

Politecnico di Torino

Dipartimento di Ingegneria Meccanica e Aerospaziale
Corso di Laurea Magistrale in Ingegneria Biomedica

Anno Accademico 2025–2026

Rilevamento di parametri da tracciati ECG: Correzione di annotazioni automatiche mediante Deep Learning e Knowledge Distillation

Relatore:
Prof. Vincenzo Randazzo

Candidata:
Maria Chiara Ierovante

Correlatori:
Prof. Eros Gian Alessandro Pasero
Dott. Ing. Alessandro Casella

A papà

Abstract

L'analisi automatizzata dell'elettrocardiogramma (ECG) è da decenni uno strumento fondamentale nella diagnostica cardiologica. Tuttavia, i software industriali attualmente in uso clinico si basano su regole geometriche e algoritmi deterministici che mostrano forti limiti nella segmentazione della ripolarizzazione ventricolare (complesso ST-T). In particolare, convenzioni di annotazione semplificate (come la coincidenza forzata tra l'offset del complesso QRS e l'onset dell'onda T adottata dal software GE MAC2000) portano a un sistematico inglobamento del tratto ST all'interno dell'onda T. Questo errore distorce pesantemente la morfologia restituita all'utente, precludendo a valle indagini cliniche fondamentali come la misurazione del sopraslivellamento o del sottoslivellamento dello ST (marker di ischemia e infarto).

In questo lavoro di tesi viene proposto un approccio basato sul Deep Learning volto a correggere tale distorsione industriale, ottenendo una segmentazione anatomica rigorosa e direttamente applicabile sui tracciati clinici reali ospedalieri. Il problema metodologico primario affrontato in questo scenario è l'assenza di *ground truth* affidabile nei dati di produzione, in quanto tutte le etichette disponibili (fornite dai dispositivi GE) sono intrinsecamente viziate dalla medesima convenzione.

Per superare questo limite, è stata impiegata una rete neurale *fully-convolutional* multicanale (EnhancedUNet) addestrata sfruttando il paradigma della *Knowledge Distillation*: un modello Teacher, istruito su un dataset pubblico validato e annotato manualmente da cardiologi esperti (LUDB), è stato utilizzato per generare delle pseudo-label sui tracciati clinici ospedalieri prodotti dal software GE MAC2000 (dataset PRIV).

Attraverso un rigoroso framework sperimentale composto da 8 fasi, si è dimostrato che la rete neurale è in grado di apprendere e correggere l'errore del software GE se messa nelle condizioni ottimali. L'evidenza principale della ricerca dimostra infatti che la distillazione risulta pienamente efficace se applicata a segmenti ECG sufficientemente lunghi (tracciati da 10 secondi), i quali forniscono alla rete il contesto temporale e la variabilità inter-battito necessari per discriminare il tratto ST dall'onda T, superando i forti limiti riscontrati analizzando il singolo battito

isolato.

Il sistema finale, ingegnerizzato e integrato in un'applicazione prototipale per l'inferenza rapida, ha dimostrato la capacità di annullare il bias geometrico industriale quantificato in -94 millisecondi, innalzando l'F1 score sull'onda T fino al 99.6%. Tale risultato allinea le misurazioni automatiche al reale giudizio clinico umano, aprendo la strada a una diagnostica automatizzata di nuova generazione per le patologie della ripolarizzazione.

Indice

Abstract **2**

Elenco degli Acronimi **10**

1 Introduzione **12**
1.1 Contesto clinico 12
1.2 Il problema dei dati GE MAC2000 13
1.3 Domanda di ricerca e obiettivi della tesi 14
1.4 Struttura della tesi 14

2 Principi di elettrocardiografia e criteri di valutazione **16**
2.1 Anatomia del cuore 16
2.2 Sistema di conduzione elettrica 17
2.3 Il sistema delle 12 derivazioni e l'esplorazione spaziale 18
2.4 Le onde dell'elettrocardiogramma 19
2.5 Principali pattern patologici sull'elettrocardiogramma 22
2.6 Standard di annotazione clinica: AAMI EC57 e tolleranza dei 150 ms 23

3 Fondamenti tecnici **25**
3.1 Reti neurali per la segmentazione di serie temporali biomediche . . 25
3.2 Architetture U-Net, natura Fully Convolutional e processing multi-scala 27
3.3 Meccanismi di attenzione (SE-Attention block) 30
3.4 Dice loss vs metriche F1/Se/P+: differenze architetturali e di scopo 31

4 Architettura e addestramento **33**
4.1 Pipeline di pre-processamento: dall'estrazione dei dati alla generazione dei tensori 33
4.2 Architettura EnhancedUNet: dettagli implementativi e parametri . 35
4.2.1 Il modulo di ingresso multi-scala 35
4.2.2 Il blocco base: struttura residua con SE innestato 37

4.2.3 La progressione dei canali lungo la rete . . . 37
4.2.4 L'upsampling e le skip connections . . . 37
4.2.5 L'output e la deep supervision . . . 39
4.2.6 Complessità e conteggio dei parametri . . . 39
4.2.7 Un confronto con lo stato dell'arte: EnhancedUNet vs. 1D-UNet base . . . 39
4.3 Knowledge distillation: correzione del bias $T_{Onset} = QRS_{Offset}$ . . . 40
4.3.1 Il problema: il bias sistematico delle annotazioni GE . . . 40
4.3.2 Principio della distillazione: teacher-student cross-domain . . . 41
4.3.3 Applicazione al formato beat mediano: Fase KD-MED . . . 42
4.3.4 Applicazione al formato timbro 10s: Fase KD-TMB . . . 42

5 I dataset: LUDB e PRIV **43**
5.1 LUDB: struttura, annotazioni manuali, split train/val/test . . . 43
5.1.1 Origine e composizione del dataset . . . 43
5.1.2 Annotazione manuale: il gold standard . . . 43
5.1.3 Struttura dei dati e formato . . . 45
5.1.4 Split train/val/test . . . 46
5.2 Il dataset PRIV: struttura, bias GE, macro-classi . . . 47
5.2.1 Origine e composizione del dataset . . . 47
5.2.2 Formato XML e parsing . . . 47
5.2.3 Il bias sistematico: T_Onset = QRS_Offset . . . 48
5.2.4 Macro-classi diagnostiche per PRIV . . . 49
5.2.5 Split train/val/test . . . 49
5.3 Limiti della ground truth GE per il tracciato timbro . . . 50
5.3.1 Proiezione dinamica con template Leave-One-Out . . . 50
5.3.2 Falsi positivi nel rilevamento e inaffidabilità su PACING . . . 51
5.4 Confronto tra le due sorgenti di dati e criticità del transfer . . . 53

6 Fasi sperimentali **56**
6.1 Fase B-LUDB: Baseline LUDB (Teacher) . . . 57
6.2 Fase B-MED: Baseline mediano . . . 58
6.3 Fase KD-MED: Distillazione mediano . . . 58
6.4 Fase DIAG-GE: Diagnosi bias GE . . . 59
6.5 Fase B-TMB: Baseline timbro . . . 59
6.6 Fase KD-TMB: Distillazione timbro . . . 60
6.7 Fase COMB: Training combinato . . . 60
6.8 Fase CTRL-TMB: Controllo timbro in-domain . . . 61
6.9 Riepilogo delle configurazioni di addestramento . . . 61

7 Dall'output della rete all'applicazione clinica **63**
7.1 Algoritmo di rilevamento dei picchi (P, Q, R, S, T, J) per macro-classe 63
7.1.1 Complesso QRS . . . 63
7.1.2 Onda P . . . 64
7.1.3 Onda T . . . 65
7.1.4 Punto J e deviazione del tratto ST . . . 65
7.2 Post-processing . . . 67
7.2.1 Pulizia della maschera: battito singolo (mediano) . . . 67
7.2.2 Pulizia della maschera - tracciato multi-battito . . . 67
7.2.3 Correzioni morfologiche puntuali . . . 70

8 Risultati e Applicazione Clinica **71**
8.1 Le quattro modalità di valutazione (Raw, Record pooled, Avg12, Lead II) . . . 71
8.1.1 Raw (per-sample) . . . 72
8.1.2 Record pooled . . . 72
8.1.3 Avg 12 derivazioni . . . 72
8.1.4 Solo Lead II . . . 73
8.1.5 Perché conviene guardare tutte e quattro insieme . . . 73
8.2 Risultati delle fasi sperimentali . . . 73
8.2.1 Confronto tra le quattro modalità di valutazione . . . 73
8.2.2 Confronto Avg12 tra i tre domini di test . . . 75
8.2.3 Confronto per onda: F1, sensibilità e valore predittivo positivo 76
8.2.4 Analisi delle prestazioni delle fasi di riferimento (B-LUDB e B-MED) . . . 78
8.2.5 Quantificazione del bias sistematico industriale (DIAG-GE) . 79
8.2.6 Efficacia della Knowledge Distillation e confronto tra strategie di addestramento . . . 80
8.2.7 Sintesi dei risultati delle singole fasi sperimentali . . . 81
8.3 Errori clinici in millisecondi . . . 82
8.3.1 Perché guardare oltre l'F1 . . . 82
8.3.2 Tabella completa . . . 82
8.3.3 Il caso B-MED: ST_Segment esattamente zero . . . 83
8.3.4 KD-MED: un F1 alto che nasconde un errore reale . . . 83
8.3.5 B-TMB: il bias trasferito, misurato in millisecondi . . . 83
8.3.6 KD-TMB e B-LUDB: precisione quasi identica . . . 83
8.3.7 COMB e CTRL-TMB . . . 84
8.4 Costruzione del Sistema Definitivo . . . 84
8.4.1 Obiettivi e metodologia di fusione . . . 84
8.4.2 La Rete di Controllo GE . . . 85
8.4.3 Selezione dei modelli per l'Ensemble Fisiologico . . . 86

8.4.4 Composizione finale dell'Ensemble . . . 86
8.4.5 Prestazioni dell'Ensemble ed effetto del post-processing . . . 87
8.4.6 Valutazione della separazione ST e relativi limiti . . . 87
8.5 Integrazione nell'app Streamlit . . . 88
8.5.1 Scopo e flusso dell'applicazione . . . 88
8.5.2 Cosa mostra l'app . . . 88
8.5.3 Verifica visiva: Controllo GE vs Ensemble Fisiologico . . . 89

9 Discussione **92**
9.1 Interpretazione dei risultati . . . 92
9.1.1 Perché la distillazione risolve il bias T_Onset . . . 92
9.1.2 Il ruolo del formato esteso nella trasferibilità . . . 93
9.1.3 Perché il training combinato non basta . . . 94
9.2 Risposta alla domanda di ricerca . . . 94
9.2.1 Sostituzione parziale, condizionata alla distillazione . . . 94
9.2.2 Scomposizione per onda . . . 95
9.2.3 Implicazioni per l'applicabilità clinica su larga scala . . . 96
9.3 Limiti dello studio . . . 96
9.3.1 Limiti dell'algoritmo di rilevamento picchi . . . 96
9.3.2 Assenza di ground truth ST su dati GE . . . 96
9.4 Sviluppi futuri . . . 97
9.4.1 Creazione di un ground truth clinico-manuale per lo ST . . . 97
9.4.2 Validazione su coorte multi-paziente più ampia . . . 97
9.4.3 Estensione della distillazione ad altre convenzioni industriali 97
9.4.4 Possibili raffinamenti dell'ensemble . . . 98

10 Conclusioni **99**
10.1 Risposta alla domanda di ricerca e Risultati Raggiunti . . . 99
10.2 Sintesi dei contributi originali . . . 100
10.3 Considerazioni finali . . . 100

Ringraziamenti **102**

Elenco delle figure

1.1 Dispositivo GE MAC2000 13

2.1 Anatomia e sistema di conduzione del cuore 18
2.2 Forme d'onda e segmenti dell'ECG 20

3.1 Rappresentazione schematica di una rete U-Net 1D standard 29

4.1 Schema d'insieme dell'architettura EnhancedUNet 36
4.2 Dettaglio del blocco base `AttBlock` (ResConvBlockSE) 38
4.3 Diagramma a blocchi della Knowledge Distillation 41

5.1 Esempio di annotazione manuale su un tracciato LUDB 45
5.2 Esempio di annotazione automatica GE MAC2000 su un tracciato PRIV 48
5.3 Applicazione del metodo di allineamento (self-template LOO) su HEALTHY e PACING 52
5.4 Confronto tra LUDB, PRIV_medio e PRIV_timbro su pazienti sani 55

7.1 Predizione finale e picchi rilevati, derivazione V5 66
7.2 Predizione finale e picchi rilevati, derivazione aVR 66
7.3 Post-processing morfologico, battito singolo 69
7.4 Post-processing morfologico, tracciato multi-battito 69

8.1 Interfaccia dell'applicazione Streamlit: Rete di Controllo B-TMB . . 90
8.2 Interfaccia dell'applicazione Streamlit: Ensemble Fisiologico 91

Elenco delle tabelle

2.1 Valori di riferimento fisiologici per le durate temporali delle componenti elettrocardiografiche in ritmo sinusale normale. 21

5.1 Corrispondenza macro-classi e keyword diagnostiche per LUDB e PRIV 44
5.2 Composizione del dataset LUDB per macro-classe e split 47
5.3 Composizione del dataset PRIV per macro-classe e split 49

6.1 Sintesi delle fasi sperimentali e delle relative architetture 57
6.2 Riepilogo delle configurazioni dei set di addestramento 62

8.1 F1 Macro nelle quattro modalità di valutazione 74
8.2 Confronto F1 Avg12 tra i tre domini di test 75
8.3 F1 Macro Avg12 sul dominio di riferimento 76
8.4 Scomposizione per onda in Se, PPV e F1 76
8.5 Errore mediano sui bordi delle onde in DIAG-GE 79
8.6 Errore medio e MAE per ciascun bordo d'onda in ms 82
8.7 Errore medio e MAE per intervalli e durate derivate in ms 82
8.8 F1 Avg12 per onda su PRIV_medio nelle sette fasi 85
8.9 Confronto F1 QRS tra B-TMB e CTRL-TMB nei tre domini 86
8.10 Composizione del sistema finale 87
8.11 Prestazioni dell'Ensemble Fisiologico nei tre domini 87

Elenco degli Acronimi

Ambito Medico, Clinico ed Elettrofisiologico

ECG	Elettrocardiogramma
AV	Atrioventricolare
SA	Senoatriale
QRS	Complesso Ventricolare (Onde Q, R, S)
ST	Segmento ST dell'Elettrocardiogramma
QT	Intervallo QT
PR	Intervallo PR
BAV	Blocco Atrioventricolare
STEMI	Infarto Miocardico con Sopraslivellamento del Tratto ST
NSTEMI	Infarto Miocardico senza Sopraslivellamento del Tratto ST

Metriche di Valutazione e Metodologia

TP	Vero Positivo (*True Positive*)
FP	Falso Positivo (*False Positive*)
FN	Falso Negativo (*False Negative*)
PPV	Valore Predittivo Positivo (*Precision*)
Se	Sensibilità (*Recall*)
F1	F1-Score (Media Armonica tra Sensibilità e PPV)
MAE	Errore Assoluto Medio (*Mean Absolute Error*)
LOO	Leave-One-Out

Deep Learning, Intelligenza Artificiale e Informatica

AI	Intelligenza Artificiale (*Artificial Intelligence*)
CNN	Convolutional Neural Network
KD	Knowledge Distillation (Distillazione della Conoscenza)
ReLU	Rectified Linear Unit
XML	Extensible Markup Language
GUI	Interfaccia Grafica Utente (*Graphical User Interface*)

Database e Standard Scientifici

LUDB	Lobachevsky University Electrocardiography Database
WFDB	WaveForm DataBase
AAMI	Association for the Advancement of Medical Instrumentation
CSE	Common Standards for Quantitative Electrocardiography
GE	General Electric (GE Healthcare)
PRIV	Dataset Clinico Proprietario

Capitolo 1

Introduzione

1.1 Contesto clinico e rilevanza della segmentazione automatica dell'ECG

L'elettrocardiogramma a 12 derivazioni è uno degli esami strumentali più diffusi in cardiologia. Non richiede preparazione del paziente, dura pochi secondi e fornisce informazioni immediate su ritmo, conduzione e ischemia. Ogni anno ne vengono registrati centinaia di milioni nel mondo, una mole di dati che nessun team di cardiologi potrebbe rileggere interamente a mano.

La segmentazione automatica del tracciato (cioè l'identificazione dei confini temporali di onda P, complesso QRS e onda T campione per campione) è il primo passo necessario per estrarre in modo algoritmico le misure cliniche fondamentali: durata del PR, larghezza del QRS, intervallo QT, morfologia del segmento ST. Senza una segmentazione affidabile, qualunque analisi a valle riposa su punti fiduciari incerti, e le misure che ne derivano ereditano quell'incertezza.

Negli ultimi anni, le reti neurali profonde hanno raggiunto prestazioni comparabili o superiori a quelle degli annotatori umani su diversi benchmark pubblici di segmentazione ECG [24]. Tuttavia, quasi tutti i risultati pubblicati condividono la stessa dipendenza: un dataset di riferimento annotato a mano da cardiologi, utilizzato sia per addestrare sia per valutare il modello. Quando quel dataset è piccolo, come spesso accade, il modello impara bene il dominio in cui è stato addestrato ma non necessariamente si trasferisce a tracciati provenienti da dispositivi diversi, con convenzioni di annotazione diverse, registrati su popolazioni diverse.

1.2 Il problema: dati clinici GE MAC2000 vs dataset di ricerca annotati manualmente

Questa tesi nasce da una situazione concreta. Da un lato, un dataset di ricerca annotato manualmente, il Lobachevsky University Database (LUDB) [18], con 200 pazienti e annotazioni indipendenti per ciascuna delle 12 derivazioni: piccolo, ma con dati verificati e affidabili. Dall'altro, un archivio clinico di oltre 2400 tracciati prodotti automaticamente da elettrocardiografi GE MAC2000 [13], con annotazioni generate dal software Marquette 12SL [12]: grande, ma con un difetto sistematico.



Figura 1.1: L'elettrocardiografo GE MAC2000, il dispositivo utilizzato in ambito clinico per l'acquisizione dei tracciati.

Tale difetto è preciso e non casuale. Nelle annotazioni GE, l'inizio dell'onda T (*T_Onset*) coincide sempre con la fine del complesso QRS (*QRS_Offset*). Il segmento ST, che in un'annotazione clinica corretta separa i due punti, non ha spazio proprio, ma viene assorbito interamente nell'onda T. Non si tratta di un errore di lettura né di un bug del software, ma di una convenzione adottata dall'algoritmo del dispositivo. Per le onde P e QRS, le annotazioni automatiche si comportano diversamente: l'errore è variabile, non sistematico, e tende a compensarsi con la

numerosità dei dati. Per la T, invece, il problema è strutturale, e più dati con la stessa convenzione non fanno che rafforzarlo.

Un modello addestrato esclusivamente su queste annotazioni imparerebbe a riprodurre fedelmente il bias del GE, non la fisiologia reale del segmento ST. Sarebbe coerente con le proprie etichette di training, ma sistematicamente sbagliato rispetto a un giudizio clinico indipendente.

1.3 Domanda di ricerca e obiettivi della tesi

La domanda che ha guidato questo lavoro è: *i dati clinici GE possono sostituire, in tutto o in parte, i dataset di ricerca annotati manualmente per addestrare un sistema di segmentazione dell'ECG?*

La risposta, come si vedrà, non è binaria. Dipende dall'onda considerata e dalla natura dell'errore presente nell'etichetta automatica. Per arrivarci, la tesi percorre un cammino sperimentale che parte dall'addestramento di una rete di segmentazione (EnhancedU-Net) su entrambi i domini, separatamente e in combinazione, per poi introdurre un meccanismo di distillazione della conoscenza: un modello "teacher", addestrato sul dataset manuale LUDB, genera pseudo-etichette corrette per i tracciati GE, e un modello "student" viene addestrato su queste etichette corrette anziché su quelle originali del dispositivo.

L'obiettivo non è solo dimostrare che la distillazione migliora la segmentazione dell'onda T, ma capire perché funziona, perché il training combinato dei due domini non funziona altrettanto bene, e quale formato di segnale (beat mediano o tracciato esteso) consente alla correzione di trasferirsi ai dati di produzione.

1.4 Struttura della tesi

La tesi è organizzata in dieci capitoli, ciascuno con un ruolo preciso nel percorso dall'inquadramento teorico alla validazione sperimentale.

Il Capitolo 2 introduce i fondamenti di elettrocardiografia necessari a comprendere il problema: l'anatomia del cuore, il sistema di conduzione, le forme d'onda dell'ECG, i principali pattern patologici e lo standard AAMI EC57 che definisce la tolleranza clinica per la valutazione della segmentazione automatica.

Il Capitolo 3 presenta i fondamenti tecnici del deep learning applicato alla segmentazione, dalle architetture encoder-decoder alla funzione di costo di Dice, fino al concetto di knowledge distillation che sarà centrale nel metodo proposto.

Il Capitolo 4 descrive l'architettura EnhancedU-Net utilizzata, il pipeline di addestramento e le scelte implementative.

Il Capitolo 5 è dedicato ai due dataset, LUDB e PRIV: la loro struttura, il bias sistematico dell'annotazione GE e il meccanismo di proiezione delle etichette sul tracciato esteso (self-template LOO).

Il Capitolo 6 documenta le otto fasi sperimentali del progetto, dalla baseline LUDB alla distillazione sul tracciato timbro, passando per il training combinato e il controllo in-domain.

Il Capitolo 7 descrive il passaggio dall'output della rete neurale all'applicazione clinica: l'algoritmo di post-processing che trasforma le maschere di segmentazione in punti fiduciari puntuali (onset, offset, picchi) utilizzabili per le misure cliniche.

Il Capitolo 8 presenta i risultati quantitativi: le tabelle di F1 per onda e per dominio, il confronto tra le fasi sperimentali, la composizione e la validazione dei due ensemble, e l'applicazione interattiva Streamlit sviluppata per l'ispezione visiva dei risultati.

Il Capitolo 9 interpreta i risultati, spiega il meccanismo della distillazione e del fallimento del training combinato, risponde alla domanda di ricerca e ne discute i limiti e gli sviluppi futuri.

Il Capitolo 10 chiude il lavoro con una sintesi dei contributi originali e le considerazioni finali.

Capitolo 2

Principi di elettrocardiografia e criteri di valutazione

2.1 Anatomia del cuore: camere, valvole, miocardio

Il cuore è l'organo muscolare principale dell'apparato circolatorio ed è caratterizzato dalla presenza di quattro camere, due atri superiori e due ventricoli inferiori, separate internamente da una parete muscolare, detta setto [32].

Un apparato valvolare garantisce la rigida unidirezionalità del flusso sanguigno all'interno di queste cavità: le valvole atrioventricolari si aprono in diastole per consentire il riempimento ventricolare e si chiudono in sistole, impedendo il reflusso durante la contrazione [15, 32, 39].

Le arterie coronarie, che si diramano direttamente dall'aorta per decorrere sulla superficie del cuore, garantiscono la perfusione dello stesso. La loro specifica distribuzione topografica fornisce le coordinate spaziali essenziali per l'elettrocardiografia. Nello specifico, l'arteria coronaria destra irrora la parete inferiore e posteriore, l'arteria discendente anteriore provvede alle regioni antero-settale e apicale, mentre l'arteria circonflessa nutre la parete antero-laterale [32].

L'assimilazione di questa precisa geometria vascolare costituisce il fondamento logico per l'analisi ingegneristica del segnale: l'occlusione di un determinato vaso altera i potenziali d'azione in regioni ben delimitate, giustificando in modo diretto la necessità di adottare criteri morfologici strettamente dipendenti dalla singola derivazione (lead-specific) [32].

Dal punto di vista tissutale, la muscolatura cardiaca ha una particolare architettura, nella quale le singole fibre muscolari presentano un'estesa ramificazione e si interconnettono l'una con l'altra, in una configurazione "testa-coda", tramite i dischi intercalari [32].

Questa complessa organizzazione (illustrata nella Figura 2.1) permette alle cellule di accoppiarsi dal punto di vista elettrico, formando una fitta rete comunicante tridimensionalmente [3]. Di conseguenza, l'intero miocardio opera come un unico grande blocco funzionale: l'onda di depolarizzazione non avviene in modo isolato procedendo rigidamente cellula per cellula, ma si propaga quasi istantaneamente in maniera fluida e continua attraverso il tessuto, garantendo così una contrazione meccanica altamente sincronizzata [32].

Questa connettività elettrica forma la base fisiologica su cui agisce il sistema di conduzione del cuore, rendendo possibile la rapida diffusione dell'impulso che si traduce graficamente nel segnale registrato dall'elettrocardiogramma [3, 32].

2.2 Il sistema di conduzione elettrica – nodo SA, nodo AV e rete di His-Purkinje

Il sistema di conduzione elettrica (schematizzato in Figura 2.1) ha il compito di iniziare l'impulso del cuore e garantire la contrazione ritmica e sincronizzata di atri e ventricoli [26, 32].

Questo sistema è composto da due principali tipologie di tessuto: i nodi, caratterizzati da una conduzione lenta e dalla generazione dell'impulso, e i componenti a conduzione rapida del sistema di conduzione ventricolare [27].

L'impulso elettrico presenta una propagazione che segue una sequenza anatomica rigida: l'onda di attivazione origina nel nodo seno-atriale (SA), viaggia attraverso le vie internodali fino a raggiungere il nodo atrioventricolare (AV), per poi immettersi nel fascio di His, biforcarsi nelle branche destra e sinistra e infine distribuirsi ai ventricoli tramite le fibre del Purkinje [26, 32, 40].

Il nodo SA funge da pacemaker fisiologico primario. Tale nodo possiede la frequenza intrinseca di depolarizzazione spontanea più elevata, andando a imporre il proprio ritmo all'interno del cuore e impedendo così alle altre stazioni di conduzione di generare un proprio ritmo autonomo [26].

Dopo che l'onda di depolarizzazione ha attraversato il tessuto atriale, converge verso il nodo AV, il quale rallenta la propagazione dell'impulso elettrico prima di lasciarlo proseguire verso i ventricoli, permettendo al segnale di attraversare il tessuto nei tempi corretti [32, 40].

L'onda di depolarizzazione, attivata nel nodo SA e propagata attraverso il miocardio, determina le deflessioni del tracciato [3]. Il transito e il ritardo dell'impulso attraverso il nodo AV e il fascio di His si traducono clinicamente nell'intervallo PR; un blocco della conduzione in queste regioni causa un progressivo allungamento o un prolungamento di tale intervallo oltre i 200 millisecondi. Subito dopo, la rapi-

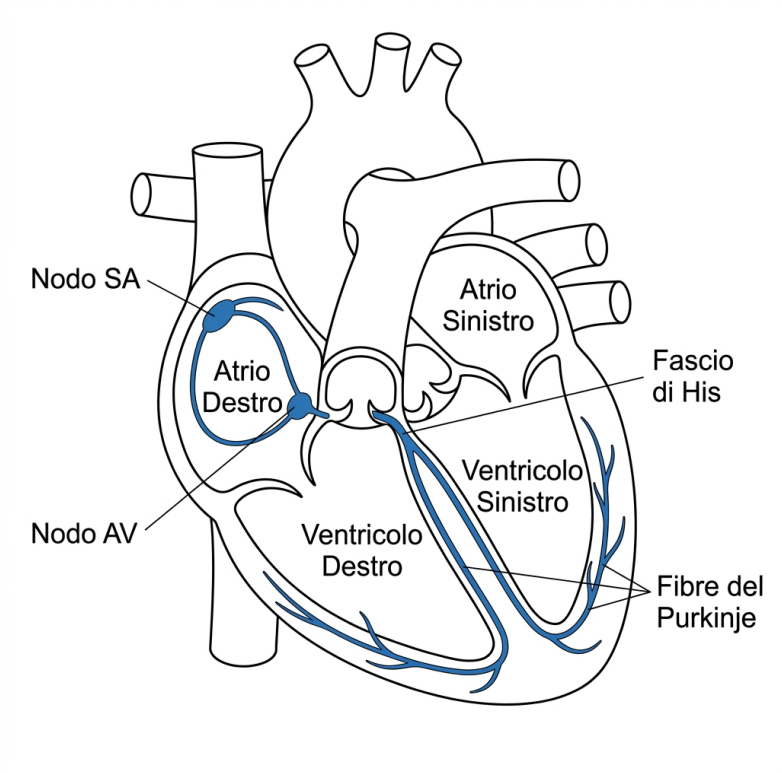


Figura 2.1: Rappresentazione onnicomprensiva dell'anatomia cardiaca. La figura illustra contestualmente sia le strutture muscolari macroscopiche e cellulari (paragrafo 2.1), sia le componenti del sistema di conduzione elettrica (nodo SA, nodo AV, fascio di His e fibre del Purkinje) sovrapposte al miocardio (paragrafo 2.2).

da attivazione e la successiva ripolarizzazione del miocardio ventricolare generano rispettivamente il complesso QRS e l'onda T, racchiusi nell'intervallo QT [32].

2.3 Il sistema delle 12 derivazioni e l'esplorazione spaziale

L'elettrocardiogramma clinico standard non si limita a registrare l'attività elettrica del cuore da un'unica prospettiva, ma utilizza un sistema di 12 derivazioni per ottenere una rappresentazione spaziale tridimensionale dell'intero miocardio. Questo approccio è indispensabile poiché il fronte d'onda elettrico è un vettore che possiede magnitudo e direzione spaziale [32].

Le 12 derivazioni diagnostiche vengono matematicamente ricavate generalmente a partire da soli 10 elettrodi fisici posizionati sul paziente: quattro sugli arti

(braccio destro, braccio sinistro, gamba sinistra e gamba destra come massa) e sei sul torace [32].

Le tre derivazioni bipolari periferiche (I, II, III) misurano la differenza di potenziale tra due specifici elettrodi degli arti, andando a formare il cosiddetto triangolo di Einthoven [19]. Dalle stesse misurazioni si ricavano, in modo matematicamente aumentato, le tre derivazioni unipolari periferiche (aVR, aVL, aVF). Nel complesso, queste 6 derivazioni esplorano l'attività cardiaca sul piano frontale.

Le 6 derivazioni precordiali (V1-V6) sono anch'esse unipolari e utilizzano i singoli elettrodi posizionati sul torace come polo positivo. Come polo di riferimento (zero teorico) impiegano invece il Terminale Centrale di Wilson (WCT), ottenuto dalla media matematica dei tre potenziali degli arti. Queste derivazioni sono deputate ad analizzare il vettore elettrico sul piano orizzontale [19].

A causa di questa diversa geometria di osservazione, lo stesso evento elettrico viene registrato con morfologie differenti a seconda del punto di vista. Ad esempio, un vettore che si avvicina all'elettrodo esplorante produrrà una deflessione positiva (come un'onda T positiva in V5), mentre lo stesso vettore produrrà una deflessione negativa se osservato da un elettrodo posizionato nel lato opposto (come il complesso negativo in aVR) [19, 32].

La necessità di analizzare un elettrocardiogramma a 12 derivazioni risiede nella capacità di localizzare spazialmente le anomalie fisiopatologiche. Un infarto miocardico acuto, ad esempio, manifesterà alterazioni evidenti del segmento ST e dell'onda T unicamente nelle derivazioni anatomicamente contigue alla regione ischemica, lasciando le altre tracce invariate o soggette a modificazioni speculari [19]. Questa disomogeneità spaziale giustifica, a livello ingegneristico, la necessità di sviluppare architetture neurali in grado di analizzare il tracciato preservando l'indipendenza delle singole derivazioni.

