Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Biomedica – Strumentazione Biomedica
a.a. 2024/2025
Sessione di Laurea Marzo 2025

Tecniche di Explainable Artificial Intelligence e Uncertainty Quantification nella Classificazione EEG delle Fasi del Sonno

Relatori:
Massimo Salvi
Silvia Seoni

Candidata:
Vittoria Oberto
302583

ABSTRACT

La classificazione delle fasi del sonno gioca un ruolo fondamentale in ambito clinico, ma l'approccio manuale tradizionale è costoso in termini di tempo ed esposto a errori causati dalla soggettività del processo. Per superare questi limiti, negli ultimi anni sono stati sviluppati metodi automatici basati su tecniche di deep learning. Tuttavia, nonostante le elevate performance, questi modelli presentano difficoltà nell'applicazione in ambito clinico per la loro natura di "scatola nera", che limita la fiducia nelle predizioni a causa della mancanza di una spiegazione del processo decisionale.

In questo lavoro, si propone un approccio innovativo che combina Uncertainty Quantification (UQ) ed Explainable Artificial Intelligence (XAI) per migliorare l'affidabilità e l'interpretabilità delle predizioni di una rete CNN+BiLSTM allenata per la classificazione automatica delle fasi del sonno.

Per affrontare le sfide legate alla classificazione delle fasi del sonno e all'impiego del deep learning in ambito clinico pratico, viene applicato il Monte Carlo Dropout (MCD) al fine di stimare l'incertezza associata alle predizioni e definita una soglia di incertezza che consente di identificare e rimuovere i campioni meno affidabili, migliorando così significativamente le metriche di performance, soprattutto della fase N1, la cui classificazione è sempre stata la più critica. Parallelamente, vengono utilizzati metodi XAI (GradCAM, GradCAM++ e ScoreCAM) per generare heatmap per la localizzazione delle caratteristiche salienti per il processo decisionale. Si è adottata una nuova versione dell'indice CO (Confidence Optimization) Score, che ha permesso di valutare la capacità discriminativa dei diversi approcci XAI.

I risultati dimostrano che il metodo ScoreCAM, in combinazione con il CO Score, è il più efficace nel distinguere tra campioni correttamente classificati e misclassificati, contribuendo a migliorare ulteriormente le performance complessive del modello. Con un'accuratezza che supera lo stato dell'arte, l'approccio proposto presenta una soluzione interpretabile e affidabile per la classificazione automatica delle fasi del sonno, aprendo nuove prospettive per l'applicazione clinica dei modelli di deep learning.

ABSTRACT

Sleep stage classification plays a crucial role in clinical settings, but the traditional manual approach is time-consuming and prone to errors due to the subjectivity of the process. To overcome these limitations, automated methods based on deep learning techniques have been developed in recent years. However, despite their high performance, these models face challenges in clinical applications due to their "black-box" nature, which limits trust in their predictions due to the lack of explainability in the decision-making process.

In this work, we propose an innovative approach that combines Uncertainty Quantification (UQ) and Explainable Artificial Intelligence (XAI) to enhance the reliability and interpretability of predictions made by a CNN+BiLSTM network trained for automatic sleep stage classification.

To address the challenges associated with sleep stage classification and the practical deployment of deep learning in clinical applications, Monte Carlo Dropout (MCD) is applied to estimate the uncertainty associated with predictions. An uncertainty threshold is then defined to identify and remove less reliable samples, significantly improving performance metrics, particularly for N1 stage classification, which has historically been the most challenging. At the same time, XAI methods (GradCAM, GradCAM++, and ScoreCAM) are employed to generate heatmaps that highlight the salient features involved in the decision-making process. A novel version of the Confidence Optimization (CO) Score is introduced to evaluate the discriminative ability of different XAI approaches.

The results demonstrate that ScoreCAM, in combination with the CO Score, is the most effective in distinguishing between correctly classified and misclassified samples, further enhancing the model's overall performance. With accuracy surpassing the state of the art, the proposed approach provides an interpretable and reliable solution for automatic sleep stage classification, paving the way for the clinical application of deep learning models.

Indice

Elenco delle figure

Elenco delle tabelle

Acronimi

XAI: Explainable Artificial Intelligence

UQ: Uncertainty Quantification

EEG: ElettroEncefaloGramma

PSG: PoliSonnoGrafia

EOG: ElettroOculoGrafia

EMG: ElettroMioGrafia

ECG: ElettroCardioGrafia

AASM: American Academy of Sleep Medicine

EDF: European Data Format

SC: Sleep Cassette

ST: Sleep Telemetry

CNN: Convolutional Neural Network

RNN: Recurrent Neural Network

BiLSTM: Bidirectional Long Short Term Memory

ReLU: Rectified Linear Unit

GPU: Graphics Processing Unit

MCD: Monte Carlo Dropout

BL: BaseLine

TTA: Test Time Augmentation

MC: MisClassificati

CC: Corretti Classificati

CAM: Class Activation Map

GradCAM: Gradient-Weighted Class Activation Mapping

AX: Aumentative Explainability

CO: Confidance Optimization

SUE: Spatial Uncertainty Estimator

AI: Artificial Intelligence

ML: Machine Learning

DL: Deep Learning

GDPR: Regolamento Generale sulla Protezione dei Dati

NCP: Performance NeuroCognitiva

W: Wake

N1, N2, N3: Non-Rapid Eye Movement Phases

REM: Rapid Eye Movement

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

ROC: Receiver Operating Characteristic

PCC: Percentage of Correct Classification

1. Introduzione

1.1 Contesto e Importanza del Lavoro

L'Intelligenza Artificiale (IA) ha rivoluzionato diversi settori, ma soprattutto ha il potenziale per rinnovare il modo in cui viene erogata l'assistenza sanitaria, grazie alla sua capacità di analizzare grandi quantità di dati, raggiungere ottime prestazioni, supportare la diagnosi e il trattamento di patologie e ridurre i costi aumentando l'accesso alle cure. Tuttavia, l'applicazione pratica di modelli di deep learning in ambito clinico incontra una sfida importante: la mancanza di interpretabilità. Molti di questi modelli, pur essendo altamente performanti, operano come "black box", rendendo difficile per i medici comprendere il processo decisionale dietro una previsione, limitando la fiducia degli utenti e ostacolando l'adozione diffusa del deep learning nel settore sanitario.

L'Explainable AI (XAI) nasce per affrontare questa problematica, agendo nell'ottica di fornire strumenti per rendere comprensibili le decisioni dei modelli AI a utenti esperti e non in ambito tecnico. La spiegabilità dei modelli di apprendimento automatico (ML) è essenziale per instaurare fiducia negli utenti e rappresenta un requisito chiave per un'implementazione sicura, equa e di successo nelle applicazioni reali.

La necessità di spiegabilità in ambito medico non è solo un'esigenza pratica ma anche etica e normativa: la trasparenza dei modelli è essenziale per garantire sicurezza, equità e conformità a regolamenti. Con l'adozione del Regolamento Generale sulla Protezione dei Dati (GDPR) da parte dell'Unione Europea nel maggio 2018, ogni cittadino ha acquisito il "diritto alla spiegazione" di una decisione algoritmica che lo riguarda [1]. La spiegabilità non è solo un diritto legale, ma anche una responsabilità con ampie implicazioni sociali. Il GDPR

stabilisce, infatti, che gli individui “hanno il diritto di non essere soggetti a una decisione basata esclusivamente su elaborazione automatizzata”.

L'interpretabilità offre molteplici benefici:

- Costruzione della fiducia nei sistemi AI da parte di medici e pazienti.
- Miglioramento delle performance dei modelli.
- Estrazione di conoscenza dai dati.
- Responsabilità etica e legale, riducendo il rischio di decisioni errate prive di giustificazione.

Esiste un compromesso tra spiegabilità e prestazioni: modelli più semplici sono più interpretabili, ma risultano meno performanti rispetto a modelli più complessi, come le reti neurali profonde. Questo equilibrio tra accuratezza e interpretabilità deve essere gestito per garantire un'adozione efficace dell'IA in ambito medico.

Un altro aspetto fondamentale per migliorare l'affidabilità dei modelli AI è la quantificazione dell'incertezza (Uncertainty Quantification, UQ). Fornire stime di incertezza nei sistemi di intelligenza artificiale è essenziale per garantire un processo decisionale sicuro in domini ad alto rischio caratterizzati da diverse fonti di dati. L'incertezza in un modello AI non equivale all'errore predittivo: un modello può fornire una previsione errata con alta certezza o una previsione corretta con bassa affidabilità e questo può risultare pericoloso in contesti critici come la diagnosi medica.

L'integrazione di XAI e UQ è fondamentale per aumentare la fiducia nei modelli AI di apprendimento automatico e profondo. Sebbene le tecniche XAI permettano di lavorare sull’interpretabilità delle decisioni del modello, esse non garantiscono l'affidabilità delle predizioni. La quantificazione dell'incertezza, invece, aiuta a identificare situazioni in cui il modello potrebbe fornire risultati poco affidabili.

Questo approccio combinato di tecniche XAI e UQ offre molteplici potenziali vantaggi: migliora il processo decisionale, fornendo informazioni su come si giunge alle previsioni, aumenta la robustezza del modello, riducendo gli errori in

condizioni di incertezza, e agevola l'interpretabilità, creando fiducia negli utenti finali (in questo caso medici e pazienti).

Un campo in cui l'integrazione di XAI e UQ può fare la differenza è la classificazione automatica delle fasi del sonno. Tradizionalmente, questa classificazione viene effettuata manualmente da esperti analizzando i tracciati EEG acquisiti durante cicli di sonno, un processo lungo e soggettivo. I modelli di deep learning offrono un'alternativa più rapida e precisa, ma la loro mancanza di interpretabilità limita la loro applicazione. La classificazione delle fasi del sonno è fondamentale per la diagnosi di disturbi del sonno e per studi clinici sulla salute dei pazienti. Tuttavia, la fiducia nelle predizioni del modello è essenziale per garantire un uso sicuro ed efficace. Integrare tecniche XAI per spiegare le decisioni del modello e metodi di UQ per valutare l'affidabilità delle predizioni rappresenta un passo fondamentale per favorire l'adozione di sistemi AI nel contesto medico.

L'IA ha il potenziale di rivoluzionare il settore sanitario. Le tecniche di Explainable AI e Uncertainty Quantification offrono soluzioni complementari per affrontare queste sfide, migliorando la trasparenza, l'affidabilità e la fiducia nei modelli AI di apprendimento automatico e profondo.

