Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Biomedica

A.A. 2025/2026

Sessione di laurea Aprile 2026

Sviluppo di un'architettura Deep Learning dual-input per la classificazione automatica delle fasi del sonno da segnale PPG

Relatore:

Prof. Massimo Salvi

Candidato:

Caterina Calvi

Correlatore:

Ing. Silvia Seoni

Sommario

Il sonno è un processo fisiologico fondamentale e la corretta classificazione delle sue fasi è essenziale per la diagnosi dei disturbi del sonno. Il riferimento diagnostico per lo sleep staging è la polisonnografia (PSG), esame basato sulla registrazione multimodale di segnali fisiologici in ambiente clinico, in cui l'elettroencefalogramma (EEG) rappresenta il principale riferimento per la determinazione degli stadi del sonno. Tuttavia, la PSG risulta onerosa, richiede personale specializzato e l'impiego di numerosi sensori che possono alterare le naturali condizioni di riposo; inoltre, lo scoring manuale dei tracciati EEG è soggetto a variabilità inter- e intra-operatore.

Per superare tali limiti, la ricerca si è orientata verso l'utilizzo di segnali alternativi e l'automatizzazione del processo di classificazione. Il segnale fotopletismografico (PPG) rappresenta un'alternativa promettente, in quanto acquisibile con dispositivi indossabili a ridotto impatto per il paziente e potenzialmente utilizzabile in contesti extra-clinici. Parallelamente, i modelli di Deep Learning (DL) consentono di automatizzare lo sleep staging, migliorandone efficienza e riproducibilità.

In questo lavoro viene proposto un approccio basato su PPG e DL per la classificazione automatica delle fasi del sonno. L'analisi è stata condotta sul Cycling Alternating Pattern (CAP) Database, includendo 84 soggetti sani e affetti da diverse patologie del sonno. L'architettura del modello prevede due rami convoluzionali residui che elaborano il segnale PPG originale e una sua versione aumentata tramite rumore controllato. Le caratteristiche estratte vengono successivamente fuse mediante un meccanismo di cross-attention e pesatura adattiva, mentre la modellazione temporale è affidata a blocchi convoluzionali dilatati. La classificazione finale degli stadi del sonno è ottenuta tramite strati convoluzionali 1D.

Il modello raggiunge un'accuratezza complessiva pari a 70% e un macro F1-score pari a 67%, con performance bilanciate tra i diversi stadi del sonno. Tali risultati devono essere interpretati alla luce di alcune limitazioni, tra cui la dimensione relativamente contenuta del dataset e l'elevata variabilità inter-soggetto, accentuata dalla presenza di differenti condizioni patologiche. Nel complesso, questa tesi pone le basi per lo sviluppo di futuri modelli per lo sleep staging automatico basato sul PPG, con l'obiettivo di migliorarne robustezza e capacità di generalizzazione.

Indice

Elenco delle figure IV

Elenco delle tabelle VI

Glossario VII

1 Introduzione 1
- 1.1 Contesto del lavoro . . . 1
- 1.2 Obiettivi . . . 2
- 1.3 Struttura della tesi . . . 3

2 Background teorico e Stato dell'arte 4
- 2.1 Fisiologia del sonno e sua classificazione . . . 4
- 2.2 La fotopletismografia . . . 6
  - 2.2.1 Principi fisici . . . 6
  - 2.2.2 Modalità di acquisizione . . . 8
- 2.3 Relazione segnale PPG e fasi del sonno . . . 10
- 2.4 Metodi di Deep Learning per l'analisi di segnali temporali . . . 10
- 2.5 Stato dell'arte . . . 12

3 Materiali e metodi 15
- 3.1 Dataset . . . 15
- 3.2 Preparazione dei dati . . . 16
  - 3.2.1 Data cleaning, Filtraggio e Downsampling . . . 16
  - 3.2.2 Costruzione delle sequenze . . . 17
- 3.3 Architettura del modello . . . 20
- 3.4 Addestramento del modello . . . 24
  - 3.4.1 Configurazione generale dell'addestramento . . . 24
  - 3.4.2 Ottimizzazione degli iperparametri . . . 25
  - 3.4.3 Strategie di classificazione . . . 25
  - 3.4.4 Protocolli sperimentali . . . 27

3.5 Valutazione delle prestazioni 28
3.5.1 Metriche di classificazione 28
3.5.2 Metriche cliniche 29

4 Risultati 31
4.1 Selezione degli iperparametri 31
4.1.1 Grid search nella classificazione a quattro classi 31
4.2 Risultati con suddivisione Train–Validation–Test 32
4.2.1 Prestazioni di classificazione 32
4.2.2 Metriche cliniche 35
4.3 Risultati con validazione LOSO 37
4.3.1 Prestazioni di classificazione 37
4.3.2 Metriche cliniche 40
4.4 Analisi delle prestazioni per gruppo clinico 42
4.5 Discussione dei risultati 48
4.5.1 Confronto tra le strategie di classificazione 48
4.5.2 Analisi delle prestazioni per classe 48
4.5.3 Valutazione delle metriche cliniche 49
4.5.4 Influenza della condizione clinica 49
4.5.5 Confronto con lo stato dell'arte 50

5 Limiti e Sviluppi futuri 51
5.1 Limiti dello studio 51
5.2 Sviluppi futuri 52

6 Conclusioni 54

Bibliografia 56

Elenco delle figure

2.1 Profondità di penetrazione della luce nei tessuti biologici in funzione della lunghezza d'onda (tratta da [11]). 7
2.2 Componenti AC e DC del segnale PPG (tratta da [12]). La figura ha scopo illustrativo e non rappresenta le reali proporzioni tra le due componenti: in condizioni tipiche di misura, la componente pulsatile AC rappresenta solo una piccola frazione ($< 10\%$) del segnale complessivo. 8
2.3 Configurazioni di acquisizione del segnale PPG: modalità in trasmittanza a sinistra e in riflettanza a destra (tratta da [13]). 9

3.1 Pipeline di preparazione dei dati, dalla registrazione polisonnografica alla costruzione delle sequenze PPG utilizzate come input del modello. 19
3.2 Schema dell'architettura utilizzata. 22
3.3 Schema delle strategie di classificazione. 27

4.1 Confusion matrix aggregata sul Test Set per la classificazione a quattro classi. 34
4.2 Confusion matrix aggregata sul Test Set per la classificazione in cascata. 34
4.3 Bland–Altman per TST sul Test Set con classificazione a quattro classi. 35
4.4 Bland–Altman per SE sul Test Set con classificazione a quattro classi. 36
4.5 Bland–Altman per TST sul Test Set con classificazione a cascata. . 36
4.6 Bland–Altman per SE sul Test Set con classificazione a cascata. . . 37
4.7 Confusion matrix media normalizzata sui soggetti nella validazione LOSO per la classificazione a quattro classi. 39
4.8 Confusion matrix media normalizzata sui soggetti nella validazione LOSO per la classificazione in cascata. 39
4.9 Grafico di Bland–Altman per il TST nella validazione LOSO con classificazione a quattro classi. 40

4.10 Grafico di Bland–Altman per la SE nella validazione LOSO con classificazione a quattro classi. 41
4.11 Grafico di Bland–Altman per il TST nella validazione LOSO con classificazione in cascata. 41
4.12 Grafico di Bland–Altman per la SE nella validazione LOSO con classificazione in cascata. 42
4.13 Distribuzione del Macro F1-score per gruppo clinico nella validazione LOSO con classificazione in cascata. 43
4.14 Distribuzione dell'errore relativo del TST nei gruppi clinici con minore variabilità. 44
4.15 Distribuzione dell'errore relativo del TST nei gruppi clinici con maggiore variabilità. 44
4.16 Esempio di segnale PPG per il soggetto rbd4 nelle diverse fasi del sonno. 45
4.17 Esempio di segnale PPG per il soggetto sdb2 nelle diverse fasi del sonno. 46
4.18 Esempio di segnale PPG per il soggetto plm6 nelle diverse fasi del sonno. 47

Elenco delle tabelle

2.1 Corrispondenza tra la classificazione degli stadi del sonno secondo R&K e la revisione AASM. 5

3.1 Distribuzione dei soggetti per condizione clinica nel CAP Sleep Database 16
3.2 Distribuzione delle epoche per stadio del sonno dopo le procedure di pulizia, filtraggio e downsampling. 17
3.3 Distribuzione dei soggetti in funzione della lunghezza delle registrazioni. 18
3.4 Distribuzione delle epoche per stadio del sonno dopo il troncamento a 1200 epoche. 18
3.5 Iperparametri esplorati durante la procedura di grid search 25

4.1 Migliori configurazioni ottenute durante la procedura di grid search, ordinate in base al coefficiente κ di Cohen sul Validation Set. 31
4.2 Metriche globali ottenute sul Validation Set. 32
4.3 Metriche globali ottenute sul Test Set. 32
4.4 Metriche per classe ottenute sul Test Set con la classificazione a quattro classi. 33
4.5 Metriche per classe ottenute sul Test Set con la classificazione a cascata. 33
4.6 Statistiche riassuntive delle metriche cliniche su Test Set. 35
4.7 Metriche globali ottenute nella configurazione LOSO. 37
4.8 Metriche per classe ottenute sui diversi fold nella configurazione LOSO per la classificazione a quattro classi. 38
4.9 Metriche per classe ottenute sui diversi fold nella configurazione LOSO per la classificazione in cascata. 38
4.10 Statistiche riassuntive delle metriche cliniche sui soggetti nella validazione LOSO. 40
4.11 Confronto tra i principali lavori basati su PPG per lo sleep staging e il modello proposto. 50

Glossario

IA
: Intelligenza Artificiale

ML
: Machine Learning

DL
: Deep Learning

PSG
: Polisonnografia

EEG
: Segnale elettroencefalografico

PPG
: Segnale fotopletismografico

R&K
: Rechtschaffen e Kales

AASM
: American Academy of Sleep Medicine

W
: Wake

NREM
: Non-Rapid Eye Movement

REM

Rapid Eye Movement

EOG

Segnale elettrooculografico

EMG

Segnale elettromiografico

ECG

Segnale elettrocardiografico

DC

Direct Current

AC

Alternating Current

LED

Light Emitting Diode

SNA

Sistema Nervoso Autonomo

LS

Light Sleep

DS

Deep Sleep

EDF

European Data Format

TST

Total Sleep Time

SE

Sleep Efficiency

CNN

Convolutional Neural Network

RNN

Recurrent Neural Network

LightGBM

Light Gradient Boosting Machine

XGBoost

eXtreme Gradient Boosting

CatBoost

Categorical Boosting

XAI

Explainable Artificial Intelligence

SVM

Support Vector Machine

RF

Random Forest

BiLSTM

Bidirectional Long Short-Term Memory

Capitolo 1

Introduzione

1.1 Contesto del lavoro

Negli ultimi decenni, l'**Intelligenza Artificiale (IA)** ha assunto un ruolo sempre più rilevante in ambito sanitario, affermandosi come un promettente strumento di supporto in numerosi contesti clinici e di ricerca. Studi recenti [1] mostrano una crescita esponenziale delle pubblicazioni scientifiche dedicate all'integrazione dell'IA in medicina, evidenziando un interesse crescente verso applicazioni che spaziano dalla diagnosi assistita all'analisi di segnali biomedici e al monitoraggio dei pazienti. In questo scenario, le tecniche di **Machine Learning (ML)** e **Deep Learning (DL)** sono oggetto di intensa attività di ricerca, con l'obiettivo di sviluppare modelli in grado di analizzare grandi quantità di dati fisiologici, migliorare l'efficienza dei processi diagnostici e ridurre il carico operativo dei professionisti sanitari.

I modelli di ML sono ampiamente studiati e applicati in ambito sperimentale; tuttavia, la loro efficacia risulta spesso limitata dalla cosiddetta "*curse of dimensionality*", ovvero dalla difficoltà nel gestire dati caratterizzati da un elevato numero di variabili. Pertanto, questi algoritmi richiedono generalmente accurate fasi di *feature engineering* e *feature selection*, finalizzate a selezionare manualmente le caratteristiche più informative dei dati, con conseguente riduzione della dimensionalità. Tale approccio rende i metodi di ML più semplici e controllabili, ma ne può limitare la capacità di operare efficacemente su dati complessi se non pre-elaborati in modo opportuno.

Il DL rappresenta un'evoluzione significativa, in quanto si basa su architetture capaci di estrarre **automaticamente** informazioni dai dati grezzi, senza la necessità di definire esplicitamente le feature da utilizzare. Grazie a questa caratteristica, i modelli di DL si sono dimostrati particolarmente efficaci nello studio di segnali biomedici complessi, spesso raggiungendo prestazioni superiori rispetto ai tradizionali metodi di ML e suscitando così un crescente interesse verso il loro impiego

in ambito di ricerca medica. In particolare, questi approcci risultano promettenti nell'analisi dei segnali fisiologici registrati durante il sonno.

In questo ambito, tra le applicazioni più rilevanti rientra la **classificazione delle fasi del sonno**. Tale processo riveste un ruolo centrale in ambito clinico, poiché la struttura del sonno influenza in modo significativo processi cognitivi, metabolici e cardiovascolari ed è strettamente legata alla diagnosi di numerosi disturbi. Attualmente, la valutazione clinica del sonno si basa prevalentemente sulla **polisonnografia (PSG)**, un esame condotto in ambiente ospedaliero che prevede l'acquisizione simultanea di diversi segnali fisiologici, tra cui l'**elettroencefalogramma (EEG)**, che costituisce il principale riferimento per la determinazione delle fasi del sonno. Sebbene sia considerata il riferimento diagnostico per lo sleep staging, la PSG è una procedura complessa che comporta un processo di analisi manuale lungo e oneroso per il personale clinico.

Per questo motivo, negli ultimi anni è cresciuto l'interesse verso lo sviluppo di sistemi in grado di automatizzare lo sleep staging. In questo contesto, diversi modelli di DL sono stati proposti per la classificazione automatica delle fasi del sonno [2] a partire dai segnali elettrofisiologici acquisiti durante la PSG, in particolare dall'EEG. Tuttavia, l'acquisizione dell'EEG e l'esecuzione della PSG rimangono procedure invasive che richiedono condizioni di registrazione controllate. Inoltre, la presenza di numerosi sensori applicati al paziente, necessari per la raccolta dei segnali, può causare disagi e rendere le condizioni di riposo non sempre rappresentative del sonno abituale del soggetto.

Di conseguenza, la ricerca si è orientata anche verso l'esplorazione di **segnali fisiologici alternativi** che possano essere acquisiti mediante dispositivi meno invasivi. Tra questi, il **segnale fotopletismografico (PPG)** rappresenta una soluzione promettente, poiché può essere registrato tramite sensori integrati in dispositivi indossabili a ridotto impatto per il paziente, come braccialetti o anelli.

In tale prospettiva, l'applicazione di modelli di DL a segnali facilmente acquisibili come il PPG rappresenta una linea di ricerca interessante nell'ambito del monitoraggio del sonno. In questo contesto si colloca il presente lavoro di tesi, che verrà descritto nei capitoli successivi.

1.2 Obiettivi

L'obiettivo principale di questa tesi è lo sviluppo e la valutazione di un'architettura DL per la classificazione automatica delle fasi del sonno a partire dal segnale PPG. L'utilizzo di tale segnale è motivato dalla possibilità di acquisirlo mediante dispositivi indossabili a ridotto impatto per il paziente, rendendo il monitoraggio del sonno meno invasivo e più adatto a contesti extra-clinici rispetto alle metodologie tradizionali basate su PSG. L'utilizzo di tecniche di DL ha invece l'obiettivo di

automatizzare il processo di classificazione degli stadi del sonno, rendendo l'analisi più efficiente e potenzialmente supportando l'attività degli specialisti sanitari.

Dal punto di vista metodologico, il lavoro si concentra sulla progettazione e ottimizzazione di un modello di DL, attraverso l'analisi dell'influenza di diversi iperparametri e la valutazione delle prestazioni mediante differenti strategie di validazione. Questo approccio consente di analizzare il comportamento del modello in diverse condizioni sperimentali e di valutarne la capacità di generalizzazione su soggetti non osservati durante l'addestramento.
Riassumendo, i principali contributi di questa tesi sono:

- Progettazione e implementazione di un modello di DL per lo sleep staging automatico basato sull'elaborazione del segnale PPG;
- Analisi dell'influenza di diversi iperparametri sul comportamento e sulle prestazioni del modello;
- Valutazione delle prestazioni del modello mediante differenti strategie di validazione, al fine di analizzarne la capacità di generalizzazione.

1.3 Struttura della tesi

Questa tesi è strutturata nei seguenti capitoli:

- **Capitolo 1 - Introduzione**: presentazione del contesto applicativo e scientifico della tesi e definizione degli obiettivi e dei principali contributi del lavoro.
- **Capitolo 2 - Background teorico e Stato dell'arte**: introduzione alla struttura del sonno e alla fotopletismografia, analisi della relazione tra segnale PPG e fasi del sonno, presentazione delle principali nozioni teoriche relative ai modelli di DL e rassegna della letteratura sullo sleep staging.
- **Capitolo 3 - Materiali e Metodi**: descrizione del dataset utilizzato, delle procedure di pre-elaborazione dei dati, dell'architettura del modello di DL adottato e delle strategie di addestramento e validazione impiegate.
- **Capitolo 4 - Risultati**: presentazione e valutazione dei risultati ottenuti nelle diverse configurazioni sperimentali.
- **Capitolo 5 - Limiti e Sviluppi futuri**: discussione dei principali limiti del lavoro svolto e individuazione di possibili sviluppi futuri.
- **Capitolo 6 - Conclusioni**: sintesi dei risultati principali e considerazioni finali sui contributi della tesi.

Capitolo 2

Background teorico e Stato dell'arte

2.1 Fisiologia del sonno e sua classificazione

Il sonno è un processo fisiologico fondamentale per il mantenimento di un'ottimale salute fisica e mentale [3]. Sia la quantità sia la qualità del sonno influenzano in modo significativo il corretto funzionamento dell'organismo: un sonno sub-ottimale è infatti associato ad alterazioni di numerosi processi biologici, tra cui la salute cardiovascolare, la funzione immunitaria e la regolazione ormonale. Durante il sonno avvengono inoltre meccanismi essenziali per il recupero energetico e per il corretto funzionamento dei sistemi cognitivi e dei processi di memoria, rendendo la struttura e la qualità del sonno elementi di grande rilevanza clinica. Alterazioni persistenti dei pattern del sonno risultano infatti frequentemente correlate allo sviluppo di diverse patologie, tra cui disturbi neurologici [4] e cardiovascolari [5].

I disturbi del sonno [6] rappresentano una delle principali cause di compromissione della normale organizzazione del riposo notturno, configurandosi come condizioni cliniche in grado di incidere in modo significativo sui processi fisiologici ad esso associati. Per questo motivo, l'analisi della struttura del sonno e la corretta classificazione delle sue fasi rivestono un ruolo centrale nella pratica clinica, poiché consentono di supportare la diagnosi e il trattamento di tali condizioni patologiche.

La classificazione delle fasi del sonno si basa sull'analisi dei segnali fisiologici acquisiti mediante polisonnografia (PSG), esame che consente la registrazione simultanea di molteplici parametri, tra cui *attività elettroencefalografica* (EEG), *elettrooculografica* (EOG), *elettromiografica* (EMG), *elettrocardiografica* (ECG), il *flusso respiratorio* e altri segnali fisiologici. L'interpretazione combinata di tali segnali viene effettuata da personale esperto, che si basa principalmente sull'analisi

del tracciato EEG per la determinazione degli stadi del sonno, seguendo criteri standardizzati definiti da linee guida internazionali.