2.4 Le onde dell'elettrocardiogramma

L'onda di depolarizzazione attivata nel nodo SA e propagata attraverso il miocardio determina le deflessioni caratteristiche del tracciato di superficie (riassunte visivamente in Figura 2.2), valutabili in modo non invasivo [2].

La prima deflessione del ciclo cardiaco è l'onda P, la quale caratterizza la depolarizzazione del miocardio atriale ed è innescata dal nodo seno-atriale (SA) [23]. Il fronte d'onda inizia nell'atrio destro, diffondendosi successivamente all'atrio sinistro. La porzione ascendente dell'onda P riflette prevalentemente l'attivazione dell'atrio destro, mentre la porzione discendente rappresenta quella dell'atrio sinistro [23].

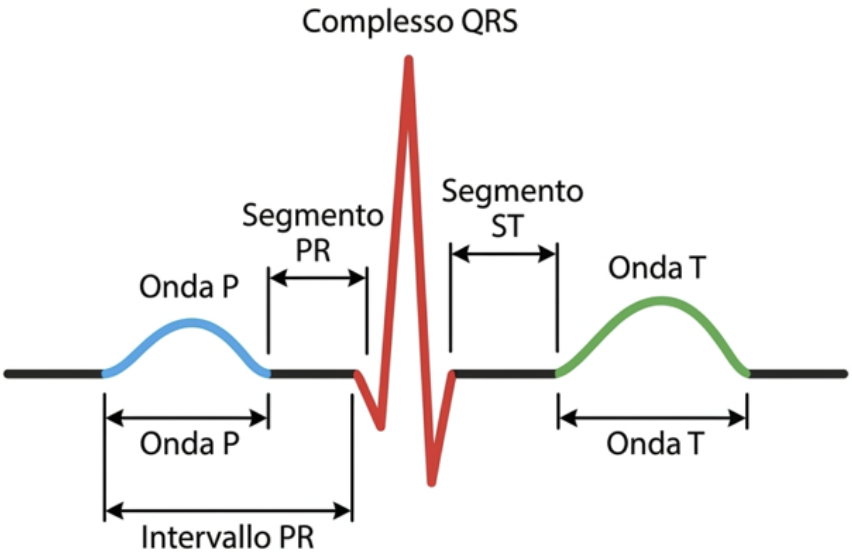


Figura 2.2: Rappresentazione schematica di un singolo battito cardiaco in ritmo sinusale normale. Sono evidenziate cromaticamente le principali deflessioni (Onda P, Complesso QRS, Onda T) e i segmenti isoelettrici di interesse clinico trattati nel testo (Intervallo PR, Segmento PR, Segmento ST).

Se non sono presenti ritardi di conduzione interatriale, l'onda P ha una durata tipica tra i 70 e i 110 ms [23]. L'onda è considerata patologica se la durata raggiunge o supera i 120ms [23, 25].

In ritmo sinusale normale, l'onda P è positiva nella derivazione II, mentre nella derivazione V1, si presenta spesso come bifasica (con una prima deflessione positiva e una seconda negativa), poiché riflette l'attivazione sequenziale dell'atrio destro seguita dall'atrio sinistro [21, 23].

Dopo l'onda P, è possibile individuare il complesso QRS, il quale riflette la depolarizzazione dei ventricoli [7]. La massa muscolare ventricolare è molto maggiore di quella atriale: di conseguenza, la sua depolarizzazione genera il segnale elettrico più ampio registrato sull'ECG [2].

Il complesso è tipicamente strutturato dal susseguirsi delle onde Q (la deflessione negativa iniziale), R (la prima deflessione positiva) e S (la deflessione negativa terminale), la cui morfologia, ampiezza e polarità variano in base alla derivazione clinica esplorata [2, 8]. La durata considerata normale per il complesso QRS è compresa tra i 75 e i 105 ms [2].

La ripolarizzazione del miocardio ventricolare la si può osservare nella successiva onda T, che conclude il ciclo elettrico. A differenza del QRS, l'onda T dura più

a lungo, perché la ripolarizzazione è un processo cellulare più lento: non segue un sistema di conduzione rapido come quello di depolarizzazione, ma si diffonde gradualmente da una cellula all'altra attraverso le giunzioni comunicanti [7]. L'onda T ha una durata media di circa 160 ms [35].

In un cuore sano, l'onda T è tipicamente asimmetrica e concordante con il complesso QRS, risultando cioè positiva nella maggior parte delle derivazioni in cui l'onda R è dominante [9].

Il complesso QRS si unisce all'onda T tramite il segmento ST [8]. A livello cellulare, tale finestra temporale corrisponde alla fase di plateau del potenziale d'azione ventricolare, periodo nel quale la maggior parte delle cellule miocardiche risulta uniformemente depolarizzata [8].

La durata media normale del segmento ST è compresa tra gli 80 e i 120 ms [32].

Componente ECG	**Durata normale (ms)**	**Significato elettrofisiologico**
Onda P	$70 - 110$	Depolarizzazione del miocardio atriale
Segmento PR	~ 80	Linea isoelettrica di base
Intervallo PR	$120 - 200$	Viaggio dell'impulso negli atri e ritardo al nodo AV
Complesso QRS	$75 - 105$	Depolarizzazione dei ventricoli
Segmento ST	$80 - 120$	Fase di plateau del potenziale d'azione ventricolare
Onda T	~ 160	Ripolarizzazione del miocardio ventricolare

Tabella 2.1: Valori di riferimento fisiologici per le durate temporali delle componenti elettrocardiografiche in ritmo sinusale normale.

Il tempo che intercorre dall'inizio della depolarizzazione atriale (l'inizio dell'onda P) fino all'inizio della depolarizzazione ventricolare (l'inizio del complesso QRS) è definito come intervallo PR [10, 32]. Questo intervallo include il tempo necessario affinché l'impulso viaggi attraverso gli atri e subisca il fisiologico ritardo di conduzione all'interno del nodo atrioventricolare (AV) [32, 38].

In condizioni fisiologiche, la durata normale dell'intervallo PR è compresa tra i 120 e i 200 millisecondi [32].

La linea isoelettrica che si trova tra la fine dell'onda P e l'inizio del complesso QRS è il segmento PR. Questo segmento è fondamentale poiché costituisce la linea di base di riferimento dell'intero ECG, a partire dalla quale si calcolano eventuali deviazioni (elevazioni o depressioni) del successivo segmento ST [10].

La durata normale del solo segmento PR è di circa 80 ms [35].

Al fine di fornire un riferimento rapido per le successive fasi di delineazione automatica del segnale, la Tabella 2.1 riassume i valori fisiologici di durata per le principali deflessioni e i segmenti isoelettrici del ciclo cardiaco appena trattati in questo paragrafo.

2.5 Principali pattern patologici sull'elettrocardiogramma

L'attività elettrica del cuore si riflette sulla superficie corporea in modo ordinato e prevedibile. Per questo motivo, la maggior parte delle alterazioni patologiche produce sull'elettrocardiogramma pattern ben riconoscibili e ricorrenti.

Un primo gruppo di alterazioni riguarda la nascita e la propagazione dell'impulso elettrico lungo il sistema di conduzione: cosa succede quando l'impulso non nasce correttamente, non si trasmette con regolarità, o percorre vie anomale. In esso possiamo individuare:

- **Assenza di onda P e ritmi caotici:** quando il nodo seno-atriale non genera l'impulso, o quando l'attivazione atriale diventa rapida e disorganizzata, l'onda P perde la sua morfologia fisiologica o scompare del tutto. È il caso della fibrillazione atriale, con il suo ritmo irregolare, ma anche dei ritmi di scappamento, in cui è un pacemaker più lento a guidare il battito [19].

- **Blocchi atrioventricolari:** se l'impulso rallenta o si blocca lungo il suo percorso verso i ventricoli, il normale ritardo al nodo AV si altera. Si va da un semplice allungamento dell'intervallo PR oltre i 200 ms (blocco di primo grado) fino alla completa dissociazione tra atri e ventricoli, dove nessun impulso riesce più a passare (blocco di terzo grado) [19, 35].

- **Allargamento del complesso QRS:** se l'impulso supera il nodo atrioventricolare ma trova un blocco lungo le branche del fascio di His, i ventricoli non si depolarizzano più insieme e velocemente attraverso le fibre di Purkinje. Il risultato è un complesso QRS allargato, di solito oltre i 120 ms (il segno tipico dei blocchi di branca destro o sinistro, o di altri ritardi della conduzione intraventricolare) [19, 35].

- **Ritmi da stimolazione artificiale:** quando la conduzione è gravemente compromessa o la frequenza cardiaca scende troppo, un pacemaker artificiale può guidare il ritmo al posto del cuore [35]. Sul tracciato si riconosce facilmente: prima di ogni battito compare un breve *spike* di stimolazione, seguito spesso da un complesso QRS slargato e dalla forma insolita [19, 35].

Un secondo raggruppamento riguarda invece le alterazioni intrinseche della morfologia dell'onda, legate a un danno strutturale, ischemico o anatomico del tessuto miocardico. In questi casi, l'anomalia non risiede nella via percorsa dall'impulso, bensì nello stato patologico del muscolo stesso:

- **Necrosi ed esiti di infarto:** un vecchio infarto lascia una cicatrice fibrosa, elettricamente inerte; da quella zona non parte più alcun impulso, e le forze elettriche vicine si allontanano senza incontrare opposizione. Questo lascia sull'ECG il segno con onde Q patologiche (durata ≥ 40 ms, ampiezza oltre il 25% dell'onda R successiva) [19, 35].
- **Danno ischemico acuto e anomalie della ripolarizzazione:** quando il cuore va in ischemia, mancano ossigeno e si accumulano scorie acide: le fasi più delicate del battito cellulare, il plateau e la ripolarizzazione, ne risentono per prime [35]. Questo genera le cosiddette "correnti di lesione", che spostano il segmento ST dalla sua linea di base. Sull'ECG si vede un sopraslivellamento se il danno attraversa tutta la parete (infarto STEMI), o un sottoslivellamento se resta più superficiale; a volte, invece, è l'onda T a invertirsi in modo anomalo e asimmetrico [19, 35].
- **Ipertrofia atriale e ventricolare:** quando le camere cardiache si ingrossano oltre il normale, generano forze elettriche più ampie del solito. Per i ventricoli, questo si vede sull'ECG come un aumento dei voltaggi del QRS; per gli atri, come una P dalla forma alterata (le classiche P mitrale o P polmonare). Nei ventricoli, agli alti voltaggi si accompagna spesso un pattern di *strain*: alterazioni secondarie del tratto ST e dell'onda T, conseguenza diretta dell'ipertrofia [19, 35].

A fare da riferimento per tutte queste alterazioni resta il tracciato normale: ritmo sinusale, tempi di conduzione regolari, voltaggi nella norma. Rientrano qui anche le varianti benigne legate all'età o al tono vagale, come la bradicardia o la tachicardia sinusale [19].

2.6 Standard di annotazione clinica: AAMI EC57 e tolleranza dei 150 ms

Lo standard ANSI/AAMI EC57 rappresenta il documento di riferimento internazionale per la validazione degli algoritmi di misurazione del ritmo cardiaco. Tale protocollo permette di utilizzare delle linee guida rigorose per poter valutare le prestazioni dei sistemi di rilevamento automatico nell'ambito dell'elaborazione del

segnale ECG, andando a stabilire dei criteri uniformi per l'identificazione dei battiti e la delineazione delle specifiche forme d'onda [20].

I segnali ECG hanno una variabilità naturale da un battito all'altro, oltre a un certo rumore di fondo e per questo una valutazione punto per punto è troppo rigida e poco realistica. Il passaggio tra la fine di un'onda e l'inizio della successiva è graduale, non netto; individuare con esattezza dove finisce un'onda (offset) e dove inizia quella dopo (onset) comporta quindi un margine di incertezza. Per gestire questo limite, in letteratura si usa comunemente una "finestra di tolleranza temporale". Nello specifico, è considerata clinicamente accettabile una discrepanza nella localizzazione dei punti fiduciari se ricade all'interno di un margine di ± 150 ms rispetto all'annotazione manuale di riferimento [6].

Si tratta di una convenzione mutuata storicamente dalle raccomandazioni CSE (Common Standards for Quantitative Electrocardiography) del 1985 e successivamente adottata come riferimento congiunto allo standard AAMI EC57 nell'analisi automatica [6].

Considerando un segnale campionato a 500 Hz, questa tolleranza temporale si traduce operativamente in $\pm$ 75 campioni. Se l'algoritmo posiziona un bordo entro questo intervallo rispetto all'annotazione di riferimento (la media dei tre cardiologi, come nel caso del dataset LUDB), la predizione viene considerata un vero positivo (True Positive). Se la distanza supera i 150 ms, la predizione viene marcata come falso positivo (False Positive) [6].

Questa distinzione concettuale tra rigidità matematica e tolleranza clinica si riflette direttamente nell'architettura di questo lavoro, imponendo una netta separazione tra la fase di addestramento (training) e quella di valutazione (evaluation). Durante l'addestramento della rete neurale, è strettamente necessario utilizzare una funzione di costo matematicamente differenziabile per consentire l'aggiornamento dei pesi tramite la backpropagation; per questo motivo, viene impiegato il coefficiente di Dice, che quantifica in modo rigido la sovrapposizione geometrica punto per punto tra la predizione e la ground truth. Tuttavia, in fase di validazione, il Dice risulta inadeguato a descrivere l'effettiva utilità clinica del modello. Pertanto, la valutazione reale del sistema si basa sulle metriche derivate dalle regole di tolleranza clinica: Sensitivity (Se), Valore Predittivo Positivo (PPV) e la loro media armonica, l'F1-score. Le implicazioni matematiche e le differenze architetturali tra la loss di Dice e le metriche di F1 verranno approfondite nel dettaglio nel successivo paragrafo 3.4 [6].

Capitolo 3

Fondamenti tecnici

Come evidenziato nel capitolo precedente, la complessa morfologia del segnale ECG e la necessità di rispettare rigorose tolleranze cliniche richiedono il superamento dei rigidi algoritmi tradizionali a favore di approcci data-driven. Questo capitolo illustra le fondamenta teoriche e architetturali su cui si basa il modello di deep learning sviluppato in questo lavoro. Si partirà dalla riformulazione del problema medico in un task di *sequence labeling*, per poi esplorare le architetture convoluzionali (U-Net) e i meccanismi di attenzione che permettono di estrarre e ricalibrare le feature temporali.

3.1 Reti neurali per la segmentazione di serie temporali biomediche

La segmentazione dell'ECG può essere formulata concettualmente come un problema di classificazione "punto-per-punto", tipicamente noto come dense labeling o sequence labeling. In tale paradigma, a ogni singolo campione temporale del segnale viene assegnata un'etichetta corrispondente a una specifica classe elettrofisiologica [34]. Questo approccio è formalmente identico alla segmentazione semantica largamente impiegata nel campo della visione artificiale, ma viene applicato e processato lungo una serie temporale unidimensionale (1D) anziché su una griglia di pixel 2D. L'inquadramento del task in questi termini costituisce il ponte teorico fondamentale che giustifica il riadattamento ai segnali temporali di potenti architetture nate originariamente per le immagini.

Per anni, il rilevamento dei punti chiave dell'ECG si è affidato a metodi di elaborazione classici: filtri adattivi, calcolo delle derivate, o soprattutto la trasformata di Wavelet. Tuttavia, anche se questi algoritmi riescono a isolare in maniera adeguata le alte frequenze del complesso QRS, se applicati a forme d'onda a bassa ampiezza (come la P o la T) si presentano limiti notevoli; dipendono tipicamente

da complessi sistemi di soglie tarate manualmente e risultano particolarmente vulnerabili al rumore di fondo, alla fisiologica variazione morfologica e ai cambiamenti nella frequenza di campionamento dell'hardware [22].

Per poter superare la rigidità di queste regole, è necessario il passaggio a un approccio basato sul deep learning e guidato dai dati, lasciando che sia l'algoritmo stesso a imparare i pattern morfologici diagnostici direttamente dai dati, anche in presenza di rumore. Al centro di questo cambiamento ci sono le reti neurali convoluzionali (CNN), qui applicate tramite convoluzioni 1D, la forma più naturale per lavorare su un segnale come l'ECG. La loro prima caratteristica utile è l'invarianza per traslazione: la rete riconosce un'onda P o un'onda T per la sua forma, indipendentemente da quando compare lungo il tracciato. C'è però un secondo requisito: per collocare ogni onda nel contesto del battito cardiaco, la rete ha bisogno di un campo recettivo, cioè di una finestra temporale, abbastanza ampio da "vedere" cosa succede intorno a ogni campione [1].

Questa operazione si traduce, in pratica, in una classificazione a quattro classi: la rete produce quattro canali di uscita, uno per ciascuna categoria (background, onda P, complesso QRS, onda T). Ogni canale ha la stessa lunghezza del segnale in ingresso e, campione per campione, restituisce quanto la rete è sicura che quel punto appartenga alla sua classe. Una funzione finale (la softmax) prende queste quattro stime e le trasforma in probabilità che competono tra loro; così, per ogni campione, vince una sola classe [34].

La letteratura recente ha dimostrato ampiamente la solidità di questo approccio, stabilendo un nuovo stato dell'arte nella delineazione automatica dell'ECG. Le architetture a risoluzioni gerarchiche, in particolare, riescono a superare i limiti dei metodi classici. Un lavoro di riferimento in questo senso è quello di Moskalenko et al. (2020): una rete completamente convoluzionale ispirata alla U-Net, capace di adattarsi a qualsiasi frequenza di campionamento e di generalizzare su diversi tipi di monitor, che raggiunge risultati notevoli nell'individuare l'inizio e la fine di ogni onda. Un importante salto di qualità è stato reso possibile grazie alla pubblicazione del Lobachevsky University Electrocardiography Database (LUDB): un database ad accesso aperto con tracciati a 12 derivazioni, i cui limiti e picchi delle onde sono stati annotati a mano da cardiologi esperti. È diventato in poco tempo lo strumento di riferimento per sviluppare e validare i nuovi delineatori basati sul deep learning [18].

Un benchmark indipendente recente (Park et al., 2025) ha confermato che le CNN restano superiori, anche rispetto ad architetture più recenti come i Transformer, sia in accuratezza che in capacità di generalizzare. Lo stesso studio, però, segnala un problema ricorrente: davanti a morfologie patologiche atipiche, le reti tendono a produrre segmenti frammentati, cioè spezzano un'onda in più pezzi invece di riconoscerla come un blocco unico. È un aspetto tutt'altro che secondario:

è proprio questa tendenza a giustificare le logiche di post-processing che verranno introdotte nel Capitolo 4, pensate apposta per ricomporre queste frammentazioni. Un altro aspetto interessante emerso dallo studio è che nel campo ECG, reti più grandi non significa reti migliori, conta di più trovare un buon equilibrio tra dimensione e capacità di generalizzare [28]. C'è però un limite di fondo nelle CNN "semplici". Le convoluzioni 1D sono efficaci nel riconoscere pattern che si ripetono nel tempo, ma una CNN pensata solo per la classificazione comprime progressivamente l'informazione nei suoi livelli più profondi, e in questo processo perde la posizione esatta di ogni campione. Per annotare con precisione al millisecondo non basta riconoscere un pattern ma bisogna anche saper dire esattamente dove si trovava nel segnale originale. Serve quindi un'architettura capace di fare entrambe le cose: estrarre l'informazione e poi ricostruirla nella sua posizione originale. È esattamente il problema che la rete U-Net, la quale verrà trattata nel paragrafo 3.2, nasce per risolvere.

3.2 Architetture U-Net, natura Fully Convolutional e processing multi-scala

Come visto nel paragrafo 3.1, un problema di *dense labeling* con questa granularità richiede un'architettura che sappia fare due cose insieme: astrarre le caratteristiche del segnale, ma anche ricollocarle esattamente dove si trovavano nel tempo. La risposta più naturale a questa doppia esigenza è la U-Net. Introdotta da Ronneberger, Fischer e Brox (2015) per la segmentazione di immagini biomediche, prende il nome dalla sua forma a "U": un percorso che scende (l'*encoder*), comprimendo il segnale e astraendone via via le caratteristiche, e uno che risale (il *decoder*), che lo ricostruisce fino a riportarlo alla risoluzione originale [31].

C'è un problema, però: un semplice encoder-decoder non basta a garantire questo livello di dettaglio. Comprimendo tutto nel collo di bottiglia centrale, la rete impara benissimo il contesto generale, ma perde per strada i dettagli più fini, proprio quelli che servono per individuare, campione per campione, dove inizia e dove finisce ciascuna onda. La U-Net risolve la problematica con le *skip connections*: dei collegamenti diretti che saltano il collo di bottiglia e portano le informazioni ad alta risoluzione direttamente dall'encoder al livello corrispondente del decoder. Il risultato è che la rete tiene insieme due cose: da un lato riconosce di che onda si tratta, dall'altro sa dire con precisione dove comincia e dove finisce [5, 36].

Passare dalle immagini bidimensionali alle serie temporali richiede però un adattamento. I principi restano gli stessi, ma le convoluzioni e il *max-pooling*, da 2D, diventano 1D, applicati lungo l'unico asse che conta qui, quello del tempo [5]. C'è un aspetto importante da sottolineare: questa architettura è a tutti gli effetti

una rete completamente convoluzionale (FCN). A differenza delle CNN pensate per la classificazione, non ha strati densi (*fully-connected*) nella parte finale; al loro posto, ha convoluzioni puntuali. Questo cambia le cose, poiché il modello smette di dipendere da una dimensione di input fissa, e può processare sequenze lunghe quanto serve, semplicemente facendo scorrere i *kernel* 1D lungo il segnale [1].

Per l'*EnhancedUNet* sviluppata in questo lavoro, questo si traduce in un vantaggio pratico: lo stesso identico modello, con gli stessi pesi e lo stesso file di *checkpoint*, può segmentare un singolo battito mediano di 600 campioni, un frammento LUDB di 3000, oppure un tracciato "timbro" intero da 5000 campioni (10 secondi), senza bisogno di ritoccare nulla. È una versatilità che tornerà utile nel Capitolo 6.

C'è poi un altro elemento centrale nell'*EnhancedUNet*: il modo in cui gestisce le diverse scale temporali del segnale fin dall'ingresso. Per fare ciò, il modulo iniziale riceve il segnale in parallelo su tre rami — il segnale intero, una versione dimezzata ($\times 2$) e una ridotta a un quarto ($\times 4$). Non è una scelta casuale: l'ECG è per natura un segnale composito, dove i pattern si muovono su scale temporali molto diverse tra loro. Riconoscere un'onda P, o un QRS stretto, richiede occhio per i dettagli più rapidi e transitori; leggere l'intero ciclo cardiaco, invece, richiede uno sguardo più ampio, orientato alle componenti lente. Queste tre rappresentazioni vengono immediatamente processate e concatenate (previo riallineamento all'identica lunghezza temporale), fondendo le diverse risoluzioni prima ancora di iniziare la discesa. Da questo punto in poi, la rete prosegue con una singola struttura encoder-decoder a 4 livelli di profondità. Ogni volta che si scende di un livello nell'encoder, la risoluzione temporale del segnale si dimezza, e con essa, raddoppia il campo recettivo dei filtri convoluzionali che seguono. È una gerarchia a cascata: più si scende, più i filtri riescono a "vedere" finestre temporali ampie dell'ECG, cogliendo le dipendenze a lungo raggio. E grazie alle *skip connections*, questo non incide sulla precisione: i dettagli fini, garantiti dai livelli più superficiali, restano comunque disponibili [5, 36].

Il *processing* multi-scala è potente, ma anche una FCN o una U-Net "base" ha il suo limite: tratta tutti i canali delle *feature map* allo stesso modo, senza distinguere quanto ciascuno conti davvero [1, 5]. Il problema è che, in pratica, non tutte le caratteristiche estratte hanno lo stesso peso diagnostico in ogni istante del ciclo cardiaco: alcune contano di più, altre meno, a seconda del momento. Serve quindi un modo per pesare dinamicamente l'importanza di ciascun canale. Ed è esattamente qui che entrano in gioco i meccanismi di attenzione, come lo *SE-Attention block*, trattato nel paragrafo 3.3 [36].

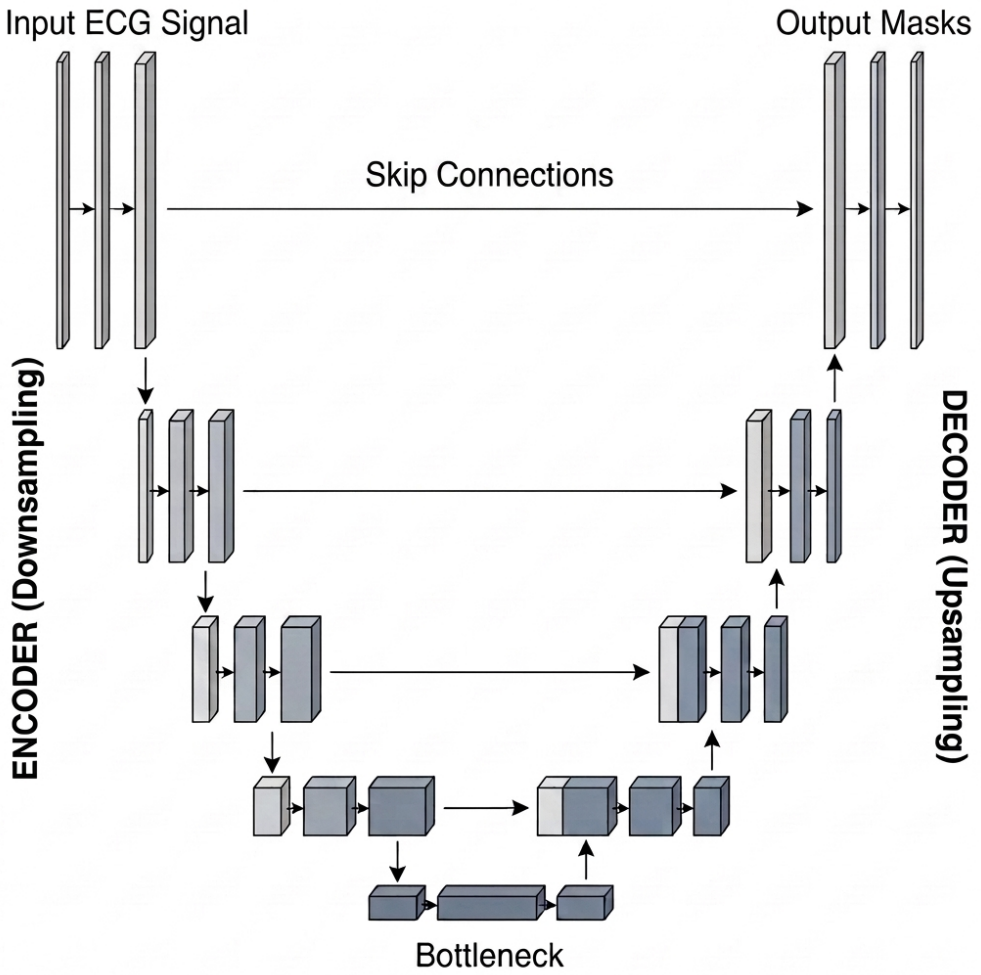


Figura 3.1: Rappresentazione schematica di una rete U-Net 1D standard per il sequence labeling. Questa topologia di base funge da fondamenta architetturale per il modello sviluppato. Sono evidenziati il segnale ECG in ingresso (Input), il percorso di contrazione (Encoder), il Bottleneck centrale e il percorso di espansione (Decoder) che porta alla predizione delle maschere probabilistiche (Output). Le Skip Connections (rappresentate dalle frecce orizzontali) trasferiscono le feature ad alta risoluzione direttamente dall'encoder al decoder dello stesso livello, permettendo il riposizionamento esatto dei pattern d'onda nel tempo.

3.3 Meccanismi di attenzione (SE-Attention block)

Come accennato alla fine del paragrafo 3.2, le reti convoluzionali hanno un limite di fondo: elaborano le mappe di feature in modo omogeneo. L'output nasce da una somma su tutti i canali, le cui dipendenze restano intrecciate tra loro; di fatto, la rete tratta ogni canale con lo stesso peso, senza distinguere quanto ciascuno conti davvero. Eppure, modellare esplicitamente queste interdipendenze aiuterebbe non poco: permetterebbe alla rete di dare più risalto alle feature realmente informative e mettere in ombra quelle meno utili. È necessario quindi un meccanismo di attenzione channel-wise, ed è qui che entra in gioco il blocco Squeeze-and-Excitation (SE), introdotto nel lavoro di Hu et al. [17].

Il blocco SE ha un obiettivo preciso: ricalibrare in modo adattivo la risposta di ogni feature. Per farlo, lavora in due passaggi:

- **Squeeze (compressione)**: i filtri convoluzionali vedono solo il proprio campo recettivo locale; non possono, da soli, cogliere cosa succede fuori da quella finestra, soprattutto nei primi livelli della rete. Lo squeeze risolve proprio questo: aggrega l'informazione spaziale (o temporale, nel caso dell'ECG 1D) di un intero canale in un unico "descrittore" globale. In pratica, basta un Global Average Pooling lungo l'asse del tempo, che riduce tutte le statistiche del canale a un solo valore scalare [17].

- **Excitation (eccitazione)**: questi valori aggregati passano poi attraverso un meccanismo di self-gating, pensato per ricalibrare l'importanza di ciascun canale. Per tenere sotto controllo la complessità e favorire la generalizzazione, il meccanismo è organizzato a collo di bottiglia: un primo strato fully-connected riduce la dimensionalità, una ReLU introduce la non linearità, e un secondo strato fully-connected la riporta alla dimensione originale. Una sigmoide finale trasforma tutto in coefficienti di importanza, uno per canale. Questi pesi vengono infine moltiplicati alle mappe di feature originali, così i canali più rilevanti vengono esaltati, e il rumore silenziato [17].

Nell'EnhancedUNet proposta in questo lavoro, si è scelto di adottare questo blocco proprio per i suoi comprovati benefici: migliora sensibilmente le prestazioni a un costo computazionale che resta quasi irrilevante [17]. Nella pratica, i blocchi SE fanno parte di moduli di attenzione più ampi (AttBlock), distribuiti lungo tutta la rete. L'attenzione entra in gioco dopo ogni convoluzione, per ricalibrare i canali, e questo avviene sistematicamente in tutti i 9 livelli principali dell'architettura: i 4 blocchi dell'encoder, il bottleneck centrale (che da solo concentra più del 33% dei parametri totali ed è il punto in cui l'astrazione è massima) e i 4 blocchi del decoder.

È inoltre doveroso precisare un dettaglio architetturale rilevante: all'interno di ogni singolo blocco di attenzione, a valle del meccanismo SE, è presente una connessione residua (*shortcut*) che somma direttamente le *feature map* in ingresso all'output ricalibrato. Questo meccanismo additivo facilita notevolmente la propagazione del gradiente durante l'addestramento, mitigando i problemi tipici delle reti profonde.

3.4 Dice loss vs metriche F1/Se/P+: differenze architetturali e di scopo

Fin qui si è esplorato come la rete impara a rappresentare le caratteristiche morfologiche e a segmentare complessi QRS, onde P e onde T. Ma sapere come funziona non basta: bisogna anche stabilire come misurarne l'accuratezza, per capire se il modello è davvero valido. Ed è qui che emerge una dicotomia di fondo nella delineazione automatizzata: da un lato serve una sovrapposizione geometrica rigida per addestrare la rete, dall'altro una valutazione orientata agli eventi, che rispecchi la tolleranza clinica reale [5, 28].