1.2 Obiettivi

L'obiettivo principale di questa tesi è esplorare il ruolo dell'Explainable Artificial Intelligence (XAI) e della quantificazione dell'incertezza (Uncertainty Quantification, UQ) per migliorare l'affidabilità e l'interpretabilità dei modelli di deep learning applicati alla classificazione delle fasi del sonno. Nonostante i progressi nei modelli basati su reti neurali profonde, il loro utilizzo in ambito clinico rimane limitato dalla scarsa trasparenza delle decisioni e dalla difficoltà nel quantificare l'incertezza delle predizioni. Questo lavoro si propone di sviluppare un approccio che combini tecniche di XAI e UQ per rendere questi modelli più robusti e applicabili nella pratica clinica.

Attraverso l'uso di tecniche avanzate di explainability, come Grad-CAM, Grad-CAM++ e ScoreCAM, e metodi di quantificazione dell'incertezza basati sul Monte Carlo Dropout, uno dei contributi chiave di questo lavoro di tesi è l'adozione di un indicatore (CO Score) in grado di discriminare adeguatamente classificazioni corrette ed errate e che possa facilitare l'integrazione dell'intelligenza artificiale nel contesto medico, nell'ottica di fornire ai medici uno strumento che consenta di interpretare meglio le decisioni del modello e di valutare il grado di affidabilità delle predizioni.

Il CO Score ha permesso di distinguere in modo efficace campioni correttamente classificati e misclassificati, contribuendo a migliorare ulteriormente le performance complessive del modello, con un'accuratezza che supera lo stato dell'arte per la classificazione automatica delle fasi del sonno.

Un ulteriore contributo di questa ricerca è il tentativo di colmare un gap scientifico nel campo dell'XAI e della UQ applicati a segnali fisiologici, un'area ancora poco esplorata rispetto alle applicazioni di queste tecniche su immagini mediche.

Riassumendo, i contributi di questo lavoro di tesi sono:

- Migliorare l'accuratezza della classificazione automatica delle fasi del sonno, con particolare attenzione alla fase N1 (fase la cui classificazione in letteratura risulta più problematica).
- Implementare metodi di quantificazione dell'incertezza (UQ) per migliorare l'affidabilità delle previsioni e la robustezza del modello.
- Applicare tecniche di explainable AI (XAI) per rendere più trasparenti le decisioni del modello e facilitarne l'adozione clinica.
- Applicazione tecniche XAI e UQ a segnali fisiologici, area ancora scarsamente esplorata.

Questa tesi è articolata nei seguenti capitoli:

- Capitolo 2 - Background e Stato dell'arte: Introduzione ai concetti base di DL (Deep Learning), XAI (Explainable Artificial Intelligence) e UQ (Uncertainty Quantification), con una rassegna della letteratura esistente sulle loro applicazioni in ambito medico nella classificazione delle fasi del sonno.
- Capitolo 3 – Materiali e Metodi: Descrizione del dataset, del modello di deep learning utilizzato, delle tecniche di explainability XAI e quantificazione dell'incertezza UQ implementate.
- Capitolo 4 - Risultati: Presentazione e discussione dei risultati ottenuti sull'efficacia delle tecniche di XAI e UQ nel migliorare l'interpretabilità e la robustezza del modello nella classificazione delle fasi del sonno.
- Capitoli 5 e 6 - Conclusioni e sviluppi futuri: Sintesi delle principali evidenze emerse dallo studio e prospettive per future ricerche volte a migliorare l'affidabilità dei modelli di deep learning in ambito clinico.

2. Background e Stato dell’arte

2.1 Classificazione delle Fasi del Sonno

Il sonno è un processo fisiologico essenziale per il benessere umano, perché influenza una vasta gamma di funzioni cognitive, emotive e fisiologiche [2]. La classificazione accurata delle fasi del sonno è fondamentale per la diagnosi e il trattamento di disturbi del sonno come l'insonnia, l'apnea ostruttiva e la narcolessia [3, 4]. Inoltre, consente di analizzare la qualità del riposo ed è importante negli studi riguardanti l'attenzione e patologie neurologiche e cardiovascolari [5].

Un esempio di applicazione importante della classificazione delle fasi del sonno è la valutazione della performance neurocognitiva (NCP). Per una corretta valutazione dell'NCP sono necessarie analisi e interpretazione accurate dei segnali elettroencefalografici (EEG) del sonno umano. La prestazione neurocognitiva (NCP) rappresenta la capacità mentale e cognitiva umana nello svolgere un compito specifico [6]. La valutazione numerica e accurata dell'NCP del soggetto è attualmente un problema aperto in diversi campi, come riabilitazione, neurologia, psicologia e psichiatria. La privazione del sonno può causare importanti rischi cognitivi nell'esecuzione di molte attività comuni come la guida o il controllo di un dispositivo generico; pertanto, il punteggio del sonno è una parte cruciale del processo. Nel ciclo del sonno, la prima fase del sonno non NREM (non-rapid eye movement) o fase N1 è la transizione tra veglia e sonnolenza e diventa rilevante per lo studio dell'NCP [7]. In questo contesto, la variabile denominata periodo di insorgenza del sonno (SOP), cioè il periodo interposto tra la veglia debole e la sonnolenza, diventa molto importante per lo studio dell'NCP. Nella classificazione degli stadi del sonno, la fase 1 del sonno non-REM (N1), considerata la prima fase del ciclo del sonno, rappresenta il centro del SOP. Per

questo motivo, una valutazione accurata delle fasi del sonno, con particolare attenzione alla fase N1, è considerata una parte cruciale del processo.

Il sonno è classificabile in diversi stadi che riflettono variazioni nell'attività cerebrale, nella respirazione, nei movimenti oculari e nel tono muscolare. La classificazione tradizionale delle fasi del sonno si basa sulla polisonnografia (PSG), considerata il gold standard per l'analisi del sonno [8]. La PSG registra simultaneamente diversi segnali fisiologici, tra cui elettroencefalogramma (EEG), elettrooculografia (EOG), elettromiografia (EMG), elettrocardiografia (ECG), ossigenazione del sangue, flusso d'aria e sforzo respiratorio. Questi segnali vengono segmentati in epoche di 30 secondi, analizzati manualmente da esperti per determinare le fasi del sonno e classificati secondo le linee guida dell'American Academy of Sleep Medicine (AASM) [9, 10] o i criteri Rechtschaffen e Kales (R&K) [11].

Secondo l'AASM le fasi del sonno includono:

- Veglia (W): Stato di coscienza prima e dopo il sonno.
- Sonno NREM:
    - N1: Fase di transizione tra veglia e sonno leggero.
    - N2: Sonno leggero stabile.
    - N3: Sonno profondo o Slow Wave Sleep (SWS).
- Sonno REM: Associato a movimenti oculari rapidi e intensa attività cerebrale.

Negli standard R&K, lo stadio N3 è diviso in due stadi indipendenti, N3 e N4.

Tuttavia, questo processo manuale di classificazione delle fasi del sonno da parte di medici esperti è dispendioso in termini di tempo e soggetto a variabilità inter- e intra-osservatore [12, 13]. Per superare questi limiti, la ricerca si è progressivamente orientata verso l’adozione di metodi automatici per la classificazione delle fasi del sonno.

I metodi automatici possono essere distinti in due principali categorie: metodi basati su tecniche di apprendimento tradizionale e metodi automatici che sfruttano l'apprendimento profondo (deep learning). I primi si basano sull'estrazione manuale di caratteristiche significative dai segnali EEG, come entropia, ampiezza ed energia. Questi approcci, sebbene efficaci, dipendono fortemente dalla scelta delle feature e dalla qualità dell'estrazione, richiedendo competenze specifiche per l'ingegnerizzazione delle caratteristiche più informative.

Il deep learning, invece, consente un'elaborazione che evita la necessità di un'estrazione manuale delle feature, grazie alla capacità delle reti neurali di apprendere direttamente dai dati grezzi [14]. Tra questi metodi, i modelli basati su reti neurali ricorrenti (RNN), reti LSTM (Long Short-Term Memory) e architetture convoluzionali (CNN) hanno dimostrato prestazioni promettenti nella classificazione automatica delle fasi del sonno [15].

Lo stato dell'arte offre diversi metodi per la classificazione delle fasi del sonno tramite apprendimento automatico profondo; tra i diversi studi a riguardo, le soluzioni che mostrano i contributi migliori sono:

- Applicazione di reti neurali convoluzionali (CNN) e reti di memoria a lungo e breve termine (LSTM) per l'estrazione di caratteristiche da trasformate tempo-frequenza (spettrogrammi) dei segnali EEG [16]
- Applicazione di reti neurali convoluzionali (CNN) e reti di memoria a lungo e breve termine bidirezionali (BiLSTM) per l'estrazione di caratteristiche da epoche di segnali EEG a canale singolo nel tempo [17]
- Applicazione di meccanismi di attenzione e rete neurale bidirezionale a memoria a lungo e breve termine (AT-BiLSTM) su segnali EEG a canale singolo [18]

- Applicazione a cascata di reti neurali ricorrenti (RNN) basate su blocchi di memoria a lungo e breve termine (LSTM), per la valutazione automatica delle fasi del sonno utilizzando segnali EEG a singolo canale. Una prima rete effettua una classificazione preliminare e una seconda rete si focalizza sulle classi più difficili da distinguere, come N1 e REM [19]

I dataset Sleep-EDF [20] e Sleep-EDFX [21] di PhysioBank rappresentano i principali benchmark per la classificazione del sonno. Questi dataset contengono registrazioni EEG segmentate in epoche di 30 secondi e annotate secondo gli standard R&K e AASM. L'uso di questi dati ha permesso di sviluppare modelli avanzati e valutare l'efficacia di diverse architetture di deep learning.

Nonostante i progressi ottenuti con i metodi basati su deep learning, esistono ancora numerose sfide nella classificazione automatica delle fasi del sonno, in particolare per la fase N1, che è spesso difficile da distinguere dalle altre fasi del sonno per diverse ragioni:

- Somiglianza con altre fasi**:** N1 presenta caratteristiche simili sia alla veglia (W) che alla fase REM, rendendo complessa la sua classificazione.
- Scarsa durata nel ciclo del sonno**:** la fase N1 ha una durata relativamente breve rispetto alle altre fasi, quindi meno rappresentata nei dataset disponibili, e varia considerevolmente tra i soggetti.
- Artefatti nei segnali EEG**:** l'EEG può essere influenzato da rumore e artefatti fisiologici, complicando ulteriormente l'identificazione accurata della fase N1

Inoltre, l'interpretabilità dei modelli di intelligenza artificiale rimane un problema aperto, rendendo difficile la loro adozione clinica su larga scala [22]. Le direzioni di ricerca, compreso questo lavoro di tesi, includono lo sviluppo di modelli per migliorare l'accuratezza della classificazione e l'applicazione di tecniche di explainable AI (XAI), per facilitare la comprensione e l'accettazione clinica di questi modelli e di quantificazione dell’incertezza, per migliorarne la robustezza.