Nel corso del tempo sono stati proposti diversi sistemi di classificazione per uniformare il processo di scoring. Il primo sistema standardizzato ampiamente adottato è stato quello introdotto da **Rechtschaffen e Kales (R&K)** [7]. Successivamente, l'**American Academy of Sleep Medicine (AASM)** ha proposto una revisione di questo sistema, introducendo una classificazione aggiornata [8] che rappresenta oggi il riferimento maggiormente adottato nella pratica clinica e nella ricerca scientifica. Secondo le linee guida AASM, le fasi del sonno comprendono:

- **Veglia (Wake, W)**: stato di coscienza che precede l'addormentamento e si manifesta al risveglio.
- Sonno **non-REM (Non-Rapid Eye Movement, NREM)**, suddiviso in:
  - **N1**: fase di transizione tra veglia e sonno
  - **N2**: fase di sonno leggero
  - **N3**: fase di sonno profondo o a onde lente
- Sonno **REM (Rapid Eye Movement)**: fase associata a intensa attività cerebrale, atonia muscolare e movimenti oculari rapidi.

Nella classificazione originaria di R&K, il sonno NREM era articolato in quattro stadi distinti (Stage 1-4). In particolare, la fase di sonno profondo era suddivisa in due stadi sulla base della percentuale di attività a onde lente presente nel tracciato EEG (Stage 3 e Stage 4). La revisione proposta dall'AASM ha accorpato tali stadi in un'unica fase, denominata N3. La corrispondenza tra i due sistemi di classificazione è riportata nella Tabella 2.1.

Tabella 2.1: Corrispondenza tra la classificazione degli stadi del sonno secondo R&K e la revisione AASM.

AASM	**R&K**
W	W
N1	Stage 1
N2	Stage 2
N3	Stage 3 + Stage 4
REM	REM

In ambito clinico, la registrazione polisonnografica viene suddivisa in segmenti temporali di durata standard pari a 30 secondi, denominati *epoche*. A ciascuna epoca viene assegnato uno stadio del sonno sulla base dei criteri AASM, attraverso un processo di **scoring manuale**. Sebbene la PSG rappresenti il riferimento diagnostico attuale, essa risulta onerosa in termini di tempo e, nonostante la presenza di linee guida standardizzate, è soggetta a **variabilità inter- ed intra-operatore** [9]. Inoltre, l'acquisizione polisonnografica richiede l'impiego di numerosi sensori e apparecchiature, con possibili ripercussioni sulle condizioni naturali di riposo del paziente. Alla luce di tali limitazioni, la ricerca si è orientata lungo due direttrici complementari:

- da un lato, l'integrazione di metodi volti ad automatizzare il processo di scoring, migliorandone riproducibilità ed efficienza;
- dall'altro, l'esplorazione di segnali fisiologici alternativi, acquisibili mediante dispositivi meno invasivi e più adatti a contesti extra-clinici.

Tra le soluzioni proposte in quest'ultima direzione, un ruolo di particolare interesse è rivestito dal segnale PPG, ottenuto mediante **fotopletismografia**, tecnica ottica non invasiva già ampiamente impiegata per il monitoraggio cardiovascolare e recentemente esplorata anche nell'ambito dell'analisi del sonno.

2.2 La fotopletismografia

2.2.1 Principi fisici

Il principio fisico alla base della fotopletismografia risiede nei fenomeni di assorbimento e diffusione della luce all'interno dei tessuti biologici [10]. La radiazione luminosa impiegata in questa tecnica opera tipicamente nel range del **rosso (∼660 nm)** e dell'**infrarosso (∼940 nm)**, lunghezze d'onda selezionate per la loro capacità di penetrare in profondità nei tessuti e per la loro diversa interazione con l'**emoglobina**, principale cromoforo responsabile dell'assorbimento ottico nel sangue. L'emoglobina presenta infatti diversi coefficienti di assorbimento, a seconda del suo stato di ossigenazione, distinguendosi tra forma ossigenata (HbO_2) e deossigenata (Hb).

In generale, all'aumentare della lunghezza d'onda aumenta la profondità di penetrazione della luce: radiazioni nel rosso e nell'infrarosso possono attraversare l'ipoderma e raggiungere arteriole e arterie più profonde. Lunghezze d'onda inferiori, come la **luce verde (∼565 nm)**, presentano invece una penetrazione più superficiale e interagiscono principalmente con il microcircolo arterioso più prossimo alla superficie cutanea, come illustrato in Figura 2.1.

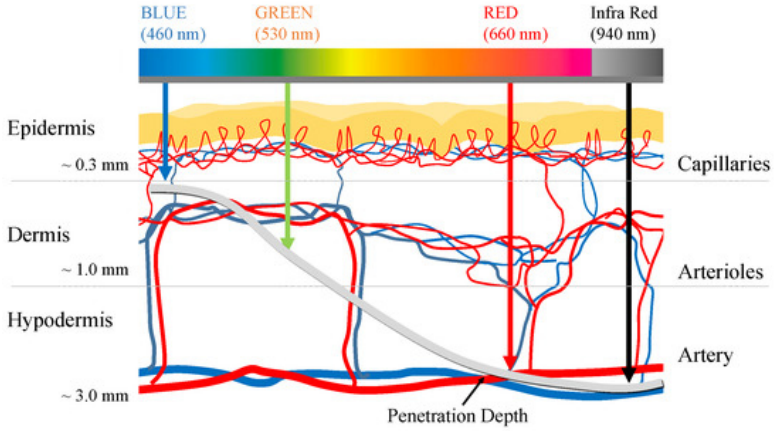

Figura 2.1: Profondità di penetrazione della luce nei tessuti biologici in funzione della lunghezza d'onda (tratta da [11]).

Quando la luce incide sul tessuto, i fotoni non percorrono un cammino rettilineo, ma subiscono fenomeni di **scattering** che ne modificano la traiettoria, determinando un cammino ottico effettivo maggiore rispetto allo spessore geometrico del tessuto attraversato. Durante la propagazione della radiazione, una parte dell'energia luminosa viene assorbita dai cromofori (principalmente emoglobina, ma anche acqua e altri costituenti tissutali), mentre un'altra parte viene diffusa a causa dei fenomeni di scattering. Solo una frazione della radiazione incidente contribuisce effettivamente al segnale rilevato dal sensore.

L'attenuazione complessiva della radiazione luminosa può essere descritta, in prima approssimazione, mediante la **legge di Lambert-Beer modificata**:

$$A = -\ln\left(\frac{I}{I_0}\right) = \varepsilon\, c\, L_{\text{eff}} \tag{2.1}$$

dove I_0 rappresenta l'intensità della luce incidente, I quella rilevata, ε il coefficiente di estinzione del cromoforo, c la sua concentrazione e L_{eff} il cammino ottico effettivo della luce all'interno del tessuto, che tiene conto dell'allungamento del percorso ottico dovuto ai fenomeni di scattering.

Poiché il volume ematico nel microcircolo arterioso varia in modo sincronizzato con il ciclo cardiaco, anche l'intensità della luce rilevata dal sensore presenta variazioni periodiche correlate alla pulsazione arteriosa. La fotopletismografia rileva quindi variazioni dell'intensità della radiazione luminosa trasmessa o riflessa dal tessuto, indirettamente associate alle variazioni del volume sanguigno.

Queste oscillazioni danno origine al segnale PPG, che può essere descritto come la sovrapposizione di una **componente continua (DC)**, associata alla struttura

tissutale e alla componente non pulsatile del sangue, e di una **componente pulsatile (AC)**, legata alle variazioni cicliche del volume sanguigno arterioso prodotte dalla propagazione dell'onda di pressione generata dal battito cardiaco (Figura 2.2). La componente AC contiene informazioni fisiologiche relative alla frequenza cardiaca, alla dinamica emodinamica e alla variabilità del sistema cardiovascolare, mentre la componente DC è influenzata principalmente dalle caratteristiche anatomiche del sito di misura e dalle condizioni di perfusione periferica.

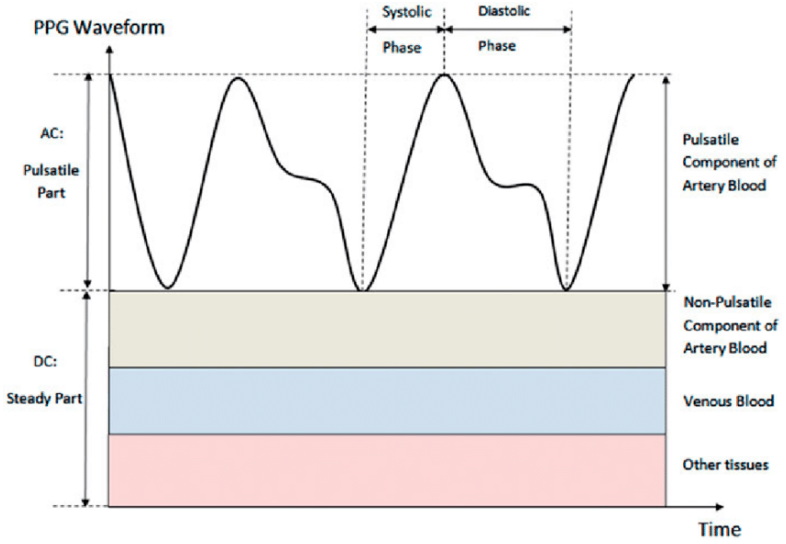

Figura 2.2: Componenti AC e DC del segnale PPG (tratta da [12]). La figura ha scopo illustrativo e non rappresenta le reali proporzioni tra le due componenti: in condizioni tipiche di misura, la componente pulsatile AC rappresenta solo una piccola frazione ($< 10\%$) del segnale complessivo.

2.2.2 Modalità di acquisizione

Alla luce dei principi fisici descritti, la misura del segnale PPG richiede l'impiego di dispositivi optoelettronici in grado di emettere radiazione luminosa e rilevarne le variazioni dopo l'interazione con il tessuto biologico. Il sistema di acquisizione è costituito da una **sorgente luminosa**, generalmente un LED, e da un **fotodiodo**, che rileva la radiazione luminosa emergente dal tessuto. I due elementi sono integrati in un unico sensore e disposti secondo configurazioni geometriche differenti, in funzione del sito anatomico e dell'applicazione, come illustrato in Figura 2.3.

Nei dispositivi tradizionalmente usati in ambito sanitario, il sensore è spesso realizzato sotto forma di clip e applicato al dito della mano, al dito del piede o al lobo auricolare. In tali configurazioni, LED e rivelatore sono posizionati su lati opposti del distretto anatomico e il segnale è ottenuto dalla luce che attraversa il

tessuto (**modalità in trasmittanza**). Questa soluzione garantisce generalmente una buona qualità del segnale ed è meno sensibile agli artefatti da movimento, risultando particolarmente adatta in ambito clinico.

Nelle applicazioni più recenti, in particolare nei dispositivi indossabili, LED e fotodiodo sono invece collocati sullo stesso piano, a breve distanza tra loro (tipicamente pochi millimetri), e il segnale viene ricavato dalla radiazione riflessa dal tessuto (**modalità in riflettanza**). Tale configurazione consente una maggiore flessibilità di posizionamento, rendendo possibile l'acquisizione del segnale in sedi anatomiche come il polso o altre regioni periferiche, più compatibili con l'utilizzo continuativo e in contesti extra-clinici.

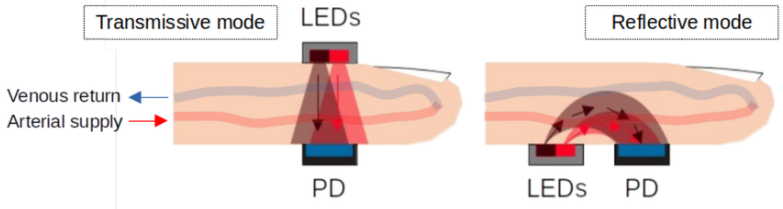

Figura 2.3: Configurazioni di acquisizione del segnale PPG: modalità in trasmittanza a sinistra e in riflettanza a destra (tratta da [13]).

Nelle configurazioni in trasmittanza si impiegano generalmente lunghezze d'onda nel rosso o nell'infrarosso, caratterizzate da maggiore profondità di penetrazione. Nelle configurazioni in riflettanza, specialmente nei dispositivi wearable, può essere utilizzata anche luce verde, la cui interazione prevalente con il microcircolo superficiale consente una buona rilevazione della pulsazione arteriosa.

Dal punto di vista strumentale, il segnale rilevato dal fotodiodo viene successivamente amplificato, filtrato e digitalizzato. Il segnale PPG presenta tipicamente una banda di frequenze compresa tra circa 0.5 e 5–10 Hz, all'interno della quale sono contenute le principali informazioni fisiologiche associate alla pulsazione cardiaca e alle variazioni emodinamiche. Per questo motivo, vengono spesso impiegati filtri passa-banda progettati per preservare tali componenti e ridurre il rumore a bassa e alta frequenza. La qualità dell'acquisizione dipende non solo dalle proprietà ottiche del tessuto, ma anche dalla stabilità meccanica del sensore e dalle condizioni operative. In particolare, la forza di contatto tra sensore e pelle rappresenta un parametro critico: una pressione insufficiente può determinare un accoppiamento instabile e un aumento del rumore, mentre una pressione eccessiva può comprimere i vasi sanguigni superficiali, riducendo l'ampiezza del segnale pulsatile [13].

2.3 Relazione tra segnale PPG e fasi del sonno

Il segnale PPG acquisito mediante le configurazioni descritte nella sezione precedente rappresenta una sorgente informativa interessante per la classificazione delle fasi del sonno, in quanto riflette diverse variazioni fisiologiche legate all'attività cardiovascolare e respiratoria. Durante il sonno, tali processi sono modulati principalmente dall'attività del **sistema nervoso autonomo (SNA)**, costituito dalle componenti **simpatico** e **parasimpatico**, che regolano in modo complementare l'attività dell'organismo. In condizioni di predominanza simpatica si osservano generalmente un aumento della frequenza cardiaca e una maggiore variabilità dei parametri fisiologici, mentre la predominanza parasimpatica è associata a una riduzione della frequenza cardiaca e a una maggiore stabilità del sistema cardiovascolare. Poiché il segnale PPG riflette le variazioni del volume sanguigno periferico sincronizzate con il ciclo cardiaco, esso risulta indirettamente correlato ai cambiamenti fisiologici che caratterizzano la struttura del sonno.

I diversi stadi del sonno presentano caratteristiche fisiologiche distinte. Durante la veglia (W) si osserva generalmente una maggiore variabilità dei parametri cardiovascolari e respiratori, associata a una più elevata attività simpatica. Con l'ingresso nelle fasi di sonno leggero N1 e N2 si verifica una progressiva riduzione della frequenza cardiaca e una maggiore regolarità dei parametri fisiologici. Nel sonno profondo N3 la frequenza cardiaca raggiunge tipicamente i valori più bassi e il sistema cardiovascolare mostra una marcata stabilità, dovuta alla predominanza dell'attività parasimpatica. Il sonno REM, al contrario, è caratterizzato da una maggiore irregolarità, con oscillazioni più marcate della frequenza cardiaca e della respirazione.

Queste variazioni fisiologiche si riflettono nella morfologia e nella dinamica temporale del segnale PPG. In particolare, caratteristiche come la variabilità dell'intervallo tra battiti consecutivi, l'ampiezza dell'onda pulsatile e la modulazione respiratoria contribuiscono a rendere il segnale PPG informativo rispetto alla struttura del sonno. L'estrazione e l'interpretazione di queste informazioni rappresentano quindi un passaggio fondamentale per lo sviluppo di metodi automatici di classificazione delle fasi del sonno basati sul PPG.

2.4 Metodi di Deep Learning per l'analisi di segnali temporali

I modelli di DL rappresentano una classe di algoritmi di apprendimento automatico basati su **reti neurali artificiali profonde**, in grado di apprendere automaticamente rappresentazioni informative a partire dai dati. A differenza dei metodi di ML, tali algoritmi consentono di elaborare direttamente i dati grezzi, riducendo la

necessità di procedure manuali di estrazione delle caratteristiche. Questa proprietà li rende particolarmente adatti all'analisi di segnali complessi, come i segnali fisiologici, nei quali le informazioni rilevanti sono spesso distribuite sia nella morfologia del segnale che nelle relazioni temporali tra campioni successivi.

L'addestramento di un modello di DL consiste nell'aggiornamento iterativo dei parametri della rete neurale al fine di minimizzare una funzione di perdita che misura la differenza tra le predizioni del modello e i valori reali. Tale processo viene generalmente effettuato mediante algoritmi di ottimizzazione basati sulla discesa del gradiente, che consentono di aggiornare i pesi del modello in modo progressivo.

Durante l'addestramento, i dati vengono suddivisi in sottoinsiemi di dimensione ridotta, denominati **batch**, e il processo viene ripetuto per più cicli completi sul dataset di training, denominati **epoche**. La scelta dei parametri che regolano il processo di addestramento, detti **iperparametri**, influisce in modo significativo sulle prestazioni del modello. Tra gli iperparametri più rilevanti rientrano il **learning rate**, che controlla l'entità degli aggiornamenti dei pesi durante l'ottimizzazione, la **dimensione del batch**, che determina il numero di campioni utilizzati per ciascun aggiornamento dei parametri, e il **numero di epoche**, che stabilisce quante volte il modello viene addestrato sull'intero training set. Una scelta appropriata di tali parametri è fondamentale per garantire una buona convergenza dell'algoritmo e una adeguata capacità di generalizzazione.

Nel caso di segnali fisiologici, i dati sono tipicamente organizzati come sequenze temporali. Per l'analisi di tali segnali vengono impiegate architetture neurali specifiche progettate per estrarre automaticamente informazioni sia dalla struttura locale del segnale sia dalle relazioni temporali tra campioni successivi. Tra le architetture più utilizzate rientrano le **reti neurali convoluzionali (CNN)**, che permettono di estrarre automaticamente caratteristiche locali mediante operazioni di convoluzione applicate lungo l'asse temporale del segnale. Le CNN risultano particolarmente efficaci nell'identificazione di pattern caratteristici nella forma d'onda dei segnali fisiologici. Per modellare le dipendenze temporali del segnale possono inoltre essere impiegate architetture specifiche per l'elaborazione di sequenze, tra cui le **reti neurali ricorrenti (RNN)**. Tali modelli consentono di descrivere l'evoluzione del segnale nel tempo e risultano particolarmente adatti all'analisi di segnali fisiologici caratterizzati da dinamiche temporali complesse.

Grazie alla loro capacità di apprendere automaticamente caratteristiche informative direttamente dai dati grezzi, i modelli di DL rappresentano uno strumento promettente per lo sviluppo di sistemi automatici di classificazione delle fasi del sonno basati su segnali fisiologici. Nel paragrafo successivo verrà fornita una panoramica dello stato dell'arte relativo allo sleep staging automatico basato su segnale PPG.