Durante l'addestramento, serve una funzione di costo continua per aggiornare i pesi tramite la discesa del gradiente. Nei task di segmentazione semantica basati su reti convoluzionali, come la U-Net, la rete classifica ogni singolo campione, confrontando le proprie mappe predette con le annotazioni di riferimento (*ground truth*) [28]. La letteratura offre diverse metriche spaziali per farlo, ma in questo lavoro si è scelto il coefficiente di Dice, definito per una singola classe c come:

$$DSC_c = \frac{2 \sum_i p_{i,c}\, g_{i,c}}{\sum_i p_{i,c} + \sum_i g_{i,c}} \tag{3.1}$$

dove $p_{i,c} \in [0, 1]$ è la probabilità predetta dalla rete (dopo softmax) che il campione i appartenga alla classe c, e $g_{i,c} \in \{0, 1\}$ è il valore corrispondente nella maschera *one-hot* di *ground truth*. La funzione di costo minimizzata durante il training, mediata sulle $C = 4$ classi (background, P, QRS, T), è quindi:

$$\mathcal{L}_{Dice} = 1 - \frac{1}{C} \sum_{c=1}^{C} DSC_c \tag{3.2}$$

Non è una scelta casuale: la Dice loss gestisce bene lo sbilanciamento delle classi, un problema molto concreto nei tracciati ECG, dove i campioni di background dominano numericamente su quelli, ben più rari, delle singole onde [5].

La Dice loss è essenziale per far convergere la rete durante l'addestramento, ma come metrica di validazione clinica non basta: non sa distinguere un errore accettabile da uno grave. Questo perché calcola una sovrapposizione geometrica,

e quindi penalizza ogni discrepanza in modo continuo, senza riconoscere nessuna soglia netta che separi "giusto" da "sbagliato" dal punto di vista diagnostico [5,28]. Per colmare questo vuoto, la letteratura si appoggia a standard internazionali come l'ANSI/AAMI EC57, che introducono l'idea di una finestra di tolleranza, un modo per decidere, in termini binari, se un algoritmo ha misurato correttamente il ritmo cardiaco o i segmenti dell'ECG. Nel caso specifico della validazione dei modelli di delineazione rispetto alle annotazioni umane, la finestra tipicamente usata è di ± 150 ms [5, 19].

In pratica, un punto di onset o offset individuato dalla rete conta come vero positivo (TP) solo se cade entro questa finestra di tolleranza rispetto all'annotazione di riferimento. Se la deviazione è maggiore, si conta un falso positivo (FP); se l'onda viene mancata del tutto, un falso negativo (FN) [5, 19]. Su questa base si costruiscono le due metriche cardine della valutazione: la *Sensitivity* (Se), che dice quante onde reali la rete è riuscita a individuare,

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

e il Valore Predittivo Positivo (PPV), che dice quanto ci si può fidare di ciò che la rete ha effettivamente segnalato,

$$PPV = \frac{TP}{TP + FP} \tag{3.4}$$

Per riassumere entrambe in un solo numero, lo standard è l'F1-score, calcolato come la loro media armonica:

$$F1 = 2 \cdot \frac{Se \cdot PPV}{Se + PPV} \tag{3.5}$$

[5, 36].

Addestramento e validazione, dunque, usano metriche diverse per un motivo preciso: la Dice loss è uno strumento matematico, pensato per essere differenziabile e ottimizzare la rete; F1, PPV e Se, invece, raccontano quanto il modello funziona davvero, secondo gli standard AAMI [36].

A questo punto il quadro teorico è completo, ed è il momento di passare dalla teoria alla pratica. Il Capitolo 4 entra nel cuore tecnico del lavoro, descrivendo l'intera spina dorsale del sistema: l'estrazione e il pre-processamento del segnale clinico grezzo, i dettagli strutturali dell'architettura EnhancedUNet e, infine, le strategie di addestramento adottate per correggere e distillare la conoscenza medica. Le fasi sperimentali e le successive operazioni di *post-processing* (che convertono le maschere della rete in referti diagnostici) saranno invece oggetto dei capitoli successivi.

Capitolo 4

Architettura e addestramento

4.1 Pipeline di pre-processamento: dall'estrazione dei dati alla generazione dei tensori

Come visto alla fine del Capitolo 3, per valutare l'architettura in modo oggettivo, e misurarne le prestazioni con metriche di tolleranza clinica, servono dati allineati con precisione temporale assoluta. Il pre-processamento è il primo anello di questa catena: il suo compito è trasformare dati clinici grezzi ed eterogenei in tensori standardizzati, pronti per essere dati in pasto alla rete.

La pipeline di questo lavoro è stata pensata per gestire due sorgenti dati profondamente diverse tra loro. Da una parte c'è LUDB: uno strumento di validazione open-access, con 200 tracciati a 12 derivazioni di 10 secondi, campionati a 500 Hz, corredati di annotazioni manuali indipendenti per onda P, complesso QRS e onda T, validate da cardiologi esperti [18]. Dall'altra c'è PRIV, il dataset clinico privato, generato direttamente dagli elettrocardiografi commerciali GE MAC2000. Come riportano le specifiche ufficiali del dispositivo, la macchina acquisisce di norma tracciati a riposo, a 12 derivazioni, della durata standard di 10 secondi, esportati, per l'archiviazione esterna, nel formato proprietario non compresso DCAR XML, con frequenza di campionamento nativa di 500 Hz o 1000 Hz a seconda della configurazione [13]. La composizione clinica delle due coorti sarà affrontata nel Capitolo 5: qui l'obiettivo è un altro, affrontare le sfide di estrazione di questo dato.

Il *parsing* dei file DCAR XML si è rivelato un passaggio ingegneristico tutt'altro che scontato. Come segnala la letteratura, General Electric non pubblica una documentazione tecnica ufficiale per il proprio formato XML; chi lo usa deve dedurre da sé il significato dei campi e la struttura gerarchica dei dati, basandosi solo sui nomi degli identificatori [30]. In questo lavoro, la difficoltà principale è stata la mutabilità dei *namespace* XML; a seconda del dispositivo o della versione *firmware*,

i file GE dichiarano namespace diversi: dalla versione `urn:ge:sapphire:dcar_1` fino a varianti più recenti come `sapphire_3`. Per questo l'algoritmo di *parsing* interroga più namespace in sequenza, con un meccanismo di *fallback* dinamico. Inoltre, un'altra complicazione risiede nel fatto che lo stesso campo diagnostico, l'inizio dell'onda P, ad esempio, può comparire sotto nomi diversi (`"Pon"` oppure `"P_Onset"`) dentro i sotto-elementi del blocco `medianTemplate`.

Oltre all'estrazione testuale, c'è stata un'altra insidia tecnica da risolvere: l'interpretazione temporale delle annotazioni. Nei file XML GE, i valori posizionali (il tag `toc`) non indicano un tempo assoluto in millisecondi, ma indici di posizione espressi in campioni nativi del dispositivo. Un errore di conversione qui avrebbe disallineato in modo irreparabile le maschere di segmentazione rispetto al segnale. Per evitarlo, indipendentemente dal fatto che il file fosse campionato a 500 Hz o 1000 Hz, si è resa necessaria una costante nel codice di parsing: `anchor_ms=500`, un valore di riferimento a cui ancorare la posizione di ogni battito individuato nel tracciato. Tenendo questo valore fisso, invece di ricalcolarlo in base alla frequenza di campionamento, la pipeline garantisce che ogni annotazione clinica (espressa in ms) venga proiettata esattamente sulla posizione del battito corrispondente, qualunque sia il *sample rate* nativo del file. Una volta risolte queste due criticità (namespace variabile e allineamento temporale) l'estrazione dei dati grezzi può dirsi completa; il passo successivo è organizzarli nei formati di input concreti che la rete andrà a processare.

Terminata l'estrazione, la pipeline organizza il segnale nei formati di input che serviranno alla rete, ciascuno dimensionato in base alla propria lunghezza in campioni. Sono tre:

- **Beat mediano (600 campioni)**: un singolo battito rappresentativo, isolato direttamente dal `medianTemplate` del sistema GE. È un formato "di produzione", compatto, che rispecchia fedelmente le convenzioni di refertazione clinica dei software commerciali.
- **Timbro (5000 campioni)**: la traccia grezza continua, estratta dalla registrazione nativa di 10 secondi, che a 500 Hz corrisponde esattamente a 5000 campioni, lo stesso formato standard usato nelle annotazioni di LUDB. È uno scenario clinico molto più realistico e complesso: qui vengono a galla le limitazioni delle convenzioni proprietarie, come quella coincidenza un po' forzata tra la fine del QRS (`QRS_Offset`) e l'inizio dell'onda T (`T_Onset`).
- **Crop dinamico (3000 campioni)**: un formato intermedio, pensato apposta per dare flessibilità e regolarità all'addestramento, ottenuto ritagliando i tracciati LUDB.

Prima di arrivare in input al modello, i tensori estratti passano per un'ultima fase di pulizia e normalizzazione. Il segnale grezzo porta con sé alcuni disturbi

tipici dell'acquisizione ECG, fluttuazioni della linea isoelettrica (*baseline wander*) e interferenza di rete a 50Hz, che vengono rimossi con un filtro passa-alto e un filtro *notch* dedicato. A quel punto, ogni canale viene standardizzato con una normalizzazione Z-score: si sottrae la media e si divide per la deviazione standard, portando il segnale a media nulla e varianza unitaria, un passaggio che serve a uniformare le forti variazioni di ampiezza tipiche dell'acquisizione clinica. Parallelamente, la pipeline applica criteri di scarto rigorosi: i canali "piatti" o i segmenti gravemente corrotti da artefatti vengono esclusi in automatico dall'addestramento, per non "inquinare" l'apprendimento della rete.

A questo punto, il processo di estrazione e standardizzazione può dirsi concluso e i tensori del segnale (X) e le rispettive maschere target (Y) sono pronti, strutturati in tutti i formati necessari.

4.2 Architettura EnhancedUNet: dettagli implementativi e parametri

Una volta completata la pipeline di pre-processamento, che trasforma i tracciati clinici grezzi nei tensori standardizzati X e Y, il passo successivo è dare forma concreta all'architettura di rete. Riprendendo i concetti teorici del Capitolo 3, la topologia dell'EnhancedUNet prende corpo attraverso una configurazione precisa di filtri, convoluzioni e strati operativi: scelte pensate per massimizzare la precisione della delineazione, tenendo però sotto controllo l'onere computazionale (lo schema d'insieme è riportato in Figura 4.1).

4.2.1 Il modulo di ingresso multi-scala

Il segnale non entra nella rete in modo diretto. Prima passa attraverso un modulo di ingresso multi-scala (`MultiScaleInput`), che espande il singolo canale in ingresso fino a un numero di filtri base fissato a $f = 12$. Questa espansione si divide in tre rami (*branch*) paralleli, ciascuno responsabile di produrre 4 canali (12/3). Il primo ramo lavora sulla risoluzione originale, con una convoluzione `kernel=5` e `padding=2`. Gli altri due operano invece su versioni sottocampionate del segnale, con fattori di riduzione $\times 2$ e $\times 4$, tramite `AvgPool1d` seguito dalla stessa convoluzione. Prima di essere concatenati per formare i 12 canali complessivi, i loro output vengono riportati alla lunghezza originale con un'interpolazione lineare. Questa fusione di *feature* multi-scala già in ingresso arricchisce la rappresentazione morfologica delle onde fin dalle primissime fasi dell'elaborazione.

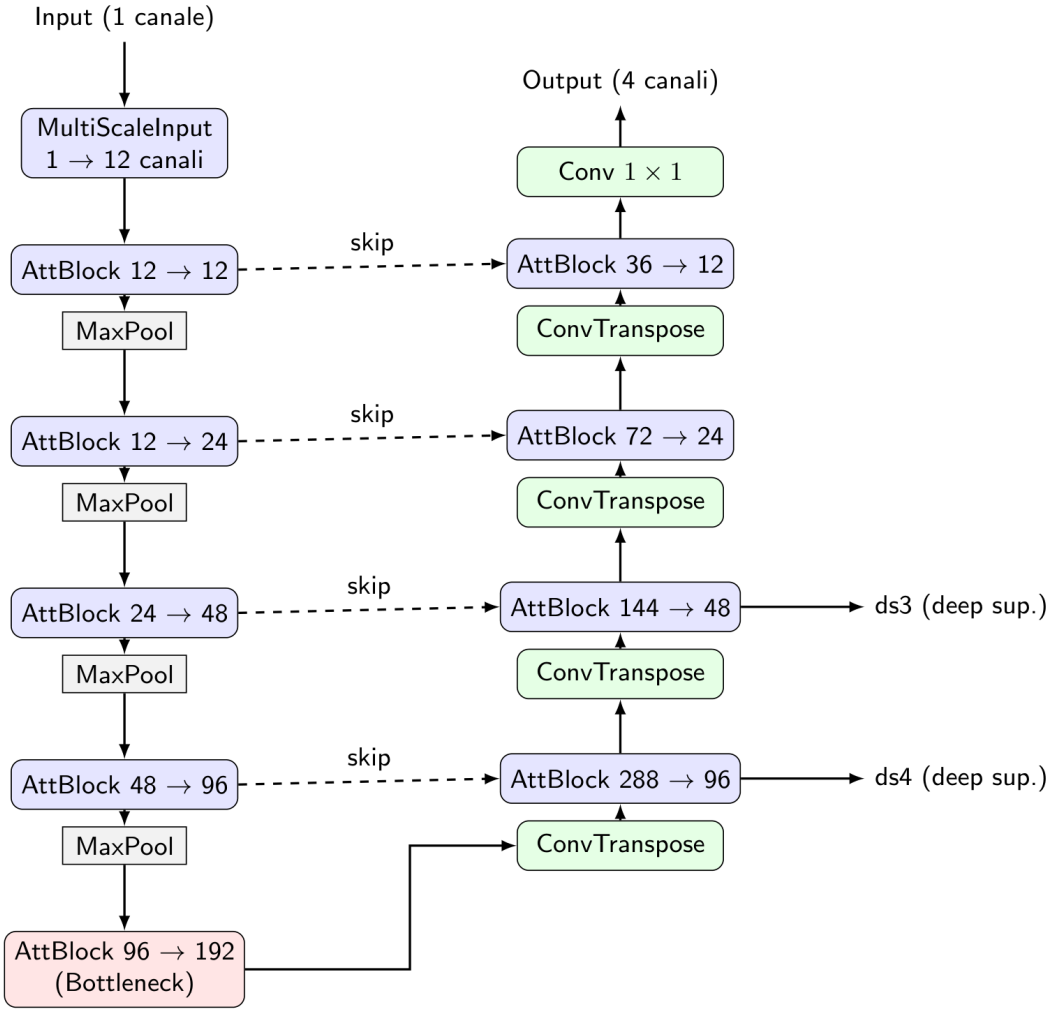


Figura 4.1: Schema d'insieme dell'architettura EnhancedUNet. La rete riceve in ingresso il segnale monodimensionale, lo elabora tramite un modulo `MultiScaleInput` *e procede attraverso una struttura encoder-decoder con* skip connections. *Sono evidenziate le teste decisionali ausiliarie (*deep supervision*) ai livelli intermedi del decoder per stabilizzare l'addestramento.*

4.2.2 Il blocco base: struttura residua con SE innestato

Il cuore computazionale della rete, usato sia in fase di contrazione (*encoder*) che di espansione (*decoder*), è un blocco che è stato chiamato `ResConvBlockSE`, o `AttBlock` (la cui struttura interna è dettagliata in Figura 4.2). Qui si fondono due idee: l'apprendimento residuo e il meccanismo di attenzione Squeeze-and-Excitation (SE), quest'ultimo già validato in letteratura (le cosiddette SENet) per ricalibrare dinamicamente le interdipendenze tra i canali [17]. Il blocco si apre con due convoluzioni unidimensionali (`kernel=9` e `padding=4`), ciascuna seguita da un `BatchNorm1d`. Il kernel di dimensione 9 non è un dettaglio casuale poiché esso serve a massimizzare il campo recettivo effettivo della rete, una scelta che si è rivelata efficace per catturare le caratteristiche morfologiche transitorie delle onde ECG, coerentemente con quanto riscontrato in recenti lavori basati su 1D-UNet [5].

Dopo la prima convoluzione arrivano una `ReLU` e un `Dropout1d` al 15%, per tenere a bada l'*overfitting*. Subito dopo la seconda convoluzione si innesta il blocco SE, un'operazione puntuale (kernel 1×1) con fattore di riduzione `reduction=4`. Infine, il collegamento residuo (*shortcut*) somma l'input iniziale all'output del blocco SE: se il numero di canali cambia lungo il percorso, lo shortcut viene adattato con una convoluzione 1×1 seguita da `BatchNorm1d`. Chiude tutto un'ultima `ReLU` sulla somma finale.

4.2.3 La progressione dei canali lungo la rete

Scendendo nell'encoder, la rete espande progressivamente la propria capacità rappresentazionale. Si parte dai 12 filtri base prodotti dal modulo d'ingresso, elaborati inizialmente a risoluzione piena da un primo `AttBlock` ($12 \to 12$). Dopodiché, intervallati da strati di *pooling*, si susseguono i restanti quattro livelli di codifica spaziale, raddoppiando il numero di canali a ogni step: $12 \to 24 \to 48 \to 96$. In fondo all'architettura, il bottleneck centrale lavora con 192 canali, e va precisato che, nel codice, è implementato a tutti gli effetti come un `AttBlock` completo: beneficia quindi anche lui della ricalibrazione dei canali (*channel-wise*) fornita dal blocco SE al suo interno [17].

4.2.4 L'upsampling e le skip connections

Risalire i quattro livelli simmetrici del decoder, per ripristinare la risoluzione originaria del segnale, è compito degli strati di deconvoluzione (`ConvTranspose1d`), configurati con `kernel=8`, `stride=2` e `padding=3`. Qui si segue una prassi ormai consolidata nelle architetture derivate dalla U-Net, ampiamente usata nella segmentazione biomedica [31]: per compensare la risoluzione persa lungo la strada, l'output di ogni deconvoluzione viene concatenato con le mappe di feature dello

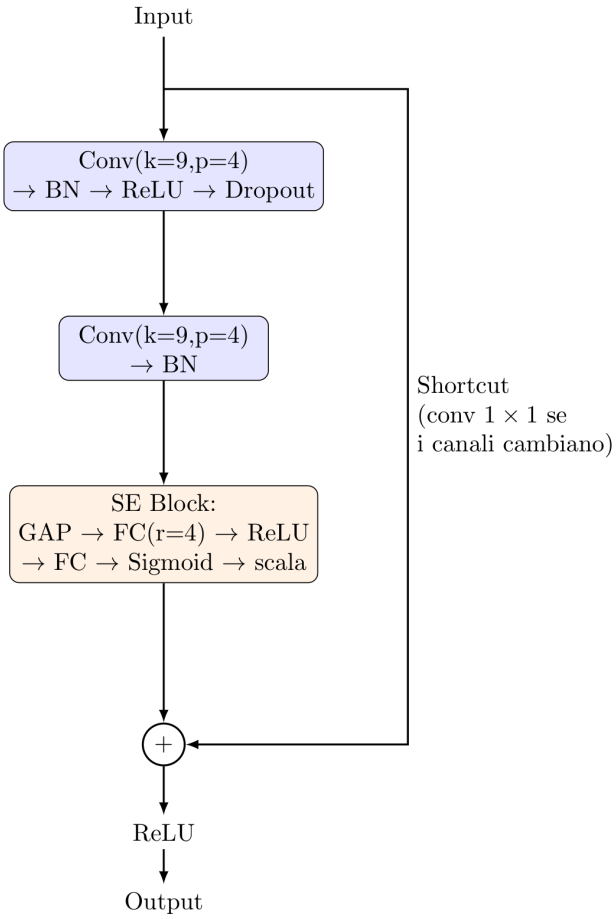


Figura 4.2: Dettaglio del blocco base `AttBlock` *(ResConvBlockSE). Il modulo integra convoluzioni 1D, normalizzazione, un meccanismo di attenzione Squeeze-and-Excitation (SE) per la ricalibrazione dinamica dei canali e un collegamento residuo (*shortcut*).*

stesso livello provenienti dall'encoder attraverso 4 distinte *skip connections* (provenienti rispettivamente dai livelli a 96, 48, 24 e dall'iniziale a 12 canali), prima di passare per l'`AttBlock` del decoder corrispondente.

4.2.5 L'output e la deep supervision

A chiudere la pipeline c'è una convoluzione terminale (1×1), che proietta le mappe in un output finale a 4 canali (le probabilità, punto per punto, di appartenere a Background, onda P, complesso QRS o onda T). C'è però un dettaglio implementativo che vale la pena sottolineare: questa architettura non produce un solo output, ma incorpora anche due teste decisionali ausiliarie (`ds4` e `ds3`), agganciate ai livelli intermedi, quelli a risoluzione minore, del decoder. È il meccanismo della *deep supervision*: fornisce un segnale di gradiente aggiuntivo, a diverse scale, durante la retropropagazione, e aiuta a stabilizzare l'addestramento nelle reti profonde, limitando la dispersione dell'errore.

4.2.6 Complessità e conteggio dei parametri

Dal punto di vista computazionale, il risultato è un modello molto leggero: appena 1.594.458 parametri in tutto (tutti allenabili), che occupano circa 6,08 MB in formato `float32`. Guardando come si distribuiscono, salta all'occhio che il bottleneck da solo pesa più di ogni altra parte: il 33,6% del totale, pari a 536.496 parametri. Il decoder segue con il 30,4%, l'upsampling con il 24,6%, mentre i quattro livelli dell'encoder si dividono una fetta via via più piccola man mano che si risale (dal livello 4 al livello 1). Input multi-scala e strato di output, invece, pesano quasi nulla, praticamente lo 0%.

4.2.7 Un confronto con lo stato dell'arte: EnhancedUNet vs. 1D-UNet base

Per dare il giusto peso a queste scelte progettuali, vale la pena confrontarle direttamente con una recente architettura 1D-UNet proposta per l'analisi del segnale ECG [5], un punto di riferimento consolidato in letteratura per la delineazione. Il modello di base condivide con la rete proposta in questo lavoro alcune impostazioni fondamentali, in particolare la struttura encoder-decoder e i parametri convoluzionali (`kernel=9`, `stride=1`, `same padding`, max-pooling `size=2`). La loro architettura, però, si ferma a 3 livelli di downsampling e non integra alcun meccanismo di attenzione, né spaziale né di canale, pur richiedendo comunque circa 1,8 milioni di parametri [5].

L'EnhancedUNet, al contrario, si spinge a 4 livelli di profondità per estrarre feature più astratte, eppure risulta paradossalmente più compatta: 1,59M di para-

metri contro 1,8M. Questa efficienza è coerente con quanto riportato per i blocchi Squeeze-and-Excitation, che migliorano la capacità rappresentazionale della rete a fronte di un incremento di parametri quasi trascurabile [17], un beneficio che, unito alla fusione multi-scala in ingresso, può aver contribuito a ottenere prestazioni comparabili con un numero di filtri inferiore rispetto a un'architettura priva di questi meccanismi.

Definita in modo trasparente l'architettura nei suoi parametri strutturali concreti, resta da stabilire la metodologia ottimale di apprendimento.

4.3 Knowledge distillation: correzione del bias $T_{Onset} = QRS_{Offset}$

Il paragrafo precedente ha descritto l'architettura EnhancedUNet nei suoi dettagli implementativi. Resta da chiarire come è stata addestrata. L'apprendimento supervisionato standard non basta: i dati clinici di partenza contengono un'anomalia strutturale che va corretta a monte. La soluzione adottata è una strategia di *Knowledge Distillation* (distillazione della conoscenza) cross-domain, descritta nel resto del paragrafo (le metriche e i valori numerici relativi a ogni fase sono riportati e discussi nel Capitolo 8).

4.3.1 Il problema: il bias sistematico delle annotazioni GE

Il problema di fondo è già stato anticipato nel Capitolo 2 e verrà quantificato nel Capitolo 8: l'algoritmo proprietario degli elettrocardiografi GE MAC2000 (dataset PRIV) forza sistematicamente la coincidenza tra la fine del complesso ventricolare e l'inizio della ripolarizzazione ($T_{Onset} = QRS_{Offset}$). Il segmento ST, di fatto, sparisce dentro l'onda T. E se una rete viene addestrata direttamente su queste etichette, impara a fare lo stesso: ignora il segmento ST, che è però proprio il tratto rilevante per diagnosticare ischemie e infarti.

Quanto pesa questo limite lo mostra la Fase DIAG-GE (descritta nel paragrafo 6.4): addestrando e valutando il modello esclusivamente nel dominio GE, le prestazioni restano vincolate a un tetto massimo, e la rete eredita fedelmente il difetto della macchina sia in termini di F1 sia di errore temporale sui confini delle onde. Va detto che la localizzazione dei confini dell'onda T è comunque, in letteratura, tra i compiti più ambigui della delineazione ECG [22]. Qui però la causa non è il rumore fisiologico del segnale, ma un bias di etichettatura imposto a monte, dalla sorgente stessa.

4.3.2 Principio della distillazione: teacher-student cross-domain

Per spezzare la dipendenza da questo bias senza dover ri-annotare manualmente migliaia di tracciati clinici, la soluzione adottata è un approccio Teacher-Student: un modello ad alte prestazioni (*Teacher*) guida l'apprendimento di un secondo modello (*Student*), tecnica dimostratasi particolarmente efficace in letteratura per ottimizzare le reti neurali applicate all'ECG [4]. Questo, in sintesi, è il principio della *Knowledge Distillation*.

Nel setup sperimentale il Teacher è un'istanza della EnhancedUNet, addestrata in Fase B-LUDB (paragrafo 6.1) sul database pubblico LUDB. La scelta non è casuale: LUDB è il gold standard del settore, con annotazioni eseguite manualmente da cardiologi, in modo indipendente per ciascuna derivazione [18]. Questo garantisce una separazione fisiologica reale tra QRS e onda T, quella che nei dati GE manca.

Una volta addestrato, il Teacher viene congelato e messo al lavoro sui tracciati grezzi del dataset PRIV, dove genera *pseudo-label* correggendo il bias originario senza bisogno di intervento manuale. Lo Student (stessa architettura EnhancedUNet) impara poi da queste pseudo-label, non dalle annotazioni XML originali della GE: la *ground truth* diventa quella del Teacher, non quella della macchina.

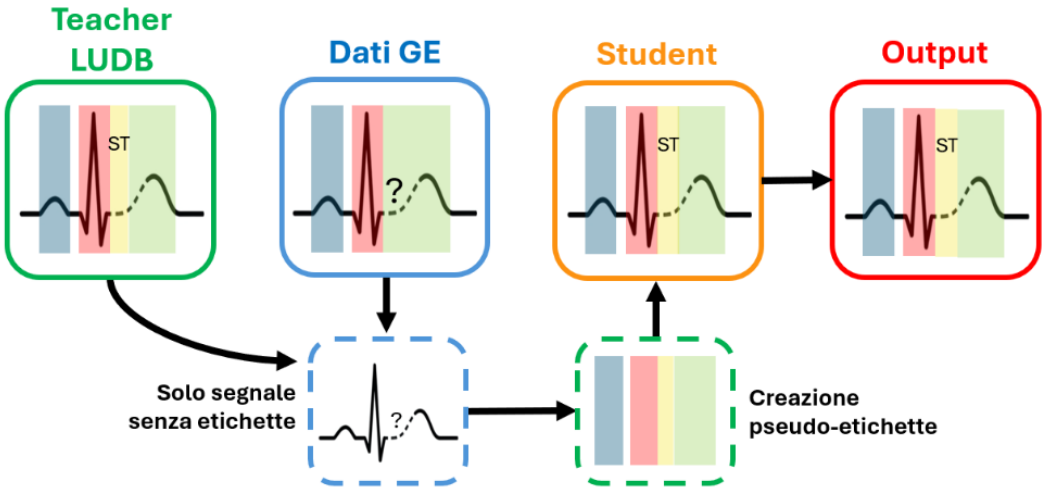


Figura 4.3: Diagramma a blocchi della Knowledge Distillation. I segnali GE (privi delle loro etichette originali) vengono passati al Teacher (precedentemente addestrato su LUDB) per generare delle pseudo-label. Lo Student viene quindi addestrato utilizzando i segnali GE in abbinamento a queste nuove etichette, imparando a separare correttamente il tratto ST-T e ignorando del tutto le annotazioni automatiche industriali.

Che il trasferimento cross-domain (LUDB↔PRIV) funzioni, nonostante le differenze in durata del segnale, tipologia di derivazioni e rumore di fondo, non è scontato. Due fattori lo rendono possibile: la capacità della U-Net di estrarre feature morfologiche invarianti al dominio [31], e la standardizzazione Z-score già applicata in fase di pre-processamento (Paragrafo 4.1), che annulla le discrepanze di ampiezza e deriva della linea isoelettrica tra le due sorgenti.

4.3.3 Applicazione al formato beat mediano: Fase KD-MED

La strategia è stata testata per prima sul formato beat mediano, tensori da 600 campioni (Fase KD-MED, paragrafo 6.3). Rispetto alla Fase B-MED (paragrafo 6.2), stesso formato, ma senza distillazione, qui lo Student, guidato dalle pseudo-label del Teacher, comincia a forzare la separazione tra QRS_{Offset} e T_{Onset}. La tendenza è positiva. Resta però un limite: la finestra temporale è breve, appena 600 campioni, 1,2 secondi, e non basta a dare alla rete il contesto necessario per sciogliere del tutto l'ambiguità del complesso ST-T.

4.3.4 Applicazione al formato timbro 10s: Fase KD-TMB

È qui che arriva il miglioramento più significativo di tutto il lavoro: applicando la distillazione al formato timbro, tracciati da 10 secondi, 5000 campioni (Fase KD-TMB, paragrafo 6.6). Con una traccia continua di questa lunghezza, i blocchi convoluzionali non vedono più solo il singolo battito, ma anche la sequenza delle onde e l'andamento della linea isoelettrica tra un battito e l'altro.

Il risultato si vede subito: addestrando lo Student sulle pseudo-label estratte da questi tracciati, l'errore sistematico crolla. Il modello si svincola così dai limiti imposti dalle annotazioni automatiche GE, e la topologia temporale del segmento ST torna coerente con quella dell'annotazione manuale.

Prima di tradurre questa strategia e questa architettura in fasi sperimentali concrete (oggetto del Capitolo 6), è però fondamentale analizzare a fondo la materia prima su cui i modelli operano: il prossimo capitolo esplorerà quindi in dettaglio i due dataset protagonisti di questo lavoro, LUDB e PRIV.

Capitolo 5

I dataset: LUDB e PRIV

5.1 LUDB: struttura, annotazioni manuali, split train/val/test

5.1.1 Origine e composizione del dataset

Il Lobachevsky University Database (LUDB) è composto da 200 registrazioni ECG a 12 derivazioni, 10 secondi ciascuna. I tracciati arrivano dall'ospedale cittadino n.5 di Nizhny Novgorod, raccolti tra il 2017 e il 2018: soggetti sani insieme a pazienti con diverse patologie cardiovascolari, alcuni dei quali portatori di pacemaker. In questa tesi LUDB gioca un ruolo preciso: è il gold standard, il dominio "teacher" della distillazione (4.3). È l'unico dataset con annotazioni manuali complete [18], ed è da qui che parte il modello che genera le pseudo-label corrette per PRIV.