2.2 Tecniche di Uncertainty Quantification (UQ) e Explainable Artificial Intelligence (XAI)

Negli ultimi anni, gli approcci di apprendimento profondo sono emersi come nuovo approccio in grado di superare i limiti della classificazione tradizionale delle fasi del sonno. Tuttavia, questi metodi agiscono come "scatole nere" [23, 24] e una loro l'applicazione responsabile in contesti clinici deve essere necessariamente accompagnata da una migliore robustezza, affidabilità, trasparenza e spiegabilità [25, 26]. A questo proposito, la combinazione di metodi di apprendimento profondo con tecniche di intelligenza artificiale spiegabile (XAI) [27] e di quantificazione dell'incertezza (UQ) [28] hanno recentemente guadagnato sempre più attenzione perché possono offrire modelli ad alte prestazioni e che vanno nella direzione dell'applicabilità clinica.

2.2.1 Explainable Artificial Intelligence (XAI)

L'Explainable AI comprende un insieme di tecniche sviluppate per interpretare il comportamento e migliorarne la trasparenza dei sistemi di intelligenza artificiale. Queste tecniche mirano a fornire informazioni su come i modelli di intelligenza artificiale effettuano previsioni o decisioni, consentendo agli esseri umani di comprendere e fidarsi del ragionamento alla base di tali risultati.

Esistono diverse categorie di tecniche XAI:

- Metodi basati sulla retropropagazione del gradiente, che calcolano la sensibilità della previsione rispetto alle singole parti dell'input (es. Saliency Maps, Grad-CAM, Grad-CAM++).
- Metodi basati sulla perturbazione dell'input, che misurano l'importanza delle feature eliminando o modificando parti dell'input (es. ScoreCAM).

- Metodi basati su *surrogate models*, che approssimano il modello di deep learning con modelli più semplici e interpretabili.

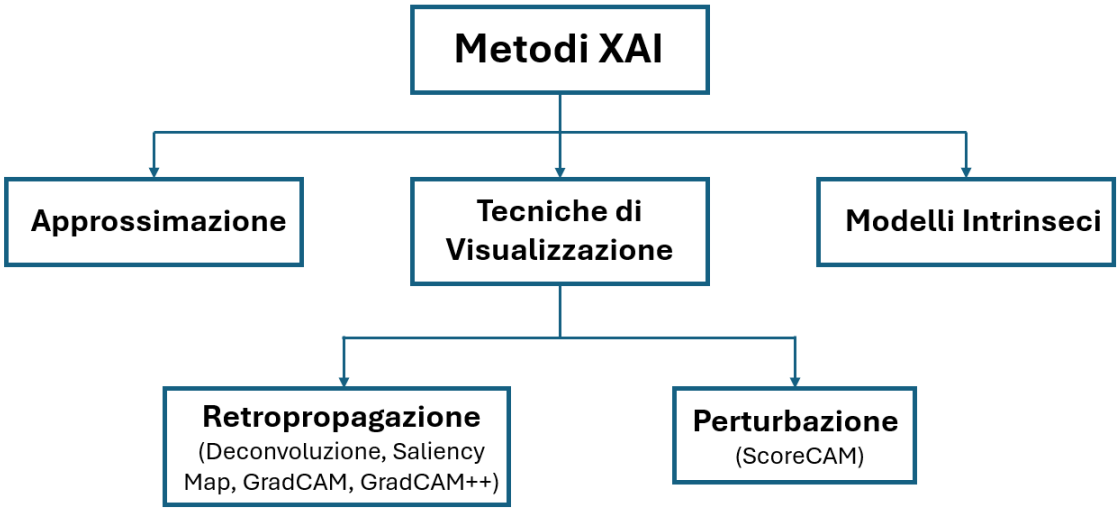

Figura 1: Panoramica metodi XAI attualmente disponibili e più comuni

Tecniche di Visualizzazione

I metodi di visualizzazione sono tra le strategie più utilizzate per interpretare il comportamento di una rete neurale; essi evidenziano tramite mappe a colori (heatmap) le regioni dell'input che contribuiscono maggiormente alla previsione. I metodi di spiegabilità di visualizzazione si dividono in metodi di retropropagazione e metodi di perturbazione. Tra i primi, le mappe di salienza rappresentano uno degli approcci più diffusi e vengono utilizzate principalmente nelle applicazioni di analisi di immagini e segnali fisiologici. Le saliency maps sono mappe che evidenziano le regioni di un'immagine o le parti di un segnale considerate rilevanti dal modello per effettuare la predizione. Tra queste tecniche, oltre alle mappe di salienza, gli approcci più attualmente attenzionati sono i **metodi basati sul gradiente**. I metodi basati sulla perturbazione sono più intensi dal punto di vista computazionale, mentre quelli basati sul gradiente possono essere rumorosi e

portare a mappe di attribuzione che mostrano contributi derivanti da caratteristiche irrilevanti. Meno utilizzate sono le tecniche di deconvoluzione. La deconvoluzione è l'inverso del processo svolto da una rete neurale convoluzionale CNN e tenta di ricreare l'input dall'attivazione dell'output.

Uno dei metodi più utilizzati per la spiegabilità è il **Gradient-weighted Class Activation Mapping** (**Grad-CAM**), che sfrutta i gradienti calcolati rispetto alle feature map dei layer convoluzionali di una rete neurale CNN per evidenziare le aree dell'input che hanno influenzato maggiormente la predizione [29]. Grad-CAM calcola i gradienti dell'output della classe target rispetto alle mappe di attivazione di un determinato livello convoluzionale. Questi gradienti vengono mediati globalmente per ottenere i pesi che indicano l'importanza di ciascuna mappa di attivazione per la classe target. Questo processo genera una heatmap con le dimensioni dello strato convoluzionale, che viene poi normalizzata e ridimensionata per una visualizzazione coerente con l'input. Gli ultimi strati convoluzionali hanno il miglior compromesso tra semantica di alto livello e informazioni spaziali dettagliate; quindi, GradCAM usa le informazioni del gradiente che fluiscono nell'ultimo strato convoluzionale della CNN per assegnare valori di importanza a ciascun neurone per una particolare classe di interesse [30].

Grad-CAM++ è una evoluzione di GradCAM che migliora la localizzazione delle regioni discriminanti utilizzando pesi più raffinati per la stima delle feature map [31]. Questo metodo utilizza le derivate di secondo e terzo ordine dei gradienti per ottenere i pesi.

Un'ulteriore variante è **ScoreCAM**, metodo XAI basato sulla perturbazione delle Class Activation Map (CAM) per generare heatmap più stabili e meno rumorose. ScoreCAM elimina la dipendenza dai gradienti, colmando il divario tra metodi basati sulla perturbazione e metodi basati su CAM, e consiste nel

mascherare/perturbare parti dell'input e osservare la variazione della previsione del modello per la classe di interesse. Per perturbare l'input, le feature map vengono sovracampionate alle dimensioni dell'input e normalizzate. Successivamente, vengono mascherate in base ai punteggi di attivazione e l'input mascherato viene passato attraverso la CNN per calcolare il punteggio di previsione, che viene utilizzato come peso per la feature map. Questo processo viene ripetuto per tutti i filtri presenti nell'ultimo strato convoluzionale e i risultati vengono combinati per ottenere la rappresentazione finale di ScoreCAM. Quindi, il risultato finale è ottenuto da una combinazione lineare di pesi e mappe di attivazione [32].

Una proprietà fondamentale che tutti i metodi di spiegabilità dovrebbero possedere è l'insensibilità alle piccole perturbazioni dell'input, una lieve modifica dell'input, che non altera la decisione della rete, non dovrebbe modificare significativamente le attribuzioni [33]; questo concetto definisce la nozione di **robustezza** del metodo XAI. Allo stesso tempo, una buona spiegazione dovrebbe essere sufficientemente sensibile da riflettere i cambiamenti indotti da un **attacco avversario**: una modifica malevola e impercettibile dell'input che porta la rete a prendere una decisione errata [34]. In un contesto ideale, le mappe di attribuzione dovrebbero rilevare gli attacchi avversari e rimanere invarianti di fronte a piccole perturbazioni innocue dell'input [35]. La robustezza può essere definita come la distanza di un punto di test dal confine decisionale più vicino, e l'aumento di questa distanza porta a un maggiore allineamento tra l'input e la sua mappa di attribuzione.

Applicazioni di XAI di interesse per questo lavoro di tesi:

- **XAI nella Classificazione delle Fasi del Sonno**

L'uso di XAI nella classificazione delle fasi del sonno è ancora limitato, ma alcuni studi hanno iniziato a sfruttare queste tecniche per comprendere meglio il comportamento dei modelli di apprendimento profondo per la classificazione delle fasi del sonno. Lo stato dell'arte offre in questo settore approcci basati sull'applicazione di GradCAM a reti neurali convoluzionali CNN allenate con lo scopo di effettuare classificazione delle fasi del sonno accompagnata dal rilevamento dei pattern di segnale EEG che influenzano maggiormente l'identificazione di ogni fase del sonno [36, 37].

- **XAI per migliorare la fiducia nella Classificazione**

Si può quantificare la qualità dei metodi di Intelligenza Artificiale Spiegabile (XAI) basati sulla stima di heatmap, valutandone l'efficacia nel migliorare la probabilità di predizione delle classi corrette. Per svincolarsi dal concetto di localizzazione come metrica per quantificare le prestazioni di XAI, dal momento che esiste una differenza tra rilevanza computazionale e rilevanza umana (ciò che gli algoritmi considerano saliente potrebbe non essere significativo per un osservatore umano), si introduce il concetto di Spiegazione Aumentativa AX. Questo metodo combina un'immagine con la relativa heatmap stimata con un metodo XAI per ottenere una probabilità più alta di predire la classe corretta. In questo modo le heatmap non si limitano a indicare le aree di attenzione della rete, ma possono anche essere strumenti utili per migliorare la fiducia predittiva. La spiegazione aumentativa (AX) offre un nuovo approccio per valutare e sfruttare i metodi XAI al fine di potenziare la trasparenza e l'affidabilità dei modelli di deep learning [38].