2.5 Stato dell'arte

Negli ultimi anni lo sviluppo di metodi automatici per la classificazione delle fasi del sonno ha ricevuto crescente attenzione nella letteratura scientifica. Tradizionalmente, molti sistemi sono stati sviluppati utilizzando il segnale EEG, che rappresenta il riferimento principale per la determinazione degli stadi del sonno e continua tuttora a essere ampiamente utilizzato negli studi di sleep staging automatico [14, 15, 16, 17]. Parallelamente, diversi lavori hanno iniziato ad esplorare l'utilizzo di segnali fisiologici alternativi, tra cui il segnale PPG, che ha ricevuto particolare attenzione grazie alla possibilità di essere acquisito mediante dispositivi indossabili e sensori a ridotto impatto per il paziente. In questo contesto, sono stati proposti numerosi approcci per la classificazione automatica degli stadi del sonno basati sul segnale PPG, che impiegano sia tecniche di ML sia modelli di DL.

Tra gli approcci basati su ML, Zhao e Sun [18] hanno proposto un metodo basato sull'estrazione di 21 caratteristiche dal segnale PPG nei domini temporale, frequenziale e non lineare, calcolate su singole epoche di 30 secondi e utilizzate come input per un classificatore **Light Gradient Boosting Machine (LightGBM)**, che assegna uno stadio del sonno a ciascuna epoca. Il modello è stato valutato su 27 soggetti del dataset CAP [19], comprendenti sia soggetti sani sia pazienti con disturbi del sonno, ottenendo un'accuratezza dell'86.3% nella classificazione a 3 stadi, del 77.1% nella classificazione a 4 stadi e del 72.2% nella classificazione a 5 stadi.

Ferdous et al. [20] hanno invece confrontato diversi classificatori di ML, tra cui Random Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost) e Categorical Boosting (CatBoost), utilizzando come input 42 feature estratte dal segnale PPG. Anche in questo caso ciascun vettore di feature rappresenta una singola epoca di segnale, alla quale viene assegnato lo stadio del sonno corrispondente. Il metodo è stato valutato su un dataset composto da 10 soggetti con disturbi respiratori del sonno e il modello con le migliori prestazioni nella classificazione a 4 stadi risulta **XGBoost**, che raggiunge un'accuratezza complessiva del 75.29%.

Smarandache et al. [21] hanno proposto un approccio basato su **RF** che combina caratteristiche statistiche, temporali e non lineari tramite una procedura di **feature fusion**. Come nei lavori precedenti, il segnale PPG viene segmentato in epoche di 30 secondi, dalle quali vengono estratte le caratteristiche utilizzate per classificare lo stadio del sonno di ciascuna epoca. Il modello è stato valutato su un dataset composto da 10 soggetti con disturbi respiratori del sonno, ottenendo un'accuratezza dell'82.56% nella classificazione a 2 stadi, del 77.79% nella classificazione a 3 stadi e del 69.20% nella classificazione a 4 stadi.

Oltre agli approcci basati su ML, numerosi studi hanno iniziato ad applicare modelli di DL alla classificazione automatica degli stadi del sonno, inizialmente

operando su singole epoche di segnale PPG analogamente ai metodi basati su feature e, successivamente, introducendo architetture in grado di elaborare sequenze di epoche per modellare le dipendenze temporali del sonno.

Tra gli approcci basati su DL, Huttunen et al. [22] hanno proposto un modello basato su una combinazione di **reti convoluzionali (CNN) e ricorrenti (RNN)** applicato direttamente al segnale PPG grezzo segmentato in epoche di 30 secondi, elaborate individualmente per lo sleep staging. L'algoritmo è stato addestrato su oltre 3000 registrazioni PSG di pazienti con sospetta apnea ostruttiva del sonno, raggiungendo un'accuratezza dell'83.3% nella classificazione a 3 stadi, del 74.1% nella classificazione a 4 stadi e del 68.7% nella classificazione a 5 stadi.

In [23] viene proposta un'architettura basata su una combinazione di **CNN e reti Bidirectional Long Short-Term Memory (BiLSTM)**, progettata per catturare sia caratteristiche locali del segnale sia dipendenze temporali tra campioni successivi. Il modello opera su singole epoche di 30 secondi ed è stato progettato per utilizzare diverse combinazioni di segnali cardiorespiratori come input. Nella configurazione che utilizza esclusivamente il segnale PPG, valutata su un dataset polisonnografico di 123 registrazioni, il metodo raggiunge un'accuratezza del 91% nella classificazione a 2 stadi, dell'84% nella classificazione a 3 stadi, del 76% nella classificazione a 4 stadi e del 74% nella classificazione a 5 stadi.

Nam et al. [24] propongono **InsightSleepNet**, un modello di DL progettato per la classificazione degli stadi del sonno a partire da segnali PPG continui. L'architettura opera su sequenze di epoche di 30 secondi estratte dal segnale PPG, permettendo di sfruttare il contesto temporale tra epoche successive. Il metodo è stato valutato su diversi dataset polisonnografici pubblici. In particolare, nella classificazione a 4 stadi sul dataset CAP, considerando un sottoinsieme di 24 soggetti, il modello raggiunge un'accuratezza dell'80.6%. Gli autori introducono inoltre una strategia di stima dell'incertezza basata su energy score per identificare e scartare le predizioni con bassa confidenza.

Un'evoluzione significativa degli approcci basati su sequenze temporali è rappresentata dal modello **SleepPPG-Net** proposto da Kotzen et al. [25]. L'architettura combina una rete convoluzionale residua per l'estrazione automatica delle caratteristiche dal segnale PPG con una Temporal Convolutional Network (TCN) per modellare le dipendenze temporali tra epoche successive. Il modello utilizza come input sequenze lunghe di epoche di 30 secondi di segnale PPG pre-processato, consentendo di sfruttare il contesto temporale del sonno su intervalli di tempo estesi. Addestrato e valutato su diversi dataset pubblici contenenti oltre 2300 soggetti, il metodo raggiunge un'accuratezza dell'84% nella classificazione a 4 stadi del sonno.

A partire da questa architettura, diversi lavori successivi hanno proposto estensioni e miglioramenti del modello. Constantin et al. [26] ne hanno valutato

l'applicabilità a segnali PPG acquisiti tramite dispositivi indossabili e hanno analizzato l'integrazione di informazioni di attività motoria (actigraphy) come input aggiuntivo. Utilizzando esclusivamente il segnale PPG acquisito da dispositivo da polso, il modello raggiunge un'accuratezza del 78.1%, che aumenta fino al 78.3% con l'integrazione dell'actigraphy.

Wang et al. [27] propongono ulteriori miglioramenti introducendo un'architettura dual-stream con meccanismo di cross-attention, che consente di combinare il segnale PPG con modalità ausiliarie derivate dal PPG stesso, come PPG aumentato tramite tecniche di data augmentation o segnali ECG sintetici. Valutato sul dataset MESA [28], il modello single-stream basato solo su PPG raggiunge un'accuratezza del 78.3%, mentre l'architettura dual-stream con PPG e PPG aumentato migliora le prestazioni fino all'83.3%.

Il confronto diretto tra i risultati riportati in letteratura risulta spesso complesso a causa delle diverse strategie di validazione e delle caratteristiche dei dataset utilizzati. In particolare, la separazione dei dati a livello di soggetto rappresenta un aspetto fondamentale per valutare la capacità di generalizzazione dei modelli su individui non osservati durante l'addestramento. In questo contesto si inserisce il presente lavoro di tesi, che propone lo sviluppo e la valutazione di un modello di DL per la classificazione automatica delle fasi del sonno a partire dal segnale PPG.

Capitolo 3

Materiali e metodi

3.1 Dataset

Il dataset utilizzato in questo lavoro è il **CAP Sleep Database** [19], disponibile sulla piattaforma PhysioNet [29]. Il database comprende 108 registrazioni polisonnografiche notturne acquisite presso il Sleep Disorders Center dell'Ospedale Maggiore di Parma. Ciascuna registrazione include diversi segnali fisiologici, tra cui segnali EEG, EOG, EMG, ECG, flusso respiratorio e saturazione di ossigeno, oltre alle annotazioni degli stadi del sonno.

Il termine *Cyclic Alternating Pattern* (CAP) indica un'attività periodica osservata nel tracciato EEG durante il sonno NREM, associata alla microstruttura del sonno e a diversi disturbi del sonno. Sebbene il database includa anche annotazioni CAP, nel presente lavoro sono state considerate esclusivamente le annotazioni relative agli stadi del sonno.

Delle 108 registrazioni disponibili, solo 85 contengono il segnale PPG. Lo scoring degli stadi del sonno è stato effettuato da neurologi esperti presso lo Sleep Center secondo lo standard di R&K, classificando epoche di durata pari a 30 s nelle classi W, Stage 1, Stage 2, Stage 3, Stage 4 e REM.

I soggetti per i quali risulta disponibile il segnale PPG comprendono 34 donne e 51 uomini, con un'età compresa tra 14 e 82 anni. La durata delle registrazioni varia tra 5.5 e 14 ore. Il dataset include sia soggetti sani sia pazienti affetti da diversi disturbi del sonno. Ciascun soggetto è identificato mediante una sigla che indica la condizione clinica di appartenenza e un numero progressivo (ad esempio `nfle5`). La distribuzione dei soggetti nelle diverse categorie è riportata nella Tabella 3.1. Le sigle utilizzate derivano dalla denominazione inglese delle patologie, in particolare PLM (Periodic Leg Movements), NFLE (Nocturnal Frontal Lobe Epilepsy), RBD (REM Behaviour Disorder) e SDB (Sleep-Disordered Breathing).

Tabella 3.1: Distribuzione dei soggetti per condizione clinica nel CAP Sleep Database

Sigla	Condizione clinica	Numero soggetti
N	Soggetti sani	4
PLM	Movimenti periodici degli arti	9
INS	Insonnia	7
NARCO	Narcolessia	4
NFLE	Epilessia frontale notturna	40
RBD	Disturbo comportamentale del sonno REM	18
SDB	Disturbi respiratori del sonno	3
Totale		**85**

3.2 Preparazione dei dati

3.2.1 Data cleaning, Filtraggio e Downsampling

Le registrazioni del CAP Sleep Database sono fornite in formato `EDF` (European Data Format), ampiamente utilizzato per l'archiviazione di registrazioni polisonnografiche multi-canale. Ciascun file `EDF` contiene diversi segnali fisiologici, generalmente campionati a frequenze differenti. Per ogni soggetto sono inoltre disponibili file di annotazione in formato `.txt`, contenenti le etichette degli stadi del sonno.

Dall'analisi degli header dei file `EDF` è emerso che il segnale PPG risulta campionato a 128 Hz per tutti i soggetti. È stato quindi individuato ed estratto il canale corrispondente, ottenendo il segnale continuo per ciascuna registrazione. Le annotazioni sono state lette dai file `.txt` e successivamente ricondotte ad una classificazione a quattro classi (W, LS, DS, REM), ottenuta aggregando Stage 1 e Stage 2 nella classe LS e Stage 3 e Stage 4 nella classe DS.

Le informazioni contenute negli header indicano inoltre che la maggior parte dei segnali PPG risultava già filtrata mediante un filtro passa-banda compreso tra 0.05 Hz e 5 Hz, coerente con il contenuto informativo tipico del segnale PPG. Per garantire uniformità tra i soggetti, è stato applicato un filtro Butterworth passa-banda 0.05–5 Hz ai segnali che non risultavano già filtrati con impostazioni analoghe. Questa operazione consente di attenuare componenti a bassissima e alta frequenza non informative, rendendo i tracciati più omogenei e comparabili. Successivamente i segnali sono stati sottoposti a downsampling da 128 Hz a $34.1\overline{3}$ Hz, riducendo la dimensionalità dei dati pur mantenendo una risoluzione temporale adeguata alla rappresentazione del PPG.

L'istante di inizio della registrazione, ricavato dall'header, è stato utilizzato per allineare temporalmente il segnale PPG e le annotazioni degli stadi del sonno. Il segnale è stato quindi suddiviso in epoche di durata pari a 30 s, ciascuna associata

allo stadio del sonno corrispondente. Sono state mantenute esclusivamente le epoche per le quali risultava disponibile un'etichetta.

A seguito di un'analisi visiva della qualità dei dati, il soggetto `nfle27` è stato escluso in quanto presentava un segnale PPG fortemente corrotto e non confrontabile con quello degli altri soggetti. Il dataset risultante è quindi composto da **84 soggetti**. Durante la stessa fase è stato inoltre osservato che il segnale del soggetto `nfle6` risultava invertito rispetto agli altri tracciati; esso è stato pertanto corretto invertendone il segno. La distribuzione delle epoche nelle diverse fasi del sonno, calcolata sull'insieme degli 84 soggetti selezionati, è riportata nella Tabella 3.2.

Tabella 3.2: Distribuzione delle epoche per stadio del sonno dopo le procedure di pulizia, filtraggio e downsampling.

Classe	**Numero epoche**	**Percentuale**
W	16125	18.8%
LS	34563	40.3%
DS	20465	23.9%
REM	14529	17%
Totale	**85682**	**100%**

Successivamente, il segnale PPG è stato standardizzato mediante normalizzazione z-score. Sono state considerate due modalità: una normalizzazione globale, in cui media e deviazione standard sono calcolate sull'insieme dei soggetti del Train Set, e una normalizzazione per soggetto, in cui tali statistiche vengono calcolate separatamente per ciascun soggetto. La modalità di normalizzazione è stata trattata come un iperparametro, come descritto nella Sezione 3.4.2.

3.2.2 Costruzione delle sequenze

Poiché il modello adottato richiede in ingresso sequenze di lunghezza fissa pari a 1200 epoche (circa 10 ore di registrazione), le epoche sono state organizzate in sequenze temporali continue per ciascun soggetto. In particolare:

- per i soggetti con un numero di epoche superiore a 1200, sono state considerate le prime 1200;
- per i soggetti con un numero inferiore, è stato applicato un padding fino al raggiungimento della lunghezza richiesta.

Il padding è stato realizzato aggiungendo valori nulli al segnale e assegnando un'etichetta fittizia alle epoche non reali, successivamente escluse dal calcolo della funzione di perdita e delle metriche mediante una maschera di validità. Questa

procedura consente di fornire al modello input di dimensione costante, preservando al contempo la sola parte informativa delle registrazioni durante l'addestramento. La distribuzione dei soggetti in funzione della lunghezza delle registrazioni è riportata nella Tabella 3.3.

Tabella 3.3: Distribuzione dei soggetti in funzione della lunghezza delle registrazioni.

Categoria	**Numero soggetti**
Meno di 1200 epoche (padding)	78
Almeno 1200 epoche (troncamento)	6
Totale	**84**

È importante osservare che il troncamento a 1200 epoche modifica leggermente la distribuzione delle classi rispetto al dataset originale, come mostrato nella Tabella 3.4. Tuttavia, lo sbilanciamento tra le diverse fasi del sonno, tipico dei problemi di sleep staging, rimane sostanzialmente invariato. Ciò è dovuto al fatto che solo un numero limitato di soggetti presenta registrazioni di durata superiore alle 1200 epoche. Nel presente lavoro non è stato effettuato un bilanciamento preliminare delle classi; sono state invece adottate specifiche strategie durante la fase di addestramento per mitigare l'effetto dello sbilanciamento.

Tabella 3.4: Distribuzione delle epoche per stadio del sonno dopo il troncamento a 1200 epoche.

Classe	**Numero epoche**	**Percentuale**
W	15362	18.3%
LS	34186	40.6%
DS	20290	24.1%
REM	14300	17%
Totale	**84138**	**100%**

Le operazioni descritte, dalla pre-elaborazione fino alla costruzione delle sequenze a lunghezza fissa, sono riassunte nella pipeline riportata in Figura 3.1.

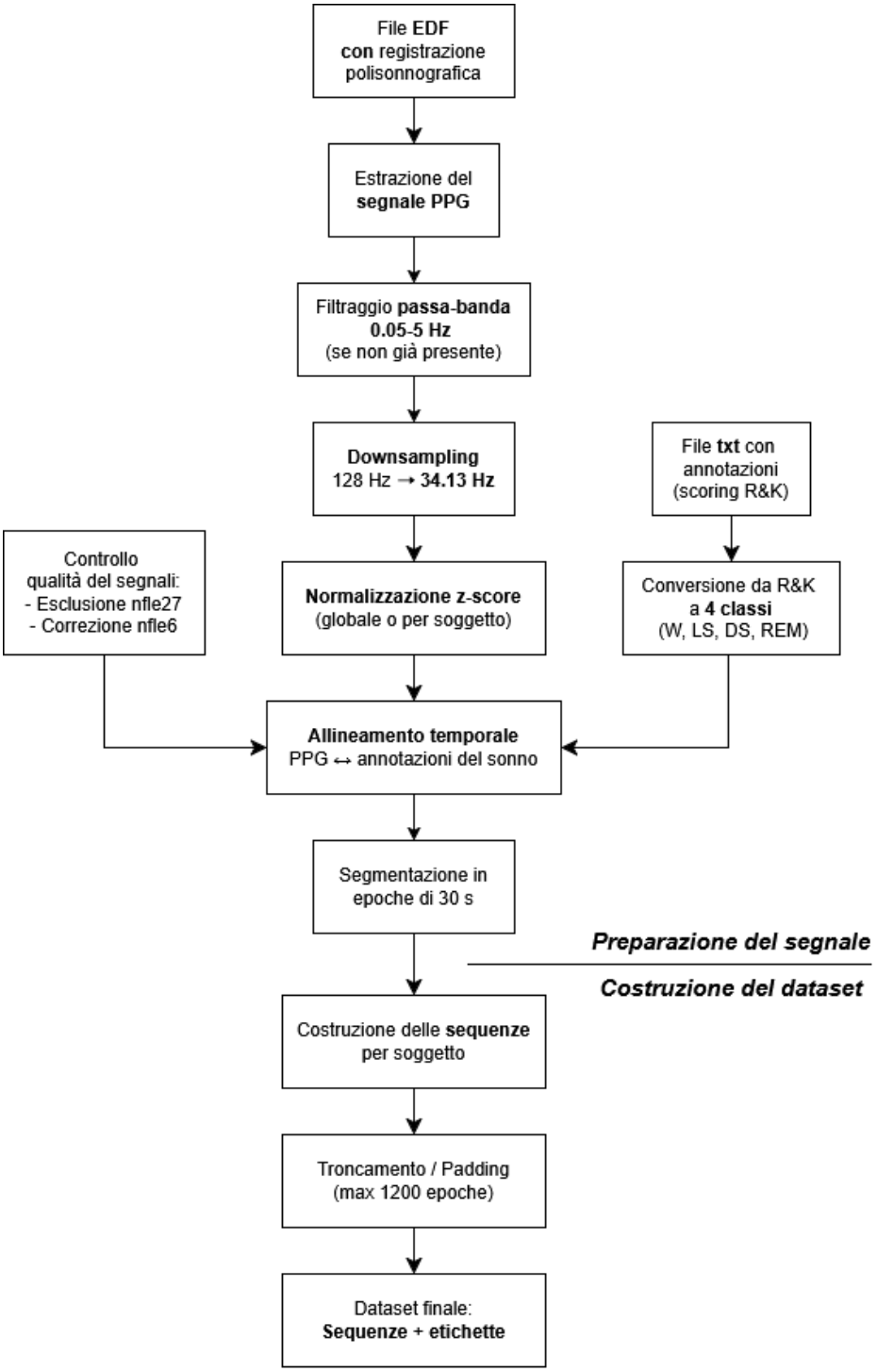

Figura 3.1: Pipeline di preparazione dei dati, dalla registrazione polisonnografica alla costruzione delle sequenze PPG utilizzate come input del modello.