5.1.2 Annotazione manuale: il gold standard

I confini e i picchi di onda P, complesso QRS e onda T sono annotati a mano da cardiologi, e in modo indipendente per ciascuna delle 12 derivazioni: non un'unica annotazione replicata su tutte le tracce, ma dodici giudizi clinici distinti per ogni record (si veda un esempio in Figura 5.1). È proprio questa indipendenza a fare la differenza: garantisce la reale separazione fisiologica tra QRS e onda T, quella che nelle annotazioni automatiche GE MAC2000 semplicemente non c'è. Ed è per questo che LUDB può correggere il bias descritto in 4.3.1. Ogni record porta con sé anche una diagnosi testuale, conservata nel file .hea di ciascun paziente (formato WFDB [14]). Da lì è stata derivata la classificazione patologica usata in questo lavoro.

#	Macro-classe	Patologie associate	LUDB — keyword (EN)	PRIV — keyword (IT)
1	**PACING**	Ritmo indotto da pacemaker artificiale (stimolazione atriale e/o ventricolare)	`pacing`, `pacemaker`	`pacing atriale`, `pacing ventricolare`, `stimolazione`, `pacemaker`
2	**NO_P**	Fibrillazione atriale, flutter atriale, ritmo giunzionale/idioventricolare (assenza di onda P riconoscibile)	`atrial fibrillation`, `atrial flutter`, `wandering atrial pacemaker`	`fibrillazione atriale`, `flutter atriale`, `ritmo giunzionale`, `ritmo idioventricolare`, `centro giunzionale`
3	**WIDE_QRS**	Blocco di branca (destro/sinistro), emiblocco, ritardo di conduzione intraventricolare	`bundle branch block`, `hemiblock`, `intraventricular conduction delay`, `aberrant conduction`	`blocco di branca`, `emiblocco`, `allargamento del qrs`, `conduzione intraventricolare`, `blocco intraventricolare`, `ritardo della conduzione`
4	**OLD_MI**	Infarto miocardico pregresso, necrosi/cicatrice, onda Q patologica	`scar formation`, `ischemia/scar/supp.nstemi`, `stemi`	`infarto`, `onda q patolog`
5	**AV_BLOCK**	Blocco atrioventricolare (I, II, III grado), blocco seno-atriale, PR corto	`av block`, `sinoatrial blockade`	`blocco a-v`, `bav`, `atrio-ventricolare`, `p-r corto`
6	**ST_T**	Alterazioni ST-T aspecifiche, ischemia, sovraccarico ventricolare, pericardite/miocardite, QT lungo	`repolarization abnormalities`, `early repolarization syndrome`, `ischemia:`, `overload`	`st aspecifiche`, `anormalità`, `ischemia`, `depressione st`, `sottoslivellamento`, `sopraslivellamento`, `ripolarizzazione`, `sovraccarico`, `pericardite`, `miocardite`, `qt lungo`
7	**HYPERTROPHY**	Ipertrofia ventricolare (sinistra/destra), ingrandimento atriale, deviazione assiale	`hypertrophy`, `left axis deviation`, `right axis deviation`	`ipertrofia`, `ingrandimento atriale`, `ivs`, `deviazione assiale sinistra`, `deviazione assiale destra`, `basso voltaggio`
8	**HEALTHY**	Ritmo sinusale normale, bradicardia/tachicardia sinusale, aritmia sinusale fisiologica	`sinus rhythm`, `sinus bradycardia`, `sinus tachycardia`, `sinus arrhythmia`	`ecg normale`, `ritmo sinusale normale`, `bradicardia sinusale`, `tachicardia sinusale`

Tabella 5.1: Corrispondenza tra le 9 macro-classi diagnostiche, le patologie cliniche associate e le keyword utilizzate per l'assegnazione automatica a partire dalle diagnosi testuali, per il dataset LUDB (diagnosi in inglese) e per il dataset PRIV (diagnosi in italiano, referto GE). L'ordine delle macro-classi segue la priorità di assegnazione: quando un record presenta più patologie contemporaneamente, viene assegnata la prima classe della lista che trova corrispondenza. La classe UNKNOWN (non riportata in tabella) raccoglie i record privi di corrispondenza con qualunque keyword, in entrambi i dataset.

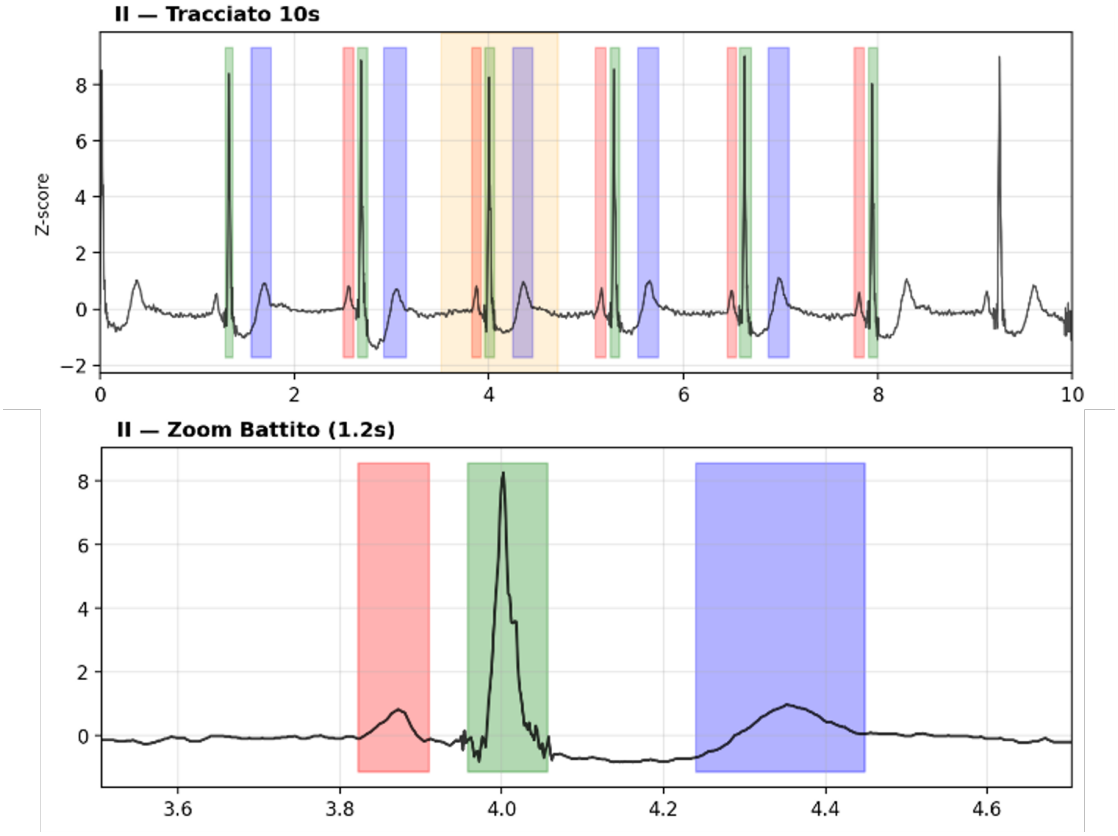


Figura 5.1: Esempio di annotazione manuale su un tracciato LUDB (derivazione II). A destra, i confini di onda P (rosso), complesso QRS (verde) e onda T (blu) su una finestra di 10 secondi; a sinistra, lo zoom su un singolo battito (1,2s), che evidenzia la netta separazione tra la fine del QRS e l'inizio dell'onda T, la stessa separazione assente nelle annotazioni automatiche GE (4.3.1, 5.2.2).

5.1.3 Struttura dei dati e formato

Il dataset originale, in formato WFDB, è stato riletto e ricodificato in un formato tensoriale. La conversione avviene in quattro passaggi:

- **Maschera di segmentazione:** per ciascuna derivazione, gli onset/offset annotati diventano una mappa campione-per-campione a quattro classi (Background, Onda P, QRS, Onda T). Sono le stesse classi usate per PRIV.
- **Macro-classe diagnostica:** la diagnosi testuale di ogni record passa attraverso un keyword-matching (in inglese, la lingua originale delle diagnosi LUDB), che la assegna a una delle nove macro-classi: HEALTHY, PACING, NO_P, WIDE_QRS, OLD_MI, AV_BLOCK, ST_T, HYPERTROPHY,

UNKNOWN. Quando un record ha più patologie insieme, vince quella con priorità più alta, secondo un ordine fisso: PACING > NO_P > WIDE_QRS > OLD_MI > AV_BLOCK > ST_T > HYPERTROPHY > HEALTHY.

- **Ricampionamento:** il segnale viene portato a 500Hz per interpolazione a spline cubica, ma solo se la frequenza originale è diversa; quando è già a 500Hz, non succede nulla.
- **Normalizzazione:** Z-score, derivazione per derivazione.

Ogni record contribuisce con 12 campioni distinti al tensore finale, uno per derivazione; sui 200 record totali, il dataset conta quindi 2400 campioni.

5.1.4 Split train/val/test

Lo split è stato fatto a livello di record, quindi, dato che ogni record è un paziente diverso, anche a livello di paziente, con le proporzioni 80/10/10 (160/20/20 record su 200 totali), stratificato per macro-classe diagnostica (le stesse nove classi usate per PRIV). Ogni record porta con sé tutte le sue 12 derivazioni, sempre nello stesso split, senza eccezioni, per questo la proporzione resta identica anche contando i singoli campioni (1920/240/240 su 2400 totali). Viene usata una stratificazione in due passaggi (usando *scikit-learn* [29]): prima si isola il 10% di test, poi si divide il restante 90% in train/val con lo stesso rapporto. Entrambe le volte la stratificazione avviene sulla macro-classe (con seed fisso per poter riprodurre esattamente lo stesso split in futuro). Il bilanciamento regge per la maggior parte delle classi. ST_T (51 record, 25,5% del totale) e WIDE_QRS (50 record, 25,0%), le due più numerose, si dividono quasi esattamente nelle proporzioni attese (80,4/9,8/9,8% e 80,0/10,0/10,0%); lo stesso vale per HYPERTROPHY (27 record, 13,5%, 77,8/11,1/11,1%). Diverso il discorso per AV_BLOCK: appena 4 record su 200, il 2,0% del totale. Con numeri così piccoli, la stratificazione non ha materialmente modo di distribuirla su tutti e tre gli split, e finisce tutta nel train (100/0/0%), non per una scelta specifica in fase di implementazione, ma perché l'allocazione proporzionale della stratificazione non lascia margine quando una classe ha solo 4 record: è un limite intrinseco quando una classe è troppo rara per essere divisa in tre parti non vuote.

			Record per split			% entro la classe		
Macro-classe	**Tot**	**% sul totale**	**Train**	**Val**	**Test**	**%Tr**	**%Val**	**%Test**
HEALTHY	25	12.5%	20	2	3	80.0%	8.0%	12.0%
PACING	12	6.0%	10	1	1	83.3%	8.3%	8.3%
NO_P	15	7.5%	12	2	1	80.0%	13.3%	6.7%
WIDE_QRS	50	25.0%	40	5	5	80.0%	10.0%	10.0%
OLD_MI	16	8.0%	12	2	2	75.0%	12.5%	12.5%
AV_BLOCK	4	2.0%	4	0	0	100.0%	0.0%	0.0%
ST_T	51	25.5%	41	5	5	80.4%	9.8%	9.8%
HYPERTROPHY	27	13.5%	21	3	3	77.8%	11.1%	11.1%
Totale	**200**	**100.0%**	**160**	**20**	**20**	**80.0%**	**10.0%**	**10.0%**

Tabella 5.2: Composizione del dataset LUDB (200 record) per macro-classe diagnostica e relativo split train/val/test. La colonna "% sul totale" indica il peso di ciascuna classe sull'intero dataset; le colonne "% entro la classe" mostrano come i record di quella classe si distribuiscono tra i tre insiemi. Lo split è stratificato per macro-classe (proporzione 80/10/10) a livello di record/paziente. Fa eccezione AV_BLOCK, la classe meno rappresentata (4 record su 200), che per la sua bassa numerosità non viene distribuita su tutti e tre gli insiemi.

5.2 Il dataset PRIV: struttura, bias GE, macro-classi

5.2.1 Origine e composizione del dataset

I tracciati PRIV arrivano da elettrocardiografi GE MAC2000 [13] in uso in ambito clinico privato, in formato XML proprietario [11]. Il dataset conta 2464 file XML, tutti prodotti dallo stesso dispositivo (un solo seriale distinto in tutto l'archivio).

5.2.2 Formato XML e parsing

Il parser prova due namespace GE in cascata: se il primo non restituisce nulla, scatta il secondo come fallback. Da ogni file XML si possono estrarre due formati di segnale:

- **Beat mediano (600 campioni):** il battito rappresentativo calcolato internamente dal GE.
- **Tracciato timbro 10 secondi (5000 campioni a 500Hz):** la striscia di ritmo continua.

Un dettaglio emerso dal codice, sul beat mediano: i valori GE segnano il "non dato" con una sentinella numerica dedicata. Qui la conversione non riempie i buchi con zeri, ma con un forward-fill: l'ultimo valore valido si ripete finché non ne arriva uno nuovo. Serve a non introdurre discontinuità artificiali nel segnale.

Le annotazioni di onset e offset (P, QRS, T) sono globali per record, non per singola derivazione. LUDB fa diversamente: lì ogni derivazione ha la propria annotazione indipendente (5.1.2). Non è solo una differenza di qualità tra i due dataset: è un'asimmetria strutturale.

5.2.3 Il bias sistematico: T_Onset = QRS_Offset

Il bias è già discusso in dettaglio in 4.3.1. Qui basta un chiarimento tecnico che aiuta a capire come si manifesta concretamente nei file: quando il campo T_Onset manca nell'XML, la conversione applica un fallback esplicito al valore di QRS_Offset. Non è un errore di lettura, è il riflesso diretto di come il software GE (guidato dall'algoritmo Marquette 12SL [12]) genera, o semplicemente omette, l'annotazione del T_Onset.

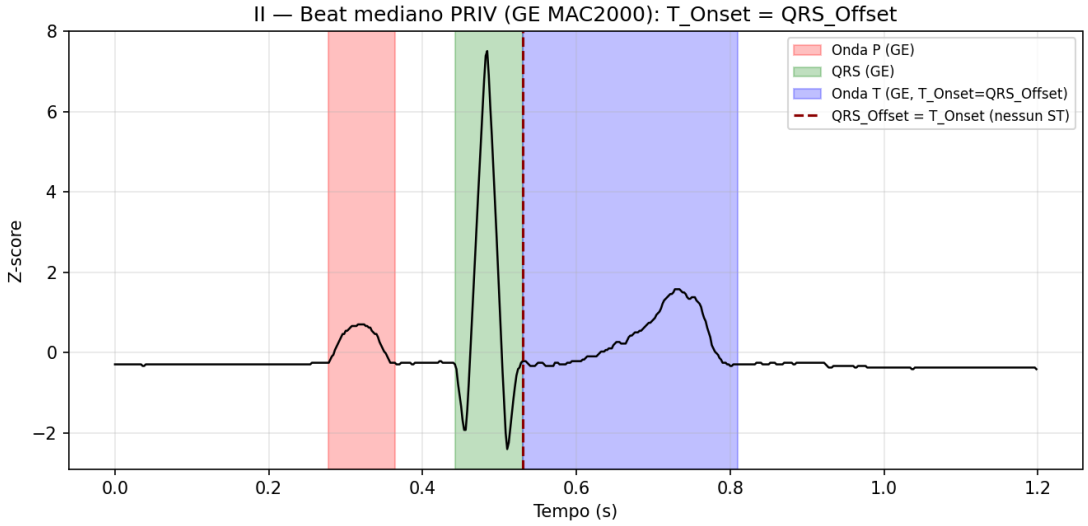


Figura 5.2: Esempio di annotazione automatica GE MAC2000 su un tracciato PRIV (derivazione II, beat mediano). La linea tratteggiata indica il punto QRS_Offset, che coincide esattamente con l'inizio dell'onda T (T_Onset): il segmento ST non ha spazio proprio nell'annotazione. Il confronto diretto con la Figura 5.1 (LUDB) rende visibile la differenza: lì i tre confini sono ben separati, qui QRS e T sono attaccati.

5.2.4 Macro-classi diagnostiche per PRIV

Anche per PRIV si è usato lo stesso schema a nove macro-classi di LUDB (HEALTHY, PACING, NO_P, WIDE_QRS, OLD_MI, AV_BLOCK, ST_T, HYPERTROPHY, UNKNOWN) con lo stesso ordine di priorità nell'assegnazione. Cambia la lingua delle keyword: qui è l'italiano, coerente con i referti diagnostici GE (Tabella 5.1).

Sul totale dei 2464 file, HEALTHY guida la distribuzione con 670 file (27,2%), seguita da WIDE_QRS (511, 20,7%) e NO_P (330, 13,4%). Poi vengono ST_T (225, 9,1%), OLD_MI (207, 8,4%), PACING (191, 7,8%) e AV_BLOCK (183, 7,4%). HYPERTROPHY è la meno numerosa tra le classi patologiche, con 141 file (5,7%); UNKNOWN chiude con appena 6 file, lo 0,2% del totale.

C'è anche un meccanismo che in LUDB non esiste: un set separato di keyword per gli artefatti rumore da elettrodi, interferenza di linea, tremore muscolare, baseline wander. Non è una macro-classe a sé, ma un flag indipendente.

5.2.5 Split train/val/test

Lo split segue lo stesso schema di LUDB, stratificato per macro-classe, 80/10/10, ma qui applicato a livello di file: 1970 in train, 247 in val, 247 in test, proporzioni 80,0/10,0/10,0% esatte. Il bilanciamento regge su tutta la linea, comprese le classi più piccole come UNKNOWN (6 file in tutto, 4/1/1) e HYPERTROPHY (141 file, 113/14/14).

			File per split			**% entro la classe**		
Macro-classe	**Tot**	**% sul totale**	**Train**	**Val**	**Test**	**%Tr**	**%Val**	**%Test**
HEALTHY	670	27.2%	536	67	67	80.0%	10.0%	10.0%
PACING	191	7.8%	153	19	19	80.1%	9.9%	9.9%
NO_P	330	13.4%	264	33	33	80.0%	10.0%	10.0%
WIDE_QRS	511	20.7%	409	51	51	80.0%	10.0%	10.0%
OLD_MI	207	8.4%	165	21	21	79.7%	10.1%	10.1%
AV_BLOCK	183	7.4%	147	18	18	80.3%	9.8%	9.8%
ST_T	225	9.1%	179	23	23	79.6%	10.2%	10.2%
HYPERTROPHY	141	5.7%	113	14	14	80.1%	9.9%	9.9%
UNKNOWN	6	0.2%	4	1	1	66.7%	16.7%	16.7%
Totale	**2464**	**100.0%**	**1970**	**247**	**247**	**80.0%**	**10.0%**	**10.0%**

Tabella 5.3: Composizione del dataset PRIV (2464 file XML) per macro-classe diagnostica e relativo split train/val/test. Split stratificato per macro-classe (proporzione 80/10/10) a livello di file XML.

5.3 Limiti della ground truth GE per il tracciato timbro

Il GE fornisce un solo set di annotazioni globali per record (onset e offset di P, QRS, T) pensate per il battito mediano. Nel tracciato timbro da 10 secondi, invece, i battiti sono molti: proiettare la stessa annotazione su tutti significa assumere che ognuno abbia esattamente la stessa forma del battito mediano, ma tale assunzione si è rivelata falsa.

Il problema è che il battito mediano GE rappresenta una sola morfologia per tracciato, quella dominante, calcolata internamente dal software. Finché tutti i battiti del tracciato condividono quella forma, proiettare le stesse finestre P/QRS/T su ognuno di essi è corretto. Il problema emerge quando lo stesso tracciato contiene battiti di forma sostanzialmente diversa tra loro. Il caso più evidente è la classe PACING. Il pacing "a domanda" interviene solo quando il ritmo intrinseco del paziente è troppo lento; nello stesso tracciato da 10 secondi convivono spesso battiti intrinseci (conduzione naturale, QRS stretto) e battiti stimolati (spike seguito da un QRS artificiale, tipicamente più largo). Il battito mediano GE ne descrive uno solo, quello prevalente, e le sue annotazioni, per quanto corrette per quel tipo di battito, sono semplicemente sbagliate se applicate all'altro. Non è un caso isolato di PACING: è la manifestazione più netta di un problema generale, che si presenta ogni volta che la morfologia battito-per-battito non è stabile quanto il GE assume implicitamente scegliendo un unico template di riferimento.

5.3.1 Proiezione dinamica con template Leave-One-Out

La soluzione adottata rovescia il punto di partenza: invece di ancorare il confronto a un template esterno (il battito mediano GE), il riferimento viene ricostruito dal tracciato stesso, dinamicamente, un battito alla volta. Per ogni posizione di battito rilevata, si estrae prima la finestra QRS usando Q_Onset/Q_Offset del GE come offset; tutte le finestre del tracciato vengono poi ricampionate alla stessa lunghezza. A questo punto, per ciascun battito si calcola un template *leave-one-out*, la mediana punto-per-punto di tutti gli altri battiti del tracciato, escludendo il battito stesso, per evitare che si "auto-convalidi", e lo si correla con quel template tramite la correlazione di Pearson, invariante a scala e offset del segnale. Se il battito è coerente con l'andamento tipico del resto del tracciato, l'annotazione viene proiettata, invece, se la sua morfologia si discosta troppo (correlazione sotto una soglia di 0,7 di default), il battito resta background, escluso dalla maschera di segmentazione.

Questo capovolge il criterio di giudizio: non si chiede più "questo battito somiglia al beat mediano del GE?", una domanda che presuppone che esista un'unica forma di riferimento valida per tutto il tracciato, ma "questo battito somiglia alla maggioranza degli altri battiti nello stesso tracciato?", una domanda che si adatta automaticamente alla morfologia effettivamente presente, qualunque essa sia, senza dipendere da un singolo template esterno potenzialmente non rappresentativo.

Un dettaglio conseguente a questo procedimento è che la maschera dipende dalla derivazione, perché la correlazione è calcolata sul segnale di quella specifica derivazione, a differenza del beat mediano, dove un'unica maschera vale per tutte le 12. Il conteggio di battiti usati ed esclusi, di conseguenza, è specifico per ogni combinazione (record, derivazione).

Applicato all'intero dataset PRIV timbro, il meccanismo esclude il 3,3% delle 330.516 coppie (battito, derivazione) totali (10.836 casi) lasciandone il 96,7% correttamente proiettate, con una correlazione media di 0,953 (mediana 0,994). L'esclusione non è uniforme tra le macro-classi. PACING ha il tasso più alto tra le classi patologiche (11,1%, più di tre volte la media globale), esattamente la conferma empirica di quanto atteso: è la classe in cui la convivenza di due morfologie nello stesso tracciato è più frequente e più marcata. All'estremo opposto, HYPERTROPHY (1,2%) e NO_P (1,7%) mostrano la massima stabilità morfologica. UNKNOWN registra il tasso di esclusione percentuale più alto (19,4%), ma essendo la classe numericamente più piccola, in termini assoluti il numero di battiti scartati non è elevato; inoltre, non avendo per definizione un pattern diagnostico riconoscibile, risulta anche la meno significativa come riferimento.

5.3.2 Falsi positivi nel rilevamento e inaffidabilità su PACING

Un limite legato al rilevamento dei battiti stessi deriva dal fatto che il meccanismo presuppone che la maggioranza dei *toc* rilevati dal GE corrisponda a battiti reali. Quando questa assunzione è valida, anche in presenza di ritmi molto irregolari (es. fibrillazione atriale), il self-template si forma correttamente e la stragrande maggioranza dei battiti viene mantenuta. In alcuni casi, però, il rilevatore di battiti individua eventi spuri, tipicamente su ritmi irregolari, dove un'oscillazione della linea di base o parte dell'onda T viene scambiata per un secondo QRS. Se questi falsi positivi restano una minoranza, il meccanismo li isola correttamente; il self-template, formato in prevalenza da battiti veri, riconosce il *toc* spurio (correlazione nettamente sotto soglia) e lo esclude, senza intaccare gli altri. Se invece i falsi positivi diventano una parte sostanziale dei toc rilevati, il self-template stesso risulta contaminato da una mescolanza di battiti veri e spuri, e il meccanismo può

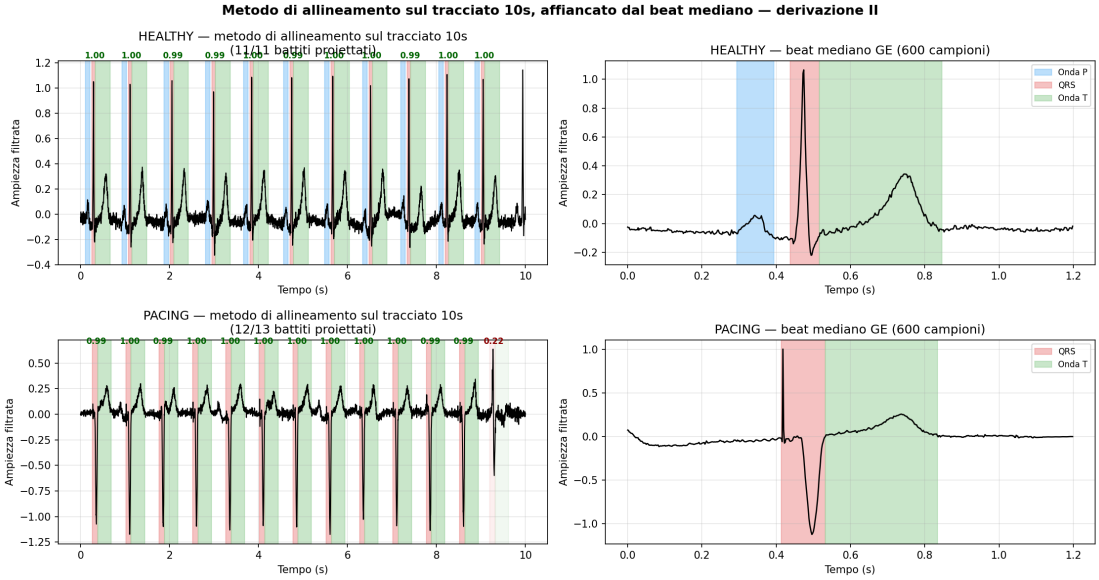


Figura 5.3: Applicazione del metodo di allineamento (self-template leave-one-out) su un tracciato HEALTHY (in alto) e su un tracciato PACING (in basso), derivazione II, affiancata dal rispettivo beat mediano GE. Nel caso HEALTHY tutti i battiti vengono proiettati correttamente (11/11). Nel caso PACING, un toc spurio (correlazione 0,22) viene correttamente escluso, mentre i 12 battiti reali restano tutti proiettati (correlazione $\geq$0,99 ciascuno).

fallire integralmente: nessun battito, nemmeno quelli reali, riesce più a correlare bene con un template che non rappresenta più nulla di coerente.

Un limite ulteriore, di natura diversa, è che in alcuni casi di PACING anche l'annotazione GE di partenza sul beat mediano non sembra affidabile. Lo spike di pacing reale può trovarsi isolato, lontano dalla finestra QRS che il GE stesso annota, un intervallo troppo ampio per rappresentare la risposta fisiologica a uno stimolo catturato. In questi casi il problema precede il self-template LOO poiché non è il meccanismo di proiezione sul tracciato timbro a fallire, è la *ground truth* di riferimento che parte già da un'annotazione dubbia. Questa instabilità degli algoritmi automatici (come il Marquette 12SL) sui ritmi stimolati è un limite ben noto e ampiamente documentato in letteratura [33, 37]; i filtri di segnale spesso non riescono a rilevare in modo affidabile gli spike a bassa ampiezza, portando il software a confondere i battiti stimolati con blocchi di branca o a omettere del tutto la stimolazione.

Questo limite, però, ha un impatto contenuto sul lavoro nel suo complesso, per due motivi:

- PACING è una classe minoritaria nel dataset PRIV (come descritto in 5.2);

anche se una parte dei suoi casi ha annotazioni del battito mediano discutibili, si tratta di una porzione piccola dell'intero dataset, non di un problema trasversale a tutte le patologie.

- La validazione di riferimento del lavoro non si basa sulle annotazioni GE, ma su LUDB, dove le stesse patologie (incluso il pacing) sono annotate manualmente da cardiologi e risultano coerenti. È proprio questo il motivo per cui LUDB resta il *gold standard* e il dominio "teacher" della distillazione (si veda 4.3): eventuali imperfezioni nella *ground truth* automatica (il dominio "student") non si propagano al criterio con cui il modello viene davvero giudicato.

5.4 Confronto tra le due sorgenti di dati e criticità del transfer

I due dataset descritti in questo capitolo non sono intercambiabili: rappresentano due modi radicalmente diversi di produrre la stessa informazione (dove iniziano e finiscono le onde P, QRS, T) e le differenze tra loro non sono solo di scala, ma di natura.

LUDB è piccolo (200 pazienti) ma affidabile: annotazione manuale, indipendente per ciascuna delle 12 derivazioni, eseguita da cardiologi. PRIV è grande (2464 file, oltre 1300 pazienti) ma automatico: le annotazioni derivano dal software 12SL del GE MAC2000, con un'unica maschera per record, non per derivazione, e con la convenzione strutturale T_Onset=QRS_Offset che elimina il segmento ST dall'annotazione (paragrafo 4.3.1). Un dataset è piccolo e corretto, l'altro è grande e sistematicamente distorto su un punto specifico ma clinicamente rilevante.

La Figura 5.4 rende visibile questa differenza su un caso semplice: un paziente sano, in tutti e tre i pannelli. Il tracciato LUDB (pannello 1) mostra P, QRS e T come tre intervalli distinti, ciascuno delimitato dal proprio giudizio clinico. Il battito mediano PRIV (pannello 2) mostra invece l'onda T che comincia esattamente dove finisce il QRS: non c'è spazio per lo ST, non perché il paziente non ne abbia uno, ma perché il software non lo separa mai. Il pannello 3, lo stesso tipo di annotazione GE, mediata dal metodo di allineamento descritto in precedenza (paragrafo 5.3), mostra come queste etichette, per quanto imperfette all'origine, possano comunque essere proiettate in modo affidabile su un tracciato di 10 secondi con più battiti.

La criticità del transfer tra le due sorgenti sta nel fatto che un modello addestrato solo su PRIV imparerebbe a riprodurre fedelmente il bias del GE, non la fisiologia reale del segmento ST, coerente per costruzione con le proprie etichette, ma sistematicamente sbagliato rispetto a un giudizio clinico indipendente. Il pro-

blema non si risolve annotando manualmente il dataset PRIV (operazione di costo elevatissimo), né correggendo euristicamente le etichette GE file per file. Si risolve trasferendo la definizione corretta da un dominio all'altro: è questa l'idea alla base della distillazione (si veda 4.3 e la formalizzazione classica in letteratura [16]), dove un modello addestrato su LUDB, che ha imparato la vera separazione ST-T dai dati manuali, genera pseudo-etichette per PRIV, sostituendo l'annotazione GE originale con una stimata secondo un criterio clinicamente corretto, senza mai richiedere la revisione manuale dei tracciati ospedalieri.