I metodi XAI non riescono a fornire una comprensione completa delle complesse relazioni all'interno dei processi decisionali alla base delle predizioni effettuate da una rete di apprendimento profondo. Affiancare a questi delle tecniche di quantificazione dell'incertezza UQ delle previsioni per stimare l'affidabilità del modello può offrire una migliore comprensione del processo.

2.2.2 Uncertainty Quantification (UQ)

La quantificazione dell'incertezza è un aspetto fondamentale per migliorare l'affidabilità dei modelli di deep learning [39]. Esistono due principali categorie di incertezza nei modelli di intelligenza artificiale:

- **Incertezza epistemica**, dovuta alla limitata conoscenza del modello, riducibile con un miglior addestramento e un dataset più ampio.
- **Incertezza aleatoria**, intrinseca nei dati stessi e non eliminabile, come il rumore nei segnali EEG.

I principali metodi di quantificazione dell'incertezza UQ nell'apprendimento automatico sono [28]:

- **Monte Carlo Dropout (MCD)**: un metodo che abilita il dropout anche in fase di inferenza, generando predizioni multiple per stimare la distribuzione delle probabilità delle classi. L'abilitazione del dropout consiste nell'attivazione e spegnimento randomico dei neuroni di un layer della rete con una certa probabilità durante la fase di training della rete per ridurre l'overfitting. Se il dropout viene abilitato anche in fase di inference, si ottengono distribuzioni di predizioni, anziché una singola previsione deterministica, con la possibilità di estrarre informazioni sull'incertezza delle predizioni. La media delle previsioni da più campioni di dropout porta a una migliore accuratezza della classificazione. È una delle tecniche più utilizzate poiché facile da implementare e molto versatile; può essere implementato nella maggior parte delle reti neurali profonde semplicemente aggiungendo livelli di dropout all'interno dell'architettura.
- **Deep Ensembles**: utilizza un insieme di modelli indipendenti per ottenere una previsione media più robusta e una misura dell'incertezza basata sulla varianza tra le predizioni.

- **Test-Time Data Augmentation (TTA)**: genera versioni modificate dell'input durante l'inferenza (in fase di test) e analizza la variazione nelle predizioni per stimare l'incertezza.

Uno dei principali indicatori di incertezza è l'**entropia normalizzata**, che misura la dispersione della distribuzione delle predizioni. Una volta quantificata l'incertezza, si può usare questa informazione per stabilire delle soglie da applicare con lo scopo di migliorare la qualità delle predizioni, rimuovendo campioni ad elevata incertezza, oppure per ottimizzare la fase di calibrazione del modello.

Applicazioni di UQ di interesse per questo lavoro di tesi:

- **UQ nella Classificazione delle fasi del Sonno**

L'integrazione di tecniche di UQ, come Monte Carlo Dropout, nei modelli di classificazione delle fasi del sonno consente di migliorare la robustezza delle predizioni, identificando epoche ad alta incertezza che possono essere riesaminate da un esperto [40].

Applicazione della combinazione di UQ e XAI

Recentemente, l'uso combinato di XAI e UQ ha portato allo sviluppo di nuovi indicatori come lo **Spatial Uncertainty Estimator (SUE)** per valutare l'affidabilità della previsione delle reti di classificazione. SUE quantifica la sovrapposizione spaziale delle caratteristiche salienti identificate con il metodo XAI Grad-CAM, offrendo un punteggio di confidenza per le previsioni di una rete di apprendimento profondo che integra meccanismi di Convolutional Neural Network (CNN) e Bidirectional Long Short-Term Memory (BiLSTM). SUE distingue accuratamente tra input classificati correttamente e classificati in modo errato, dimostrando il potenziale della combinazione di tecniche XAI e UQ per migliorare l'utilizzo dei metodi di apprendimento profondo [41].

L’adozione di metodi XAI e UQ rappresenta un passo essenziale per colmare il divario tra le reti neurali profonde e l’applicabilità clinica. Le tecniche XAI migliorano l’interpretabilità dei modelli, mentre la quantificazione dell’incertezza fornisce una misura della loro affidabilità; combinando queste due strategie, è possibile sviluppare modelli più trasparenti e robusti.

Collocazione del Lavoro Svolto

Lavori su EEG	**Classificazione fasi del sonno**			
	Anno	**Primo Autore**	**Modello**	**Input**
[16]	2022	Chengfan Li	CNN+LSTM	Spettrogramma
[17]	2022	Hisham ElMoaqet	CNN+BiLSTM	Segnale nel tempo
[18]	2021	Mingyu Fu	AT-BiLSTM	Segnale nel tempo
[19]	2019	Michielli Nicola	Cascata di 2 LSTM+RNN	Segnale nel tempo

Tabella 1: Stato dell'Arte. Lavori su EEG - Classificazione fasi del sonno

Lavori su EEG	**Applicazione Tecniche XAI e UQ**			
	Anno	**Primo Autore**	**Tecnica XAI**	**Tecnica UQ**
[36]	2024	Shivam Sharma	GradCAM	\
[37]	2023	Fernando Vaquerizo-Villar	GradCAM	\
[40]	2021	Luigi Fiorilli	\	Monte Carlo Dropout

Tabella 2: Stato dell'Arte. Lavori su EEG - Applicazione Tecniche XAI e UQ alla Classificazione delle fasi del sonno

La maggior parte degli studi sulla quantificazione dell'incertezza e sull’interpretabilità dei modelli di apprendimento profondo si è concentrata sulle immagini mediche, mentre le applicazioni ai segnali fisiologici rimangono ancora limitate [27, 28].

Questo lavoro di tesi non solo applica le tecniche appena descritte al segnale EEG, settore abbastanza inesplorato, ma propone un approccio nuovo poiché non esistono attualmente studi che combinino XAI e UQ per la classificazione delle fasi del sonno e che siano così performanti.

Il contributo innovativo del metodo proposto è quello di utilizzare misure di incertezza, saliency map e indicatori di efficacia dei metodi analizzati per stabilire quantitativamente quali sono più adatti all'applicazione esplorata e determinare soglie in grado di distinguere predizioni corrette da errate e migliorare significativamente le performance del modello.

Questa tesi si colloca quindi all'avanguardia della ricerca nel settore, cercando di integrare spiegabilità e quantificazione dell'incertezza per migliorare la trasparenza e l'affidabilità dei modelli di deep learning in ambito clinico.

3. Materiali e Metodi

3.1 Dataset

I Dataset Sleep-EDF [20] e Sleep-EDFX [21] sono due set di dati pubblici di PhysioBank. Il set di dati Sleep-EDF contiene registrazioni di segnali polisonnografici PSG di 8 soggetti. Sleep-EDFX è una versione estesa di Sleep-EDF, che ha 197 registrazioni PSG dopo due estensioni nel 2013 e nel 2018. In entrambi i set di dati, i file con nomi che iniziano con SC indicano soggetti sani e i file con nomi che iniziano con ST indicano soggetti con lieve difficoltà ad addormentarsi. Ogni registrazione contiene segnali EEG (Pz-Oz, Fpz-Cz) campionati a 100 Hz e segnali EOG. Gli esperti hanno etichettato ogni epoca di 30 s in sei fasi del sonno in base allo standard R&K [8]: veglia, S1–S4 e REM. Sulla base dello standard AASM [9], S1 e S2 sono stati registrati come N1 e N2, S3 e S4 sono stati uniti in una fase del sonno N3.

Dataset Sleep-EDF

Il Sleep-EDF Database è una raccolta di registrazioni del sonno provenienti da soggetti sani, disponibile su PhysioNet. Le registrazioni sono presentate in formato EDF (European Data Format), uno standard per l'archiviazione di registrazioni polisonnografiche. I file con estensione .rec e .hyp contengono rispettivamente le registrazioni originali e i loro ipnogrammi, entrambi in formato EDF. Ogni file EDF include un'intestazione con informazioni sul paziente, sulla registrazione e sui segnali acquisiti. Le registrazioni sono state ottenute da soggetti di sesso maschile e femminile, caucasici, di età compresa tra 21 e 35 anni.

Esse includono:

- EEG (FpzCz e PzOz) e EOG orizzontale, campionati a 100 Hz.
- Le registrazioni sc* contengono anche EMG sottomentale (involucro), flusso oro-nasale, temperatura corporea rettale e marcatori di eventi, campionati a 1 Hz.
- Le registrazioni st* includono EMG sottomentale campionato a 100 Hz e marcatori di eventi campionati a 1 Hz.

Gli ipnogrammi sono stati valutati manualmente. Le fasi del sonno sono codificate nei file come: W (veglia), 1, 2, 3, 4 (stadi NREM), R (REM), M (movimento), 'unscored' (non valutati)

Le registrazioni sc* sono ottenute da volontari sani durante 24 ore nella loro normale vita quotidiana, quelle st* sono ottenute in ospedale da soggetti con lieve difficoltà ad addormentarsi, ma comunque sani, utilizzando un sistema di telemetria in miniatura con alta qualità del segnale.

Dataset Sleep-EDFX

Una versione ampiamente estesa del database, contenete 197 registrazioni polisonnografiche PSG, è ora disponibile e raccomandata per nuovi studi. Le Registrazioni PSG contengono segnale EEG (Fpz-Cz e Pz-Oz), segnale EOG orizzontale, segnale EMG sottomentale, respirazione oro-nasale e temperatura corporea rettale. Gli ipnogrammi contengono anche le annotazioni delle fasi del sonno abbinate ai PSG, valutate manualmente da tecnici esperti. Le fasi includono: W (Wake, veglia), R (REM, Rapid Eye Movement), 1 (S1 Non-Rapid Eye Movement), 2 (S2 Non-Rapid Eye Movement), 3 (S3 Non-Rapid Eye Movement), 4 (S4 Non-Rapid Eye Movement), M (tempo di movimento) e ? (non valutati). I file PSG sono in formato EDF, mentre gli ipnogrammi sono in formato EDF+ con intestazioni che specificano genere ed età del paziente. Sulla base dello standard

AASM, S1 e S2 sono stati successivamente annotati come N1 e N2, mentre S3 e S4 sono stati uniti in una fase del sonno N3.

Registrazioni specifiche:

- 153 SC* file (Sleep Cassette): ottenuti da uno studio sugli effetti dell'età sul sonno in soggetti sani (25-101 anni).
    - Due PSG si circa 20 ore registrati durante due giorni consecutivi nelle case dei soggetti.
    - Frequenza di campionamento: 100 Hz per EEG/EOG, 1 Hz per EMG, temperatura rettale e marcatori di eventi.
- 44 ST* file (Sleep Telemetry): ottenuti da uno studio sugli effetti del temazepam sul sonno in 22 soggetti caucasici sani con lieve difficoltà ad addormentarsi.
    - PSG di circa 9 ore, registrati in ospedale per due notti con un sistema di telemetria in miniatura.