3.3 Architettura del modello

Il modello utilizzato in questo lavoro si basa su una delle architetture proposte in [30], per la quale è disponibile un'implementazione open-source su GitHub. Nel lavoro originale, tale architettura è stata sviluppata per la classificazione automatica delle fasi del sonno a partire dal segnale PPG e prevede una configurazione a due ingressi, costituiti dal segnale PPG originale e da una sua versione augmentata. In questa tesi il modello è stato applicato al CAP Sleep Database e adattato alle caratteristiche di questo dataset.

L'architettura adottata è di tipo **multi-stream sequence-to-sequence** ed è progettata per elaborare sequenze temporali lunghe, corrispondenti all'intera registrazione notturna (circa 10 ore). Il modello segue una struttura modulare composta da tre componenti principali:

- due encoder convoluzionali paralleli, impiegati per l'estrazione delle caratteristiche dal segnale PPG;
- un modulo di fusione basato su meccanismi di cross-attention, utilizzato per combinare le informazioni provenienti dai due ingressi;
- una serie di blocchi convoluzionali temporali, deputati alla modellazione delle dipendenze lungo l'intera sequenza e alla produzione della classificazione finale delle epoche.

Il modello riceve in ingresso due versioni dello stesso segnale PPG continuo: il segnale originale, ottenuto dalla fase di pre-elaborazione descritta in precedenza, e una sua versione perturbata mediante aggiunta di rumore sintetico controllato. Questa struttura a doppio flusso consente al modello di apprendere rappresentazioni complementari dello stesso segnale fisiologico, migliorando la robustezza rispetto a rumore e artefatti tipici delle acquisizioni reali.

Nel presente lavoro, le perturbazioni introdotte sono state mantenute di intensità contenuta, in considerazione della dimensione ridotta del dataset utilizzato (84 soggetti), significativamente inferiore rispetto a quello impiegato in [30], che comprende più di 2000 registrazioni. Tale scelta può essere interpretata come una forma di **augmentazione strutturale del segnale**, in cui due versioni leggermente differenti dello stesso segnale vengono fornite simultaneamente al modello. L'obiettivo è verificare se il meccanismo di cross-attention sia in grado di estrarre informazioni complementari a partire da diverse variazioni dello stesso segnale PPG, anche in assenza di sorgenti informative aggiuntive.

Inoltre, il rumore viene applicato al segnale PPG pre-processato prima della fase di normalizzazione, in modo da introdurre perturbazioni controllate che vengono successivamente ridimensionate dalla standardizzazione, senza alterare in

modo significativo la morfologia del segnale. Il segnale perturbato viene ottenuto combinando diverse tipologie di variazione:

- rumore gaussiano additivo, che introduce piccole fluttuazioni casuali nel segnale;
- baseline drift a bassa frequenza, modellato come una componente sinusoidale lenta;
- perturbazioni impulsive, generate in modo casuale, che producono variazioni locali nel segnale.

La generazione del rumore viene effettuata separatamente per ciascun soggetto. Questo approccio garantisce la riproducibilità delle perturbazioni e introduce al tempo stesso una variabilità controllata tra i diversi segnali.
Il modello è rappresentato nella Figura 3.2 ed è stato inizializzato utilizzando i seguenti parametri principali:

- **numero di classi di output**: 4 (W, LS, DS, REM);
- **dimensione dello spazio delle caratteristiche**: $d_{model} = 256$;
- **numero di teste del meccanismo di attenzione**: $n_{heads} = 8$;
- **numero di blocchi di fusione basati su cross-attention**: 3.

La figura fornisce una rappresentazione schematica dell'architettura complessiva del modello. Nel seguito, i principali blocchi funzionali vengono descritti in dettaglio, illustrandone il ruolo nel processo di elaborazione del segnale e nella produzione della classificazione finale.

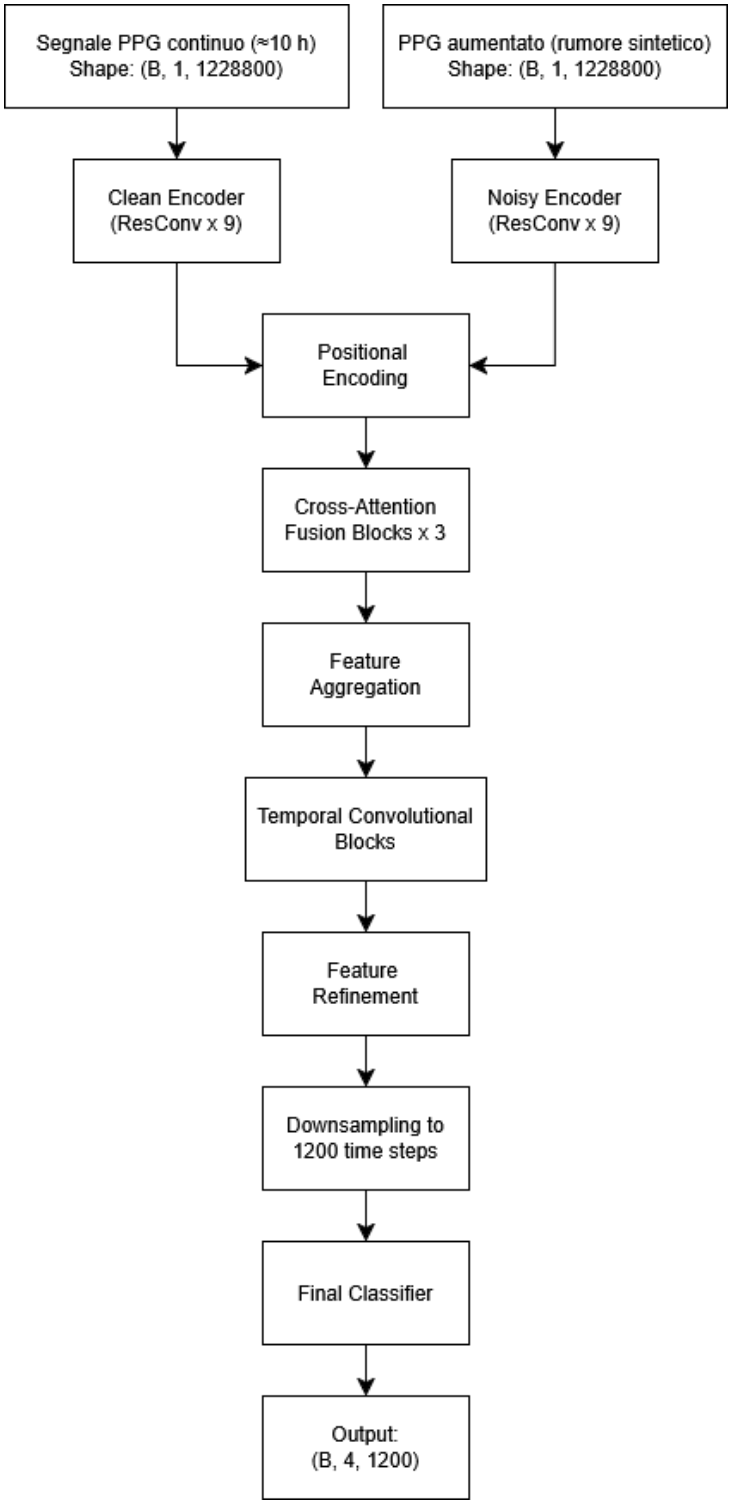

Figura 3.2: Schema dell'architettura utilizzata.

Convolutional Encoders (Encoder convoluzionali paralleli)
Ciascuna delle due versioni del segnale viene elaborata da un encoder convoluzionale indipendente, composto da nove *Residual Convolutional Blocks*. Ogni blocco utilizza convoluzioni monodimensionali con connessioni residue e operazioni di pooling con stride pari a 2. Questa struttura consente una progressiva riduzione della risoluzione temporale del segnale e un incremento del numero di canali fino a $d_{model} = 256$, permettendo di estrarre rappresentazioni gerarchiche delle caratteristiche locali e semi-globali dal segnale.

Positional Encoding (Codifica posizionale)
Le rappresentazioni prodotte dai due encoder vengono arricchite mediante positional encoding, che introduce informazioni sull'ordine temporale dei campioni all'interno della sequenza. Questo passaggio è necessario poiché i meccanismi di attenzione utilizzati nelle fasi successive non incorporano intrinsecamente informazioni sulla posizione temporale degli elementi della sequenza.

Cross-Attention Fusion Blocks (Blocchi di fusione con Cross-Attention)
Le caratteristiche estratte dai due rami vengono combinate mediante tre blocchi consecutivi di Multi-Head Cross-Attention. In ciascun blocco, le rappresentazioni dei due segnali vengono utilizzate alternativamente come *query*, *key* e *value*, consentendo al modello di apprendere relazioni tra il segnale PPG originale e la sua versione perturbata.

Feature Aggregation (Aggregazione delle caratteristiche)
Le rappresentazioni fuse vengono ulteriormente elaborate mediante un modulo di *Adaptive Modality Weighting*, che stima dinamicamente un peso relativo per ciascun ingresso in funzione della sua informatività. Le feature pesate vengono quindi integrate mediante convoluzioni 1×1, che consentono di consolidare l'informazione multimodale riducendo al contempo la dimensionalità delle rappresentazioni.

Temporal Convolutional Blocks (Blocchi convoluzionali temporali)
Le feature aggregate vengono elaborate mediante una sequenza di *Temporal Convolutional Blocks* con kernel di dimensione 7 e dilatazioni crescenti (1, 2, 4, 8). L'utilizzo di convoluzioni dilatate consente di ampliare il campo recettivo senza aumentare significativamente il numero di parametri, permettendo al modello di catturare dipendenze temporali a breve e lungo termine lungo l'intera sequenza.

Feature Refinement (Raffinamento delle caratteristiche)
Le rappresentazioni vengono ulteriormente raffinate mediante strati convoluzionali aggiuntivi, migliorando la qualità delle feature prima della fase di classificazione.

Temporal Downsampling (Riduzione temporale)
Un'operazione di downsampling temporale riduce la sequenza a 1200 passi temporali, ciascuno corrispondente a un'epoca di 30 s.

Final Classifier (Classificatore finale)
Il classificatore finale, implementato mediante convoluzioni 1×1, produce un output di dimensioni $(B, 4, 1200)$, dove B è la dimensione del batch. Le probabilità di appartenenza alle classi vengono ottenute applicando una funzione di attivazione softmax lungo la dimensione delle classi.

3.4 Addestramento del modello

3.4.1 Configurazione generale dell'addestramento

Poiché il problema affrontato consiste nella classificazione multi-classe delle fasi del sonno, è stata utilizzata come funzione di perdita la **Cross-Entropy Loss**. Per mitigare gli effetti dello sbilanciamento tra le diverse classi, sono stati introdotti **pesi di classe** inversamente proporzionali alla loro frequenza nel Train Set, in modo da penalizzare maggiormente gli errori sulle classi meno rappresentate. È stata inoltre valutata l'introduzione del **label smoothing**, tecnica che consente di regolarizzare la funzione di perdita riducendo l'eccessiva fiducia del modello nelle predizioni.

L'ottimizzazione dei parametri della rete è stata effettuata mediante l'algoritmo **AdamW**, scelto per la sua capacità di adattare dinamicamente il learning rate durante il processo di ottimizzazione.

Per migliorare la stabilità dell'addestramento, è stato applicato un meccanismo di **gradient clipping**, limitando la norma del gradiente a un valore massimo pari a 1.0, così da prevenire fenomeni di esplosione del gradiente.

Per favorire una migliore convergenza, è stato inoltre utilizzato uno **scheduler del learning rate**. In particolare, è stato impiegato il metodo *ReduceLROnPlateau*, con fattore di riduzione pari a 0.5 e parametro *patience* pari a 5 epoche. Lo scheduler monitora il valore del **coefficiente kappa di Cohen** calcolato sul Validation Set e riduce il learning rate qualora tale metrica non mostri miglioramenti.

Infine, per ridurre il rischio di overfitting, è stato adottato un meccanismo di **early stopping**: il processo di addestramento viene interrotto qualora il coefficiente kappa di Cohen sul Validation Set non mostri miglioramenti per 15 epoche consecutive.

3.4.2 Ottimizzazione degli iperparametri

La configurazione ottimale degli iperparametri del modello è stata individuata mediante una procedura di **grid search**, finalizzata a valutare l'influenza di diverse scelte di addestramento e tecniche di pre-elaborazione sulle prestazioni del modello. In particolare, sono state esplorate diverse combinazioni di iperparametri, riportati nella Tabella 3.5, relativi sia al processo di training che alla normalizzazione dei dati. Per ciascuna configurazione, il modello è stato addestrato sul Train Set e valutato sul Validation Set.

La selezione della configurazione migliore è stata effettuata sulla base delle prestazioni ottenute sul Validation Set, utilizzando come metrica principale il **coefficiente kappa di Cohen**, scelto in quanto particolarmente adatto alla valutazione di problemi con classi sbilanciate e coerente con quanto riportato nell'articolo di riferimento [30].

Tabella 3.5: Iperparametri esplorati durante la procedura di grid search

Parametro	**Valori testati**
Normalization	global, subject wise
Learning rate	10^{-5}, 5×10^{-5}, 10^{-4}
Batch size	2, 4
Weight decay	10^{-5}, 10^{-4}
Class weights	Yes, No
Label smoothing	0, 0.05

La procedura di ottimizzazione è stata condotta utilizzando la configurazione Train–Validation–Test nel caso di classificazione a 4 classi, scelta come riferimento per la selezione degli iperparametri. Una volta individuata la configurazione migliore sul Validation Set, gli iperparametri sono stati mantenuti fissi in tutti gli esperimenti successivi, al fine di garantire coerenza tra i diversi protocolli sperimentali. Questa impostazione consente di analizzare separatamente l'effetto delle strategie di classificazione e dei protocolli di validazione sulle prestazioni del modello, mantenendo invariata la configurazione del modello.

3.4.3 Strategie di classificazione

La prima strategia prevede una **classificazione diretta a 4 classi**. Il modello riceve in ingresso le due sequenze di segnale PPG e produce in uscita, per ciascuna epoca, una delle quattro classi considerate: W, LS, DS e REM. Il modello viene quindi addestrato a discriminare simultaneamente tutte le classi del problema.

Questa strategia corrisponde all'approccio adottato in [30]. Nell'ambito dell'analisi del sonno, la classificazione multi-classe diretta risulta spesso efficace poiché

il modello può sfruttare il contesto temporale per apprendere non solo le caratteristiche delle singole classi, ma anche le transizioni tipiche tra i diversi stadi del sonno.

La seconda strategia utilizza invece un **approccio gerarchico a cascata**. Nel primo stadio, il modello riceve in ingresso le due sequenze e assegna a ciascuna epoca una delle tre classi W, NREM o REM. In questa fase, le classi LS e DS vengono aggregate nella classe NREM. Nel secondo stadio, un ulteriore modello viene addestrato per distinguere esclusivamente tra le classi LS e DS. Anche nel secondo stadio, il modello riceve in ingresso l'intera sequenza temporale, preservando quindi il contesto delle epoche circostanti. L'architettura utilizzata è la stessa impiegata nel primo stadio, ma i due modelli vengono addestrati separatamente per i rispettivi compiti di classificazione. Durante l'addestramento del secondo modello, la funzione di perdita viene calcolata considerando unicamente le epoche appartenenti alle classi LS e DS, così da concentrare l'apprendimento sulla discriminazione tra questi due stadi.

Questo secondo approccio è motivato dal fatto che LS e DS rappresentano gli stadi più difficili da distinguere, poiché condividono caratteristiche fisiologiche simili. La suddivisione del problema in due passaggi consente quindi di semplificare il compito di classificazione, permettendo al secondo modello di specializzarsi nella discriminazione tra queste due classi.

Per quanto riguarda il processo decisionale finale:

- nel caso della classificazione diretta a 4 classi, la predizione deriva dal vettore di probabilità prodotto dalla funzione softmax, selezionando la classe con probabilità massima;
- nel caso dell'approccio a cascata, il secondo modello viene applicato esclusivamente alle epoche che il primo modello classifica come NREM, realizzando quindi un processo decisionale di tipo hard gating tra i due stadi della classificazione.

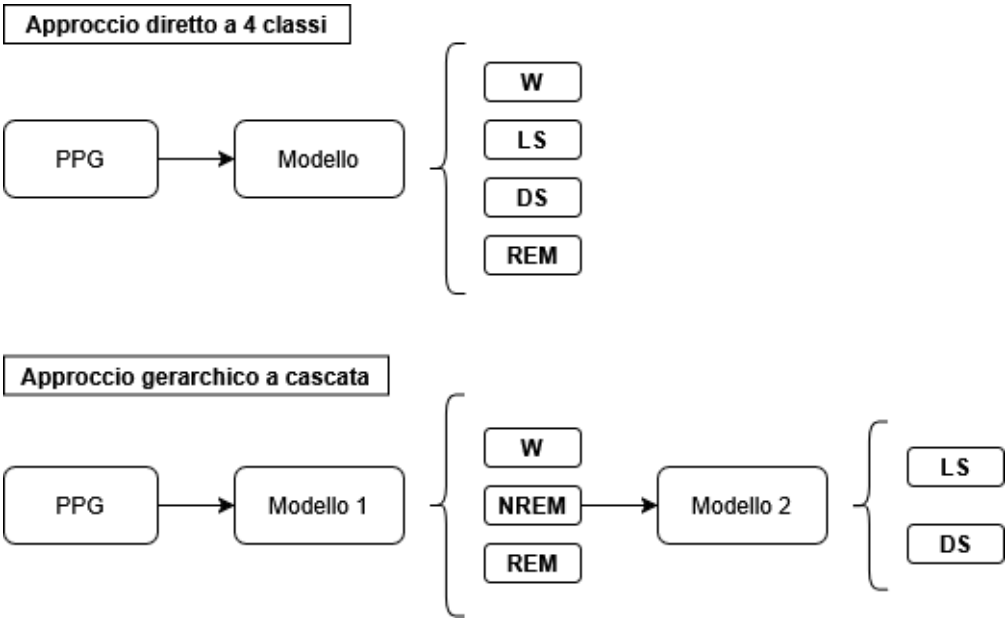

Figura 3.3: Schema delle strategie di classificazione.

3.4.4 Protocolli sperimentali

Validazione Train-Validation-Test
Una prima valutazione delle prestazioni del modello è stata effettuata utilizzando una suddivisione del dataset in Train Set, Validation Set e Test Set. In questa configurazione, il Train Set viene utilizzato per l'addestramento del modello, mentre il Validation Set viene impiegato per monitorare le prestazioni durante l'allenamento e per la selezione degli iperparametri. Il Test Set è invece utilizzato esclusivamente per la valutazione finale delle prestazioni del modello.

La suddivisione del dataset è stata effettuata a livello di soggetto, in modo che tutte le epoche appartenenti allo stesso paziente fossero assegnate allo stesso sottoinsieme. Questa scelta consente di evitare fenomeni di data leakage e di valutare il modello su soggetti non osservati durante la fase di addestramento. Inoltre, la suddivisione è stata effettuata in modo stratificato rispetto alle diverse condizioni cliniche presenti nel dataset, al fine di mantenere una distribuzione il più possibile equilibrata delle patologie nei diversi sottoinsiemi.

Validazione Leave-One-Subject-Out (LOSO)
Oltre alla suddivisione Train–Validation–Test, è stato adottato anche il protocollo di validazione Leave-One-Subject-Out (LOSO), comunemente utilizzato negli studi di analisi del sonno per valutare la capacità di generalizzazione dei modelli tra soggetti diversi.