Il capitolo si chiude quindi su un punto che percorre tutta la tesi: LUDB e PRIV non sono due dataset equivalenti da sommare, ma due ruoli distinti, uno di riferimento (*teacher*, *gold standard*), l'altro di dominio applicativo (*student*, dati reali di produzione), e il valore del lavoro sta proprio nel trasferire l'affidabilità del primo al volume del secondo. Nel prossimo capitolo (Capitolo 6) si vedrà come questo trasferimento è stato effettivamente implementato, snodandosi attraverso le otto fasi sperimentali del progetto.

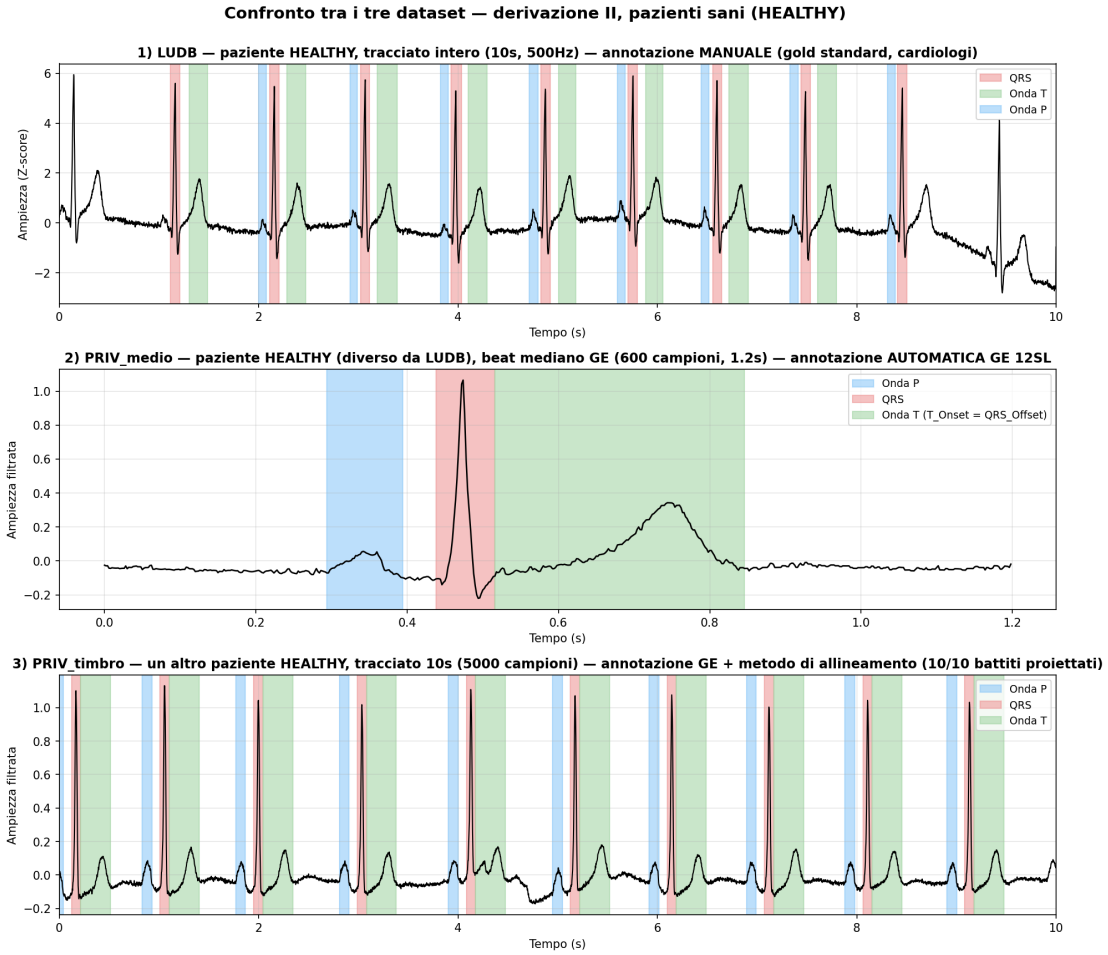


Figura 5.4: Confronto tra i tre dataset utilizzati nella tesi, tutti su pazienti sani (HEALTHY), derivazione II. (1) LUDB: tracciato intero da 10 secondi con annotazione manuale di cardiologi indipendenti per ciascuna derivazione (paragrafo 5.1.2), le bande P/QRS/T sono nettamente separate. (2) PRIV_medio: beat mediano GE MAC2000 (600 campioni), con annotazione automatica del software 12SL (si nota la convenzione T_Onset=QRS_Offset) (paragrafo 4.3.1), che assorbe il segmento ST nell'onda T. (3) PRIV_timbro: tracciato completo da 10 secondi (5000 campioni) di un altro paziente sano, con le stesse annotazioni GE proiettate su ogni battito tramite il metodo di allineamento (paragrafo 5.3); tutti i 10 battiti rilevati vengono proiettati correttamente (correlazione $\geq 0,7$).

Capitolo 6

Fasi sperimentali

Le otto fasi sperimentali di questo capitolo condividono tutte la stessa architettura di base, ovvero la EnhancedUNet (paragrafo 4.2), e la stessa famiglia di metriche di valutazione (F1 con tolleranza AAMI EC57 di 150ms su onset/offset, paragrafo 2.6). Quello che cambia da una fase all'altra è su cosa il modello viene addestrato: quale dominio (LUDB o PRIV), quale formato di segnale (beat mediano 600 campioni, crop LUDB 3000 campioni, o tracciato timbro 5000 campioni) e, soprattutto, quali etichette vede durante il training, ovvero quelle originali del dataset oppure le pseudo-label generate da un altro modello tramite distillazione (paragrafo 4.3).

Per orientarsi tra le fasi, ognuna di esse ha un nome che porta un prefisso che ne indica la natura:

- **B- (baseline):** addestrata direttamente sulle etichette originali del proprio dominio, senza alcuna correzione.
- **KD- (knowledge distillation):** addestrata sulle pseudo-label generate dal teacher LUDB, non sulle etichette originali.
- **CTRL- (controllo):** usata per isolare un effetto specifico (tipicamente il confronto in-domain puro).

Da ogni fase si ottiene un checkpoint indipendente, poi valutato nel Capitolo 8 su più domini di test (LUDB, PRIV_medio, PRIV_timbro) e con più modalità di aggregazione (per singolo campione, per paziente, media sulle 12 derivazioni, solo derivazione II), per capire quanto le prestazioni dipendano dal dominio di valutazione piuttosto che dal modello in sé. In questo capitolo, invece, il focus resta sulla logica metodologica di ciascuna fase (cosa fa e perché), lasciando i numeri completi al capitolo dei risultati.

Fase	Nome esteso	Sigla	Cosa fa
1	Baseline LUDB (Teacher)	**B-LUDB**	training su LUDB, gold standard, diventa il teacher della distillazione
2	Baseline mediano	**B-MED**	baseline PRIV su beat mediano (600 campioni), in-domain, produzione
3	Distillazione mediano	**KD-MED**	knowledge distillation su formato mediano
4	Diagnosi bias GE	**DIAG-GE**	fase diagnostica, quantifica il bias $T_{Onset} = QRS_{Offset}$
5	Baseline timbro	**B-TMB**	PRIV timbro raw (5000 campioni), lower bound senza distillazione
6	Distillazione timbro	**KD-TMB**	knowledge distillation su timbro
7	Training combinato	**COMB**	LUDB GT + pseudo-label PRIV insieme
8	Controllo timbro in-domain	**CTRL-TMB**	training/test entrambi su PRIV timbro, controllo

Tabella 6.1: Sintesi delle otto fasi sperimentali descritte nel capitolo, con la relativa nomenclatura. Le fasi si differenziano per dataset di origine, formato in input e tipologia di etichette (ground truth originale vs pseudo-label distillate).

6.1 Fase B-LUDB: Baseline LUDB (Teacher)

La Fase B-LUDB (paragrafo 6.1) addestra l'EnhancedUNet direttamente sulle annotazioni manuali di LUDB (paragrafo 5.1), usando il formato crop da 3000 campioni. Split e stratificazione seguono lo schema descritto precedentemente per il dataset: 80/10/10 a livello di record/paziente, stratificato per macro-classe diagnostica (paragrafo 5.1). Qui però lo split non è a livello di paziente-file GE come per PRIV, ma coincide con il paziente stesso poiché ogni record LUDB rappresenta già un paziente univoco.

Questa fase ha un duplice ruolo nella tesi: il primo è quello di *baseline* di riferimento; essendo addestrata e valutata sull'unico dataset con *ground truth* manuale, rappresenta il limite superiore di ciò che il modello può imparare quando le etichette di partenza sono corrette, un termine di paragone per tutte le fasi successive che invece lavorano con annotazioni automatiche GE.

Il secondo ruolo, più importante per il seguito del lavoro, è quello di *teacher* della distillazione (paragrafo 4.3). Il checkpoint migliore di questa fase viene congelato (i suoi pesi non vengono più aggiornati) e usato in inferenza sui tracciati PRIV, per generare le *pseudo-label* che sostituiscono le etichette GE originali nelle Fasi KD-MED (paragrafo 6.3) e KD-TMB (paragrafo 6.6). È precisamente perché la Fase B-LUDB (paragrafo 6.1) ha imparato la separazione ST-T corretta, quella che l'annotazione GE non fornisce mai (paragrafo 4.3.1), che il suo output

può essere usato come sostituto affidabile della *ground truth* mancante sul dominio clinico.

6.2 Fase B-MED: Baseline mediano

La Fase B-MED (paragrafo 6.2) addestra l'EnhancedUNet sul dataset PRIV nel formato beat mediano (600 campioni, paragrafo 5.2), usando direttamente le etichette automatiche del GE MAC2000, senza alcuna distillazione o correzione. Lo split segue la proporzione 80/10/10 stratificato per macro-classe, a livello di file (paragrafo 5.2).

In questa fase è possibile osservare il punto di partenza *in-domain*: è il modello che si otterrebbe seguendo l'approccio più diretto e meno costoso possibile, ovvero addestrare su dati clinici reali, fidandosi delle annotazioni che il software diagnostico fornisce già, senza introdurre altre fonti (LUDB) né altri meccanismi (distillazione). Il battito mediano è inoltre il formato più vicino a quello di produzione ed è quindi ciò che un sistema che analizza referti GE già strutturati vedrebbe normalmente in input.

Il ruolo di questa fase nel percorso della tesi è duplice. Da un lato, stabilisce insieme alla Fase B-LUDB (paragrafo 6.1) i due riferimenti dell'intero lavoro (paragrafo 5.2): uno addestrato su dati corretti (LUDB), l'altro su dati clinici reali ma strutturalmente distorti (PRIV). Dall'altro, è il termine di paragone diretto per la Fase KD-MED (paragrafo 6.3): il confronto tra le due fasi è ciò che permette di misurare quanto la distillazione migliori, o meno, le prestazioni sullo stesso formato di input.

6.3 Fase KD-MED: Distillazione mediano

La Fase KD-MED (paragrafo 6.3) applica per la prima volta il meccanismo di distillazione (paragrafo 4.3) al formato battito mediano. Il *teacher* della Fase B-LUDB (paragrafo 6.1) genera *pseudo-label* sui tracciati PRIV_medio, e un nuovo modello (stessa architettura) viene addestrato su queste *pseudo-label* anziché sulle etichette GE originali. Lo *student* non vede mai l'annotazione automatica del software diagnostico: impara esclusivamente a replicare il comportamento del *teacher*.

Il confronto diretto con la Fase B-MED (paragrafo 6.2) (stesso formato, stesso dominio, unica differenza le etichette di training) è ciò che permette di isolare l'effetto della distillazione da quello del formato di input. Come già discusso (paragrafo 4.3), il risultato è una tendenza positiva ma parziale: lo *student* comincia a forzare la separazione tra QRS_{Offset} e T_{Onset}, ma la brevità della finestra (600

campioni, 1,2 secondi) limita la quantità di contesto disponibile per risolvere pienamente l'ambiguità del complesso ST-T. Questo limite, non del meccanismo di distillazione in sé, ma del formato su cui viene applicato, è il motivo diretto per cui il lavoro prosegue verso il formato timbro, con la Fase B-TMB (paragrafo 6.5) e la Fase KD-TMB (paragrafo 6.6): una finestra più lunga offre al modello lo spazio necessario per apprendere pattern temporali che il beat mediano, da solo, non può fornire.

6.4 Fase DIAG-GE: Diagnosi bias GE

A differenza di tutte le altre fasi di questo capitolo, la Fase DIAG-GE non addestra alcun modello. È una fase puramente diagnostica, il cui unico scopo è misurare quanto le annotazioni automatiche del GE si discostino dalla vera separazione ST-T che LUDB fornisce come *gold standard*.

Il meccanismo è un confronto diretto: si simula il comportamento dell'algoritmo GE (la sua convenzione $T_{Onset} = QRS_{Offset}$, paragrafo 4.3.1) applicandolo alle stesse registrazioni di LUDB, per cui esiste anche l'annotazione manuale di riferimento. In questo modo il confronto non è tra due dataset diversi (LUDB vs PRIV, che differiscono anche per provenienza dei pazienti e caratteristiche del segnale), ma tra due modi di annotare lo stesso identico tracciato: quello che farebbe un cardiologo, e quello che farebbe il software GE se venisse applicato a quel tracciato.

Il risultato quantifica in modo diretto il problema descritto qualitativamente nel paragrafo 4.3.1, l'errore sistematico sul T_{Onset} e la sovrastima della durata dello ST che ne consegue, oltre al tetto massimo di prestazione raggiungibile restando fedeli all'annotazione GE. I valori esatti sono riportati e discussi nel Capitolo 8.

Lo scopo di questa fase nel percorso della tesi è quello di misurare il problema, prima di provare a risolverlo. È questa quantificazione a rendere necessaria la distillazione (Fase KD-MED, paragrafo 6.3, e Fase KD-TMB, paragrafo 6.6); senza di essa, l'affermazione che "il GE non separa lo ST dalla T" resterebbe qualitativa. Con questa fase, diventa un errore misurabile, direttamente confrontabile con il risultato ottenuto dopo la distillazione.

6.5 Fase B-TMB: Baseline timbro

La Fase B-TMB addestra l'EnhancedUNet sul dataset PRIV nel formato tracciato timbro (5000 campioni, paragrafo 5.2), usando le etichette GE proiettate su ogni battito tramite il metodo di allineamento descritto in precedenza (paragrafo 5.3), ma senza alcuna distillazione. È l'equivalente della Fase B-MED (paragrafo 6.2)

sul formato timbro: stesso principio (training diretto su etichette automatiche GE), formato diverso.

Il ruolo di questa fase è duplice. Da un lato è il *lower bound* esplicito per il formato timbro: rappresenta cosa si ottiene usando questo formato più ricco di contesto, ma senza correggere il bias sistematico delle etichette GE (paragrafo 4.3.1); quindi eredita comunque il problema $T_{Onset} = QRS_{Offset}$, semplicemente su una finestra più lunga. Dall'altro, è il termine di paragone diretto per la Fase KD-TMB (paragrafo 6.6): il confronto tra le due, a parità di formato e dominio, isola l'effetto della distillazione esattamente come il confronto Fase B-MED / Fase KD-MED (paragrafo 6.3) lo isolava sul beat mediano.

6.6 Fase KD-TMB: Distillazione timbro

La Fase KD-TMB applica lo stesso meccanismo di distillazione della Fase KD-MED (paragrafo 6.3), ma sul formato timbro. Il *teacher* della Fase B-LUDB (paragrafo 6.1) genera *pseudo-label* sui tracciati PRIV_timbro (5000 campioni), e lo *student* le apprende senza mai vedere le etichette GE originali.

È qui che la combinazione tra distillazione e formato esteso produce il risultato centrale della tesi (paragrafo 4.3). Con una finestra di 10 secondi, il modello ha accesso non solo al singolo battito ma alla sequenza delle onde e all'andamento della linea isoelettrica inter-battito, il contesto che mancava alla Fase KD-MED (paragrafo 6.3). Il risultato è una riduzione drastica dell'errore sistematico sul T_{Onset} rispetto alle etichette GE grezze (quantificate in Fase DIAG-GE, paragrafo 6.4). Il modello smette di riprodurre il bias del software diagnostico e converge verso la definizione corretta di separazione ST-T appresa dal *teacher* su LUDB. I valori numerici completi, errore residuo, F1 Macro, confronto con Fase DIAG-GE (paragrafo 6.4) e Fase KD-MED (paragrafo 6.3), sono riportati nel Capitolo 8.

Il confronto a tre tra la Fase B-TMB (paragrafo 6.5) (*baseline* timbro), la Fase KD-MED (paragrafo 6.3) (distillazione senza contesto esteso) e la Fase KD-TMB (distillazione con contesto esteso), è ciò che permette di attribuire il miglioramento non a un singolo fattore isolato, ma alla sinergia tra i due: né la distillazione da sola (Fase KD-MED, paragrafo 6.3, formato corto) né il formato timbro da solo (Fase B-TMB, paragrafo 6.5, nessuna distillazione) ottengono lo stesso risultato.

6.7 Fase COMB: Training combinato

La Fase COMB esplora una strategia alternativa alla distillazione in due stadi: invece di usare il *teacher* LUDB per generare *pseudo-label* su cui addestrare uno *student* separato, i dati di LUDB (con la loro annotazione manuale originale) e i

dati di PRIV_timbro (con *pseudo-label* o etichette GE, a seconda della configurazione) vengono uniti direttamente in un unico set di *training*, e un solo modello viene addestrato su questa combinazione.

L'idea di fondo è economicamente più semplice della distillazione: un solo passaggio di *training*, un solo modello, senza il passaggio intermedio di generare e salvare le *pseudo-label* di un *teacher* congelato.

6.8 Fase CTRL-TMB: Controllo timbro in-domain

La Fase CTRL-TMB addestra l'EnhancedUNet sul dataset con struttura PRIV_timbro e lo valuta sullo stesso dominio (*training* e *test* entrambi su PRIV_timbro) senza alcuna distillazione. È, per costruzione, molto simile alla Fase B-TMB (paragrafo 6.5): stesso formato, stesse etichette GE di partenza, nessuna correzione del bias.

La differenza tra le due fasi sta nel ruolo metodologico che svolgono nel percorso della tesi. La Fase B-TMB (paragrafo 6.5) serve da *lower bound* diretto per isolare l'effetto della distillazione (confronto con la Fase KD-TMB, paragrafo 6.6); la Fase CTRL-TMB serve invece da controllo sperimentale, per rispondere a una domanda diversa: quanto delle prestazioni osservate nelle altre fasi dipende dal fatto che il modello viene valutato nel proprio dominio nativo di addestramento, invece che trasferito su un dominio diverso (ad esempio LUDB)?

La Fase CTRL-TMB (paragrafo 6.8) rende esplicito questo scenario (valutazione rigorosamente *in-domain*) così che il confronto con le fasi valutate anche su LUDB (dominio esterno, senza alcun rischio di *leakage*) permetta di distinguere un miglioramento reale da un artefatto dello split. I valori numerici di questo confronto sono riportati nel Capitolo 8.

6.9 Riepilogo delle configurazioni di addestramento

A conclusione del capitolo, la Tabella 6.2 offre un riepilogo delle sette fasi sperimentali che prevedono l'addestramento di un modello (è esclusa la Fase DIAG-GE, in quanto puramente diagnostica). Questa vista d'insieme permette di seguire facilmente come varia la composizione dei set di addestramento, validazione e test per ciascuna fase, un aspetto cruciale per interpretare correttamente i risultati che verranno discussi nel Capitolo 8.

Fase	Descrizione	Train	Val	Test	Ruolo nel percorso
B-LUDB	Baseline LUDB	LUDB	LUDB	LUDB	*Baseline* di riferimento (limite superiore) e *teacher* della distillazione per le fasi successive.
B-MED	Baseline PRIV_medio	PRIV_medio	PRIV_medio	PRIV_medio	Punto di partenza *in-domain*. Stabilisce il riferimento su annotazioni GE distorte e funge da paragone diretto per isolare l'effetto della distillazione.
KD-MED	Distillazione medio	PRIV_medio	PRIV_medio	LUDB	Prima applicazione della distillazione. Mostra un miglioramento solo parziale, evidenziando come la brevità della finestra limiti il contesto necessario a risolvere l'ambiguità ST-T.
B-TMB	Timbro raw (lower bound)	PRIV_timbro	PRIV_timbro	LUDB	*Lower bound* esplicito per il formato esteso (eredita il bias GE). Termine di paragone per misurare l'efficacia della distillazione sul tracciato da 10s.
KD-TMB	Distillazione timbro	PRIV_timbro	PRIV_timbro	LUDB	Risultato centrale della tesi: la sinergia tra distillazione e contesto esteso azzera il bias GE, permettendo al modello di riprodurre la corretta separazione ST-T.
COMB	Combined PRIV+LUDB	LUDB + PRIV_timbro	LUDB + PRIV_timbro	LUDB	Esplora una strategia alternativa ed economicamente più semplice rispetto alla distillazione, unendo direttamente i due domini in un unico addestramento.
CTRL-TMB	PRIV_timbro in-domain	PRIV_timbro	PRIV_timbro	PRIV_timbro	Controllo sperimentale valutato rigorosamente *in-domain*, per verificare che i miglioramenti osservati non siano artefatti dovuti al cambio di dominio nel test.

Tabella 6.2: Schema riassuntivo delle configurazioni di Train, Validation e Test per le diverse fasi sperimentali di addestramento.

Capitolo 7

Dall'output della rete all'applicazione clinica

Il Capitolo 6 ha delineato le strategie di addestramento utilizzate per ottenere una segmentazione accurata del segnale elettrocardiografico, culminando con i modelli addestrati tramite *knowledge distillation*. Tuttavia, le maschere grezze prodotte dalla rete neurale indicano solamente l'estensione temporale delle onde (P, QRS, T) e non costituiscono ancora un formato sufficiente per la refertazione cardiologica. Questo capitolo illustra i passaggi algoritmici necessari per colmare la distanza tra l'output continuo del modello e la diagnostica: si parte dalla localizzazione precisa dei singoli picchi e del punto J (Sezione 7.1), per poi passare a un livello di *post-processing* (Sezione 7.2) incaricato di ripulire la predizione da eventuali incertezze o artefatti fisiologicamente implausibili.

7.1 Algoritmo di rilevamento dei picchi (P, Q, R, S, T, J) per macro-classe

Una volta che la rete di segmentazione ha individuato gli intervalli grezzi di P, QRS e T, resta un problema distinto: localizzare i picchi puntuali dentro quegli intervalli. L'algoritmo non applica un criterio unico, ma seleziona un *detector* diverso a seconda della macro-classe diagnostica del tracciato, perché la morfologia attesa di ciascuna onda cambia sostanzialmente da un quadro clinico all'altro [12].

7.1.1 Complesso QRS

Vale sempre una regola invariante: i tre punti Q, R e S devono esistere come indici distinti e rispettare l'ordine temporale $Q < R < S$. La linea di base è calcolata come la mediana dell'intero segnale nella finestra.

Nel rilevamento *standard*, usato per le classi HEALTHY, NO_P e per le classi senza un criterio dedicato, l'algoritmo individua il massimo assoluto del tratto come candidato R e il minimo assoluto come candidato S. Se R si trova prima di S, allora si mantiene l'attribuzione individuata, ma se S è molto maggiore rispetto a R (sotto una soglia di *default* pari al 15% dell'ampiezza di S), il complesso viene riclassificato come QS. Nelle derivazioni V1-V3, dove un'onda S profonda potrebbe generare un falso innalzamento del tratto ST successivo, l'algoritmo fissa prima la posizione della S su quel minimo e cerca poi Q ed R nella porzione di segnale che la precede.

Per la classe OLD_MI si utilizza la stessa logica di rilevamento, ma con la soglia di riconoscimento QS aumentata da 0,15 a 0,25, un criterio più permissivo, coerente con la morfologia da necrosi. La classe HYPERTROPHY usa il rilevamento *standard* senza modifiche.

Per la classe WIDE_QRS l'algoritmo cerca esplicitamente una morfologia frammentata; dopo aver individuato un massimo iniziale (R) e la valle successiva (S), esplora la porzione destra del complesso alla ricerca di un secondo picco positivo (R'). Se R' raggiunge almeno il 30% dell'ampiezza di R, viene registrato come componente distinta, e la posizione della S viene ricalcolata nella valle tra R e R' [2].

Per la classe AV_BLOCK, prima della ricerca del QRS, l'algoritmo esplora i campioni iniziali della finestra alla ricerca di uno *spike*, un picco breve e intenso, confermato tramite un filtro mediano. Se lo *spike* viene identificato, la finestra di ricerca viene ristretta escludendolo, e la ricerca del QRS prosegue sul tratto ripulito con il criterio *standard*. Per la classe PACING il rilevamento dello *spike* usa soglie più permissive e intercetta anche *spike* bifasici [19]. Lo *spike* e i tre campioni successivi vengono esclusi dall'analisi, e il tratto residuo viene classificato come QS, rS, qR o Rs.

Quando il complesso è un QS puro, i criteri sopra descritti tenderebbero a far collassare Q ed R sullo stesso campione. Un modulo dedicato interviene in questi casi posizionando la R sulla "spalla" del segnale tra Q ed S, così da mantenere sempre tre indici distinti.

7.1.2 Onda P

Il criterio applica una rimozione della deriva della linea di base e una finestratura pensata per evitare che i picchi rilevati si concentrino ai bordi della finestra.

Il rilevamento *standard* si basa su una mappa di polarità attesa per ciascuna derivazione (positiva in II, negativa in aVR, e così via). Se è presente un'onda di polarità opposta con ampiezza superiore al 10% di quella attesa, l'algoritmo seleziona quest'ultima invece della polarità di *default*. Nelle derivazioni V1, III e

aVL, dove l'onda P è spesso poco visibile, si alza la soglia minima richiesta per riconoscere un picco, così da ridurre i falsi positivi dovuti al rumore.

Per la classe AV_BLOCK il criterio forza la ricerca di un'onda P positiva specificamente in II, III e aVF. Per NO_P, se la finestra individuata dalla rete è più stretta di 30 ms o l'escursione del segnale non è distinguibile dal rumore di fondo, l'algoritmo restituisce l'assenza dell'onda P; se un'onda viene comunque trovata, viene marcata con un'indicazione di verifica, perché potrebbe trattarsi di un'onda *f* di fibrillazione piuttosto che di una vera P.

7.1.3 Onda T

Il rilevamento *standard*, usato per tutte le classi eccetto ST_T, individua il picco con la polarità attesa per quella derivazione, dopo aver ripulito il segnale da eventuali derive della linea di base.

Per la classe ST_T l'algoritmo cambia approccio: estrae tutti i massimi e minimi locali nel tratto, e confronta il picco più ampio in assoluto con il picco più ampio di polarità attesa. Se il primo supera il secondo di un fattore 1.1, viene selezionato il picco di polarità opposta a quella attesa, permettendo a una T invertita di prevalere [9]. L'algoritmo cerca poi una seconda componente di polarità opposta al picco principale: se la trova, e la sua ampiezza raggiunge almeno il 40% di quella del picco principale, la registra come componente bifasica secondaria.

7.1.4 Punto J e deviazione del tratto ST

A partire dalla posizione della S (o della R, se la S non è disponibile), l'algoritmo osserva il primo 15% dell'intervallo compreso tra la fine del QRS e l'inizio dell'onda T. In questa porzione cerca un picco locale; se non lo trova, usa il valore massimo osservato in quell'intervallo. La deviazione del tratto ST viene poi calcolata come differenza, in millivolt, tra il valore nel punto J e la linea di base misurata all'inizio del QRS: un valore positivo corrisponde a un sopraslivellamento, uno negativo a un sottoslivellamento [8].

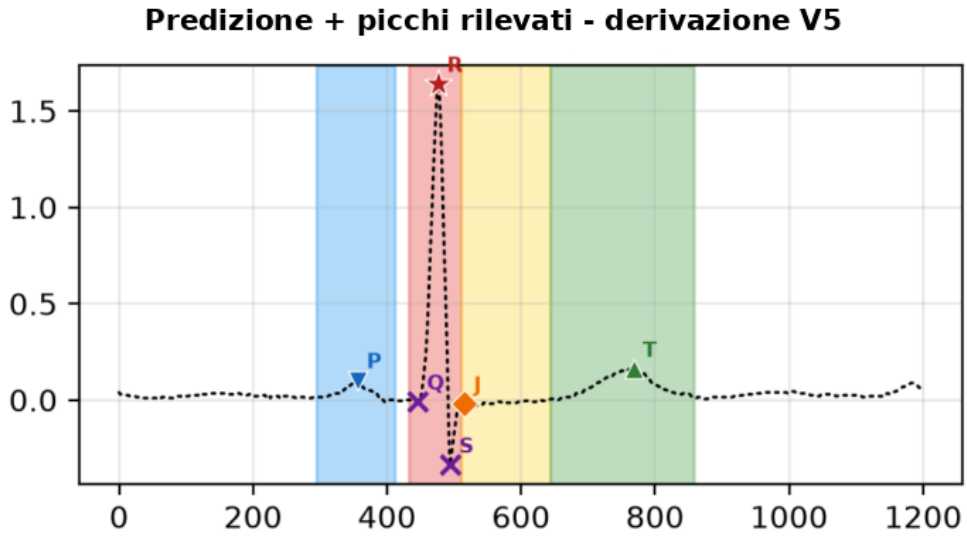


Figura 7.1: Predizione finale e picchi rilevati, derivazione V5. Le bande colorate indicano gli intervalli P (blu), QRS (rosso), ST (giallo) e T (verde) prodotti dall'ensemble; i marker indicano i picchi puntuali (P, Q, R, S, J, T) individuati dall'algoritmo di rilevamento.

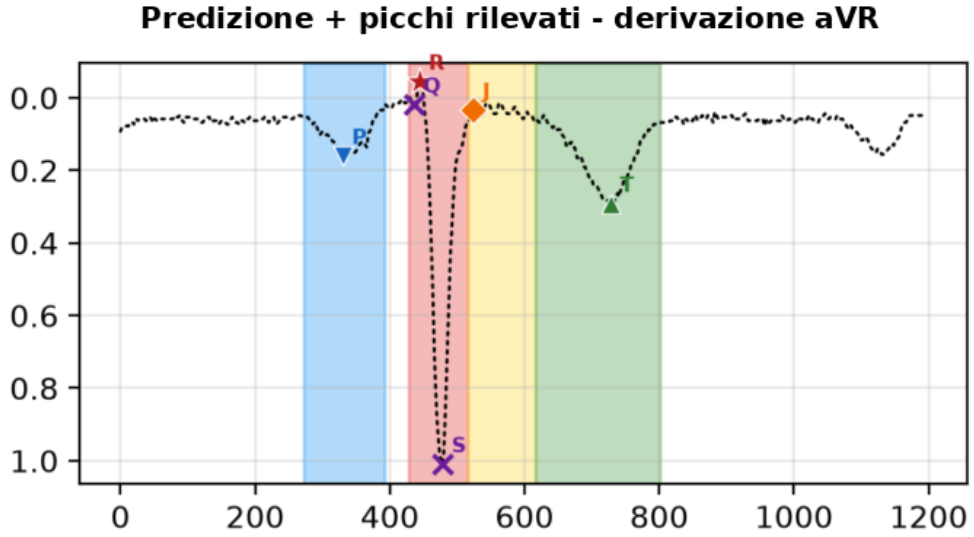


Figura 7.2: Predizione finale e picchi rilevati, derivazione aVR. Stesso schema della Figura 7.1, su una derivazione a polarità invertita per convenzione anatomica: R positiva ma di piccola ampiezza, S marcatamente negativa, T invertita.