In questo lavoro di tesi, tra tutti i dati e segnali disponibili nel dataset Sleep-EDFX, è stata utilizzata una sottoporzione di registrazioni ed è stato selezionato il segnale EEG Fpz-Cz come singolo canale e diviso in epoche di 30 secondi per effettuare la classificazione delle fasi del sonno tramite apprendimento profondo. Viene scelto il canale EEG Fpz-Cz perché i complessi K e i fusi del sonno (schemi tipici dello stadio N2) e le onde vertex sharp (tipiche dello stadio N1) possono essere registrati nelle regioni cerebrali centrali/frontali, secondo le linee guida AASM.

3.2 Preprocessing dei dati

Il preprocessing dei dati rappresenta una fase fondamentale per garantire che i segnali EEG utilizzati nell'addestramento del modello siano omogenei, di qualità e adeguatamente strutturati per l'analisi automatizzata. Questo processo include operazioni di selezione delle epoche da utilizzare per l'allenamento, bilanciamento delle classi del dataset, filtraggio del segnale EEG e suddivisione in training, validation e test set.

3.2.1 Data Cleaning

Tra tutti i soggetti e le registrazioni per ogni soggetto disponibili del dataset Sleep-EDFX, in questo lavoro se ne utilizza una sottoporzione, definita su criteri di omogeneità delle caratteristiche della popolazione su cui i segnali sono stati acquisiti e su criteri di qualità e completezza dei dati.

Nel dataset originale, alcune acquisizioni contengono segnali di qualità insufficiente o presentano artefatti dovuti a movimenti o disturbi tecnici. Per garantire un dataset più omogeneo, è stata selezionata una sottoporzione del dataset completo, escludendo le registrazioni con qualità compromessa. Questa scelta si basa su studi precedenti [19] che evidenziano l'importanza di una pulizia accurata dei dati prima dell'analisi.

Sulla base di queste considerazioni, i soggetti selezionati sono giovani, con età compresa tra i 26 e i 33 anni, sani, maschi e femmine, e di essi si è deciso di selezionare dodici registrazioni (SC4001E0, SC4002E0, SC4011E0, SC4012E0, SC4021E0, SC4031E0, SC4051E0, SC4061E0, SC4112E0, SC4122E0, SC4131E0, SC4182E0) relative a 10 soggetti. Gli altri soggetti sono stati esclusi dall'analisi poiché le loro registrazioni contenevano movimenti e epoche non riportate.

Uno dei problemi principali nella classificazione delle fasi del sonno è lo sbilanciamento delle classi, che può influenzare significativamente le prestazioni del modello. In particolare, la fase N1 è sottorappresentata, costituendo solo circa il 3% del dataset totale, mentre la fase W (veglia) è sovrarappresentata ed è la più facilmente classificabile. Senza un'adeguata strategia di bilanciamento, il modello tenderebbe a favorire le classi più frequenti, penalizzando la capacità di riconoscere correttamente la fase N1.

Delle fasi N1, N2, N3, R sono state selezionate tutte le epoche dei soggetti di interesse, mentre per la fase W, essendo la fase più rappresentata e la più facilmente riconoscibile, si è deciso di ridurne la numerosità in modo da essere confrontabile con quella delle fasi meno rappresentate. Oltre a questa prima strategia grossolana di bilanciamento delle classi, si prenderà atto della problematica anche in fase di allenamento attribuendo in modo specifico ad ogni classe un peso in base alla sua rappresentazione all'interno dell'intero dataset di addestramento.

In seguito a queste operazioni di selezione dei soggetti e delle registrazioni, si ottengono le seguenti numerosità di epoche di 30 secondi di segnale EEG per ogni fase del sonno:

Fase del sonno	W	N1	N2	N3	REM	Totale
Numero epoche	840	980	4969	1733	1866	10388

Tabella 3: Numero di epoche delle fasi del sonno nel set di dati finale

3.2.2 Filtraggio del segnale EEG

Prima di essere utilizzati per l'addestramento, i segnali EEG vengono sottoposti a un filtraggio passabanda con banda passante compresa tra 0.3 e 45 Hz, al fine di rimuovere le componenti di frequenza indesiderate che potrebbero interferire con l'analisi del sonno. Questo intervallo è comunemente utilizzato negli studi EEG per preservare le frequenze di interesse associate alle diverse fasi del sonno, eliminando il rumore ad alta frequenza e derive a bassa frequenza.

3.2.3 Suddivisione in Training Set, Validation Set e Test Set

Dopo la fase di preprocessing, il dataset è stato suddiviso in tre sottoinsiemi:

- Training set (80%): utilizzato per l'addestramento del modello.
- Validation set (10%): impiegato per monitorare le prestazioni del modello durante l'addestramento e prevenire il fenomeno dell'overfitting.
- Test set (10%): utilizzato per la valutazione finale delle prestazioni del modello.

	Training Set	Validation Set	Test Set
Numero Epoche	8327 (10 registrazioni)	967 (1 registrazione)	1094 (1 registrazione)

Tabella 4: Numero di epoche nella suddivisione del Dataset in Training, Validation e Test Set

Questa suddivisione garantisce un equilibrio tra una consistente quantità di dati disponibili per l'apprendimento e la capacità di valutare il modello su dati mai visti prima per fornire una misura delle prestazioni del modello.

3.3 Architettura del modello e Allenamento

3.3.1 Scelta del Modello e Descrizione dell'Architettura

Per la classificazione delle fasi del sonno tramite Deep Learning, è stata scelta un'architettura basata su una combinazione di **reti neurali convoluzionali (CNN) e reti Long Short-Term Memory bidirezionali (BiLSTM)**.

Questa soluzione di architettura è motivata dalla capacità delle componenti di cui si articola di estrarre caratteristiche diverse ma importanti per la classificazione delle fasi del sonno e dalla necessità di estrarre sia caratteristiche spaziali dal segnale EEG che informazioni sequenziali sulle transizioni tra le fasi del sonno:

1. **CNN** per l'estrazione delle caratteristiche locali e spaziali dal segnale EEG.
2. **BiLSTM**, che permette di catturare la relazione temporale tra epoche adiacenti, migliorando la capacità del modello di apprendere informazioni sequenziali sulle transizioni tra le fasi del sonno.

Motivazione della Scelta di CNN+BiLSTM

L'uso di sole CNN per la classificazione del sonno ha dimostrato buone capacità di estrazione delle caratteristiche [42], ma tende a ignorare le transizioni temporali tra le fasi. Per questo motivo, molti studi recenti hanno combinato CNN con modelli ricorrenti per migliorare l'accuratezza nella stadiazione del sonno. Le reti neurali ricorrenti (RNN) sono una classe di reti che eccellono nella gestione di dati di serie temporali e sono adatte per informazioni sequenziali, ma le RNN tradizionali non sono in grado di rilevare dipendenze a lungo raggio. Questo problema è affrontato dalla rete Long Short-Term Memory (LSTM), che è una versione espansa di RNN. Le reti di memoria a lungo e breve termine bidirezionali BiLSTM sono particolarmente efficaci in questo contesto perché considerano la

dipendenza temporale tra epoche consecutive, catturando sia informazioni passate che future, sono progettate per preservare le informazioni a lungo termine [43, 44]. Queste caratteristiche rendono le reti BiLSTM particolarmente adatte all'analisi di segnali fisiologici, come EEG, per i quali le transizioni tra le fasi del sonno sono fondamentali per una classificazione accurata.

Studi precedenti [18] hanno dimostrato che la combinazione di CNN per l'estrazione delle feature e LSTM per la modellazione temporale migliora significativamente le prestazioni nella classificazione automatica delle fasi del sonno.

Descrizione Architettura del Modello

L'architettura del modello è composta da:

- **Blocchi convoluzionali**: più livelli convoluzionali con attivazione ReLU e pooling per ridurre la dimensionalità ed estrarre le caratteristiche.
- **Layer BiLSTM**: due strati LSTM bidirezionali per catturare relazioni di transizione tra epoche consecutive.
- **Dropout**: strati di dropout per migliorare la generalizzabilità e ridurre l'overfitting.
- **Strato denso finale**: strato completamente connesso con funzione di attivazione softmax per la classificazione nelle cinque fasi del sonno.

Input e Output della rete neurale profonda DNN usata per l'allenamento:

- **Input**: epoche di 30 secondi (frequenza di campionamento 100 Hz => 3000 campioni per ogni epoca) di segnale EEG con dimensioni 3000x1.
- **Output**: classificazione multi-classe (0=W, 1=N1, 2=N2, 3=N3, 4=REM).

L'architettura del modello proposto è progettata con una sequenza di 6 strati convoluzionali 1D (1D Conv), seguiti da due strati BiLSTM, e tre successivi strati densi. I primi due strati Dense impiegano la funzione di attivazione Rectified Linear Unit (ReLU), mentre lo strato finale utilizza la funzione di attivazione Softmax per generare probabilità.