In questa configurazione, il modello viene addestrato utilizzando i dati di tutti i soggetti tranne uno, che viene utilizzato come insieme di test. Il processo viene ripetuto iterativamente per ciascun soggetto del dataset, in modo che ogni paziente venga utilizzato una volta come Test Set. Le prestazioni finali del modello sono quindi ottenute calcolando la media delle metriche di valutazione sulle diverse iterazioni. Questo protocollo consente di stimare in modo più realistico la capacità del modello di generalizzare a nuovi soggetti non osservati durante l'addestramento.

3.5 Valutazione delle prestazioni

3.5.1 Metriche di classificazione

La valutazione delle prestazioni del modello è stata effettuata utilizzando diverse metriche comunemente adottate nei problemi di classificazione multi-classe. L'impiego di più indicatori consente di ottenere una valutazione più completa, considerando sia le prestazioni globali sia la capacità dle modello di discriminare correttamente le diverse classi. In particolare, sono state considerate le seguenti metriche:

- **Accuracy**: rappresenta la proporzione di predizioni corrette rispetto al numero totale di campioni:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

  dove TP indica i veri positivi, TN i veri negativi, FP i falsi positivi e FN i falsi negativi. Nel contesto multi-classe, tali quantità sono da intendersi secondo uno schema one-vs-all. Sebbene l'accuracy fornisca una misura intuitiva delle prestazioni complessive del modello, essa può risultare meno informativa in presenza di dataset sbilanciati.

- **Precision**: misura la percentuale di predizioni corrette tra tutti i campioni assegnati a una determinata classe:

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

  Questa metrica indica quindi quanto siano affidabili le predizioni del modello per una specifica classe ed è particolarmente rilevante nei contesti in cui il costo dei falsi positivi è elevato.

- **Recall**: nota anche come *sensitivity*, misura la capacità del modello di identificare correttamente i campioni appartenenti a una determinata classe:

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

  Un valore elevato di recall indica che il modello è in grado di riconoscere correttamente la maggior parte dei campioni appartenenti alla classe considerata.

- **F1-score**: rappresenta la media armonica tra precision e recall:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.4}$$

  Nel presente lavoro viene considerato il Macro F1-score, ottenuto come media degli F1-score delle singole classi. Questa metrica risulta particolarmente utile in presenza di classi sbilanciate, in quanto attribuisce lo stesso peso a tutte le classi.

- **Cohen's Kappa**: misura il grado di accordo tra le predizioni del modello e le etichette reali, tenendo conto anche dell'accordo atteso per puro caso:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.5}$$

  dove p_o rappresenta la probabilità di accordo osservato tra predizioni e etichette reali, mentre p_e indica la probabilità di accordo atteso per caso. Questa metrica è ampiamente utilizzata nei problemi di sleep staging per valutare la qualità della classificazione.

- **Confusion Matrix**: è una matrice quadrata in cui ciascuna riga rappresenta la classe reale e ciascuna colonna la classe predetta. Essa consente di analizzare nel dettaglio le prestazioni del modello, evidenziando in quali casi le diverse fasi del sonno vengono confuse tra loro.

3.5.2 Metriche cliniche

Oltre alle metriche di classificazione, sono state considerate anche alcune metriche di rilevanza clinica comunemente utilizzate nell'analisi del sonno. al fine di valutare l'impatto degli errori di classificazione sulla stima di parametri clinicamente significativi. In particolare, sono stati calcolati:

- **Total Sleep Time (TST)**: rappresenta il tempo totale trascorso in sonno durante la registrazione. Viene calcolato come la somma delle epoche classificate come stadi di sonno:

$$TST = \frac{N_{\text{Sleep}} \cdot T_{\text{epoch}}}{60} \tag{3.6}$$

  dove N_{Sleep} indica il numero di epoche classificate come sonno (LS, DS o REM) e T_{epoch} la durata di ciascuna epoca, pari a 30 s. Il risultato è espresso in minuti.

- **Sleep Efficiency (SE)**: rappresenta la frazione di tempo effettivamente trascorso in sonno rispetto alla durata totale della registrazione ed è definita come:

$$SE = \frac{TST}{T_{\text{Total}}} \tag{3.7}$$

  dove T_{Total} è il tempo totale di registrazione in minuti. Questa metrica fornisce una misura sintetica della qualità complessiva del sonno.

Nel presente lavoro tali metriche sono state calcolate sia a partire dalle annotazioni di riferimento sia dalle predizioni del modello, consentendo di valutare l'errore nella stima dei parametri clinici derivati.

Capitolo 4

Risultati

4.1 Selezione degli iperparametri

4.1.1 Grid search nella classificazione a quattro classi

La configurazione ottimale degli iperparametri è stata individuata mediante una procedura di grid search condotta nella configurazione Train–Validation–Test con classificazione diretta a quattro classi. In particolare, sono state valutate diverse combinazioni di parametri di addestramento e di normalizzazione del segnale, selezionando la configurazione migliore sulla base del coefficiente κ di Cohen calcolato sul Validation Set. Tale metrica è stata scelta in quanto particolarmente adatta in presenza di dataset sbilanciati, poiché tiene conto anche dell'accordo atteso per caso. La Tabella 4.1 riporta le cinque configurazioni che hanno ottenuto le migliori prestazioni durante la procedura di ottimizzazione.

Tabella 4.1: Migliori configurazioni ottenute durante la procedura di grid search, ordinate in base al coefficiente κ di Cohen sul Validation Set.

Norm.	**LR**	**Batch size**	**WD**	**CW**	**Label smoothing**	κ
global	10^{-4}	2	10^{-4}	Yes	0.05	**0.474**
subject wise	10^{-4}	2	10^{-5}	Yes	0.0	0.463
subject wise	10^{-4}	2	10^{-4}	No	0.05	0.457
subject wise	10^{-4}	2	10^{-5}	No	0.0	0.451
global	10^{-4}	2	10^{-5}	No	0.05	0.450

Nella tabella, LR indica il learning rate, WD il weight decay e CW l'utilizzo dei *class weights* nella funzione di perdita. La configurazione che massimizza il coefficiente κ utilizza normalizzazione globale, learning rate pari a 10^{-4}, batch size pari a 2, weight decay pari a 10^{-4}, pesi di classe nella funzione di perdita e label

smoothing pari a 0.05. Tale configurazione è stata quindi adottata in tutti gli esperimenti successivi, i cui risultati sono riportati nelle sezioni seguenti.

Le differenze tra le configurazioni migliori risultano relativamente contenute, suggerendo una buona stabilità del modello rispetto alle variazioni degli iperparametri considerati. Si osserva inoltre che le configurazioni che includono i pesi di classe nella funzione di perdita tendono a ottenere prestazioni leggermente migliori, verosimilmente grazie a una migliore gestione dello sbilanciamento tra le classi.

4.2 Risultati con suddivisione Train–Validation–Test

4.2.1 Prestazioni di classificazione

In questa sezione vengono confrontate le prestazioni del modello nella configurazione Train–Validation–Test, considerando sia la classificazione diretta a quattro classi sia l'approccio a cascata.

Tabella 4.2: Metriche globali ottenute sul Validation Set.

Strategia	**Accuracy**	**Precision**	**Recall**	**F1-score**	κ
4 classi	0.625	0.635	0.628	0.630	0.474
Cascata	0.611	0.630	0.619	0.622	0.454

Le prestazioni sul Validation Set sono riportate nella Tabella 4.2. Le due strategie mostrano risultati complessivamente comparabili, con valori leggermente superiori per la classificazione a quattro classi in tutte le metriche considerate. In particolare, la classificazione diretta raggiunge un'accuracy pari a 0.625 e un coefficiente κ pari a 0.474, mentre l'approccio a cascata ottiene rispettivamente 0.611 e 0.454.

Tabella 4.3: Metriche globali ottenute sul Test Set.

Strategia	**Accuracy**	**Precision**	**Recall**	**F1-score**	κ
4 classi	0.618	0.622	0.627	0.624	0.470
Cascata	0.597	0.605	0.593	0.597	0.436

Le prestazioni sul Test Set, riportate nella Tabella 4.3, confermano lo stesso andamento osservato sul Validation Set. Anche in questo caso la classificazione a quattro classi ottiene valori superiori rispetto all'approccio a cascata, con accuracy

pari a 0.618 contro 0.597 e coefficiente κ pari a 0.470 contro 0.436. Nel complesso, i risultati evidenziano una buona coerenza tra Validation e Test Set.

Tabella 4.4: Metriche per classe ottenute sul Test Set con la classificazione a quattro classi.

Classe	**Precision**	**Recall**	**F1-score**
W	0.647	0.672	0.659
LS	0.591	0.586	0.588
DS	0.663	0.627	0.644
REM	0.589	0.624	0.606

Tabella 4.5: Metriche per classe ottenute sul Test Set con la classificazione a cascata.

Classe	**Precision**	**Recall**	**F1-score**
W	0.620	0.642	0.630
LS	0.565	0.625	0.594
DS	0.693	0.557	0.617
REM	0.541	0.548	0.545

Le prestazioni per classe sul Test Set sono riportate nelle Tabelle 4.4 e 4.5. In entrambe le strategie si osservano differenze tra le classi, con valori generalmente più elevati per W e DS rispetto a LS e REM. Nel caso della classificazione a quattro classi, ad esempio, la classe W raggiunge un F1-score pari a 0.659, mentre LS si attesta a 0.588; nella classificazione in cascata, i valori corrispondenti sono pari a 0.630 per W e 0.594 per LS. Nel complesso, le due strategie mostrano profili prestazionali simili, pur con alcune variazioni nelle singole classi.

Per un'analisi più dettagliata degli errori di classificazione, le confusion matrix aggregate sul Test Set per le due strategie sono riportate in Figura 4.1 e Figura 4.2. In entrambi i casi, si osserva una maggiore concentrazione dei valori lungo la diagonale principale, indicativa di una buona capacità di classificazione complessiva. Le principali confusioni si verificano tra le classi LS e DS, mentre W e REM risultano generalmente meglio discriminate.

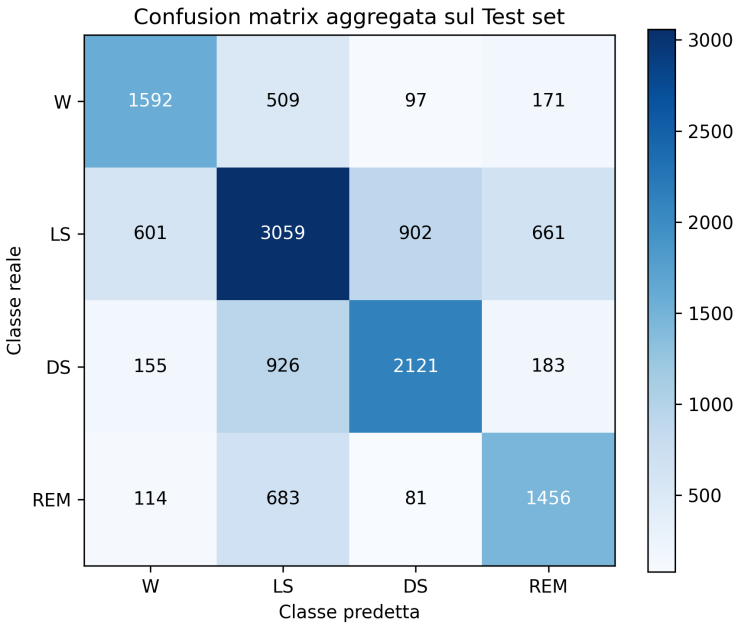

Figura 4.1: Confusion matrix aggregata sul Test Set per la classificazione a quattro classi.

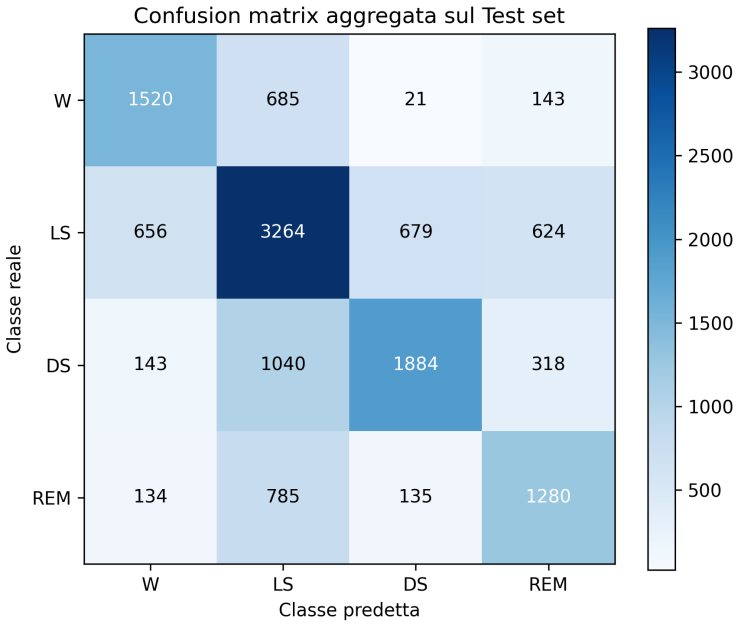

Figura 4.2: Confusion matrix aggregata sul Test Set per la classificazione in cascata.

4.2.2 Metriche cliniche

Tabella 4.6: Statistiche riassuntive delle metriche cliniche su Test Set.

Strategia	Metrica	Manuale	Automatica	Errore relativo
4 classi	TST	390.8 ± 67.9 min	387.5 ± 75.6 min	0.2 ± 15.4 %
4 classi	SE	0.818 ± 0.093	0.813 ± 0.122	0.2 ± 15.4 %
Cascata	TST	390.8 ± 67.9 min	387.8 ± 79.2 min	-0.6 ± 18.1 %
Cascata	SE	0.818 ± 0.093	0.816 ± 0.140	-0.6 ± 18.1 %

Le statistiche riassuntive delle metriche cliniche sul Test Set sono riportate nella Tabella 4.6. I valori stimati risultano complessivamente in accordo con quelli di riferimento per entrambe le strategie. In particolare, per la classificazione a quattro classi il TST medio stimato è pari a 387.5 min rispetto a 390.8 min del riferimento manuale, mentre la SE media è pari a 0.813 rispetto a 0.818. Anche per l'approccio a cascata si osservano differenze contenute, con valori pari a 387.8 min per il TST e 0.816 per la SE. Si osserva inoltre che l'errore relativo risulta identico per TST e SE. Questo comportamento è atteso, in quanto SE è definita come rapporto tra TST e durata totale della registrazione, risultando quindi direttamente dipendente dalla stima del TST.

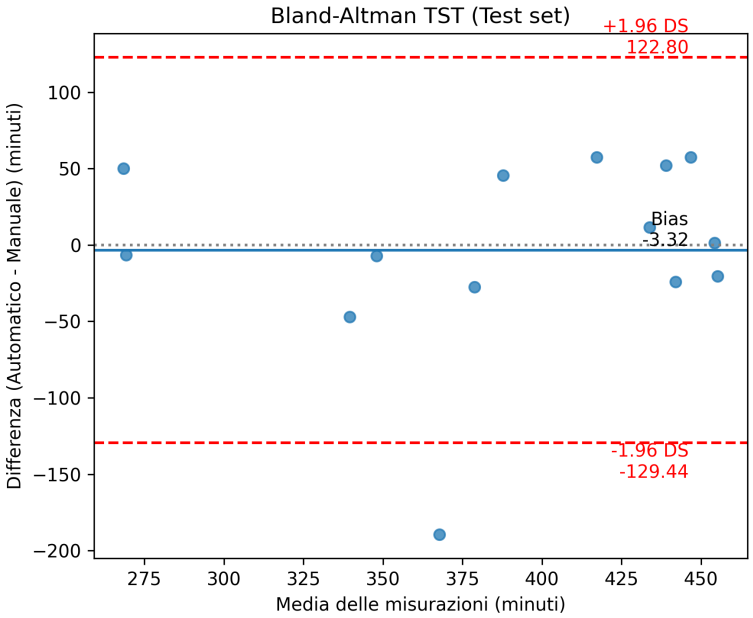

Figura 4.3: Bland–Altman per TST sul Test Set con classificazione a quattro classi.

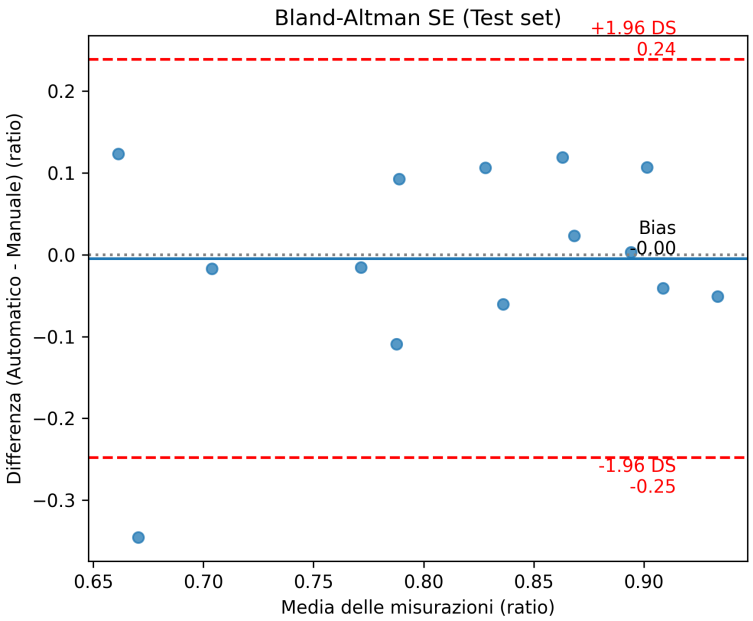

Figura 4.4: Bland–Altman per SE sul Test Set con classificazione a quattro classi.

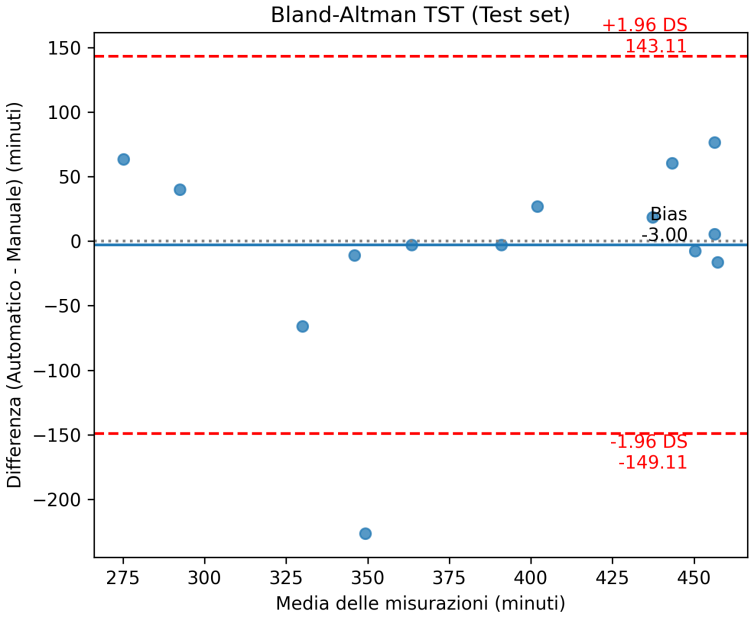

Figura 4.5: Bland–Altman per TST sul Test Set con classificazione a cascata.