7.2 Post-processing

Una volta che la rete di segmentazione produce una maschera grezza, questa deve essere ripulita prima che i picchi possano essere estratti (Sezione 7.1) e i risultati mostrati clinicamente. Il post-processing agisce su tre livelli distinti: la pulizia della maschera sul battito singolo, la pulizia sul tracciato multi-battito, e le correzioni morfologiche puntuali già descritte nella Sezione 7.1.

7.2.1 Pulizia della maschera: battito singolo (mediano)

Sul beat mediano (600 campioni, un solo battito per costruzione) la maschera predetta viene ripulita in tre passaggi consecutivi, applicati a ciascuna onda separatamente.

La chiusura dei gap unisce frammenti della stessa onda separati da un vuoto breve, dovuto tipicamente a rumore che spezza momentaneamente la predizione della rete: fino a 30 ms per il QRS, 50 ms per la P, 80 ms per la T. Poiché ci si aspetta esattamente una sola occorrenza di ciascuna onda in un battito isolato, il codice individua tutte le componenti connesse per classe e tiene solo la più lunga, scartando le altre. Un filtro di durata minima elimina infine i frammenti troppo corti per essere plausibili: 40 ms per QRS e P, mentre per la T la soglia sale a 80 ms, coerentemente con il fatto che sia l'onda più larga del ciclo cardiaco. Due vincoli fisiologici *hard* chiudono la pulizia: la T non può iniziare prima della fine del QRS, e la P non può estendersi oltre l'inizio del QRS; ogni frammento che viola questi vincoli viene azzerato.

7.2.2 Pulizia della maschera - tracciato multi-battito

Sui tracciati lunghi e continui (LUDB, o il tracciato timbro 10 s) la logica cambia natura: qui sono legittimi molti battiti, quindi non si può semplicemente tenere "la componente più lunga" come nel caso singolo, bisogna preservare ogni battito valido e scartare solo ciò che è realmente spurio.

La chiusura dei gap segue soglie leggermente diverse rispetto al caso singolo (QRS $\leq$ 30 ms, P $\leq$ 20 ms, T $\leq$ 80 ms), e il filtro di durata minima è molto più permissivo: scarta solo frammenti minuscoli (10 ms per QRS e P), mentre per la T non è imposta alcuna durata minima.

Da qui in avanti la logica sfrutta le proprietà refrattarie del cuore. Il periodo refrattario ventricolare impone che due complessi QRS non possano susseguirsi a meno di 250 ms di distanza: se accade, il codice confronta i due segmenti e scarta quello più corto, ritenendolo rumore residuo piuttosto che un vero secondo battito. Lo stesso principio si applica alla P con un periodo refrattario di 200 ms. Un'onda T viene classificata come "orfana", e quindi eliminata, se non esiste un QRS la

cui fine ricada nei 1200 ms precedenti al suo inizio, con una tolleranza di 100 ms (50 campioni) oltre l'inizio della T stessa: questo margine serve specificamente ai QRS allargati (blocchi di branca, *pacing*), dove la rete tende a far sovrapporre leggermente l'inizio della T con il tratto finale del QRS ampio, un comportamento che senza questa tolleranza verrebbe penalizzato come se la T non avesse alcun QRS precedente. Infine, se tra due complessi QRS consecutivi vengono rilevate più onde P, si mantiene solo l'ultima, quella immediatamente precedente al QRS successivo, scartando le altre come duplicati o artefatti.

Un caso limite di questa regola riguarda il primo e l'ultimo battito di ciascuna finestra di analisi: se il QRS di riferimento di una T cade fuori dalla finestra stessa (ad esempio un battito che inizia appena prima dell'estremo della finestra, il cui QRS risulta quindi troncato e invisibile all'algoritmo), la regola non trova alcun QRS valido e scarta una T in realtà legittima. Nella pipeline finale (Sezione 8.4.5), questo caso viene gestito escludendo dal vincolo il primo e l'ultimo battito rilevato in ciascuna finestra, la cui T viene mantenuta indipendentemente dalla distanza dal QRS più prossimo.

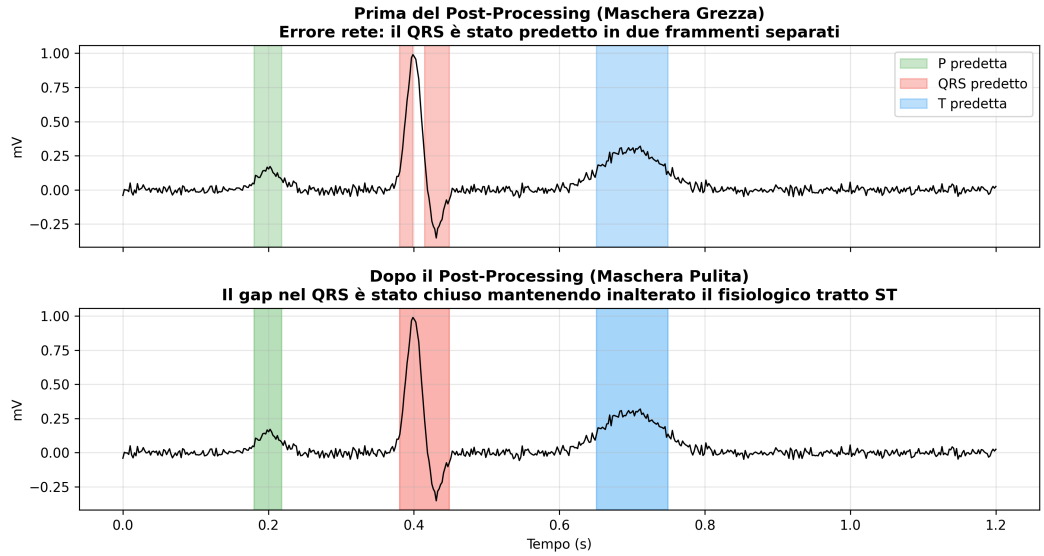


Figura 7.3: Esempio di post-processing morfologico applicato alla segmentazione di un battito mediano isolato. Nel pannello superiore è visibile l'output grezzo della rete neurale, in cui la maschera del complesso QRS (in rosso) risulta frammentata a causa di una momentanea perdita di confidenza del modello. Nel pannello inferiore, l'algoritmo di chiusura dei gap interviene unendo i due frammenti ravvicinati, restituendo una segmentazione fisiologicamente coerente senza intaccare le porzioni limitrofe, come il tratto ST e l'onda T.

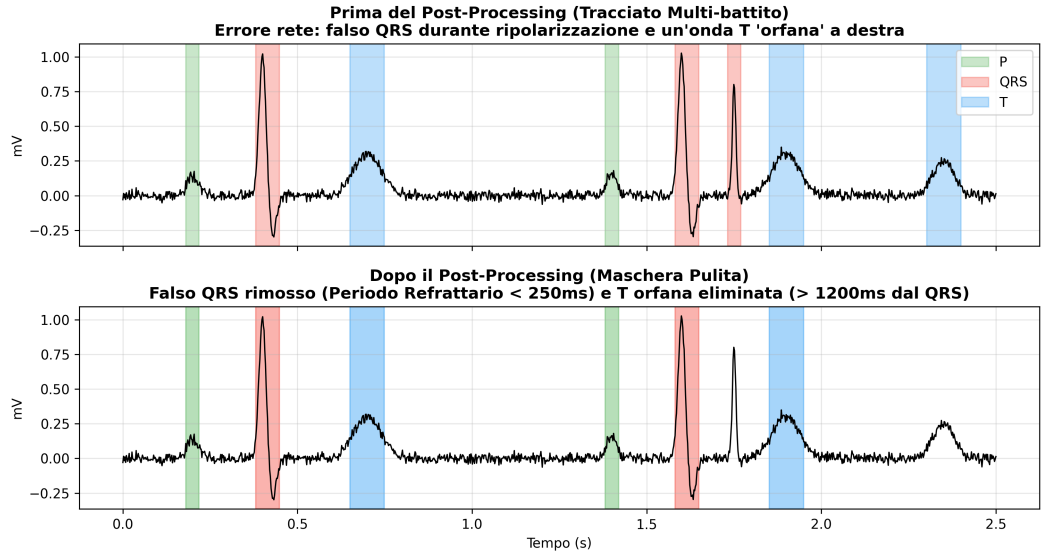


Figura 7.4: Intervento del post-processing su un tracciato multi-battito continuo. Il pannello superiore mostra la predizione grezza inficiata da due artefatti: un QRS spurio rilevato durante la fase di ripolarizzazione (violazione del periodo refrattario di 250 ms*) e un'onda T "orfana" individuata in assenza di un complesso ventricolare associato. Nel pannello inferiore, l'applicazione delle regole cliniche temporali ha permesso all'algoritmo di scartare automaticamente le maschere fittizie, conservando unicamente la sequenza P-QRS-T dei due battiti reali.*

7.2.3 Correzioni morfologiche puntuali

Un'ultima categoria di correzioni non agisce sulla maschera nel suo complesso, ma sui singoli picchi geometrici estratti al suo interno, la stessa logica già descritta in dettaglio nella Sezione 7.1: separare Q, R e S nei complessi QS puri, applicare detrending e tapering prima di localizzare P e T, e isolare lo spike di pacing dalla sua derivata altissima e concentrata in pochi campioni. Questa fase non pulisce la maschera, ma la traduce in coordinate cliniche precise, in millisecondi e millivolt.

In sintesi, la predizione grezza della rete passa sempre attraverso la pulizia della maschera, quella per il battito singolo o quella per il tracciato continuo, a seconda del formato, prima di essere utilizzata: solo a valle di questa pulizia entrano in gioco le regole geometriche della Sezione 7.1 per l'estrazione fine dei picchi e dei punti clinici.

Capitolo 8

Risultati e Applicazione Clinica

8.1 Le quattro modalità di valutazione (Raw, Record pooled, Avg12, Lead II)

Prima di leggere qualsiasi tabella di risultati, è necessario chiarire un punto che altrimenti renderebbe i numeri di questo capitolo difficili da confrontare tra loro: lo stesso identico modello, applicato agli stessi dati, può restituire F1 sensibilmente diverse a seconda di come i campioni vengono aggregati per calcolare la metrica. Non è un artefatto statistico da correggere, ma una scelta metodologica che risponde a domande diverse, e va dichiarata esplicitamente ogni volta che si legge un numero.

Per ciascuna onda $w \in \{\text{P, QRS, T}\}$ e ciascun punto $b \in \{\text{onset, offset}\}$, un punto predetto è conteggiato come vero positivo (TP) se cade entro la tolleranza AAMI EC57 [18, 22] (± 75 campioni a 500 Hz, cioè ± 150 ms) dal punto corrispondente nella *ground truth*; altrimenti è un falso positivo (FP) se non trova corrispondenza in nessun punto reale, o il punto reale resta un falso negativo (FN) se nessuna predizione lo raggiunge entro la tolleranza. Da questi conteggi si calcolano sensibilità, valore predittivo positivo e F1:

$$\begin{aligned} \text{Sensibilità} &= \frac{TP}{TP + FN} \\ \text{PPV} &= \frac{TP}{TP + FP} \\ \text{F1} &= \frac{2 \cdot \text{Sensibilità} \cdot \text{PPV}}{\text{Sensibilità} + \text{PPV}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \end{aligned} \tag{8.1}$$

Quello che cambia tra le quattro modalità è a quale livello i conteggi TP, FP, FN vengono accumulati prima di applicare queste formule.

8.1.1 Raw (per-sample)

I conteggi vengono accumulati su tutto il test set indipendentemente da paziente o derivazione:

$$TP_{\text{Raw}} = \sum_{i=1}^{N} \sum_{d=1}^{12} TP_{i,d} \tag{8.2}$$

dove i è l'indice del paziente, d l'indice della derivazione, N il numero di pazienti nel test set (analogamente per FP e FN). L'F1 finale si calcola una sola volta su questi totali globali. È la modalità più severa poiché un singolo battito anomalo pesa esattamente quanto un errore sistematico ripetuto su tutto il dataset. Questa strategia corrisponde alla metrica "Gross" storicamente raccomandata dalle linee guida e applicata nella letteratura di riferimento [22, 36].

8.1.2 Record pooled

Per ciascun paziente i, i conteggi delle sue 12 derivazioni vengono sommati in un *pool* paziente-specifico, e l'F1 si calcola su quel *pool* prima di mediare tra pazienti:

$$TP_i = \sum_{d=1}^{12} TP_{i,d} \quad \Longrightarrow \quad F1_i = \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \tag{8.3}$$

$$F1_{\text{Record pooled}} = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{8.4}$$

A differenza della modalità Raw, qui ogni paziente contribuisce con lo stesso peso alla media finale, indipendentemente da quanti campioni contenga il suo tracciato. Nota in letteratura come statistica "Average" [22], questa metrica evita che record molto lunghi dominino le prestazioni complessive del modello [36].

8.1.3 Avg 12 derivazioni

L'aggregazione avviene prima della decodifica, non dopo. Sia $\hat{y}_t^{(i,d)}$ l'output softmax del modello al campione t per il paziente i, derivazione d (le 4 componenti corrispondono a Background/P/QRS/T). Si calcola prima la media delle probabilità tra le 12 derivazioni:

$$\bar{\hat{y}}_t^{(i)} = \frac{1}{12} \sum_{d=1}^{12} \hat{y}_t^{(i,d)} \tag{8.5}$$

e solo su questa media si applica l'*argmax* per ottenere una sola maschera di consenso per paziente:

$$M_t^{(i)} = \text{argmax}_{c \in \{0,1,2,3\}} \left(\bar{\hat{y}}_{t,c}^{(i)} \right) \tag{8.6}$$

I conteggi TP, FP, FN (e di conseguenza l'F1) si calcolano su questa singola maschera per paziente, non sulle 12 maschere individuali. È concettualmente diversa dalla modalità Record: lì si decodificano 12 maschere e si accumulano i loro errori; qui, invece, si media l'incertezza del modello prima di decidere (un approccio di *late fusion / soft voting*). In questo modo, le derivazioni in cui il modello è meno sicuro pesano meno rispetto a quelle in cui è molto sicuro, sfruttando la ridondanza spaziale delle 12 derivazioni con logiche di post-processing dimostrate efficaci sulle reti neurali [5].

8.1.4 Solo Lead II

Caso particolare della modalità Raw ristretto a una singola derivazione fissata ($d = \text{II}$):

$$TP_{\text{Lead II}} = \sum_{i=1}^{N} TP_{i,\text{II}} \tag{8.7}$$

con F1 calcolata sui conteggi risultanti. Non è una media né un *pooling* tra derivazioni: rappresenta il risultato che si otterrebbe se il sistema venisse usato con una sola derivazione disponibile (tipicamente la DII, storicamente la più informativa per l'analisi del ritmo [19]). Valutare questa modalità isolata è oggi uno standard cruciale per simulare scenari di monitoraggio ambulatoriale e con dispositivi indossabili *single-lead* [6].

8.1.5 Perché conviene guardare tutte e quattro insieme

Nessuna delle quattro modalità è "quella giusta" in assoluto: raccontano aspetti diversi dello stesso modello. Un valore alto in Avg12 ma basso in Raw suggerisce che il modello beneficia molto dalla ridondanza tra derivazioni, ma fatica quando deve segmentare una singola derivazione isolata. Un valore alto in Solo Lead II ma più basso nelle altre modalità indica il contrario: buone prestazioni sulla derivazione più informativa per il ritmo, ma meno robustezza sulle altre. Per questo, nelle tabelle che seguono, tutte e quattro le modalità vengono riportate fianco a fianco per ciascuna fase, invece di ridurre il confronto a un singolo numero.

8.2 Risultati delle fasi sperimentali

8.2.1 Confronto tra le quattro modalità di valutazione

La Tabella 8.1 riporta l'F1 Macro per ciascuna fase, nelle quattro modalità descritte nella Sezione 8.1, valutata sul dominio di riferimento di ciascuna. DIAG-GE non compare in questa tabella; non essendo un modello addestrato, ma un confronto

diretto tra una simulazione della convenzione GE e la *ground truth* manuale, non produce le quattro modalità di aggregazione; la sua metrica è intrinsecamente una valutazione per singolo campione, discussa a parte nella Sezione 8.2.5.

Fase	**Raw**	**Record**	**Avg12**	**Lead II**
B-LUDB	95.79 %	97.35 %	98.34 %	98.62 %
B-MED	92.19 %	97.01 %	98.37 %	96.67 %
KD-MED	96.05 %	96.05 %	97.98 %	97.52 %
B-TMB	91.96 %	93.56 %	93.43 %	94.07 %
KD-TMB	97.66 %	96.37 %	98.34 %	98.49 %
COMB	97.28 %	97.30 %	98.61 %	98.24 %
CTRL-TMB	96.71 %	96.71 %	97.01 %	96.47 %

Tabella 8.1: F1 Macro per ciascuna fase nelle quattro modalità di valutazione, misurata sul dominio di riferimento di ciascuna.

Il modello B-LUDB mostra il salto più contenuto tra le quattro modalità (95.79 % → 98.62 %, meno di 3 punti): essendo valutato sullo stesso dominio manuale coerente su cui è stato addestrato, risulta già stabile a livello di singolo campione.

Per la baseline B-MED si osserva invece l'escursione più ampia in assoluto (92.19 % → 98.37 %, quasi 6 punti). Come emerge dalla scomposizione per onda (Sezione 8.2.3), il calo nella modalità Raw è concentrato sull'onda P: pochi campioni isolati e rumorosi incidono pesantemente in una metrica non aggregata, mentre vengono filtrati efficacemente non appena si effettua la media tra le derivazioni o si raggruppano i campioni per paziente.

La configurazione distillata KD-MED presenta un comportamento intermedio (96.05 % → 97.98 %, circa 1,9 punti), più contenuto rispetto a B-MED poiché la distillazione corregge parzialmente l'instabilità sull'onda P ereditata dalla baseline.

La baseline sul tracciato esteso B-TMB evidenzia una variazione minima tra tutte le fasi (91.96 % → 94.07 %, poco più di due punti): il formato a 10 secondi offre già di per sé sufficiente contesto temporale affinché la segmentazione del singolo campione risulti robusta, indipendentemente dalla successiva strategia di aggregazione.

Nel caso di KD-TMB emerge un comportamento peculiare: il valore in Record pooled (96.37 %) risulta inferiore sia a Raw (97.66 %) sia ad Avg12 (98.34 %), in controtendenza rispetto all'andamento consueto. Tale dinamica indica che, mediando per singolo paziente invece che sul pool globale, alcuni tracciati specifici penalizzano la media aggregata, evidenziando una moderata variabilità inter-paziente piuttosto che un limite sistematico del modello.

La variante con training combinato COMB mostra una dinamica analoga a B-LUDB (escursione contenuta nell'ordine di 1,3 punti), coerentemente con l'inclusione diretta del dominio manuale nel processo di ottimizzazione; il valore raggiunto in modalità Avg12 (98.61 %) costituisce il massimo assoluto tra tutte le fasi analizzate.

Infine, il controllo sperimentale CTRL-TMB si mantiene entro un intervallo ristretto di appena 1,4 punti (96.47 % – 97.01 %), confermando l'elevata stabilità intrinseca conferita dal formato a tracciato esteso.

8.2.2 Confronto Avg12 tra i tre domini di test

Da qui in avanti si userà solo la modalità Avg12, per coerenza con Moskalenko et al. (2020) e con il resto della tesi. La Tabella 8.2 riporta l'F1 Avg12 per ciascuna onda, su tre domini: LUDB (gold standard manuale), `PRIV_medio` (beat mediano GE) e `PRIV_timbro` (tracciato 10 s GE).

	LUDB			**PRIV_medio**			**PRIV_timbro**		
Fase	**P**	**QRS**	**T**	**P**	**QRS**	**T**	**P**	**QRS**	**T**
B-LUDB	96,1%	99,3%	99,6%	70,9%	99,4%	83,4%	87,5%	96,9%	91,7%
B-MED	95,5%	98,9%	83,1%	95,5%	99,6%	99,9%	92,8%	97,6%	95,1%
KD-MED	96,5%	98,5%	98,9%	71,5%	99,2%	83,6%	86,6%	96,9%	90,5%
B-TMB	96,8%	97,4%	86,1%	93,6%	99,8%	99,8%	96,1%	98,7%	98,4%
KD-TMB	96,1%	99,3%	99,6%	72,4%	98,6%	85,0%	88,8%	97,0%	92,2%
COMB	96,5%	99,6%	99,6%	65,3%	98,0%	79,7%	91,7%	97,1%	95,8%
CTRL-TMB	96,8%	97,4%	85,8%	91,9%	99,8%	99,4%	94,8%	98,5%	97,8%

Tabella 8.2: F1 Avg12 per ciascuna onda sui tre domini di test. In grassetto il valore massimo di ciascuna colonna.

Il valore 99.9 % (B-MED, `PRIV_medio`, onda T) non indica una segmentazione perfetta della ripolarizzazione, ma riflette la natura degenere dell'etichetta in questo dominio: poiché *T_Onset* coincide sempre esattamente con *QRS_Offset* per convenzione GE, il modello deve solo imparare a far iniziare la classe T dove finisce la classe QRS, un pattern unico, mai contraddetto nel *ground truth* di training. Il crollo della stessa metrica a 83.1 % quando il modello viene valutato su LUDB (dove *T_Onset* è un punto clinico indipendente) conferma che l'alto valore su `PRIV_medio` non misura la capacità del modello di individuare la vera fine dello ST, ma solo la sua capacità di riprodurre la convenzione di etichettatura. Questa stessa convenzione è quantificata direttamente, senza alcun modello, nella Sezione 8.2.5.

Sul dominio metodologicamente corretto per ciascuna fase, l'F1 Macro Avg12 risulta riportato in Tabella 8.3.

Fase	Dominio di riferimento	F1 Macro
B-LUDB	LUDB	98.34 %
B-MED	PRIV_medio	98.37 %
KD-MED	LUDB	97.98 %
B-TMB	PRIV_timbro	97.71 %
KD-TMB	LUDB	98.34 %
COMB	LUDB	98.61 %
CTRL-TMB	PRIV_timbro	97.01 %
DIAG-GE	LUDB	93.23 %

Tabella 8.3: F1 Macro in modalità Avg12 valutata sul dominio di riferimento metodologicamente corretto per ciascuna fase.

8.2.3 Confronto per onda: F1, sensibilità e valore predittivo positivo

Guardare le tre onde separatamente, invece che solo l'F1 Macro aggregato, rivela un quadro molto più informativo: le tre onde non condividono affatto lo stesso comportamento, né nell'F1 né nella sua scomposizione in sensibilità (Se) e valore predittivo positivo (PPV), come illustrato in Tabella 8.4.

	Onda P			Complesso QRS			Onda T		
Fase	**Se**	**PPV**	**F1**	**Se**	**PPV**	**F1**	**Se**	**PPV**	**F1**
B-LUDB	98,4%	93,9%	96,1%	99,3%	99,3%	99,3%	99,9%	99,3%	99,6%
B-MED	99,0%	92,3%	95,5%	99,9%	99,2%	99,6%	99,9%	99,9%	99,9%
KD-MED	98,4%	94,7%	96,5%	97,8%	99,3%	98,5%	99,3%	98,6%	98,9%
B-TMB	95,2%	98,4%	96,8%	94,9%	99,9%	97,4%	83,9%	88,5%	86,1%
KD-TMB	98,4%	93,9%	96,1%	98,5%	99,9%	99,3%	99,9%	99,3%	99,6%
COMB	99,9%	93,3%	96,6%	99,9%	99,3%	99,6%	99,9%	99,3%	99,6%
CTRL-TMB	99,1%	90,8%	94,8%	99,2%	97,7%	98,5%	99,2%	96,5%	97,8%
DIAG-GE	95,4%	95,0%	95,2%	96,4%	100,0%	98,2%	82,1%	90,9%	86,3%

Tabella 8.4: Scomposizione per onda (P, QRS, T) delle prestazioni in Sensibilità (Se), Valore Predittivo Positivo (PPV) e F1.

Onda QRS

È l'onda più stabile in assoluto: non scende mai sotto il 96.9 % di F1 in nessuna fase, in nessun dominio. Il complesso QRS è l'evento elettricamente più marcato del battito, ampiezza maggiore e transizione più ripida, e questo lo rende relativamente facile da segmentare indipendentemente da come il modello è stato addestrato o su quale dominio viene valutato. Anche DIAG-GE, che non è un modello ma una semplice simulazione geometrica della convenzione GE, raggiunge un PPV del 100.0 % su questa onda: il punto di ancoraggio della simulazione (*QRS_Onset*) è per costruzione quello meglio definito nella *ground truth* manuale.

Onda T

È l'onda con la variabilità più estrema, e il pattern è diagnostico di per sé: ogni volta che una fase viene valutata su un dominio diverso da quello per cui le sue etichette di training erano pensate, la T crolla. B-MED passa da 99.9 % (`PRIV_medio`, proprio dominio, valore da leggere con la cautela discussa nella Sezione 8.2.2) a 83.1 % (LUDB), un calo di 17 punti, la firma numerica del bias $T_Onset = QRS_Offset$. Le fasi distillate (KD-MED e KD-TMB) non mostrano questo crollo quando valutate su LUDB (98.9 % e 99.6 % rispettivamente), ma lo mostrano ancora se valutate sul proprio dominio nativo PRIV, con l'eccezione cruciale di KD-TMB, che raggiunge un compromesso più equilibrato (92.2 % su `PRIV_timbro`) rispetto a KD-MED (90.5 % su `PRIV_medio`) e soprattutto rispetto a B-MED di partenza. B-TMB e CTRL-TMB, che condividono lo stesso formato ma nessuna distillazione, si comportano in modo speculare a B-MED: alte sul proprio dominio nativo (98.4 % e 97.8 %) e basse su LUDB (86.1 % e 85.8 %).

La simulazione DIAG-GE isola l'onda T come il nucleo del problema, a prescindere dall'interferenza di un modello addestrato. L'F1 pari a 86.3 % (Se 82.1 %, PPV 90.9 %) è quasi identico a quello di B-TMB su LUDB (86.1 %) e di CTRL-TMB (85.8 %): la simulazione diretta della convenzione GE, senza alcun training, produce lo stesso livello di errore che i modelli addestrati su etichette GE riproducono fedelmente. È la controprova più diretta che l'errore sulla T non è un limite dell'architettura neurale, ma un vincolo intrinseco dell'etichetta da cui la rete apprende.

Guardando Se e PPV, sia B-TMB sia DIAG-GE mostrano il quadro più delicato su questa onda: divari contenuti tra Se e PPV (83.9 %/88.5 % e 82.1 %/90.9 % rispettivamente), non il pattern "tanti falsi positivi, pochi falsi negativi" che si osserva altrove, ma un errore bilanciato su entrambi i fronti. È coerente con il fatto che qui la T include ancora lo ST. Il criterio, umano o appreso, non sa bene dove porre il confine, quindi sbaglia in entrambe le direzioni, non solo per eccesso.

Onda P

Il pattern è opposto a quello della T: qui è B-LUDB, addestrata solo su dati manuali, a soffrire di più fuori dal proprio dominio, scendendo al 70.9 % su `PRIV_medio`, il valore più basso in assoluto in tutta la tabella. Le fasi addestrate anche su dati clinici reali (B-MED, B-TMB, CTRL-TMB) restano invece sopra il 91 % su tutti e tre i domini. Questo suggerisce che il problema della P non è un bias sistematico di etichettatura come per la T, ma una questione di robustezza al rumore e alla variabilità morfologica reale dei tracciati clinici, un problema che il solo dominio manuale, piccolo e pulito, non espone al modello durante il training.

Un secondo pattern emerge quasi ovunque nella scomposizione Se/PPV: la sensibilità sull'onda P supera quasi sempre il PPV, spesso di parecchi punti, arrivando fino a un divario di 8,4 punti in CTRL-TMB (99.1 % contro 90.8 %). Il modello tende dunque a sovra-segmentare la P: individua quasi tutti i punti reali, ma a costo di produrre più falsi positivi di quanti sarebbero necessari. DIAG-GE costituisce l'unico caso, unitamente a B-TMB, in cui tale rapporto risulta pressoché bilanciato (95.4 % contro 95.0 %): la geometria fissa della simulazione, priva dell'incertezza tipica di un modello appreso, non introduce lo stesso bias verso il falso positivo.

Sintesi del confronto per onda

Il QRS non ha bisogno di correzioni particolari, in nessuna fase e nemmeno nella simulazione diagnostica. La P ha bisogno di essere esposta a dati clinici reali con rumore, non solo a etichette pulite, e tende sistematicamente a un eccesso di falsi positivi piuttosto che di falsi negativi. La T ha bisogno della vera separazione ST-T, che solo il dominio manuale fornisce nativamente e che la distillazione trasferisce al dominio clinico, e DIAG-GE dimostra che questo bisogno esiste indipendentemente da qualunque modello, essendo già presente nella sola geometria dell'etichetta GE.

8.2.4 Analisi delle prestazioni delle fasi di riferimento (B-LUDB e B-MED)

Le due fasi di riferimento raggiungono F1 Macro molto vicine sul proprio dominio (98.34 % contro 98.37 %), ma il numero aggregato nasconde la differenza reale già isolata nella Sezione 8.2.3: B-LUDB crolla sulla P fuori dominio, B-MED crolla sulla T fuori dominio. Interessante anche il terzo dominio: B-LUDB su `PRIV_timbro` ottiene un F1 macro intermedio (92.0 %), migliore del crollo su `PRIV_medio`, segno che il formato più lungo offre al modello addestrato sul solo dominio manuale più

contesto per generalizzare, anche senza mai aver visto un tracciato clinico reale in training.

8.2.5 Quantificazione del bias sistematico industriale (DIAG-GE)

L'esperimento DIAG-GE consente di quantificare il problema alla radice, prima ancora di procedere all'addestramento di una rete neurale. Non si tratta di un modello di deep learning, bensì di una simulazione geometrica deterministica della convenzione industriale GE (*T_Onset* fissato a *QRS_Offset*, con finestre per P, QRS e T dimensionate sulle durate mediane osservate nei tracciati clinici reali). Tale logica viene applicata direttamente al *gold standard* manuale di LUDB, consentendo di isolare con esattezza l'errore generato dalla convenzione di annotazione, escludendo qualunque distorsione legata alla capacità di generalizzazione dei modelli.

Il risultato più rilevante riguarda i singoli bordi delle onde, non solo l'F1 aggregato (Tabella 8.5).

Bordo	**Errore mediano**	**Giudizio clinico**
P_Onset	−24.0 ms	Buono
P_Offset	−14.0 ms	Molto Buono
QRS_Onset	0.0 ms	Eccellente
QRS_Offset	6.0 ms	Molto Buono
T_Onset	−94.0 ms	**Fuori Tolleranza**
T_Offset	38.0 ms	Accettabile

Tabella 8.5: Errore mediano sui singoli bordi delle onde nella simulazione DIAG-GE valutata rispetto al gold standard manuale LUDB.

Il contrasto è evidente: QRS_Onset è pressoché perfetto, e anche P_Offset, P_Onset e T_Offset restano in un range clinicamente accettabile o migliore. È solo il T_Onset a cadere fuori tolleranza, isolando con precisione dove nasce il problema; non è un errore diffuso su tutta l'annotazione GE, ma un errore concentrato esattamente nel punto che la convenzione *T_Onset* = *QRS_Offset* distorce per costruzione. La durata dell' ST che questa convenzione assorbe sistematicamente nella T ha una mediana di 99.9 ms, pressoché l'intero segmento che il giudizio clinico riconoscerebbe come tratto ST viene inglobato nell'onda T.