- **Layer convoluzionali e di pooling**:
    - **Conv1D** con filtri crescenti: 8, 16, 24, 36, 48, 56
    - **Kernel size**: 55, 41, 33, 21, 9, 3
    - **MaxPooling1D** per riduzione della dimensionalità
- **Dropout**:
    - Tre strati di dropout con probabilità *p=0.3* per ridurre overfitting
- **BiLSTM**:
    - Due livelli con 10 e 5 unità nascoste rispettivamente, per modellare la dipendenza tra epoche consecutive
- **Strati densi**:
    - 50 e 20 neuroni con funzione di attivazione *ReLU*
    - Output finale con 5 neuroni e funzione di attivazione *Softmax*

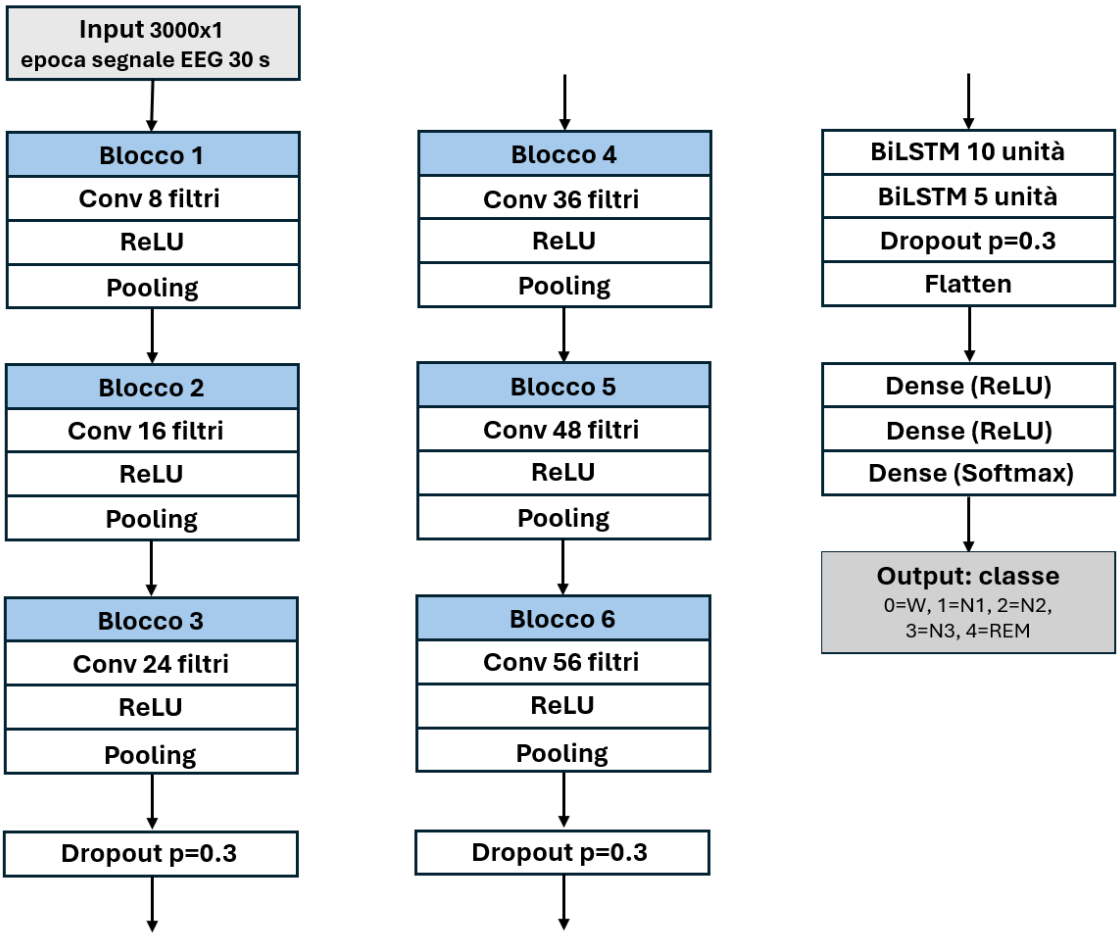

Figura 2: Schema architettura rete CNN+BiLSTM

3.3.2 Fase di Training e Ottimizzazione

La fase di training del modello di deep learning CNN+BiLSTM è stata progettata con l'obiettivo di massimizzare la capacità di generalizzazione del modello, minimizzando il rischio di overfitting. Questa sezione descrive i parametri di training, le tecniche di ottimizzazione e i metodi utilizzati per migliorare le prestazioni del modello durante l'addestramento; di seguito si riporta la configurazione finale, risultato di un'attenta operazione di fine tuning dei parametri della rete in modo da ottenere delle buone performance di partenza per gli step successivi di XAI e UQ.

I principali parametri configurati per l'allenamento del modello includono:

- **Funzione di perdita**: *Categorical Crossentropy*, scelta perché la classificazione delle fasi del sonno è un problema multi-classe.
- **Ottimizzatore**: *Adam* (Adaptive Moment Estimation), selezionato per la sua capacità di adattare dinamicamente il learning rate e migliorare la convergenza del modello.
- **Learning rate**: 0.00001, un valore basso per garantire un aggiornamento graduale dei pesi e ridurre il rischio di oscillazioni e divergenza.
- **Batch size**: 20, bilanciato per garantire un compromesso tra efficienza computazionale e stabilità del training.
- **Numero di epoche**: 500, definito per permettere al modello di apprendere efficacemente le caratteristiche dei dati senza incorrere in underfitting.
- **Class Weight**: assegnazione di pesi alle classi in base alla loro numerosità all'interno del dataset. Questo permette di compensare la sottorappresentazione della fase N1 nel dataset e migliorare la capacità del modello di riconoscerla.

Questa configurazione di parametri permette al modello di bilanciare accuratezza, stabilità e generalizzabilità, risultando efficace nel task di classificazione automatica delle fasi del sonno.

Per migliorare l'allenamento, sono state adottate diverse strategie di ottimizzazione:

- **Early Stopping e Model Checkpoint**: Per evitare l'overfitting, il training è stato monitorato utilizzando la metrica della *val_loss* (perdita sul validation set), salvando il modello all'epoca intermedia con le migliori prestazioni e terminando l'allenamento se non si ottengono ulteriori miglioramenti per un determinato numero di epoche.
- **Dropout**: Tre livelli di dropout con probabilità *p=0.3* sono stati introdotti per ridurre l'overfitting e migliorare la capacità di generalizzazione del modello.
- **Bilanciamento delle classi**: La fase N1, meno rappresentata nel dataset, ha ricevuto un peso maggiore nella funzione di perdita per evitare che il modello favorisca le classi più frequenti.

Implementazione dell'Allenamento

Il modello è stato implementato in **Python 3.10** utilizzando la libreria di deep learning **TensorFlow 2.11.0**.

I dettagli implementativi per l'esecuzione del training della rete del modello sono riportati nella sezione Appendici A.

3.3.3 Valutazione del modello

Dopo la fase di training, il modello viene analizzato utilizzando diverse metriche di valutazione per compiti di classificazione multiclasse. L'obiettivo principale di questa analisi è quantificare la bontà del modello nella classificazione delle fasi del sonno e osservare eventuali criticità, in particolare per la fase N1, che in letteratura è la più difficile da classificare. È importante riportare diverse metriche di valutazione per fornire informazioni complete ed esaustive sulla capacità del modello di svolgere il task richiesto poiché ogni metrica può portare informazioni parziali su quanto il modello sia adatto alla classificazione.

Metriche di Valutazione

Le prestazioni del modello vengono misurate utilizzando le seguenti metriche:

- **Accuracy**: misura la percentuale di campioni correttamente classificati rispetto al totale dei campioni.

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

  dove:

  - TP (True Positive) = numero di campioni positivi classificati correttamente.
  - TN (True Negative) = numero di campioni negativi classificati correttamente.
  - FP (False Positive) = numero di campioni negativi classificati erroneamente come positivi.
  - FN (False Negative) = numero di campioni positivi classificati erroneamente come negativi.

L'accuracy è una metrica utile quando il dataset è bilanciato, ovvero quando tutte le classi hanno una quantità simile di campioni. Tuttavia, se una classe è molto più rappresentata delle altre, l'accuracy può essere fuorviante perché un modello potrebbe ottenere un'alta accuracy semplicemente classificando correttamente i campioni della classe più rappresentata.

- **Precision**: misura la percentuale di predizioni corrette tra tutte quelle fatte per una classe.

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

  La precision è importante quando il costo di un falso positivo è elevato.

- **Recall (Sensitivity)**: misura la capacità del modello di identificare correttamente tutti i campioni appartenenti a una classe.

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

  La recall è fondamentale quando il costo di un falso negativo è elevato.

- **F1-score**: è la media armonica tra precision e recall e rappresenta un compromesso tra le due metriche.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

  È utile quando si ha un dataset sbilanciato e il modello potrebbe avere alta precisione ma basso recall. L'F1-score aiuta a valutare meglio le prestazioni complessive.

- **Confusion Matrix**: rappresenta una tabella in cui le predizioni del modello vengono confrontate con le etichette reali per ciascuna classe. Aiuta a identificare eventuali bias del modello verso determinate classi. Per la definizione di confusion matrix, sulla diagonale si trovano i veri positivi (TP).

L’analisi delle performance del modello viene effettuata sui tre diversi set di dati:

- **Training set**: per valutare la capacità del modello di apprendere dai dati di addestramento.
- **Validation set**: utilizzato per la scelta degli iperparametri e per il monitoraggio dell’overfitting.
- **Test set**: utilizzato per valutare le performance finali su dati mai visti durante il training.

Valutazione del modello

- **Andamento di Accuracy e Loss durante l’allenamento**: viene rappresentato il comportamento della funzione di costo e dell’accuracy nel corso delle epoche di training, permettendo di identificare eventuali problemi di overfitting o underfitting.

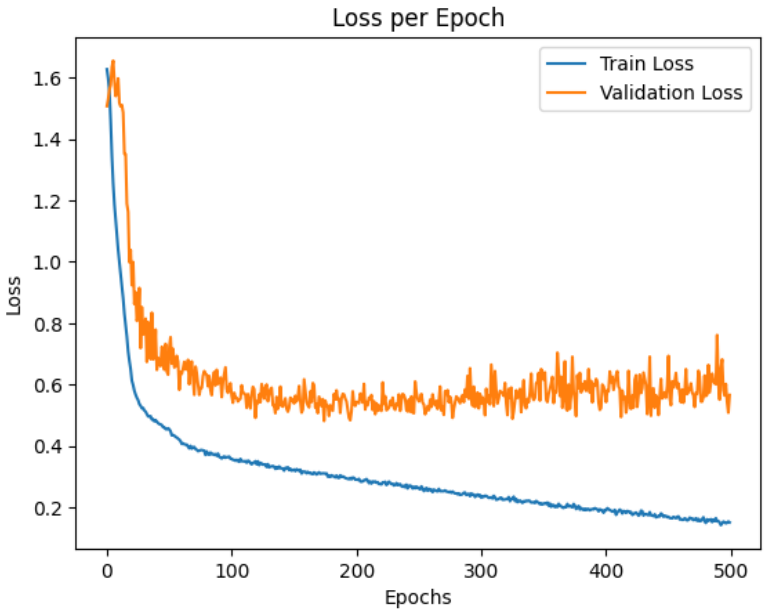

Figura 3: Andamento Loss Function durante l’allenamento

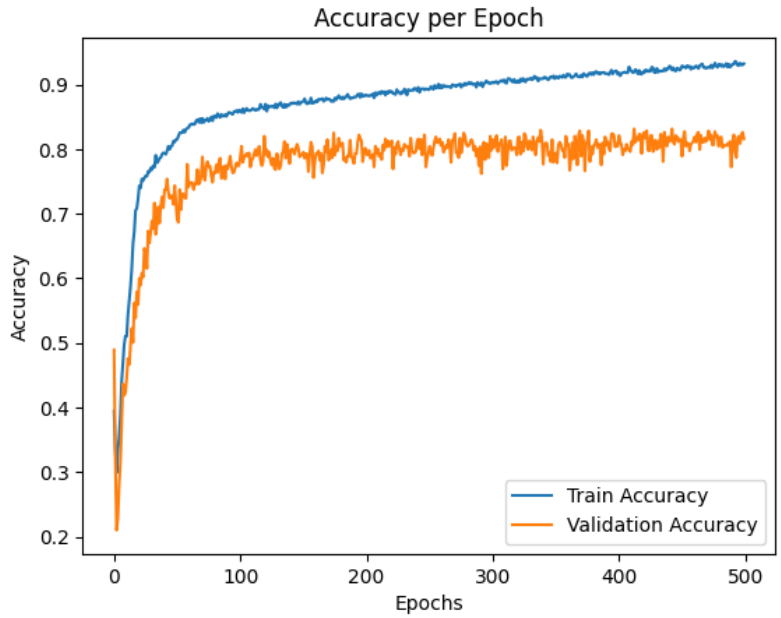

Figura 4: Andamento Accuracy durante l'allenamento

- **ROC Curve**: per valutare le capacità discriminative del modello nelle diverse classi. Approccio one-vs-rest (0: W, 1: N1, 2: N2, 3: N3, 4: R).