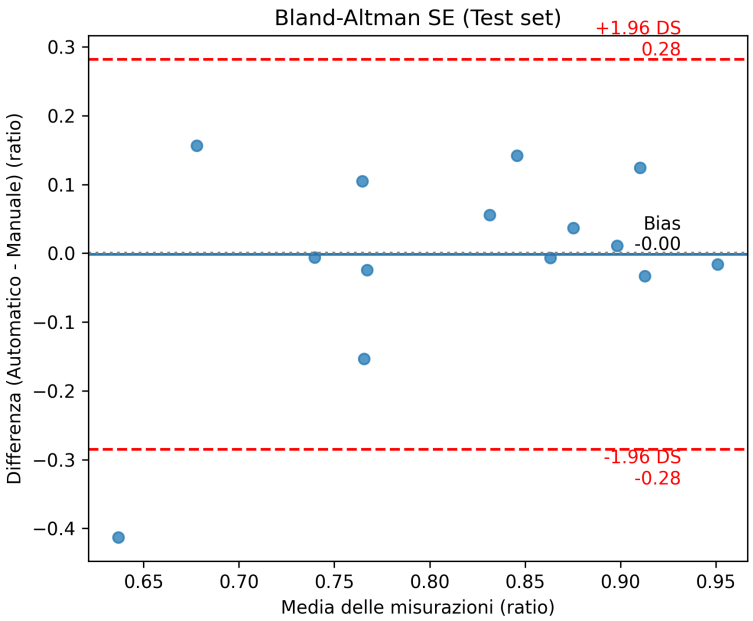

Figura 4.6: Bland–Altman per SE sul Test Set con classificazione a cascata.

I grafici di Bland–Altman relativi alle metriche cliniche sul Test Set sono riportati nelle Figure 4.3–4.6. In particolare, le Figure 4.3 e 4.4 si riferiscono alla classificazione a quattro classi, mentre le Figure 4.5 e 4.6 riguardano l'approccio a cascata. In tutti i casi le differenze tra valori stimati e di riferimento risultano distribuite attorno allo zero, con la maggior parte dei punti compresa all'interno dei limiti di accordo definiti da ±1.96 deviazioni standard, senza evidenti pattern sistematici.

4.3 Risultati con validazione LOSO

4.3.1 Prestazioni di classificazione

In questa sezione vengono riportate le prestazioni del modello nella configurazione di validazione LOSO, considerando sia la classificazione a quattro classi sia l'approccio a cascata.

Tabella 4.7: Metriche globali ottenute nella configurazione LOSO.

Strategia	**Accuracy**	**Precision**	**Recall**	**F1-score**	κ
4 classi	0.708 ± 0.121	0.693 ± 0.126	0.716 ± 0.125	0.682 ± 0.136	0.579 ± 0.159
Cascata	0.719 ± 0.113	0.699 ± 0.130	0.713 ± 0.118	0.686 ± 0.133	0.592 ± 0.156

Le metriche globali ottenute nella configurazione LOSO sono riportate nella Tabella 4.7. Entrambe le strategie mostrano prestazioni complessivamente buone, con valori di accuracy superiori a 0.70 e valori di κ prossimi a 0.6. In particolare, la classificazione a quattro classi raggiunge un'accuracy media pari a 0.708 e un coefficiente κ pari a 0.579, mentre l'approccio a cascata ottiene valori pari rispettivamente a 0.719 e 0.592. Le deviazioni standard osservate nelle diverse metriche evidenziano una variabilità delle prestazioni tra i soggetti.

Tabella 4.8: Metriche per classe ottenute sui diversi fold nella configurazione LOSO per la classificazione a quattro classi.

Classe	**Precision**	**Recall**	**F1-score**
W	0.728 ± 0.199	0.776 ± 0.198	0.728 ± 0.173
LS	0.732 ± 0.130	0.615 ± 0.184	0.651 ± 0.160
DS	0.646 ± 0.190	0.761 ± 0.199	0.682 ± 0.176
REM	0.666 ± 0.243	0.712 ± 0.262	0.665 ± 0.235

Tabella 4.9: Metriche per classe ottenute sui diversi fold nella configurazione LOSO per la classificazione in cascata.

Classe	**Precision**	**Recall**	**F1-score**
W	0.736 ± 0.190	0.755 ± 0.192	0.712 ± 0.161
LS	0.721 ± 0.152	0.646 ± 0.160	0.673 ± 0.143
DS	0.666 ± 0.189	0.764 ± 0.186	0.696 ± 0.172
REM	0.673 ± 0.262	0.687 ± 0.295	0.663 ± 0.262

Le prestazioni per classe nella configurazione LOSO sono riportate nelle Tabelle 4.8 e 4.9. Si osservano differenze tra le diverse classi, con valori di recall generalmente più elevati per W e DS. Nel caso della classificazione a quattro classi, ad esempio, il recall medio raggiunge 0.776 per W e 0.761 per DS, mentre la classe LS si attesta a 0.615. Anche nella classificazione in cascata LS presenta valori inferiori rispetto a W e DS, mentre REM mostra una maggiore variabilità, come evidenziato dalle deviazioni standard più elevate.

Per un'analisi più dettagliata degli errori, le confusion matrix medie normalizzate nella configurazione LOSO sono riportate in Figura 4.7 e Figura 4.8. In entrambi i casi si osserva una buona concentrazione dei valori lungo la diagonale principale. Le principali confusioni si verificano tra le classi LS e DS, mentre le classi W e REM risultano generalmente meglio discriminate. Le due strategie mostrano un comportamento complessivamente simile nella distribuzione degli errori.

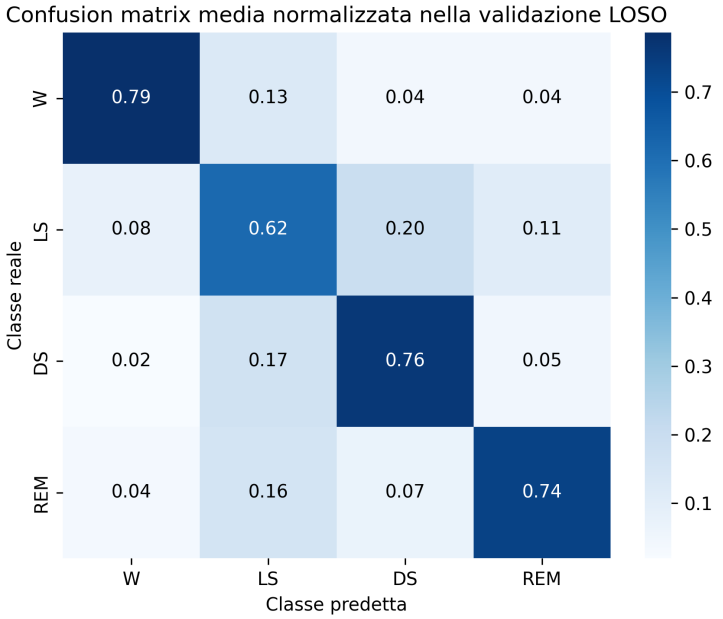

Figura 4.7: Confusion matrix media normalizzata sui soggetti nella validazione LOSO per la classificazione a quattro classi.

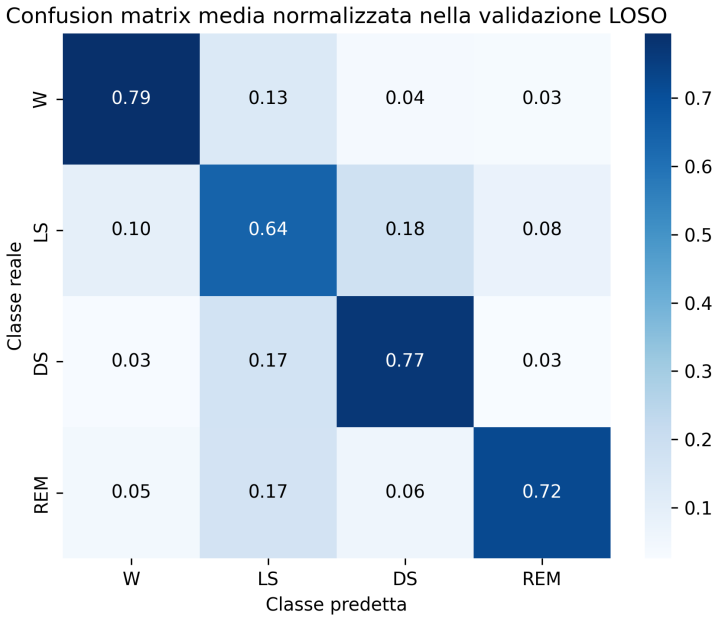

Figura 4.8: Confusion matrix media normalizzata sui soggetti nella validazione LOSO per la classificazione in cascata.

4.3.2 Metriche cliniche

Tabella 4.10: Statistiche riassuntive delle metriche cliniche sui soggetti nella validazione LOSO.

Strategia	Metrica	Manuale	Automatica	Errore relativo
4 classi	TST	409.3 ± 81.8 min	407.6 ± 80.5 min	0.1 ± 7.5 %
4 classi	SE	0.819 ± 0.142	0.816 ± 0.144	0.1 ± 7.5 %
Cascata	TST	409.3 ± 81.8 min	401.1 ± 94.8 min	1.7 ± 14.4 %
Cascata	SE	0.819 ± 0.142	0.803 ± 0.174	1.7 ± 14.4 %

Le statistiche riassuntive delle metriche cliniche nella configurazione LOSO sono riportate nella Tabella 4.10. I valori stimati risultano complessivamente in accordo con quelli di riferimento, con differenze contenute tra le stime automatiche e le annotazioni manuali. Per la classificazione a quattro classi, il TST medio passa da 409.3 min a 407.6 min e la SE da 0.819 a 0.816; nella classificazione in cascata i valori corrispondenti sono pari a 401.1 min per il TST e 0.803 per la SE. Si osserva inoltre che l'errore relativo risulta identico per TST e SE. Anche in questo caso tale comportamento è atteso, poiché SE dipende direttamente dalla stima di TST.

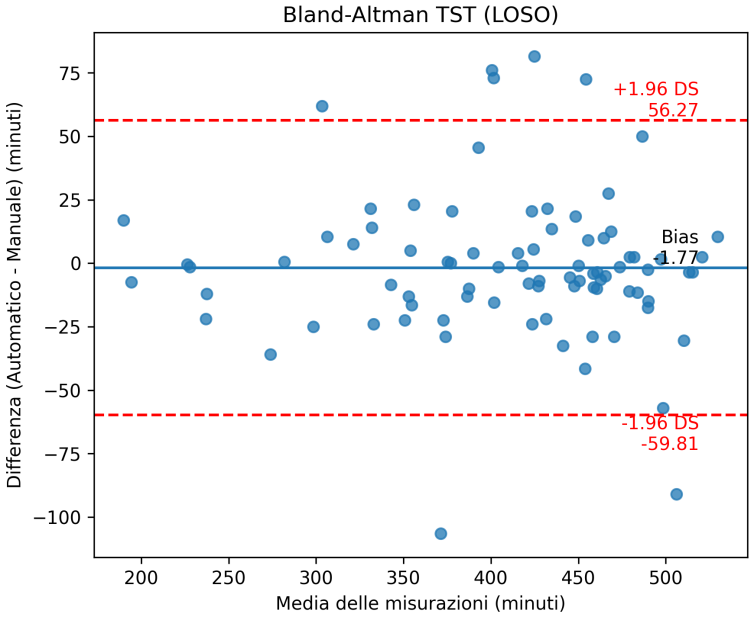

Figura 4.9: Grafico di Bland–Altman per il TST nella validazione LOSO con classificazione a quattro classi.

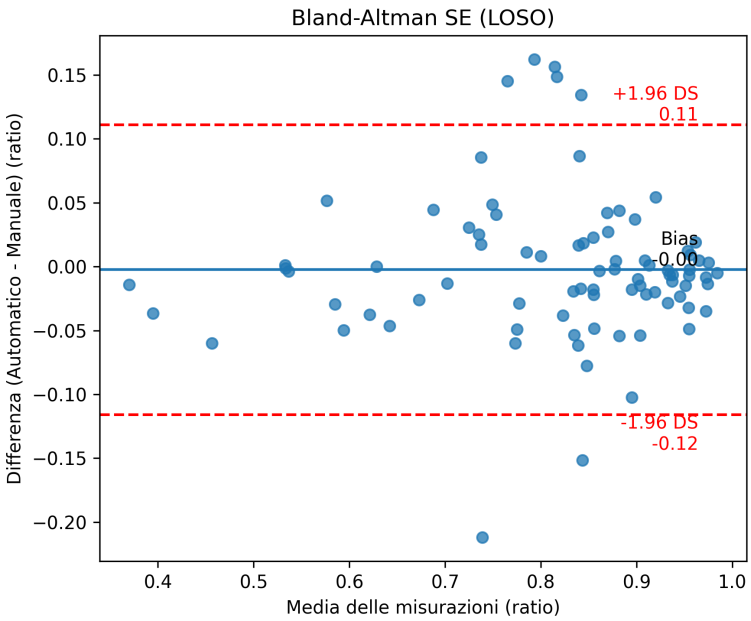

Figura 4.10: Grafico di Bland–Altman per la SE nella validazione LOSO con classificazione a quattro classi.

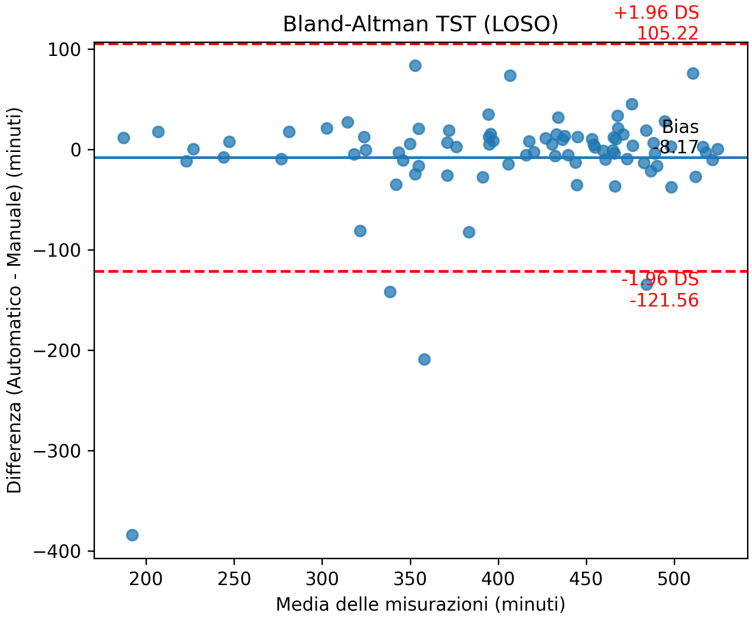

Figura 4.11: Grafico di Bland–Altman per il TST nella validazione LOSO con classificazione in cascata.

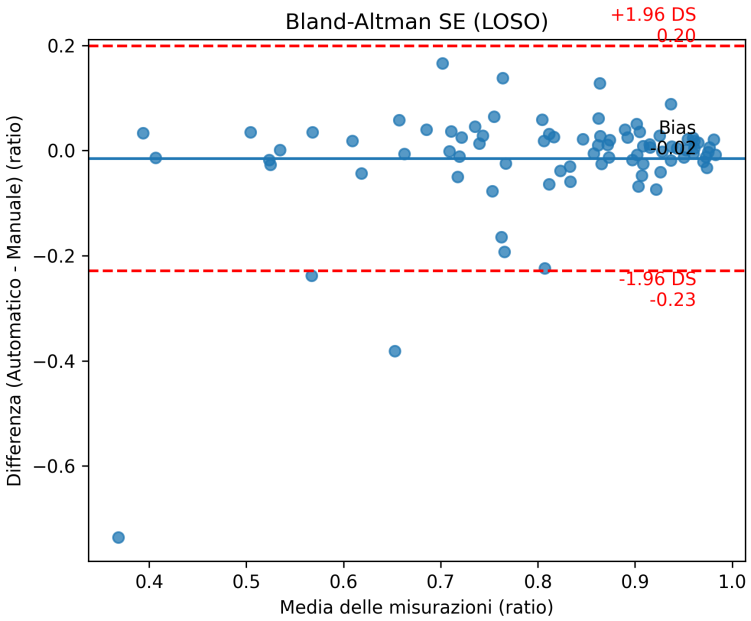

Figura 4.12: Grafico di Bland–Altman per la SE nella validazione LOSO con classificazione in cascata.

I grafici di Bland–Altman relativi alle metriche cliniche nella configurazione LOSO sono riportati nelle Figure 4.9, 4.10, 4.11 e 4.12. In tutti i casi le differenze tra valori stimati e di riferimento risultano distribuite attorno allo zero, indicando l'assenza di un bias sistematico rilevante. Per entrambe le metriche si osserva che la maggior parte delle osservazioni ricade all'interno dei limiti di accordo definiti da ± 1.96 deviazioni standard. Rispetto alla configurazione Train–Validation–Test, si evidenzia tuttavia una maggiore dispersione dei punti, in particolare per il TST, coerente con la maggiore variabilità tra soggetti propria della validazione LOSO.

4.4 Analisi delle prestazioni per gruppo clinico

Per analizzare l'influenza della condizione clinica sulle prestazioni del modello, i soggetti sono stati raggruppati in base alla patologia di appartenenza. L'analisi è stata condotta nella configurazione con validazione LOSO e classificazione in cascata, che ha mostrato le prestazioni globali più elevate.

Per ciascun soggetto è stato considerato il Macro F1-score come metrica riassuntiva delle prestazioni. La Figura 4.13 ne mostra la distribuzione per gruppo clinico: i boxplot evidenziano mediana e variabilità, mentre i punti indicano le prestazioni dei singoli soggetti.

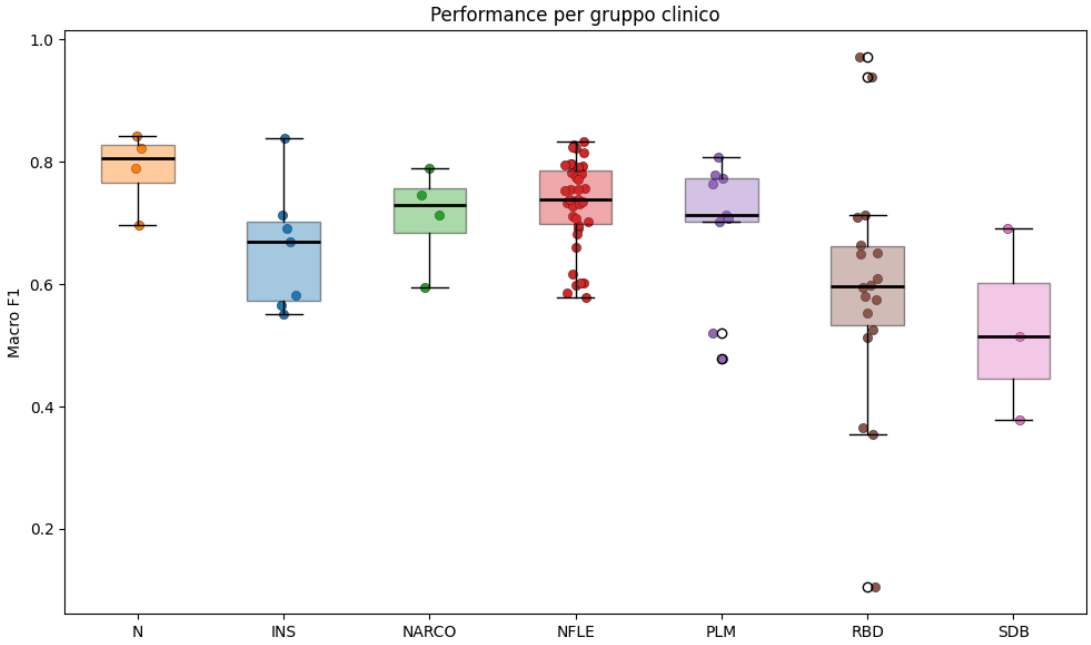

Figura 4.13: Distribuzione del Macro F1-score per gruppo clinico nella validazione LOSO con classificazione in cascata.