Questo stesso fenomeno, misurato qui in modo puro e diretto, è ciò che B-TMB e CTRL-TMB mostrano indirettamente quando vengono addestrate sulle etichette GE senza distillazione. B-TMB e CTRL-TMB condividono infatti formato ed

etichette di partenza, entrambe senza distillazione: l'unica differenza è il dominio di valutazione. CTRL-TMB *in-domain* raggiunge 97.01 % di F1 Macro; B-TMB sullo stesso identico formato ma valutata sul dominio manuale scende a 93.43 %, un calo di 3,6 punti dovuto al solo trasferimento *cross-domain*. Come già evidenziato nella Sezione 8.2.3, il crollo si concentra sull'onda T: 98.4 % (B-TMB sul proprio dominio nativo, dove include ancora lo ST) contro 86.1 % (stessa fase valutata sul dominio manuale), un valore quasi identico all'86.3 % di DIAG-GE, a conferma che il modello addestrato sulle etichette GE non fa altro che riprodurre fedelmente il bias già misurato nella sola convenzione di etichettatura.

8.2.6 Efficacia della Knowledge Distillation e confronto tra strategie di addestramento

L'analisi dell'impatto della Knowledge Distillation sul formato a singolo battito mediano (passaggio da B-MED a KD-MED) evidenzia un comportamento asimmetrico tra i domini di valutazione. Quando testato sul dominio manuale LUDB, il modello distillato mostra un netto recupero di accuratezza sull'onda T, con l'F1 che sale dall'83.1 % della baseline B-MED al 98.9 %. Tuttavia, quando KD-MED viene valutato nel proprio dominio clinico nativo (`PRIV_medio`), l'F1 per l'onda T si attesta all'83.6 %, valore sostanzialmente sovrapponibile a quello della baseline di partenza (83.1 %). Tale evidenza dimostra che la distillazione applicata al solo battito mediano isolato, privo di contesto temporale e variabilità morfologica inter-battito, consente alla rete di riprodurre la corretta separazione ST-T esclusivamente se interrogata con il medesimo formato dei dati manuali, senza però riuscire a trasferire e applicare spontaneamente tale conoscenza ai tracciati clinici reali di produzione.

Al contrario, la distillazione condotta sul tracciato esteso a 10 secondi (passaggio da B-TMB a KD-TMB) rappresenta il risultato centrale e più significativo dell'intero lavoro sperimentale. Sul dominio manuale LUDB, l'F1 per l'onda T compie un incremento decisivo, passando dall'86.1 % di B-TMB al 99.6 % di KD-TMB, un valore che eguaglia perfettamente il limite superiore stabilito dal modello di riferimento addestrato su sole annotazioni umane (B-LUDB a 99.6 %) e supera di oltre 13 punti il limite teorico imposto dalla pura convenzione industriale (86.3 % in DIAG-GE). L'elemento più rilevante risiede nel fatto che tale beneficio non rimane confinato al dominio manuale, ma si riflette positivamente anche sul dominio clinico nativo `PRIV_timbro`: in questo contesto l'F1 calcolato rispetto alle etichette GE scende dal 98.4 % (in cui B-TMB riproduceva acriticamente l'errore includendo lo ST nell'onda T) al 92.2 % di KD-TMB. Questa apparente flessione metrica è in realtà la prova empirica del successo del paradigma di distillazione: il modello ha appreso come correggere l'etichetta di training, discriminando auto-

nomamente il tratto ST dall'onda T anche in assenza di annotazioni manuali nei tracciati di produzione.

Un'ulteriore strategia indagata consiste nell'addestramento combinato multimodale (COMB), in cui i dataset LUDB e clinico vengono uniti direttamente nella stessa procedura di ottimizzazione. Sebbene tale approccio consenta di raggiungere prestazioni eccezionali sul dominio manuale (F1 Macro pari a 98.61 %, F1 per QRS e T pari a 99.6 % e una sensibilità del 99.9 % su tutte le onde, a fronte di zero falsi negativi sul test set manuale), esso manifesta un grave decadimento prestazionale quando applicato ai dati clinici reali. Sul dominio `PRIV_medio`, infatti, COMB subisce un crollo pronunciato sia sull'onda P (65.3 %, il valore peggiore tra tutte le configurazioni) sia sull'onda T (79.7 %). La co-ottimizzazione simultanea su distribuzioni disomogenee porta la rete a privilegiare il dominio manuale più pulito, riducendo la tolleranza al rumore clinico e confermando la netta superiorità della Knowledge Distillation a due stadi rispetto al semplice accorpamento dei dataset.

8.2.7 Sintesi dei risultati delle singole fasi sperimentali

Il percorso metodologico completo, dalla quantificazione del bias alla sua correzione, è sintetizzato in modo inequivocabile dalla dinamica dell'onda T: un errore mediano di -94.0 ms su *T_Onset* introdotto dalla sola convenzione GE rispetto al gold standard manuale (DIAG-GE); un F1 dell'86.1 % raggiunto dalla baseline addestrata su tali etichette estese (B-TMB); e il raggiungimento del 99.6 % di F1 dopo l'applicazione della Knowledge Distillation (KD-TMB), rendendo le predizioni indistinguibili dal riferimento manuale.

Il dato più significativo risiede nel fatto che solo la distillazione sul formato esteso si dimostra efficace anche sul proprio dominio nativo (da 98.4 % con lo ST ancora incluso a una segmentazione che lo separa correttamente), a differenza della distillazione sul beat mediano, che migliora solo se giudicata sul dominio manuale ma resta sostanzialmente invariata sul proprio dominio di produzione.

Il confronto per onda (Sezione 8.2.3) aggiunge un secondo livello di lettura: il QRS non richiede alcun intervento correttivo in nessuna fase; la P richiede esposizione a dati clinici reali con rumore; la T necessita della distillazione sul formato esteso, l'unico intervento in grado di superare il vincolo imposto dalla convenzione industriale originariamente misurato da DIAG-GE. È questa combinazione, formato esteso più distillazione a due stadi, a costituire la base metodologica per l'ensemble finale (Sezione 8.4).

8.3 Errori clinici in millisecondi

8.3.1 Perché guardare oltre l'F1

L'F1 con tolleranza AAMI (150 ms) risponde a una domanda binaria: il punto predetto cade abbastanza vicino al punto vero da essere considerato corretto? Ma un errore di 5 ms e un errore di 140 ms ricevono esattamente lo stesso trattamento in quella metrica, purché entrambi restino sotto soglia. Per questo, oltre all'F1, vale la pena guardare direttamente l'errore in millisecondi tra il punto predetto e il punto reale, calcolato battito per battito e poi aggregato per paziente in modalità Avg12, sugli stessi checkpoint già verificati nella Sezione 8.2.

La Tabella 8.6 riporta l'errore medio e l'errore assoluto medio (MAE) in millisecondi per ciascun bordo d'onda, su tutte e sette le fasi.

8.3.2 Tabella completa

Fase	P_Onset	P_Offset	QRS_Onset	QRS_Offset	T_Onset	T_Offset
B-LUDB	$-62,6$ / $151,5$	$-53,3$ / $166,7$	$+1,9$ / $6,1$	$+11,4$ / $13,4$	$-0,4$ / $7,6$	$+21,6$ / $21,8$
B-MED	$-5,5$ / $7,5$	$+12,5$ / $13,9$	$-1,8$ / $2,4$	$+1,9$ / $3,6$	$+1,9$ / $3,6$	$+4,3$ / $6,6$
KD-MED	$-12,6$ / $102,7$	$+0,5$ / $110,0$	$+48,6$ / $52,4$	$+61,3$ / $63,7$	$-2,8$ / $8,6$	$+20,8$ / $22,4$
B-TMB	$-112,5$ / $114,0$	$-97,6$ / $124,3$	$+86,5$ / $93,3$	$+98,1$ / $100,9$	$+224,9$ / $344,1$	$+368,7$ / $369,9$
KD-TMB	$-63,4$ / $151,4$	$-52,7$ / $167,7$	$+1,6$ / $5,4$	$+11,5$ / $14,3$	$-2,9$ / $8,7$	$+23,7$ / $24,1$
COMB	$-108,3$ / $113,2$	$-97,9$ / $121,9$	$-1,3$ / $5,7$	$+15,1$ / $15,1$	$-7,5$ / $9,9$	$+29,5$ / $29,5$
CTRL-TMB	$-36,7$ / $45,2$	$-22,1$ / $52,9$	$-6,5$ / $20,3$	$-0,3$ / $20,7$	$-2,9$ / $17,6$	$-6,0$ / $21,5$

Tabella 8.6: Errore medio / Errore Assoluto Medio (MAE) in millisecondi per ciascun bordo d'onda (modalità Avg12).

E gli intervalli e le durate derivate sono riportati in Tabella 8.7.

Fase	PR_Interval	QT_Interval	P_Duration	QRS_Duration	ST_Segment
B-LUDB	$+64,2$ / $152,6$	$+19,7$ / $22,7$	$+9,4$ / $30,8$	$+9,5$ / $14,1$	$-11,8$ / $14,8$
B-MED	$+3,8$ / $6,8$	$+6,1$ / $7,9$	$+17,9$ / $19,1$	$+3,7$ / $5,0$	$0,0$ / $0,0$
KD-MED	$+63,4$ / $151,0$	$-27,8$ / $66,8$	$+13,2$ / $23,5$	$+12,7$ / $17,7$	$-64,1$ / $66,7$
B-TMB	$+203,5$ / $206,0$	$+282,2$ / $283,2$	$+15,0$ / $35,4$	$+11,6$ / $15,4$	$+126,8$ / $268,8$
KD-TMB	$+64,6$ / $153,3$	$+22,1$ / $23,3$	$+10,6$ / $32,1$	$+9,9$ / $15,1$	$-14,4$ / $16,4$
COMB	$+106,8$ / $112,1$	$+30,8$ / $30,8$	$+10,4$ / $31,1$	$+16,4$ / $17,4$	$-22,6$ / $23,6$
CTRL-TMB	$+35,9$ / $46,1$	$+0,5$ / $14,9$	$+14,7$ / $20,8$	$+6,3$ / $7,1$	$-2,7$ / $3,1$

Tabella 8.7: Errore medio / Errore Assoluto Medio (MAE) in millisecondi per intervalli e durate derivate (modalità Avg12).

8.3.3 Il caso B-MED: ST_Segment esattamente zero

Per B-MED, l'errore su `ST_Segment` è esattamente pari a 0,0/0,0, senza alcuna variazione. Tale esito non è casuale: su questo dominio *T_Onset* coincide sempre con *QRS_Offset* per costruzione, quindi la differenza tra i due punti, predetta e reale, è matematicamente nulla per definizione, non per merito dell'apprendimento del modello. È la medesima osservazione già evidenziata in termini di F1 (Sezione 8.2.2, valore 99.9 %), qui confermata anche a livello di singolo millisecondo.

8.3.4 KD-MED: un F1 alto che nasconde un errore reale

Un caso particolarmente significativo riguarda il comportamento di KD-MED sul complesso QRS: *QRS_Onset* presenta un errore medio di 48.6 ms (MAE 52.4 ms) e *QRS_Offset* di 61.3 ms (MAE 63.7 ms), ordini di grandezza peggiori rispetto a B-LUDB o KD-TMB, che restano entrambe sotto i 15 ms su questi medesimi bordi. Ciononostante, valutando l'F1 QRS di KD-MED nella Sezione 8.2.3, il valore rimane elevato (98.5 %), poiché uno scostamento di 50–60ms rientra pienamente nella tolleranza AAMI di 150ms e viene quindi conteggiato come predizione corretta alla pari di un errore di pochi millisecondi. Rappresenta l'evidenza più diretta del limite intrinseco delle metriche basate su soglia di tolleranza: l'accuratezza temporale clinica e l'F1 convenzionale possono delineare quadri qualitativi differenti per il medesimo modello.

8.3.5 B-TMB: il bias trasferito, misurato in millisecondi

La baseline B-TMB evidenzia gli errori più marcati dell'intera tabella, in particolare su *T_Onset* (224.9 ms, MAE 344.1 ms) e *T_Offset* (368.7 ms, MAE 369.9 ms), valori che superano abbondantemente la tolleranza AAMI stessa. Si tratta della medesima distorsione già quantificata in modo diretto da DIAG-GE (Sezione 8.2.5, *T_Onset* mediano di −94.0 ms), qui manifestata attraverso una rete neurale che ha assimilato tali etichette sul formato esteso senza interventi di correzione. Il segno positivo riflette la direzione del calcolo (predetto meno reale) ed è pienamente coerente con l'origine del bias. Da notare anche l'intervallo `QT_Interval`, che eredita direttamente tale scostamento (282.2 ms), e il `PR_Interval`, dilatato dall'errore sull'onda P (203.5 ms).

8.3.6 KD-TMB e B-LUDB: precisione quasi identica

I modelli KD-TMB e B-LUDB esibiscono profili di errore marcatamente simili tra loro su quasi tutti i bordi: *QRS_Onset* (1.6 ms contro 1.9 ms), *T_Onset* (−2.9 ms

contro -0.4 ms), *T_Offset* (23.7 ms contro 21.6 ms). Si ottiene così la controprova, espressa in millisecondi, di quanto osservato tramite l'F1: a seguito della distillazione sul tracciato esteso, il modello raggiunge una precisione clinica sostanzialmente indistinguibile da quella del modello addestrato esclusivamente su annotazioni manuali.

Un aspetto rilevante è che sia B-LUDB sia KD-TMB mostrano un errore più esteso su *P_Onset* e *P_Offset* (attorno a $-53/-63$ms, con MAE superiore a 150ms), a fronte di un F1 sulla P superiore al 96 %. Ciò si spiega con la natura dell'onda P, caratterizzata da minore ampiezza e contorni elettrici più sfumati, che genera una distribuzione degli errori con coda lunga: un numero contenuto di casi complessi incrementa sia la media sia il MAE, pur preservando un F1 soddisfacente dato che la maggior parte dei battiti rimane entro la soglia di tolleranza.

8.3.7 COMB e CTRL-TMB

Il modello COMB si colloca in una posizione intermedia sulla P (circa $-98/-108$ms, similmente a B-TMB) mantenendo un QRS estremamente preciso (*QRS_Onset* -1.3 ms) e una T controllata (*T_Onset* -7.5 ms); tale andamento riflette l'inclusione diretta del dominio manuale nel training, che stabilizza QRS e T ma non risolve le difficoltà sul canale P riscontrate nelle configurazioni basate su formato esteso.

La configurazione CTRL-TMB registra errori generalmente più contenuti sulla P ($-36,7/-22,1$ms) rispetto alle altre fasi timbro, ma con una maggiore dispersione sui bordi (divario accentuato tra errore medio e MAE), imputabile all'elevata numerosità ed eterogeneità del proprio test set clinico `PRIV_timbro`.

L'analisi in millisecondi conferma quanto emerso tramite l'F1: il QRS risulta il complesso più accurato in quasi tutte le configurazioni, a eccezione di KD-MED in cui l'errore reale è mascherato dall'ampia tolleranza; la P evidenzia scostamenti maggiori dovuti a morfologie complesse; la T illustra il contrasto più netto, con B-TMB che supera i 300ms di MAE sullo stesso bordo in cui KD-TMB scende sotto i 10ms, quantificando alla scala temporale clinica il passaggio dall'etichetta viziata alla sua correzione.

8.4 Costruzione del Sistema Definitivo

8.4.1 Obiettivi e metodologia di fusione

Il sistema finale è stato progettato per rispondere a due finalità cliniche distinte. Da un lato, la necessità di disporre di una rete di controllo capace di massimizzare la fedeltà e le prestazioni sui file GE di produzione, replicandone accuratamente le

etichette per avere una baseline aziendale. Dall'altro, l'obiettivo primario della tesi: realizzare un'architettura avanzata in grado di operare una separazione anatomica rigorosa tra segmento ST ed onda T sui medesimi tracciati clinici.

Ogni combinazione onda-modello è stata verificata con una ricerca esaustiva: fondere realmente le maschere prodotte da ciascuna combinazione possibile di fasi (una per onda), applicare il post-processing, e misurare l'F1 sul risultato complessivo (anziché stimarlo dai valori isolati di ciascuna fase, perché le maschere si sovrascrivono in fusione e il post-processing giudica un canale guardando i bordi degli altri).

8.4.2 La Rete di Controllo GE

La ricerca esaustiva su `PRIV_medio` ha inizialmente evidenziato come la fase B-MED (addestrata su battito singolo) fosse estremamente competitiva per l'onda P. La scomposizione dell'F1 Avg12 di ciascuna fase per onda sullo stesso dominio è riportata in Tabella 8.8.

Fase	**F1 P**	**F1 QRS**	**F1 T**
B-LUDB	70.9 %	99.4 %	83.4 %
B-MED	95.5 %	99.6 %	99.9 %
KD-MED	71.5 %	99.2 %	83.6 %
B-TMB	93.6 %	99.8 %	99.8 %
KD-TMB	72.4 %	98.6 %	85.0 %
COMB	65.3 %	98.0 %	79.7 %
CTRL-TMB	91.9 %	99.8 %	99.4 %

Tabella 8.8: F1 Avg12 per onda sul dominio `PRIV_medio` per ciascuna fase.

Sull'onda P, B-MED domina con margine ampio (95.5 % contro il secondo migliore, B-TMB, a 93.6 %): è la fase addestrata e validata direttamente su questo formato, con le migliori prestazioni sulla sola onda in cui le altre fasi faticano di più (crolli fino al 65–72% per le fasi orientate al formato timbro o alla distillazione). Sul complesso QRS, B-TMB e CTRL-TMB pareggiano al vertice (99.8 %), entrambe nettamente sopra B-MED (99.6 %). Sull'onda T, B-MED è ancora la migliore (99.9 %), coerente con il fatto che qui l'etichetta segue la convenzione GE (*T_Onset* = *QRS_Offset*, Sezione 8.2.2) e la fase addestrata direttamente su quelle etichette la riproduce più fedelmente di qualunque altra.

Il pareggio sul QRS tra B-TMB e CTRL-TMB va sciolto con ulteriori evidenze, poiché su `PRIV_medio` le due fasi sono identiche fino alla seconda cifra decimale, come mostrato in Tabella 8.9.

Dominio	F1_QRS — B-TMB	F1_QRS — CTRL-TMB	Differenza
LUDB	97.38 %	97.38 %	0,00
PRIV_medio	99.80 %	99.80 %	0,00
PRIV_timbro	98.65 %	98.45 %	**+0,20**

Tabella 8.9: Confronto dell'F1 QRS tra B-TMB e CTRL-TMB nei tre domini di test.

Solo estendendo il confronto al dominio `PRIV_timbro` emerge una differenza, piccola ma univoca, a favore di B-TMB sul complesso QRS. Inoltre, l'analisi trasversale rivela che la fase B-TMB garantisce una stabilità media eccellente sull'onda P (raggiungendo il 96.8% su LUDB e il 96.1% su `PRIV_timbro`). Sulla base di questi risultati, e per il principio di parsimonia architetturale, si è stabilito che per replicare il comportamento della macchina GE (baseline) non è necessario alcun ensemble: la singola rete B-TMB è l'estrattore ottimo e sufficiente.

8.4.3 Selezione dei modelli per l'Ensemble Fisiologico

Se la B-TMB è sufficiente a replicare le etichette industriali GE, essa tuttavia fallisce nel recupero della vera anatomia cardiaca (onda T), copiandone il bias. Per risolvere questo problema clinico, è necessario costruire un'architettura dedicata: l'Ensemble Fisiologico. Per le onde P e QRS, la scelta ricade naturalmente sulla B-TMB per la sua robustezza al rumore. Al contrario, il canale T va selezionato valutando il dominio LUDB, l'unico in cui è presente un vero segnale manuale per la separazione dello ST. In questo caso non si tratta di un artefatto, bensì del comportamento clinico desiderato. KD-TMB, sviluppata appositamente tramite distillazione per questo scopo (Sezione 8.2.6), costituisce la scelta corretta per l'onda T.

8.4.4 Composizione finale dell'Ensemble

L'architettura definitiva del sistema si riduce quindi a due configurazioni operative, schematizzate nella Tabella 8.10: una rete di controllo per l'uso standard, e un Ensemble a 2 reti per l'analisi fisiologica avanzata.

Entrambe le configurazioni lavorano esclusivamente sul formato continuo a 10 secondi. L'Ensemble Fisiologico condivide i canali P e QRS con la rete di controllo, delegando unicamente il tratto ST/onda T alla rete distillata KD-TMB. Questa scelta ottimizza drasticamente le risorse computazionali, richiedendo al massimo due modelli operativi in parallelo.

Canale	Rete di Controllo GE	Ensemble Fisiologico
P	B-TMB	B-TMB
QRS	B-TMB	B-TMB
T	B-TMB	KD-TMB

Tabella 8.10: Composizione del sistema finale: Rete di Controllo GE vs Ensemble Fisiologico.

8.4.5 Prestazioni dell'Ensemble Fisiologico ed effetto del post-processing

La Tabella 8.11 sintetizza le prestazioni dell'Ensemble Fisiologico (P=B-TMB, QRS=B-TMB, T=KD-TMB) sui tre domini di validazione, quantificando l'impatto delle regole di post-processing (PP).

Dominio	Modalità	F1 P	F1 QRS	F1 T	F1 Macro
LUDB (Fisiologia)	Senza PP	96.77 %	97.38 %	99.64 %	97.93 %
	Con PP	96.77 %	97.38 %	96.63 %	96.93 %
PRIV_medio (Singolo battito)	Senza PP	93.60 %	99.80 %	87.32 %	93.57 %
	Con PP	94.29 %	99.60 %	98.58 %	97.49 %
PRIV_timbro (Tracciato 10s)	Senza PP	96.07 %	98.65 %	93.87 %	96.20 %
	Con PP	96.32 %	98.65 %	95.92 %	96.96 %

Tabella 8.11: Impatto del post-processing (PP) sull'Ensemble Fisiologico nei tre domini di validazione.

L'analisi mostra come il post-processing apporti un miglioramento netto sui domini clinici rumorosi (`PRIV_medio` e `PRIV_timbro`), dove l'onda T e l'onda P beneficiano enormemente della stabilizzazione contestuale (l'F1 Macro su `PRIV_medio` sale dal 93.57% al 97.49%). Sul dominio ideale LUDB, il PP causa invece una lieve flessione sull'onda T, coerente con il fatto che il meccanismo è progettato per regolarizzare dati clinici sporchi e può risultare troppo stringente su tracciati manuali fisiologicamente perfetti. In ogni caso, l'Ensemble conferma un recupero straordinario dell'onda T (99.64%) garantendo al contempo robustezza clinica.

8.4.6 Valutazione della separazione ST su file GE e relativi limiti

Poiché nessun dataset clinico di produzione (`PRIV_medio`, `PRIV_timbro`) presenta una separazione reale tra tratto ST ed onda T (l'annotazione GE fissa per

convenzione $T_Onset = QRS_Offset$), l'unico approccio per valutare quantitativamente il canale T dell'Ensemble su tali dati consiste nel confrontarlo con le pseudo-etichette generate dal modello teacher B-LUDB. Questa metodologia rappresenta una valida stima ma non una prova assoluta, in quanto misura la coerenza rispetto a un modello anziché a una validazione clinica autonoma. L'unica verifica verso un vero *ground truth* manuale rimane quella effettuata su LUDB, dove tuttavia l'ensemble non può essere valutato *end-to-end* per via del bias di dominio precedentemente esposto.

In parallelo si rende indispensabile un controllo di tipo qualitativo basato sull'ispezione visiva diretta dei tracciati GE reali, sovrapponendo le bande P, QRS, ST e T al segnale originale per accertare la plausibilità clinica della segmentazione, un controllo qualitativo che nessun numero aggregato può sostituire.

8.5 Integrazione nell'app Streamlit

8.5.1 Scopo e flusso dell'applicazione

Per rendere verificabile in modo diretto, non solo numerico, il comportamento dell'ensemble sui file GE reali, è stata sviluppata un'applicazione interattiva in Streamlit (disponibile all'indirizzo `ecganalyzer.streamlit.app`) che permette di caricare un tracciato, eseguire l'intera pipeline della tesi, e ispezionarne visivamente il risultato.

L'app accetta due sorgenti dati: file GE (`.xml`), il formato di produzione su cui è pensato l'uso reale del sistema, e file LUDB (`WFDB`), utili per confrontare la predizione con una vera *ground truth* manuale quando disponibile. L'utente può scegliere quale modello usare per generare le maschere testando liberamente una qualsiasi delle reti addestrate (sebbene l'attenzione clinica ricada sulla rete di controllo B-TMB e sull'Ensemble Fisiologico). Inoltre, è possibile regolare direttamente la soglia di correlazione del meccanismo *self-template* LOO (Sezione 5.3, di default 0,70) usato per allineare le annotazioni sul tracciato timbro, e selezionare quale delle 12 derivazioni visualizzare.

8.5.2 Cosa mostra l'app

Per la derivazione selezionata, l'app presenta tre pannelli sovrapposti sullo stesso tracciato da 10 secondi: la *ground truth*, la predizione diretta senza post-processing, e la predizione dopo il post-processing (Sezione 7.2), permettendo un confronto visivo immediato dell'effetto della pulizia della maschera, lo stesso tipo di verifica condotta numericamente nella Sezione 8.4.5.

Al di sotto, un secondo confronto isola un singolo battito, affiancando la *ground truth* alla predizione con i picchi puntuali (P, Q, R, S, T, punto J) individuati dall'algoritmo di rilevamento descritto nella Sezione 7.1.

Un pannello laterale riporta infine la diagnosi del paziente: le macro-classi GE attivate secondo il criterio a priorità della Tabella 5.1, insieme agli *statement* diagnostici originali estratti direttamente dal file XML, per un riscontro immediato tra il quadro clinico dichiarato e la segmentazione prodotta.

8.5.3 Verifica visiva: Controllo GE vs Ensemble Fisiologico

Per valutare qualitativamente il comportamento del sistema su un tracciato GE reale di un paziente sano (`HEALTHY.xml`, derivazione I, con tutti i 10 battiti inclusi dalla proiezione mediana LOO), le Figure 8.1 e 8.2 presentano l'interfaccia dell'applicazione interattiva in orientamento orizzontale a tutta pagina, garantendo la massima leggibilità di tutti i dettagli del tracciato, dei picchi puntuali e del quadro diagnostico.

La Figura 8.1 mostra il risultato ottenuto con la Rete di Controllo B-TMB: l'onda T (banda verde) parte immediatamente in corrispondenza del bordo finale del QRS ($T_Onset = QRS_Offset$), riproducendo con totale fedeltà la convenzione industriale del costruttore MAC2000.

La Figura 8.2 mostra invece lo stesso tracciato analizzato dall'Ensemble Fisiologico: l'algoritmo isola ed evidenzia in giallo la banda del segmento ST fisiologico, offrendo una rappresentazione clinicamente distinta ed essenziale per la valutazione cardiologica della ripolarizzazione ventricolare.

In entrambi i casi, il confronto visivo tra la predizione diretta e quella dopo il post-processing (pannelli 2 e 3) conferma l'elevata stabilità dell'inferenza, fornendo un riscontro qualitativo coerente con le valutazioni quantitative.

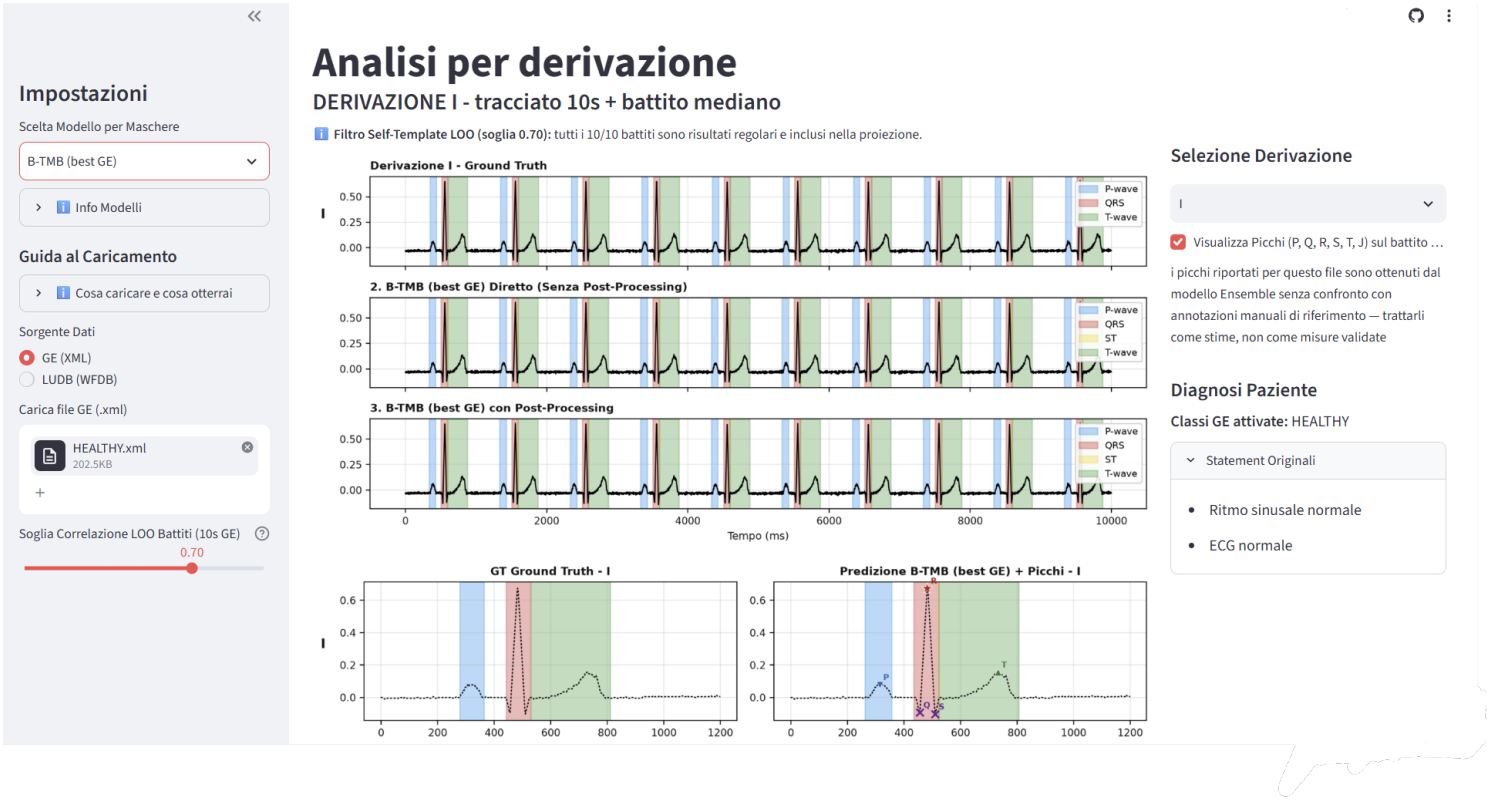


Figura 8.1: Schermata dell'applicazione Streamlit (`ecganalyzer.streamlit.app`) per l'ispezione visiva di un tracciato GE reale (`HEALTHY.xml`, derivazione I) mediante la Rete di Controllo B-TMB. L'onda T (banda verde) inizia direttamente al termine del QRS (T_Onset = QRS_Offset), rispecchiando la convenzione industriale GE.

