a te che, anche se non presente fisicamente, sei parte della mia anima giorno per giorno: saremo sempre insieme, e so che mi guardi anche quando non ti penso. Sei stata la forza della mia vitalità, non solo un aiuto nei momenti difficili. Non voglio dilungarmi, ti voglio bene.

Y un agradecimiento especial es para ti, España: un año seguido en el que mi vida ha cambiado positivamente y muy rápido, conociendo gente y participando en experiencias que no creía que existieran. Me sorprende, y nunca olvidaré, lo que las personas de allí llevan dentro del corazón: son generosas y amables. Gracias, Dario, por tu hospitalidad y por compartir conmigo tus experiencias por todo el mundo; nunca había conocido a un hombre así: eres un superhéroe. Gracias, España, y gracias, N., por dejarme entrar en vuestro corazón.

Infine, grazie a tutte le persone che sono venute: la vostra presenza abbraccia calorosamente la celebrazione di questa giornata.

Grazie.