Si osserva che il modello raggiunge prestazioni complessivamente buone nella maggior parte dei gruppi clinici, con valori medi generalmente compresi tra 0.65 e 0.80. Tuttavia, alcuni gruppi, in particolare RBD e SDB, mostrano una maggiore variabilità e la presenza di outlier, indicando una maggiore eterogeneità nelle prestazioni tra i soggetti.

Per approfondire tale variabilità, è stata analizzata la distribuzione dell'errore relativo del TST a livello di singolo soggetto. Poiché l'errore relativo della SE risulta direttamente dipendente dalla stima del TST, è stato riportato unicamente quest'ultimo. Per facilitare la lettura dei risultati, i gruppi clinici sono stati distinti in due insiemi: da un lato i gruppi con minore variabilità (N, INS, NARCO, NFLE), dall'altro quelli caratterizzati da maggiore dispersione (PLM, RBD, SDB).

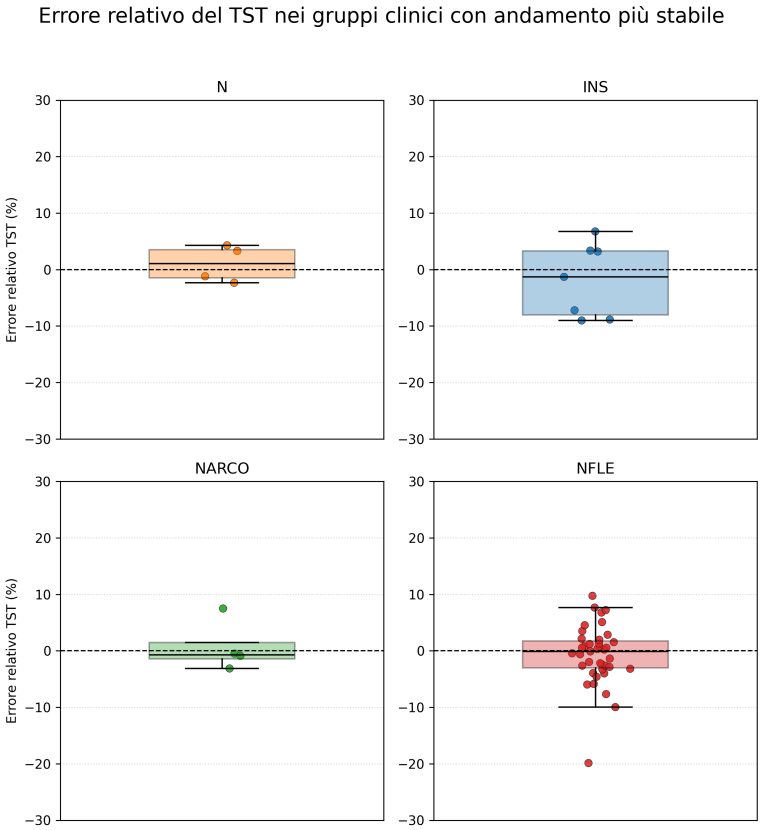

Figura 4.14: Distribuzione dell'errore relativo del TST nei gruppi clinici con minore variabilità.

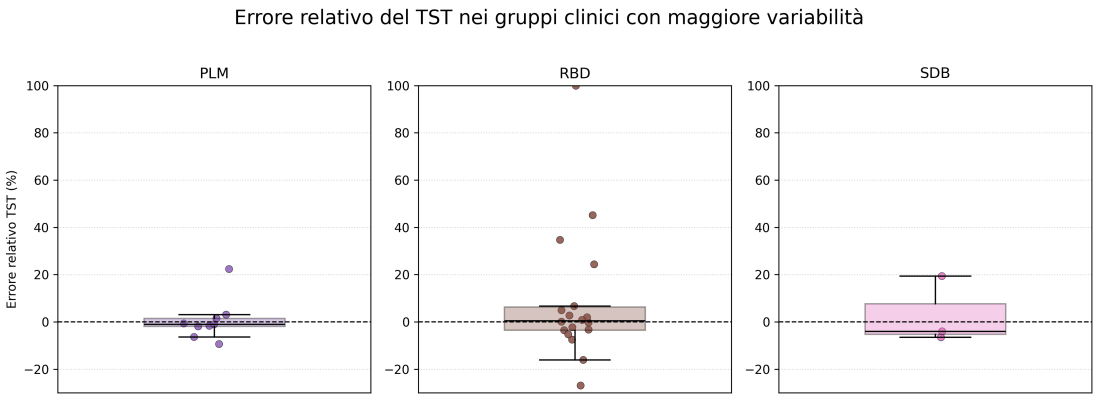

Figura 4.15: Distribuzione dell'errore relativo del TST nei gruppi clinici con maggiore variabilità.

Come mostrato in Figura 4.14, nei gruppi con minore variabilità gli errori risultano generalmente contenuti e distribuiti attorno allo zero, indicando una buona concordanza tra stime automatiche e riferimento manuale.

Al contrario, la Figura 4.15 evidenzia una maggiore dispersione nei gruppi PLM, RBD e SDB. In particolare, il gruppo RBD mostra la presenza di valori estremi, mentre SDB presenta una variabilità più marcata rispetto ai gruppi più stabili.

Per comprendere meglio le origini della variabilità osservata nei diversi gruppi clinici, sono stati analizzati alcuni casi rappresentativi di soggetti che presentano prestazioni particolarmente critiche.

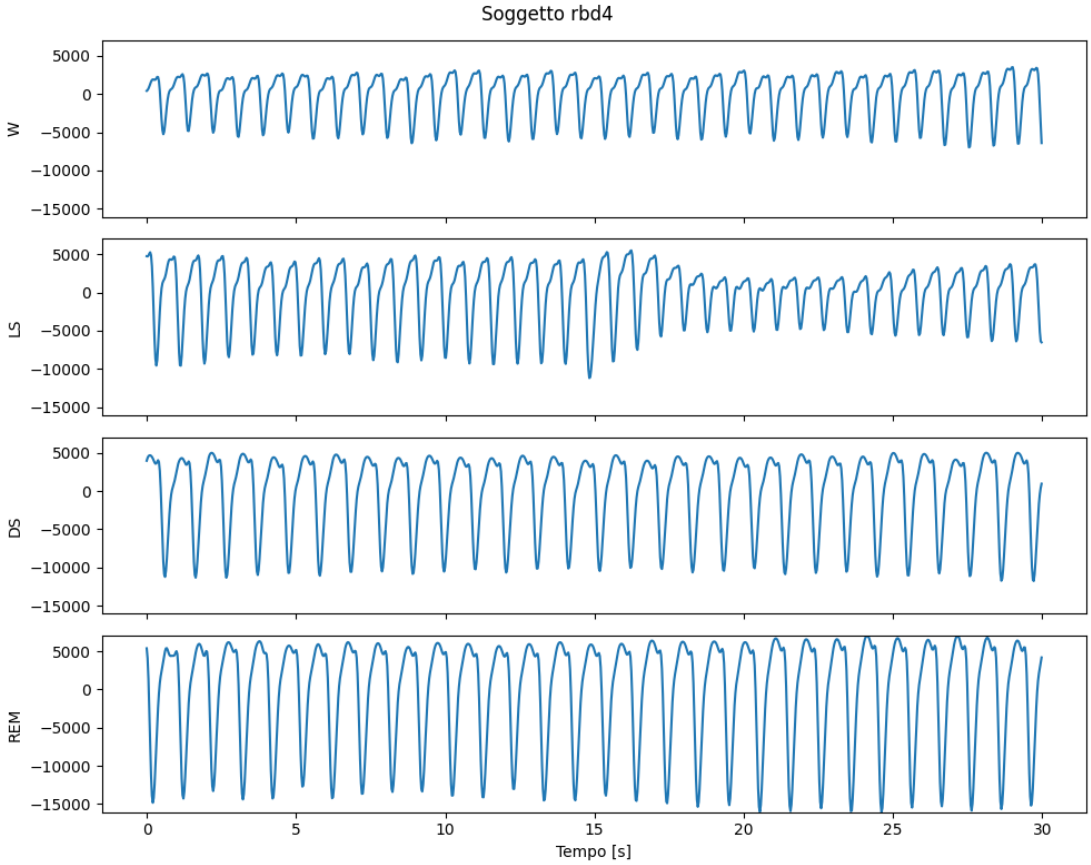

Figura 4.16: Esempio di segnale PPG per il soggetto rbd4 nelle diverse fasi del sonno.

Un primo esempio è riportato in Figura 4.16, relativo a un paziente appartenente al gruppo RBD. In questo caso, il segnale PPG si presenta regolare e poco rumoroso, ma mostra una marcata somiglianza tra le diverse fasi del sonno. Le sequenze associate alle classi W, LS, DS e REM risultano infatti caratterizzate da andamenti molto simili in termini di forma e periodicità, rendendo difficile individuare pattern distintivi utili alla discriminazione tra le classi.

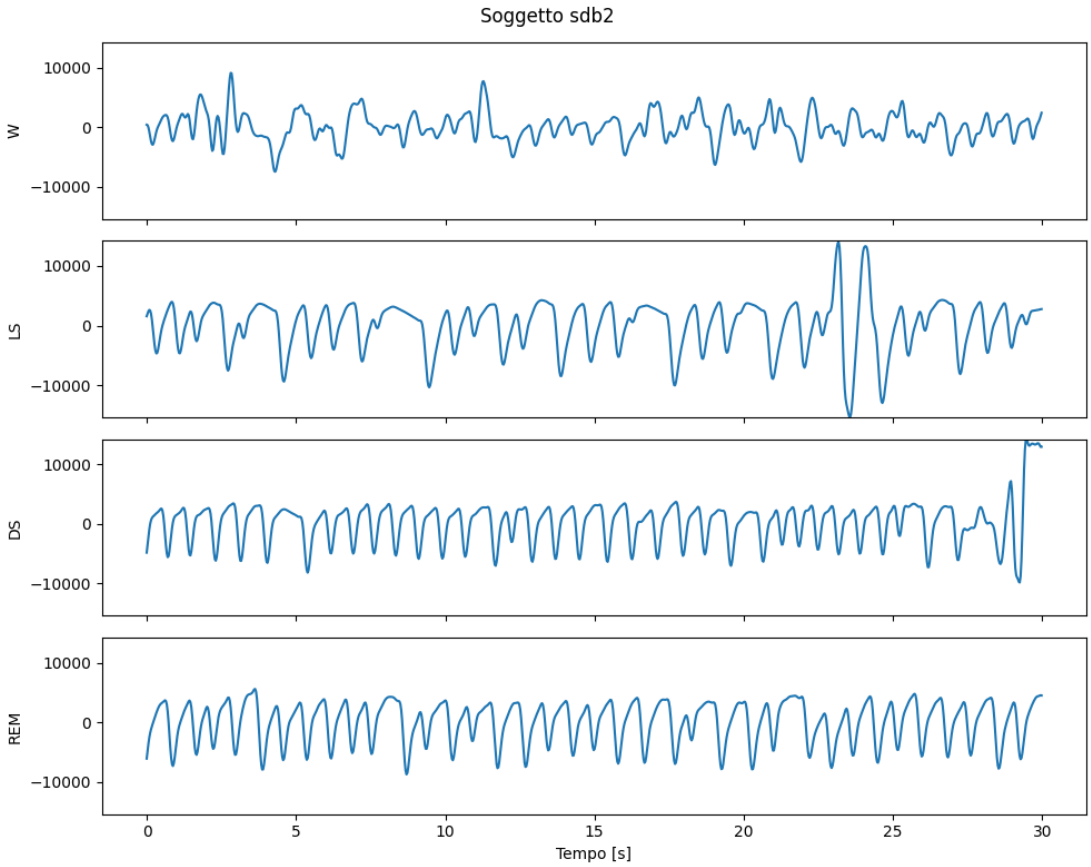

Figura 4.17: Esempio di segnale PPG per il soggetto sdb2 nelle diverse fasi del sonno.

Un secondo esempio è riportato in Figura 4.17, relativo a un soggetto appartenente al gruppo SDB. In questo caso, il segnale risulta caratterizzato da una marcata instabilità, con la presenza di rumore e artefatti evidenti. Si osservano inoltre irregolarità nella forma d'onda e variazioni significative nell'ampiezza del segnale nel corso del tempo, che rendono difficile l'identificazione di pattern consistenti tra le diverse epoche.

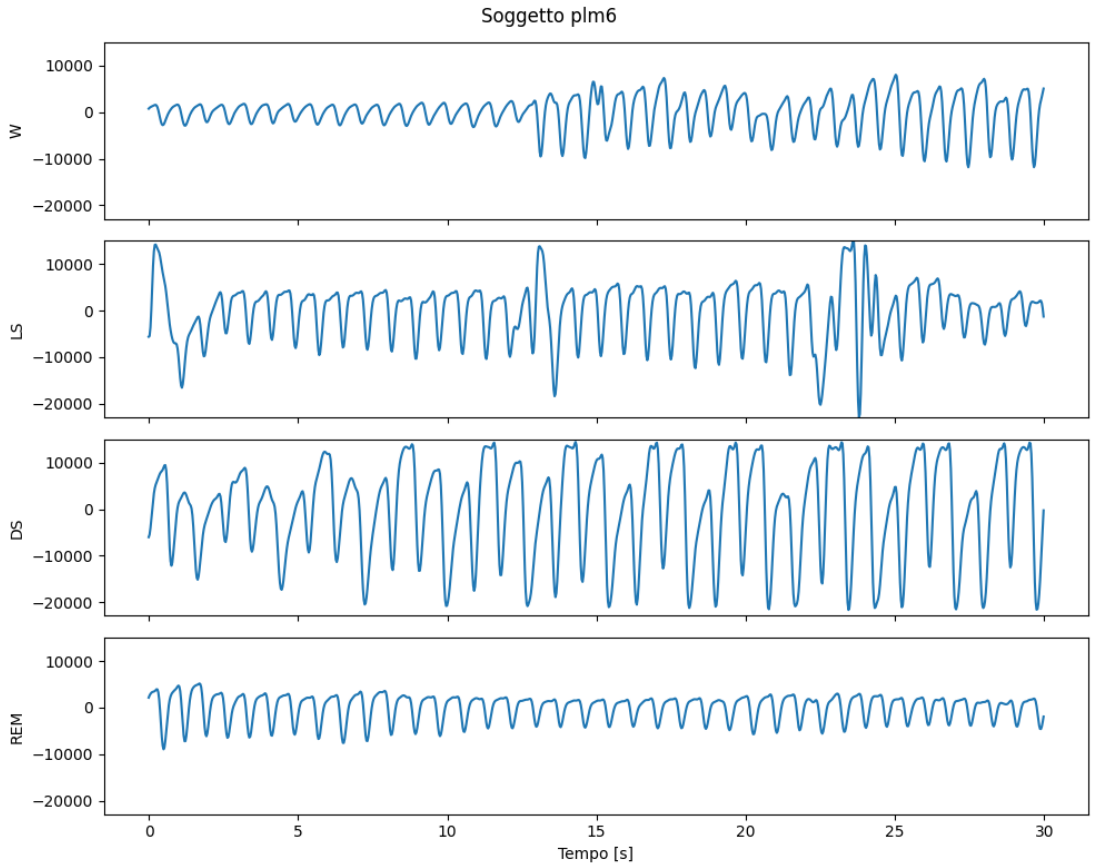

Figura 4.18: Esempio di segnale PPG per il soggetto plm6 nelle diverse fasi del sonno.

Un ulteriore esempio è mostrato in Figura 4.18, relativo a un soggetto del gruppo PLM. In questo caso, il segnale PPG evidenzia una significativa instabilità nel tempo, con cambiamenti nella forma e nell'ampiezza tra diverse porzioni della registrazione, suggerendo la presenza di pattern non stazionari.

Nel complesso, l'analisi dei casi critici evidenzia come le prestazioni del modello possano essere influenzate da diverse caratteristiche del segnale PPG, tra cui la similarità tra le classi, la presenza di rumore e artefatti e la variabilità intra-soggetto. Queste osservazioni risultano coerenti con la variabilità evidenziata nell'analisi per gruppo clinico.

4.5 Discussione dei risultati

In questo paragrafo vengono discussi i risultati ottenuti nei diversi esperimenti, con l'obiettivo di interpretarne il significato e metterli in relazione con quanto riportato in letteratura.

Nel complesso, il modello mostra prestazioni soddisfacenti nel problema dello sleep staging a partire dal segnale PPG, sia nella configurazione Train–Validation–Test sia nella validazione LOSO. In entrambe le configurazioni, le metriche globali indicano una buona capacità discriminativa, con valori di accuratezza e macro F1-score comparabili tra le due strategie di classificazione considerate. La validazione LOSO evidenzia tuttavia una maggiore variabilità tra i diversi soggetti, sottolineando l'importanza di considerare la variabilità inter-soggetto nella valutazione delle prestazioni.

4.5.1 Confronto tra le strategie di classificazione

I risultati ottenuti mostrano che il confronto tra le due strategie di classificazione dipende dalla configurazione di validazione considerata. Nella configurazione Train–Validation–Test, la classificazione diretta a quattro classi presenta prestazioni leggermente superiori, come riportato in Tabella 4.3. Al contrario, nella validazione LOSO la classificazione in cascata mostra valori lievemente migliori (Tabella 4.7).

Questa differenza, sebbene contenuta, suggerisce che l'efficacia delle due strategie sia influenzata dal contesto sperimentale. In particolare, nella validazione LOSO, che espone il modello a una maggiore variabilità tra soggetti, l'approccio gerarchico potrebbe favorire una migliore gestione della complessità del problema.

La classificazione in cascata consente infatti di suddividere il problema in sottocompiti più semplici, permettendo al modello di apprendere rappresentazioni più efficaci per la distinzione tra le diverse fasi del sonno. Questo aspetto risulta particolarmente rilevante nei casi in cui alcune classi presentano caratteristiche fisiologiche simili, come osservato per le fasi LS e DS, come evidenziato anche dalle confusion matrix (Figure 4.1, 4.2, 4.7 e 4.8).

4.5.2 Analisi delle prestazioni per classe

L'analisi delle metriche per classe evidenzia come le prestazioni del modello non siano uniformi tra le diverse fasi del sonno, come riportato nelle Tabelle 4.4, 4.5, 4.8 e 4.9. In particolare, le classi W e DS mostrano valori medi più elevati, mentre LS risulta la classe più difficile da riconoscere.

Questa differenza può essere attribuita alla natura fisiologica delle fasi del sonno: le fasi non REM presentano caratteristiche più simili tra loro, rendendo più complessa la loro discriminazione a partire dal segnale PPG. Di conseguenza,

la confusione osservata tra LS e DS rappresenta un comportamento atteso e coerente con il problema affrontato, come evidenziato anche nelle confusion matrix (Figure 4.1, 4.2, 4.7 e 4.8).