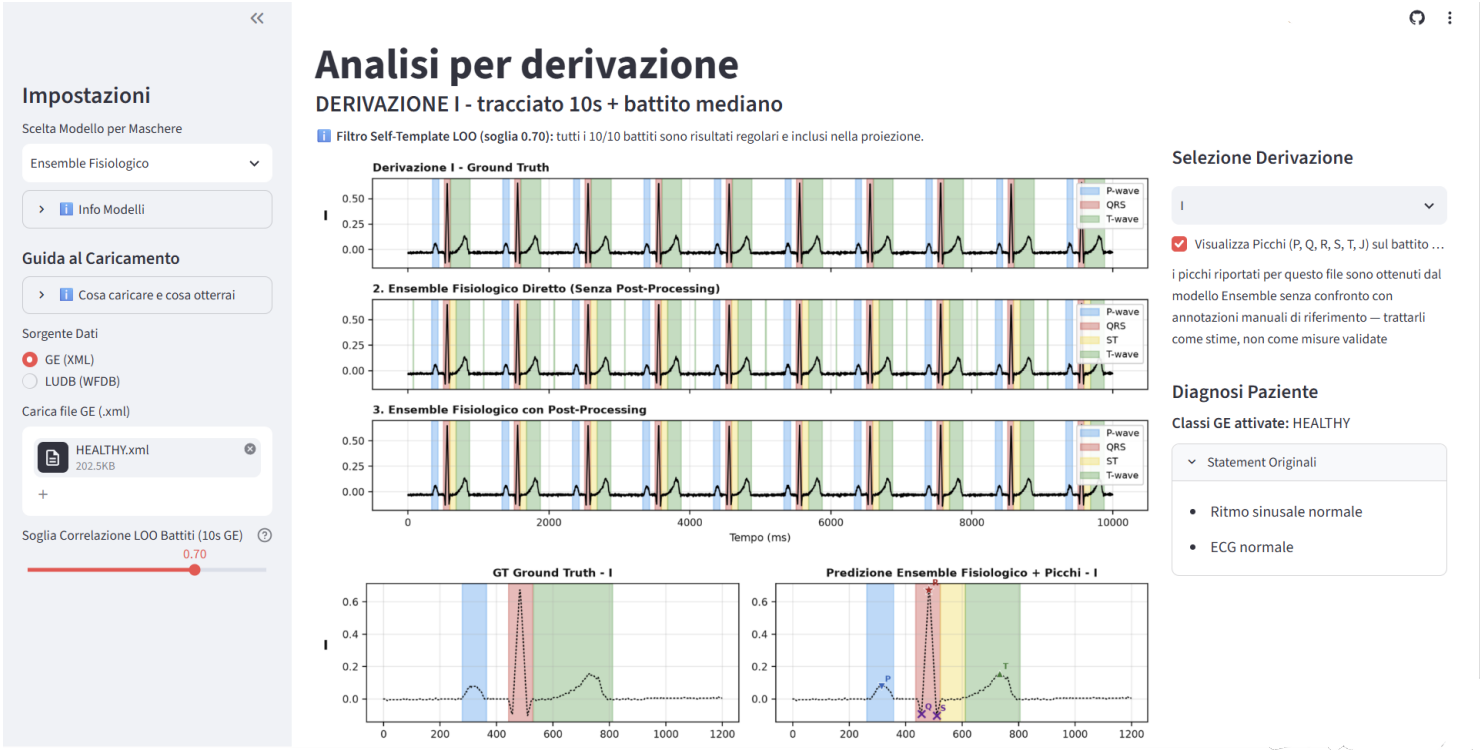


Figura 8.2: Schermata dell'applicazione Streamlit (`ecganalyzer.streamlit.app`) per l'ispezione visiva dello stesso tracciato GE (`HEALTHY.xml`, derivazione I) mediante l'Ensemble Fisiologico. L'algoritmo isola ed evidenzia in giallo la banda del segmento ST fisiologico.

Capitolo 9

Discussione

I capitoli precedenti hanno presentato i numeri: le tabelle di F1, gli errori in millisecondi, la composizione dell'Ensemble e il comportamento del post-processing. Questo capitolo cambia prospettiva. L'obiettivo non è riepilogare i risultati, ma interpretarli: capire perché la distillazione funziona e il training combinato no, rispondere alla domanda di ricerca che ha motivato l'intero lavoro, e dichiarare con onestà dove il metodo proposto incontra i propri limiti.

9.1 Interpretazione dei risultati

9.1.1 Perché la distillazione risolve il bias T_Onset

Il punto di partenza, già quantificato nella Sezione 8.2.5, è che il problema dell'etichetta GE non sta in un rumore di annotazione che si potrebbe compensare con più dati, ma è un legame deterministico tra *T_Onset* e *QRS_Offset*, privo di varianza, identico in ogni singolo esempio del dataset. Per un ottimizzatore basato su discesa del gradiente, questo legame rappresenta la soluzione a costo minimo, poiché il modello impara a copiare il bordo del QRS anziché rilevare le caratteristiche morfologiche reali dell'onda T (ad esempio la pendenza di risalita o il segmento isoelettrico che la precede) perché la scorciatoia funziona su ogni esempio di training senza eccezioni, e nessun segnale nella loss lo penalizza per averla seguita.

Quando si introduce un teacher addestrato esclusivamente su LUDB, le pseudo-etichette che questo genera rompono quel legame in modo mirato. Lo Student non può più risolvere il compito copiando il QRS, perché le nuove etichette collocano il bordo della T in una posizione che varia da battito a battito e che non coincide con la fine del complesso ventricolare. Costretto ad abbandonare la scorciatoia, il modello finisce per apprendere gli stessi tratti su cui si basa un annotatore umano

esperto: la transizione dal segmento isoelettrico alla fase di risalita dell'onda di ripolarizzazione. Non si tratta, quindi, di un generico trasferimento di conoscenza dal dominio manuale a quello clinico: è la sostituzione chirurgica di un segnale di supervisione che conteneva una scorciatoia con uno che non la contiene, l'unico modo, nella configurazione sperimentale testata, per forzare la rete a imparare il confine corretto.

9.1.2 Il ruolo del formato esteso nella trasferibilità

Spiegare perché la distillazione funzioni non basta però a giustificare la scelta del formato; anche la distillazione sul battito mediano (KD-MED) rimuove la scorciatoia, eppure i risultati della Sezione 8.2.6 mostrano che il miglioramento resta confinato al dominio manuale e non si trasferisce ai dati automatici. Il motivo va cercato nella distribuzione di variabilità che il modello incontra durante l'addestramento.

Il battito mediano è un segnale a varianza collassata per costruzione. Ogni esempio di training presenta alla rete la stessa morfologia pulita, priva delle fluttuazioni battito-per-battito che caratterizzano un tracciato reale. Nessun jitter del punto di campionamento, nessuna deriva della linea di base, nessuna differenza tra un battito e il successivo. Il gradiente che ne deriva non porta mai con sé l'informazione "questo confine può spostarsi di qualche campione da un battito all'altro", e la rete non ha alcuna spinta nel costruire un rilevatore di bordo robusto a una variabilità che non ha mai visto. La regola che impara è quindi astratta: sa dove posizionare il bordo della T *in media*, ma non sa adattarsi alle piccole irregolarità di un battito specifico registrato in condizioni reali.

Il tracciato timbro, al contrario, mostra alla rete decine di battiti reali per ciascun esempio di training, con jitter naturale, micro-variazioni morfologiche e deriva della baseline. A questo si aggiunge un effetto meno ovvio ma altrettanto importante: il meccanismo *self-template* LOO usato per generare le etichette sul tracciato timbro (Sezione 5.3) introduce a sua volta una variabilità nella posizione dei bordi annotati, perché la correlazione tra il battito mediano e ciascun battito reale non è mai esattamente uguale. Il rilevatore di bordo che ne risulta è stato costretto, già durante il training, a generalizzare su una distribuzione di variabilità molto simile a quella che incontrerà in produzione su tracciati mai visti, ed è questa la ragione per cui il formato esteso trasferisce la correzione anche al dominio nativo, mentre il battito mediano no.

Non si tratta, dunque, di una differenza nella quantità di segnale a disposizione del modello, ma nel tipo di distribuzione su cui il modello viene ottimizzato: stretta e artificialmente stabile nel primo caso, larga e rappresentativa delle condizioni reali nel secondo.

9.1.3 Perché il training combinato non basta

Alla luce di quanto discusso nei paragrafi precedenti, viene naturale chiedersi perché la soluzione più semplice, unire i due domini in un unico addestramento come fatto nella Fase COMB, non funzioni, nonostante sulla carta il modello abbia accesso sia al segnale fisiologico di LUDB sia al grande volume di dati clinici GE.

Il problema nasce a livello di ottimizzazione. In un training combinato, ogni minibatch mescola esempi con due convenzioni di etichettatura diverse per la stessa onda T: quella fisiologica di LUDB, dove il bordo della T è clinicamente indipendente dal QRS, e quella convenzionale di GE, dove i due punti coincidono per definizione. Poiché la loss è additiva sul batch, il gradiente totale rispetto ai pesi del canale T risulta dalla somma di due componenti che tirano verso bordi fisicamente diversi per pattern morfologici anche molto simili tra loro. Il punto di convergenza non è una media sensata tra le due posizioni, ma un compromesso instabile che, una volta decodificato con argmax in fase di inferenza, non corrisponde al bordo corretto per nessuno dei due domini.

I numeri della Tabella 8.2 illustrano bene la conseguenza pratica: COMB ottiene il valore più alto in assoluto su LUDB (F1 Macro 98.61 %), un risultato che inizialmente sembrava promettente ma che, verificato sui domini clinici, si è rivelato un artefatto di composizione del training set: LUDB è inclusa direttamente nell'addestramento, e il piccolo dataset manuale finisce per pesare in modo sproporzionato nell'ottimizzazione. Su `PRIV_medio`, lo stesso modello crolla sull'onda P al 65.3 % e sulla T al 79.7 %, valori inferiori a quelli della distillazione a due stadi.

È importante distinguere questo fallimento da un semplice problema di generalizzazione fuori dominio. Qui i due domini *sono* nel training set, ma portano segnali di supervisione incompatibili per lo stesso parametro. La distillazione evita questo conflitto alla radice: ogni esempio di training dello Student porta un solo segnale di supervisione, quello del teacher, e il segnale originale dell'etichetta GE non entra mai in competizione diretta con quello manuale durante la stessa fase di ottimizzazione.

9.2 Risposta alla domanda di ricerca

9.2.1 Sostituzione parziale, condizionata alla distillazione

La domanda che ha motivato l'intera tesi, se i dati clinici GE possano sostituire i dataset di ricerca annotati manualmente per la segmentazione dell'ECG, non ammette una risposta binaria, e sarebbe fuorviante tentarne una. Ciò che i risultati dei Capitoli 8 e 9 mostrano, nella configurazione sperimentale testata, è

uno spostamento nel ruolo del dataset manuale: da input diretto e ricorrente dell'addestramento a strumento di calibrazione utilizzato una sola volta.

In termini pratici, il significato di questo spostamento è che il costo di annotazione manuale diventa un costo fisso sostenuto una volta, annotare con cura qualche centinaio di tracciati come quelli di LUDB, e non un costo variabile che si ripete ogni volta che il sistema deve applicare a un nuovo lotto di pazienti, un nuovo ospedale o anni di archivio storico. Il costo marginale di ciascuna applicazione successiva tende a zero, perché il teacher già addestrato genera le pseudo-etichette senza intervento umano.

9.2.2 Scomposizione per onda

La sostituibilità non è omogenea tra le tre onde, e la differenza non dipende dalla quantità di dati disponibili ma dalla natura dell'errore presente nell'etichetta automatica.

Per l'onda P e il complesso QRS, l'errore delle etichette GE si comporta come rumore statistico: variabile, non sistematico, e compensabile per numerosità. I dati GE sono pienamente sufficienti a produrre modelli competitivi, come confermano i valori di F1 QRS mai inferiori al 96.89 % su nessuna delle nove classi patologiche in nessun dominio (Sezione 8.2.3). Non servono dati annotati manualmente per queste due onde.

Per l'onda T, la situazione è qualitativamente diversa. L'errore non è rumore: è un bias strutturale, una scorciatoia deterministica che nessuna quantità aggiuntiva di dati con la stessa convenzione può correggere. Anzi, più esempi con lo stesso bias sistematico rafforzano la scorciatoia anziché diluirla. È un principio che si applica in modo generale: qualunque convenzione industriale che introduca un legame deterministico artificiale tra due punti clinicamente indipendenti, non solo quella di GE su T_Onset, sarebbe soggetta allo stesso limite, indipendentemente dal volume del dataset.

Occorre però aggiungere una precisazione che i numeri aggregati non rendono immediatamente visibile: la correzione prodotta dalla distillazione è forte quanto il teacher che la genera. LUDB copre circa 200 pazienti, e la distillazione trasferisce solo la conoscenza che il teacher stesso ha imparato a generalizzare su quel campione. Se una morfologia rara o estrema fosse sottorappresentata in LUDB, ad esempio un particolare pattern di ripolarizzazione precoce osservato solo in una piccola frazione della popolazione, il teacher non saprebbe generare pseudo-etichette affidabili, e lo Student erediterebbe quel limite senza che nulla, nella pipeline, lo segnali esplicitamente.

9.2.3 Implicazioni per l'applicabilità clinica su larga scala

Lo spostamento del costo di annotazione a monte rende il metodo scalabile in modo asimmetrico: costoso una volta, quasi gratuito per ogni applicazione successiva. Questo lo rende particolarmente adatto ai contesti con grandi archivi storici di ECG annotati automaticamente ma mai rivisti manualmente, lo scenario tipico di un reparto di cardiologia con anni di tracciati GE accumulati.

Due cautele sono però necessarie: il fatto che la distillazione risolva il bias $T_Onset = QRS_Offset$ della convenzione GE MAC2000 non implica che risolva automaticamente bias analoghi di altri produttori, poiché se quei sistemi introducono scorciatoie diverse, ad esempio su intervalli o bordi diversi dallo ST, sarebbe necessario prima verificare che il bias esista, poi ripetere la stessa logica di correzione con un teacher appropriato, senza assumere che il modello attuale la applichi per trasferimento implicito.

Inoltre, l'intero percorso descritto in questa tesi è una validazione algoritmica, condotta confrontando le predizioni del modello con un *ground truth* manuale (LUDB) o con le etichette di produzione (PRIV). Una validazione clinica propriamente detta, in cui un cardiologo esperto valuta la segmentazione prodotta dal sistema su un campione di tracciati reali giudicando la plausibilità dei bordi in rapporto al quadro clinico del singolo paziente, non è stata condotta e resta un passo necessario prima di qualunque utilizzo diagnostico diretto.

9.3 Limiti dello studio

9.3.1 Limiti dell'algoritmo di rilevamento picchi

L'algoritmo di rilevamento dei picchi puntuali descritto nella Sezione 7.2 dipende da soglie e finestre temporali euristiche, ad esempio la finestra di ricerca del punto J rispetto all'offset del QRS, o la soglia di ampiezza minima per distinguere un'onda P bifida dal rumore di fondo. Queste soglie sono state calibrate sui pattern osservati nelle nove macro-classi disponibili e non hanno una giustificazione fisiologica universale. Un tracciato con una morfologia ai margini di quanto osservato durante la calibrazione (frequenze cardiache estreme, artefatti da movimento intensi, derivazioni con rapporto segnale-rumore particolarmente basso) può produrre rilevamenti dei picchi puntuali meno affidabili anche quando la segmentazione P/QRS/T sottostante è corretta.

9.3.2 Assenza di ground truth ST su dati GE

Nessun file GE di produzione possiede un'annotazione manuale della vera separazione ST-T: la convenzione industriale la esclude per definizione. La validazione

del canale T dell'Ensemble Fisiologico sui dati di produzione si basa quindi sul confronto con le pseudo-etichette generate dal teacher, una misura di coerenza interna al sistema, non una verifica clinica indipendente. L'unico riscontro contro un vero *ground truth* manuale resta LUDB, dove però l'ensemble non è valutabile *end-to-end* per il bias di dominio discusso nella Sezione 9.1.3. È un limite circolare, intrinseco alla natura stessa del problema affrontato: la separazione ST che si vuole validare non esiste nelle etichette del dominio su cui il sistema è progettato per lavorare.

9.4 Sviluppi futuri

9.4.1 Creazione di un ground truth clinico-manuale per lo ST

L'ostacolo metodologico più stringente affrontato in questo lavoro è stata l'impossibilità di validare numericamente in modo *end-to-end* l'Ensemble Fisiologico sui tracciati ospedalieri. Come discusso nella Sezione 9.3.2, l'assenza di un reale *ground truth* per la separazione ST nel dominio GE impedisce una verifica quantitativa diretta sul rumore reale. Uno sviluppo futuro cruciale consisterebbe nel sottoporre un sottoinsieme limitato e rappresentativo di tracciati industriali (ad esempio i tracciati del test set `PRIV_medio`) a una revisione manuale mirata da parte di un pool di cardiologi esperti. Questo permetterebbe di creare un test set *gold standard* ibrido: tracciati reali di produzione (con tutti gli artefatti clinici del caso) accoppiati a etichette manuali fisiologicamente corrette per il tratto ST, rendendo finalmente possibile quantificare in modo diretto e inequivocabile il successo della Knowledge Distillation.

9.4.2 Validazione su coorte multi-paziente più ampia

Per poter estendere il lavoro proposto, sarebbe utile ampliare la validazione a coorti multi-centro, provenienti da ospedali e popolazioni diverse, per verificare se il bias $T_Onset = QRS_Offset$ e la sua correzione tramite distillazione si comportano allo stesso modo su popolazioni con prevalenze patologiche diverse da quelle osservate nel dataset PRIV.

9.4.3 Estensione della distillazione ad altre convenzioni industriali

Il meccanismo descritto nella Sezione 9.1.1 non è specifico all'onda T né al costruttore GE. Qualunque convenzione industriale che introduca un legame determini-

stico artificiale tra due punti clinicamente indipendenti è, in linea di principio, correggibile con lo stesso approccio. Andrebbe verificato se altri produttori di elettrocardiografi introducono bias analoghi su intervalli diversi dallo ST, e se un singolo teacher addestrato su LUDB può fungere da riferimento fisiologico condiviso per più convenzioni industriali, invece di richiedere una distillazione dedicata per ciascuna.

9.4.4 Possibili raffinamenti dell'ensemble

L'Ensemble attuale fonde le onde con una regola fissa e una priorità statica tra i canali (Sezione 8.4.4). Un possibile sviluppo è sostituire questa regola con una fusione pesata dinamicamente sulla base della confidenza di ciascun modello per il singolo campione (ad esempio sfruttando la distribuzione del softmax o tecniche di stima della varianza predittiva come il Monte Carlo Dropout) così da ridurre automaticamente il peso di un canale nei casi in cui il modello è strutturalmente meno affidabile, invece di trattare ogni campione con lo stesso schema fisso indipendentemente dalla difficoltà del caso.

Capitolo 10

Conclusioni

10.1 Risposta alla domanda di ricerca e Risultati Raggiunti

La domanda da cui è partito questo lavoro chiedeva se ampi archivi di dati clinici reali, etichettati automaticamente dai software commerciali (come GE MAC2000), potessero sostituire i dataset di ricerca annotati manualmente (come LUDB) per l'addestramento di modelli di segmentazione ECG profonda. La risposta emersa è articolata, ma porta a un importante risultato applicativo.

Per l'onda P e il complesso QRS, la sostituzione è non solo possibile, ma altamente efficace. Il rumore presente nelle etichette automatiche non è sistematico e viene facilmente compensato dalla rete neurale grazie all'elevata numerosità dei dati clinici. La rete addestrata esclusivamente su dati GE (Fase B-TMB) raggiunge prestazioni di eccellenza, con un F1 score sul QRS costantemente superiore al 98.6 % e prestazioni sull'onda P (F1 96.8 %) che superano nettamente i modelli addestrati su soli dati manuali.

Per la ripolarizzazione ventricolare (onda T e tratto ST), la sostituzione diretta fallisce a causa di un grave bias geometrico introdotto dal costruttore: la coincidenza forzata $T_Onset = QRS_Offset$, che ingloba sistematicamente l'intero tratto ST all'interno dell'onda T (errore mediano quantificato in -94 ms). Tuttavia, questo ostacolo è stato superato con totale successo. Attraverso il paradigma della *Knowledge Distillation* applicato su tracciati estesi da 10 secondi (Fase KD-TMB), la rete neurale è riuscita a spezzare il vincolo deterministico industriale, imparando a ignorare le etichette errate di training. Il risultato è un modello finale che ha ripristinato l'esatta separazione fisiologica ST-T, raggiungendo un F1 score sull'onda T del 99.6 % sul *ground truth* manuale, annullando di fatto il bias di 94 ms e rendendo le predizioni automatiche indistinguibili dal giudizio di un cardiologo esperto.

10.2 Sintesi dei contributi originali

Il lavoro presentato in questa tesi ha permesso di raggiungere i seguenti traguardi:

1. Quantificazione del Bias Industriale: È stata misurata e dimostrata la natura degenerata della convenzione GE MAC2000 sulla ripolarizzazione (Capitolo 5, Sezione 5.2.3).

2. Correzione tramite Knowledge Distillation: Progettazione e validazione di una pipeline a due stadi che corregge il bias clinico senza richiedere la costosa e impraticabile revisione manuale di migliaia di tracciati ospedalieri (Capitoli 6 e 9).

3. Il ruolo cruciale del contesto temporale: Dimostrazione inedita che la distillazione risulta pienamente efficace per la correzione morfologica solo se applicata a segmenti ECG estesi (tracciato da 10 secondi), fornendo alla rete la variabilità inter-battito necessaria per il trasferimento di dominio (Sezione 9.1.2).

4. L'Ensemble Fisiologico: Sviluppo di un'architettura dedicata a due reti in grado di estrarre un'onda T anatomicamente corretta preservando la stabilità clinica sulle onde P e QRS (Capitolo 8).

5. Post-Processing e Ingegnerizzazione: Implementazione di un algoritmo a regole fisiologiche per ripulire le maschere spurie e sviluppo di un'applicazione web interattiva (Streamlit) capace di eseguire l'inferenza *end-to-end* su file XML di produzione, offrendo una *second opinion* visiva immediata e precisa.

10.3 Considerazioni finali

Questa tesi è partita dall'identificazione di un limite pratico grave nella strumentazione cardiologica odierna: l'impossibilità di usare software industriali legacy per un'indagine automatizzata e fine del tratto ST. L'approccio sviluppato non si è limitato a studiare il problema da un punto di vista teorico, ma ha fornito una soluzione ingegneristica completa, robusta e immediatamente applicabile al flusso di lavoro ospedaliero.

L'impiego sinergico di dataset eterogenei ha dimostrato come l'Intelligenza Artificiale, se guidata da una profonda conoscenza del dominio fisiologico, possa spingersi ben oltre i limiti intrinseci dei dati su cui viene addestrata. Il sistema duale realizzato, composto da una Rete di Controllo aderente allo standard industriale e da un Ensemble Fisiologico per la vera correzione anatomica, non solo

annulla il bias sistematico di partenza, ma raggiunge metriche di accuratezza elevatissime sull'onda T (F1 99.6 %), paragonabili all'accordo inter-osservatore dei medici specialisti. Questo lavoro pone così fondamenta solide per lo sviluppo di una diagnostica algoritmica di nuova generazione per le patologie cardiache.

Ringraziamenti

Bibliografia

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[2] Maarten Blondeel, Tomas Robyns, Rik Willems, and Bert Vandenberk. Ventricular depolarization abnormalities and their role in cardiac risk stratification: A narrative review. *Reviews in Cardiovascular Medicine*, 26(1):25921, 2025.

[3] Radomir Chabiniok, Vicky Y. Wang, Myrianthi Hadjicharalambous, Liya Asner, Jack Lee, Maxime Sermesant, Ellen Kuhl, Alistair A. Young, Philippe Moireau, Martyn P. Nash, Dominique Chapelle, and David A. Nordsletten. Multiphysics and multiscale modelling, data-model fusion and integration of organ physiology in the clinic: ventricular cardiac mechanics. *Interface Focus*, 6:20150083, 2016.

[4] Ekansh Chauhan, Swathi Guptha, Likith Reddy, and Bapi Raju. Lrh-net: A multi-level knowledge distillation approach for low-resource heart network. In *MICCAI Workshop on Distributed, Collaborative, and Federated Learning*, pages 178–188. Springer, 2022.

[5] Zhenqin Chen, Mengying Wang, Meiyu Zhang, Wei Huang, Hanjie Gu, and Jinshan Xu. Post-processing refined ecg delineation based on 1d-unet. *Biomedical Signal Processing and Control*, 79:104106, 2023.

[6] Amsalu Tomas Chuma, Chunzhuo Wang, Bart Vanrumste, Ahmed Saeed Youssef, Melkamu Hunegnaw Asmare, Desalew Mekonnen Kassie, and Jens-Uwe Voigt. Performance assessment of ecg delineators on single-lead wearable ambulatory data. *medRxiv*, 2026. Preprint posted March 26, 2026.

[7] Chantal E. Conrath and Tobias Opthof. Ventricular repolarization: An overview of (patho)physiology, sympathetic effects and genetic aspects. *Progress in Biophysics and Molecular Biology*, 92:269–307, 2006.

[8] Erwin Christian de Bliek. St elevation: Differential diagnosis and caveats. a comprehensive review to help distinguish st elevation myocardial infarction from nonischemic etiologies of st elevation. *Turkish Journal of Emergency Medicine*, 18:1–10, 2018.

[9] ECG Waves. The t wave: Physiology, variants and ecg features. Consultato il: 26-07-2026.

[10] EKG Echo. Overview of the ecg waves, deflections, intervals and durations. Consultato il: 26-07-2026.

[11] GE Healthcare. *GE Cardiology Open XML Reference Manual*, 2012.

[12] GE Healthcare. *Marquette 12SL ECG Analysis Program - Physician's Guide*, 2015.

[13] GE Healthcare. *MAC 2000 Resting ECG Analysis System Operator's Manual*. General Electric Company, 2020.

[14] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[15] Lorenzo Guidi, Elisabetta Rosellini, Gaia Riccio, and Maria Grazia Cascone. Advancements and challenges in tissue-engineered heart valves: Integrating biomechanics, biomaterials, and biomimetic design for functional maturity. *Biomimetics*, 11(3):185, March 2026.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[18] Alena I Kalyakulina, Igor I Yusipov, Viktor A Moskalenko, Alexander V Nikolskiy, Konstantin A Kozlov, Konstantin A Kosonogov, Nikolai Yu Zolotykh, and Mikhail V Ivanchenko. LUDB: a new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, 8:188718–188731, 2020.

[19] Fred M. Kusumoto, Mark H. Schoenfeld, Coletta Barrett, James R. Edgerton, Kenneth A. Ellenbogen, Michael R. Gold, Nora F. Goldschlager, Robert M. Hamilton, José A. Joglar, Robert J. Kim, Richard Lee, Joseph E. Marine, Christopher J. McLeod, Keith R. Oken, Kristen K. Patton, Cara N. Pellegrini, Kimberly A. Selzman, Annemarie Thompson, and Paul D. Varosy. 2018 ACC/AHA/HRS guideline on the evaluation and management of patients with bradycardia and cardiac conduction delay: A report of the american college of cardiology/american heart association task force on clinical practice guidelines and the heart rhythm society. *Circulation*, 140(8):e382–e482, 2019.

[20] Ruri Lee, Byungmun Kang, DongHyeon Kim, and DaeEun Kim. Correlation based dynamic time warping for ecg waveform. *Applied Sciences*, 16(5):2369, 2026.

[21] Life in the Fast Lane (LITFL). P wave ecg library. Consultato il: 26-07-2026.

[22] Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna. A wavelet-based ECG delineator: Evaluation on standard databases. *IEEE Transactions on Biomedical Engineering*, 51(4):570–581, 2004.

[23] Antonio Michelucci, Giuseppe Bagliani, Andrea Colella, Paolo Pieragnoli, Maria Cristina Porciani, Gianfranco Gensini, and Luigi Padeletti. P wave assessment: State of the art update. *Cardiac Electrophysiology Review*, 6:215–220, 2002.

[24] Viktor Moskalenko, Nikolai Zolotykh, and Grigory Osipov. Deep learning for ECG segmentation. In *Advances in Neural Computation, Machine Learning, and Cognitive Research III*, pages 246–254. Springer, 2020.

[25] Kamran Namjouyan, Ervin Sejdić, Mark S. Link, Antonio Pelliccia, Benjamin Glicksberg, Natalia Trayanova, and Chayakrit Krittanawong. Precision medicine for electrocardiogram interpretation: Clinical relevance, challenges, and advances. *Reviews in Cardiovascular Medicine*, 26(12):47007, 2025.

[26] Santosh K. Padala, José-Angel Cabrera, and Kenneth A. Ellenbogen. Anatomy of the cardiac conduction system. *Pacing and Clinical Electrophysiology*, 2020. Accepted Article.

[27] David S. Park and Glenn I. Fishman. The cardiac conduction system: development, function and therapeutic targets. *Nature Reviews Cardiology*, 23:303–323, May 2026.

[28] JaeHo Park, TaeJun Park, Gyurin Kim, JiHyun Lee, Jeong Min Son, Joonmyoung Kwon, and Yong-Yeon Jo. Benchmarking ECG delineation using deep neural network-based semantic segmentation models. In *Proceedings of Machine Learning Research*, volume 287, pages 1–25. Conference on Health, Inference, and Learning (CHIL), 2025.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[30] Teodoru R. Popa and Andrei C. Mocanu. Medical data storage, visualization and interpretation: A case study using a proprietary ecg xml format. *Annals of the University of Craiova, Series: Automation, Computers, Electronics and Mechatronics*, 8:44–49, 2011.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[32] Yasar Sattar and Lovely Chhabra. Electrocardiogram. StatPearls [Internet], June 2023.

[33] J. Schläpfer and H. J. Wellens. Computer-interpreted electrocardiograms: Benefits and limitations. *Journal of the American College of Cardiology*, 70(9):1183–1192, 2017.

[34] Iana Sereda, Sergey Alekseev, Aleksandra Koneva, Roman Kataev, and Grigory Osipov. ECG segmentation by neural networks: Errors and correction. *arXiv preprint arXiv:1812.10386*, 2018.

[35] Akshay Kumar Sharma and Kyung Ki Kim. Real-time ecg signal acquisition and processing using labview. *Journal of Sensor Science and Technology*, 29(3):162–171, 2020.

[36] Guilherme A. L. Silva, Pedro H. L. Silva, Gladston J. P. Moreira, Vander L. S. Freitas, Jadson C. Gertrudes, and Eduardo J. S. Luz. A systematic review of ecg arrhythmia classification: Adherence to standards, fair evaluation, and embedded feasibility. *arXiv preprint arXiv:2503.07276*, 2025.

[37] H. Smulyan. The computerized ecg: Friend and foe. *The American Journal of Medicine*, 132(2):153–160, 2019.

[38] Teaching Medicine. Tutorial: Waveforms - pr interval. Consultato il: 26-07-2026.

[39] Jorge D. Zhingre Sanchez, Michael G. Bateman, Jason L. Quill, Alexander J. Hill, and Paul A. Iaizzo. *The Anatomy and Function of the Atrioventricular Valves*, pages 3–26. Springer International Publishing, Cham, 2023.

[40] Yiqiu Zhou. Excitation contraction coupling in hypertrophy and failing heart cells. In *Proc. E3S Web of Conferences (ICEPE 2021)*, volume 271, page 03008, 2021.