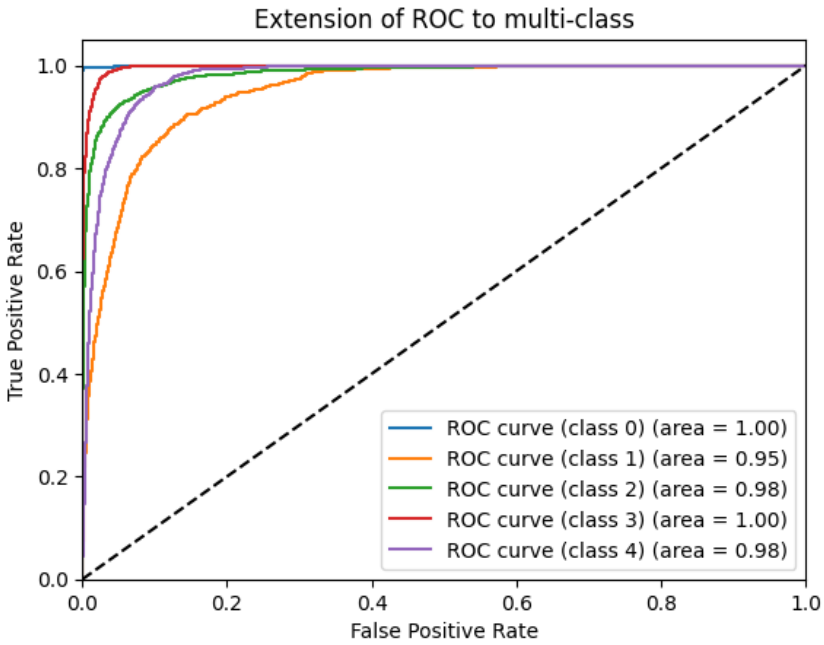

Figura 5: ROC Curve Classe per Classe

- **Confusion Matrix per Training, Validation e Test Set**: per analizzare gli errori di classificazione e identificare le classi più problematiche.