4.5.3 Valutazione delle metriche cliniche

Oltre alle prestazioni di classificazione, i risultati mostrano una buona concordanza tra le stime automatiche e i valori di riferimento per le metriche cliniche considerate (TST e SE), come riportato nelle Tabelle 4.6 e 4.10.

I grafici di Bland–Altman (Figure 4.3, 4.4, 4.5, 4.6, 4.9, 4.10, 4.11 e 4.12) evidenziano un bias medio contenuto e una distribuzione delle differenze centrata attorno allo zero, indicando l'assenza di errori sistematici rilevanti. Tuttavia, nella configurazione LOSO si osserva una maggiore dispersione, in particolare per il TST, suggerendo una variabilità più elevata tra soggetti.

È inoltre importante osservare che l'errore relativo della SE risulta direttamente legato a quello del TST, poiché la SE è calcolata come rapporto tra TST e tempo totale di registrazione. Di conseguenza, l'analisi del TST risulta sufficiente per descrivere il comportamento complessivo delle metriche cliniche.

4.5.4 Influenza della condizione clinica

L'analisi per gruppo clinico evidenzia che le prestazioni del modello non sono uniformi tra le diverse condizioni, come mostrato in Figura 4.13. In particolare, alcuni gruppi, come RBD e SDB, mostrano una maggiore variabilità e la presenza di soggetti con prestazioni significativamente inferiori. Questo comportamento è coerente anche con la maggiore dispersione osservata nell'errore relativo del TST (Figure 4.14 e 4.15).

L'analisi dei casi critici ha permesso di individuare diverse possibili cause di tale variabilità. In alcuni soggetti, come rbd4 (Figura 4.16), il segnale PPG si presenta regolare ma caratterizzato da una elevata similarità tra le diverse fasi del sonno, rendendo difficile l'individuazione di pattern distintivi. In altri casi, come per il soggetto sdb2 (Figura 4.17), la presenza di rumore e artefatti, unita a irregolarità nella forma e nell'ampiezza del segnale, compromette la qualità delle informazioni disponibili per la classificazione. Infine, nel caso plm6 (Figura 4.18), la marcata variabilità intra-soggetto e i cambiamenti nel tempo del segnale suggeriscono la presenza di pattern non stazionari, che rendono più complesso l'apprendimento di rappresentazioni stabili.

Questi risultati suggeriscono che la variabilità osservata nelle prestazioni non è riconducibile unicamente a limiti del modello o a fenomeni di overfitting, ma riflette anche la complessità intrinseca del segnale e delle condizioni cliniche considerate. In particolare, alcune patologie possono influenzare la qualità del segnale PPG: ad

esempio, condizioni associate a movimenti durante il sonno, come RBD e PLM, possono introdurre artefatti e discontinuità nel segnale, mentre disturbi respiratori come SDB possono determinare una maggiore instabilità fisiologica. Questi fattori contribuiscono ad aumentare la variabilità delle caratteristiche del segnale e, di conseguenza, la difficoltà del problema di classificazione.

4.5.5 Confronto con lo stato dell'arte

Il confronto con i principali lavori presenti in letteratura, riportato in Tabella 4.11, mostra che le prestazioni ottenute nel presente lavoro risultano complessivamente in linea con quelle dei metodi basati su segnale PPG per la classificazione a quattro classi.

Tabella 4.11: Confronto tra i principali lavori basati su PPG per lo sleep staging e il modello proposto.

Studio	**Metodo**	**Dataset**	**Validazione**	**Accuracy**
[18]	LightGBM	CAP (27 subj.)	10-fold CV	77.1%
[20]	XGBoost	10 subj. (SDB)	70/30 split (x5)	75.3%
[21]	RF	10 subj. (SDB)	10-fold CV	69.2%
[22]	CNN+RNN	>3000 subj.	Train/Val/Test	74.1%
[24]	InsightSleepNet	CAP (24 subj.)	4-fold CV	80.6%
[25]	SleepPPGNet	>2300 subj.	Train/Val/Test	84%
Questo lavoro	Dual-input SleepPPGNet	CAP (84 subj.)	LOSO	**71.9%**

È tuttavia importante osservare che il confronto diretto tra i diversi studi risulta complesso, a causa delle differenze nei dataset utilizzati, nelle popolazioni considerate e nelle strategie di validazione adottate. In particolare, molti dei lavori presenti in letteratura utilizzano schemi di validazione di tipo hold-out o cross-validation, mentre nel presente lavoro è stata adottata una validazione LOSO, che rappresenta un protocollo più restrittivo e maggiormente orientato alla valutazione della capacità di generalizzazione su soggetti non osservati.

Inoltre, il dataset considerato include soggetti affetti da diverse condizioni cliniche, introducendo un ulteriore livello di variabilità rispetto a dataset più omogenei. Questi aspetti contribuiscono ad aumentare la complessità del problema e devono essere tenuti in considerazione nell'interpretazione delle prestazioni ottenute.

Nel complesso, i risultati ottenuti confermano la fattibilità dello sleep staging a partire dal segnale PPG e risultano coerenti con quanto riportato in letteratura, pur evidenziando le sfide ancora aperte legate alla variabilità inter-soggetto e alla qualità del segnale.

Capitolo 5

Limiti e Sviluppi futuri

5.1 Limiti dello studio

Nonostante i risultati ottenuti dimostrino la validità dell'approccio proposto, il presente lavoro presenta alcune limitazioni che devono essere considerate nell'interpretazione delle prestazioni del modello.

Un primo aspetto riguarda la dimensione e la distribuzione del dataset utilizzato. Il CAP Sleep Database, pur includendo soggetti affetti da diverse patologie del sonno, presenta un numero complessivo di registrazioni relativamente limitato, soprattutto se confrontato con i dataset utilizzati in alcuni lavori recenti basati su DL. Inoltre, la distribuzione delle classi risulta sbilanciata, con una predominanza della fase LS, come evidenziato nella Tabella 3.4. Questo aspetto può influenzare il processo di apprendimento del modello, rendendo più complessa la corretta classificazione delle classi meno rappresentate.

Un ulteriore elemento rilevante è rappresentato dall'elevata variabilità intersoggetto osservata, emersa in modo evidente nella validazione LOSO. Tale variabilità è in parte riconducibile alle differenze fisiologiche individuali, ma risulta fortemente influenzata anche dalla presenza di diverse condizioni cliniche all'interno del dataset. Come discusso nella Sezione 4.4, alcuni gruppi patologici mostrano una dispersione maggiore delle prestazioni e la presenza di soggetti con risultati significativamente inferiori alla media.

In questo contesto, è importante osservare come la variabilità delle prestazioni non sia attribuibile unicamente a limiti del modello o a fenomeni di overfitting, ma sia strettamente legata alla complessità intrinseca del segnale PPG e alle condizioni di acquisizione. In particolare, alcune patologie del sonno sono associate a una maggiore instabilità del segnale, dovuta, ad esempio, alla presenza di movimenti durante il sonno o a una ridotta qualità del contatto tra sensore e pelle nei dispositivi indossabili. Questi fattori possono introdurre artefatti e degradare la qualità del

segnale, rendendo più difficile l'estrazione di caratteristiche informative da parte del modello.

Un'ulteriore limitazione riguarda la natura delle etichette utilizzate per l'addestramento e la valutazione del modello. Le annotazioni degli stadi del sonno sono infatti derivate dall'analisi del segnale EEG, che rappresenta il riferimento clinico per lo sleep staging, mentre il modello opera esclusivamente sul segnale PPG. Di conseguenza, esiste un disallineamento intrinseco tra il segnale utilizzato come input e il processo fisiologico su cui si basano le etichette, che può introdurre una componente di incertezza nel problema di classificazione.

Infine, il modello è stato sviluppato e valutato su un dataset prevalentemente composto da soggetti patologici. Sebbene ciò rappresenti un aspetto rilevante dal punto di vista clinico, limita la possibilità di generalizzare direttamente i risultati a popolazioni più ampie o a soggetti sani acquisiti in contesti diversi.

5.2 Sviluppi futuri

Alla luce delle limitazioni discusse, il lavoro svolto apre diverse prospettive per sviluppi futuri.

In primo luogo, un'estensione naturale consiste nella validazione del modello su dataset più ampi e bilanciati, includenti un numero maggiore di soggetti e una distribuzione più uniforme delle classi. Questo permetterebbe di migliorare la robustezza del modello e di valutarne in modo più approfondito la capacità di generalizzazione.

Un ulteriore sviluppo riguarda l'introduzione di metriche di qualità delle epoche, finalizzate a quantificare l'affidabilità del segnale PPG a livello locale. L'integrazione di tali informazioni potrebbe consentire al modello di pesare diversamente le epoche in base alla loro qualità, riducendo l'impatto di segmenti rumorosi o affetti da artefatti.

Dal punto di vista delle annotazioni, un possibile miglioramento consiste nell'aumentare l'affidabilità delle etichette attraverso strategie di cross-validazione tra operatori. Considerata la variabilità inter- ed intra-operatore nello scoring del sonno, l'utilizzo di annotazioni validate da più esperti potrebbe contribuire a ridurre l'incertezza associata al ground truth.

Un'altra direzione di sviluppo riguarda l'estensione del modello a sequenze di lunghezza variabile. Nel presente lavoro, l'utilizzo di sequenze a lunghezza fissa richiede operazioni di troncamento o padding che possono introdurre una perdita di informazione o una rappresentazione non ottimale delle registrazioni. L'adozione di architetture più flessibili potrebbe consentire di sfruttare in modo più completo l'informazione temporale disponibile.

Infine, un aspetto particolarmente promettente consiste nell'integrazione multimodale di segnali fisiologici. L'aggiunta di informazioni provenienti da altre sorgenti, come segnali di movimento, ECG o parametri respiratori, potrebbe migliorare la capacità discriminativa del modello, in particolare nelle fasi del sonno caratterizzate da maggiore ambiguità fisiologica.

Nel complesso, questi sviluppi potrebbero contribuire a migliorare ulteriormente le prestazioni del modello e a rendere più robusto l'approccio proposto, favorendone l'applicazione in contesti reali e su popolazioni eterogenee.

Capitolo 6

Conclusioni

In questo lavoro di tesi è stato affrontato il problema della classificazione automatica delle fasi del sonno a partire dal segnale PPG, con l'obiettivo di sviluppare un modello basato su DL in grado di ridurre l'invasività del monitoraggio del sonno e automatizzare il processo di sleep staging. In particolare, l'utilizzo del segnale PPG, rappresenta una valida alternativa ai metodi tradizionali, mentre l'impiego di tecniche di IA permette di ridurre il carico di lavoro associato all'annotazione manuale.

I risultati ottenuti mostrano come sia possibile raggiungere prestazioni soddisfacenti utilizzando esclusivamente il segnale PPG, con valori di accuratezza e Macro F1-score complessivamente in linea con quanto riportato in letteratura. Il confronto tra le due strategie di classificazione considerate evidenzia inoltre come l'approccio a cascata e quello diretto a quattro classi presentino comportamenti simili, con differenze contenute che dipendono dal contesto di validazione.

Un elemento centrale emerso dallo studio riguarda tuttavia il ruolo della variabilità tra soggetti e tra condizioni cliniche. In particolare, la validazione LOSO e l'analisi per gruppo clinico hanno evidenziato come le prestazioni del modello non siano uniformi, ma dipendano in modo significativo dalle caratteristiche del segnale e dalla patologia del soggetto. I risultati mostrano infatti una maggiore difficoltà nei gruppi caratterizzati da segnali più variabili o meno distintivi, come nel caso di disturbi associati a movimenti durante il sonno o a instabilità fisiologica.

L'analisi dei casi critici ha ulteriormente confermato questo aspetto, evidenziando come la difficoltà del modello non sia riconducibile unicamente alla sua capacità di apprendimento, ma sia fortemente influenzata dalla qualità e dalla natura del segnale PPG. In presenza di segnali rumorosi, non stazionari o con elevata similarità tra le diverse fasi del sonno, la distinzione tra gli stadi risulta intrinsecamente più complessa.

In questo contesto, anche le metriche cliniche (TST e SE) mostrano una buona concordanza con i valori di riferimento, pur riflettendo la stessa variabilità osservata

a livello di classificazione. Questo risultato conferma che la qualità della stima dipende non solo dal modello, ma anche dalle caratteristiche del segnale di partenza.

Nel complesso, i risultati ottenuti confermano la possibilità di utilizzare il segnale PPG per lo sleep staging automatico, pur evidenziando la presenza di sfide ancora aperte, in particolare legate alla robustezza del modello rispetto alla variabilità dei segnali e alla complessità delle condizioni cliniche. Il lavoro svolto rappresenta quindi un passo verso lo sviluppo di sistemi di monitoraggio del sonno meno invasivi e più facilmente integrabili in dispositivi indossabili, aprendo la strada a ulteriori sviluppi in ambito di sleep medicine e tecnologie per la salute.

Bibliografia

[1] Y. Xie, Y. Zhai e G. Lu. «Evolution of artificial intelligence in healthcare: a 30-year bibliometric study». In: *Frontiers in Medicine* Volume 11 (2025).

[2] Tayab Uddin Wara, Ababil Hossain Fahad, Adri Shankar Das e Md Mehedi Hasan Shawon. «A systematic review on sleep stage classification and sleep disorder detection using artificial intelligence». In: *Heliyon* Volume 11, Issue 12 (2025).

[3] Navya Baranwal, Phoebe K. Yu e Noah S. Siegel. «Sleep physiology, pathophysiology, and sleep hygiene». In: *Progress in Cardiovascular Diseases* Volume 77 (2023).

[4] Bishir Muhammed et al. «Sleep Deprivation and Neurological Disorders». In: *BioMed Research International* (2020).

[5] Sum-Ping O. e Geng YJ. «Impact of Sleep on Cardiovascular Health: A Narrative Review». In: *Heart Mind* Volume 6 (2022).

[6] *International Classification of Sleep Disorders – Third Edition, Text Revision (ICSD-3-TR)*. American Academy of Sleep Medicine. 2023.

[7] A. Rechtschaffen e A. Kales. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington, DC: U.S. Government Printing Office, 1968.

[8] American Academy of Sleep Medicine. *Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Latest version available online. Darien, IL, USA: American Academy of Sleep Medicine, 2023.

[9] V. Muto et al. «Inter- and Intra-expert Variability in Sleep Scoring: Comparison Between Visual and Automatic Analysis». In: *Sleep* 41 (2018).

[10] John Allen. «Photoplethysmography and its application in clinical physiological measurement». In: *Physiological Measurement* 28.3 (2007), R1–R39.

[11] Seungwoo Han, Daeho Roh, Jihyun Park e Hyun Shin. «Design of Multi-Wavelength Optical Sensor Module for Depth-Dependent Photoplethysmography». In: *Sensors* 19.24 (2019).

[12] P Mohan, Nagarajan Velmurugan e J Vignesh. «Spot measurement of heart rate based on morphology of PhotoPlethysmoGraphic (PPG) signals». In: *Journal of medical engineering technology* 41 (set. 2016), pp. 1–10.

[13] Joan Lambert Cause, Ángel Solé Morillo, Juan C. García-Naranjo e Johan Stiens Bruno da Silva. «The Impact of Contact Force on Signal Quality Indices in Photoplethysmography Measurements». In: *Applied Sciences* 14 (giu. 2024), p. 5704.

[14] Yang C., Li B., He Y. e Zhang Y. «LWSleepNet: A lightweight attention-based deep learning model for sleep staging with singlechannel EEG». In: *DIGITAL HEALTH* 9 (2023).

[15] Aozora Ito e Toshihisa Tanaka. «SleepSatelightFTC: A Lightweight and Interpretable Deep Learning Model for Single-Channel EEG-Based Sleep Stage Classification». In: *IEEE Access* 13 (2025), pp. 46263–46272.

[16] Henri Korkalainen et al. «Accurate Deep Learning-Based Sleep Staging in a Clinical Population With Suspected Obstructive Sleep Apnea». In: *IEEE Journal of Biomedical and Health Informatics* 24.7 (2020), pp. 2073–2081.

[17] Fu M, Wang Y, Chen Z, Li J, Xu F, Liu X e Hou F. «Deep Learning in Automatic Sleep Staging With a Single Channel Electroencephalography». In: *Frontiers in Physiology* 12 (2021).

[18] Xiangfa Zhao e Guobing Sun. «A Multi-Class Automatic Sleep Staging Method Based on Photoplethysmography Signals». In: *Entropy* 23 (2021).

[19] M. G. Terzano et al. «Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep». In: *Sleep Medicine* 2.6 (2001), pp. 537–553.

[20] Tasnim Ferdous, Reshad Ul Karim, Abrar Samin, Sammam Mahdi e Aniqua Nusrat Zereen. «Improved Photoplethysmography-Based Four-Stage Sleep Classification with Explainable AI-Driven Machine Learning». In: *2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE)* (2024), pp. 117–122.

[21] F. Smarandache, S. Akula, S. I. Alzahrani, F. Arslan e A. Ijaz. «PPG-Based Sleep Stage Classification Using Pulse Wave Feature Fusion and Explainable AI». In: *Engineering, Technology and Applied Science Research* 15 (ott. 2025), pp. 27640–27645.

[22] Riku Huttunen, Timo Leppänen, Brett Duce, Arie Oksenberg, Sami Myllymaa, Juha Töyräs e Henri Korkalainen. «Assessment of obstructive sleep apnea-related sleep fragmentation utilizing deep learning-based sleep staging from photoplethysmography». In: *Sleep* 44 (ott. 2021).

[23] Kianoosh Kazemi, Arash Abiri, Yongxiao Zhou, Amir Rahmani, Rami N. Khayat, Pasi Liljeberg e Michelle Khine. «Improved sleep stage predictions by deep learning of photoplethysmogram and respiration patterns». In: *Computers in Biology and Medicine* 2 (2024).

[24] Borum Nam, Beomjun Bark, Jeyeon Lee e In Kim. «InsightSleepNet: the interpretable and uncertainty-aware deep learning network for sleep staging using continuous Photoplethysmography». In: *BMC Medical Informatics and Decision Making* 24 (feb. 2024).

[25] Kevin Kotzen, Peter H. Charlton, Sharon Salabi, Lea Amar, Amir Landesberg e Joachim A. Behar. «SleepPPG-Net: A Deep Learning Algorithm for Robust Sleep Staging From Continuous Photoplethysmography». In: *IEEE Journal of Biomedical and Health Informatics* 27.2 (2023), pp. 924–932.

[26] Constantin L. et al. «PPG-Based Sleep Staging Using SleepPPGNet: Extension to Wearables, Improvements, Limitations». In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2024* (lug. 2024).

[27] Jiawei Wang, Yu Guan, Chen Chen, Ligang Zhou, Laurence T. Yang e Sai Gu. «On Improving PPG-Based Sleep Staging: A Pilot Study». In: (lug. 2025). DOI: 10.48550/arXiv.2508.02689.

[28] J. Chung, M. Goodman, T. Huang, M. L. Wallace, P. L. Lutsey, J. T. Chen, C. Castro-Diehl, S. Bertisch e S. Redline. «Multi-dimensional sleep and mortality: The Multi-Ethnic Study of Atherosclerosis». In: *Sleep* 46 (2023).

[29] Ary L. Goldberger et al. «PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals». In: *Circulation* 101.23 (2000), e215–e220.

[30] Jiawei Wang, Yu Guan, Chen Chen, Ligang Zhou, Laurence T Yang e Sai Gu. «On Improving PPG-Based Sleep Staging: A Pilot Study». In: *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2025, pp. 1640–1644.