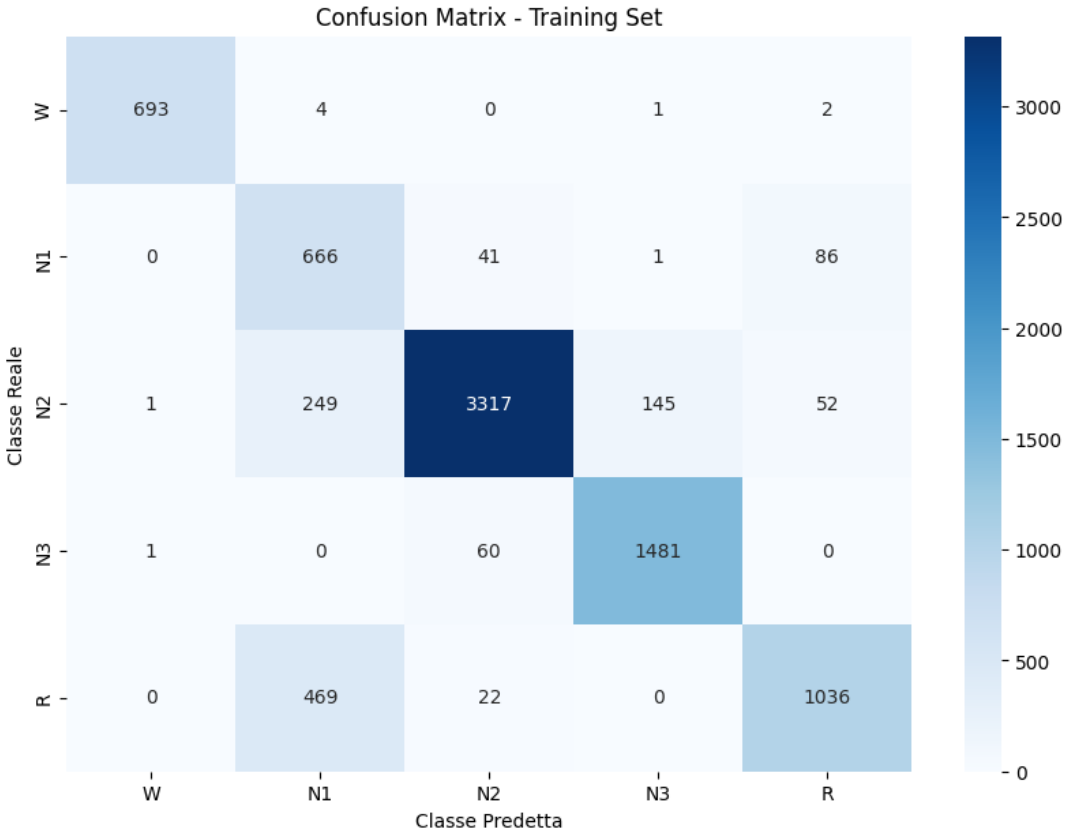

Figura 6: Confusion Matrix - CNN+BiLSTM - Training Set

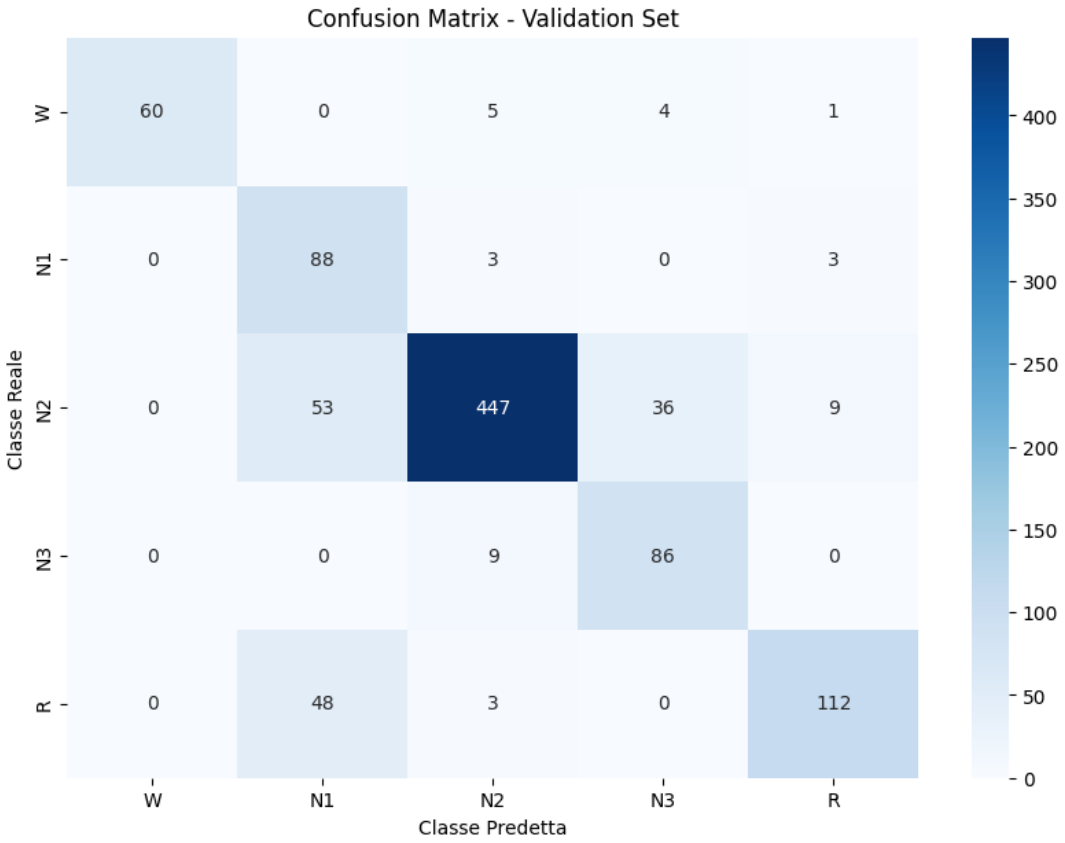

Figura 7: Confusion Matrix - CNN+BiLSTM - Validation Set

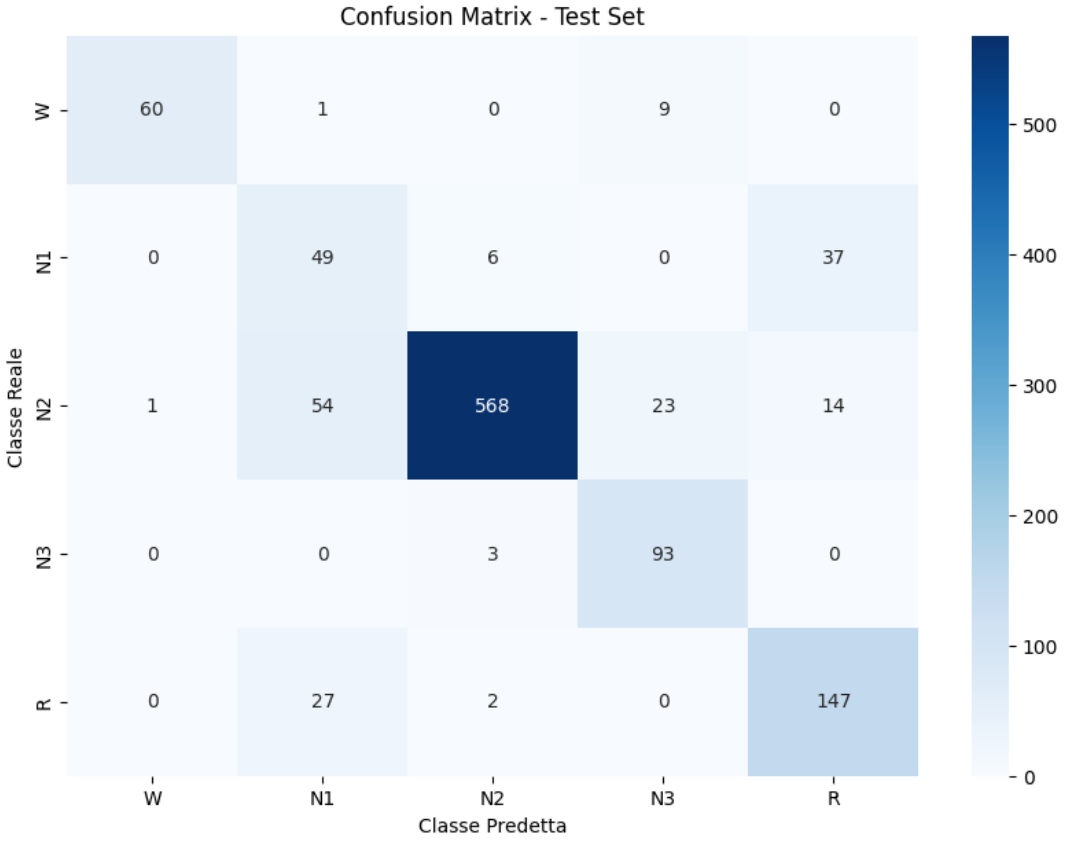

Figura 8: Confusion Matrix - CNN+BiLSTM - Test Set

- **Precision, Recall e F1-score per ogni classe.**

Training set	Precision	Recall	F1-Score
W	1.00	0.99	0.99
N1	0.48	0.84	0.61
N2	0.96	0.88	0.92
N3	0.91	0.96	0.93
REM	0.88	0.68	0.77

Tabella 5: Precision, Recall e F1-score per ogni classe - Training Set

Validation set	Precision	Recall	F1-Score
W	1.00	0.86	0.92
N1	0.47	0.94	0.62
N2	0.96	0.82	0.88
N3	0.68	0.91	0.78
REM	0.90	0.69	0.78

Tabella 6: Precision, Recall e F1-score per ogni classe - Validation Set

Test set	Precision	Recall	F1-Score
W	0.98	0.86	0.92
N1	0.37	0.53	0.44
N2	0.98	0.86	0.92
N3	0.74	0.97	0.84
REM	0.74	0.84	0.79

Tabella 7: Precision, Recall e F1-score per ogni classe - Test Set

- **Loss e Accuracy in fase di Inference**

	Train	Val	Test
Loss	0.349	0.492	0.421
Accuracy	0.864	0.820	0.838

Tabella 8: Loss Function e Accuracy in fase di Inference

Queste analisi permettono di valutare in modo completo le performance del modello proposto.

Confronto con lo Stato dell'Arte

Per contestualizzare le performance del modello rispetto ai lavori esistenti, viene fornita una tabella comparativa tra il modello proposto in questo lavoro di tesi e i modelli di classificazione automatica delle fasi del sonno più performanti presenti in letteratura.

Si può osservare che le performance ottenute dal modello proposto sono confrontabili con quelle migliori presenti in letteratura sugli studi di classificazione automatica delle fasi del sonno tramite metodi di apprendimento profondo.

Metodo	Precision					Recall					F1-score					PCC				
	W	N1	N2	N3	REM	W	N1	N2	N3	REM	W	N1	N2	N3	REM	W	N1	N2	N3	REM
1. [17]	0.81	0.55	0.78	0.76	0.71	0.86	0.38	0.86	0.65	0.69	0.83	0.45	0.82	0.70	0.70					
2. [18]	0.86	0.45	0.88	0.88	0.77	0.89	0.26	0.89	0.89	0.82										
3. [45]	0.86	0.52	0.77	0.86	0.80	0.87	0.31	0.84	0.81	0.82	0.86	0.39	0.81	0.84	0.82	0.73	0.28	0.49	0.76	0.68
4. [7]																0.95	0.61	0.89	0.92	0.84
Proposto	0.98	0.38	0.98	0.74	0.74	0.86	0.53	0.86	0.97	0.84	0.92	0.44	0.92	0.84	0.79	0.97	0.30	0.85	0.73	0.65

Tabella 9: Confronto Perfromance Stato dell'Arte - Metodo Proposto

Un'attenzione particolare viene riservata alla fase N1, che risulta sempre la più difficile da classificare. Nelle successive fasi di applicazione delle tecniche XAI e UQ al modello allenato si agirà nell'ottica di migliorare queste performance di partenza.

3.4 Quantificazione dell’incertezza UQ

Importanza della Quantificazione dell'Incertezza

Nell'ambito del deep learning, la quantificazione dell'incertezza (Uncertainty Quantification - UQ) rappresenta un aspetto fondamentale per migliorare l'affidabilità delle predizioni del modello. La principale motivazione per l'adozione di tecniche di UQ è la capacità di identificare campioni con alta incertezza, consentendo di filtrare o trattare in modo diverso le previsioni meno affidabili. Questo approccio porta a un incremento delle prestazioni complessive, riducendo il numero di predizioni errate attraverso la rimozione di campioni ad alta incertezza.

3.4.1 Monte Carlo Dropout

Uno dei metodi più efficaci per stimare l'incertezza in modelli di deep learning è il **Monte Carlo Dropout (MCD)** [28], una tecnica che consiste nell'attivare i layer di dropout anche in fase di inferenza. Il dropout, originariamente introdotto per ridurre l'overfitting durante il training, consiste nello spegnimento casuale di neuroni con una determinata probabilità. Applicato durante il test, permette di generare multiple predizioni per lo stesso input, creando così una distribuzione di probabilità invece di una singola previsione deterministica.

Questa tecnica presenta diversi vantaggi:

- **Semplicità**: può essere facilmente implementata abilitando livelli di dropout.
- **Versatilità**: si applica in qualsiasi architettura di deep learning senza modifiche sostanziali alla struttura della rete, ma semplicemente aggiungendo livelli di dropout.

- **Incremento dell'affidabilità**: consente di ottenere più predizioni da un solo input e migliorare così l'accuratezza media del modello riducendo l'incertezza.

Una volta ottenuta la distribuzione delle predizioni attraverso MCD, è possibile estrarre metriche utili per quantificare l'incertezza del modello. In questo studio si utilizza l'**entropia normalizzata** come metrica per la valutazione dell'incertezza.

3.4.2 Entropia Normalizzata

L'entropia è una misura della dispersione della distribuzione delle probabilità associate alle predizioni. Se la distribuzione è altamente concentrata su una singola classe, l'entropia sarà bassa, indicando alta confidenza. Al contrario, una distribuzione più uniforme tra le classi suggerisce alta incertezza. L'entropia normalizzata viene definita come:

$$H_{norm} = -\sum_{i=1}^{C} p_i \frac{\log(p_i)}{\log(C)}$$

Dove:

- C è il numero totale di classi;
- p_i è la probabilità predetta per la classe i.

Questa misura è normalizzata rispetto al numero di classi, permettendo un confronto equo tra modelli con differenti numeri di output.

Una volta quantificata l'incertezza delle predizioni del modello, si possono stabilire soglie per filtrare campioni con alta incertezza, migliorando così la qualità delle predizioni e ottimizzando la calibrazione del modello.

3.4.3 Calibrazione del modello

Importanza della fase di calibrazione

La fase di calibrazione del modello ha lo scopo di determinare il valore ottimale della probabilità di dropout p da applicare durante l'inferenza per l'applicazione Monte Carlo Dropout (MCD) e la quantificazione dell'incertezza (UQ). L'obiettivo è massimizzare la capacità del modello di distinguere tra campioni correttamente classificati (CC) e misclassificati (MC).

Durante la fase di training, un valore di p=0.5 è comunemente utilizzato per prevenire l'overfitting. Tuttavia, in fase di inferenza, un dropout troppo elevato potrebbe introdurre una variabilità eccessiva nelle predizioni, compromettendone l'affidabilità e rendendo meno efficace l'operazione di quantificazione dell'incertezza nell'ottica di migliorare le performance del modello grazie alla rimozione dei campioni meno affidabili. Per questa ragione, si rende necessaria una fase di calibrazione per trovare il valore di p più appropriato.

Processo di calibrazione e selezione della probabilità di dropout ottimale

La calibrazione viene eseguita testando diversi valori della probabilità di dropout (p = 0.05, 0.1, 0.3, 0.5) e analizzando l'entropia normalizzata delle predizioni ottenute con l'applicazione di MCD. L'intero processo è articolato nei seguenti passaggi:

1. **Inference baseline**
    - Il modello proposto allenato viene utilizzato per generare predizioni baseline (dropout non abilitato) sul validation set e sul test set.
    - Le predizioni baseline vengono confrontate con i target reali per identificare i campioni CC e MC. Se la classe predetta in baseline è uguale alla classe target, il campione è considerato CC; altrimenti, è MC.

2. **Inference con Monte Carlo Dropout**
    - Il modello viene applicato con MCD attivo per ogni input e con 10 iterazioni, ottenendo così 10 predizioni MCD per ogni input
    - Il dropout viene attivato con diverse probabilità (p=0.05, 0.1, 0.3, 0.5).
3. **Calcolo dell'entropia normalizzata**
    - L'entropia normalizzata viene calcolata per ogni campione, basandosi sulle probabilità softmax ottenute nelle 10 iterazioni MCD e associate alla classe predetta in baseline.
    - L'entropia normalizzata rappresenta la misura dell'incertezza associata alla predizione del modello e viene calcolata per ogni valore di p.
4. **Suddivisione dei campioni in CC e MC**
    - I campioni vengono divisi in CC e MC sulla base delle predizioni baseline e del confronto con i target reali.
5. **Analisi e visualizzazione dei risultati**
    - L'entropia normalizzata per CC e MC viene rappresentata attraverso boxplot, suddivisa per classe e per valore di ρ.
    - Il valore ottimale di ρ viene scelto osservando quale configurazione massimizza la discriminazione tra CC e MC, minimizzando l'entropia nei CC e massimizzandola nei MC.

Il processo di calibrazione si è svolto in un primo momento sul validation set e poi si è verificata la generalizzabilità delle considerazioni che ne sono derivate con l'esecuzione sul test set.

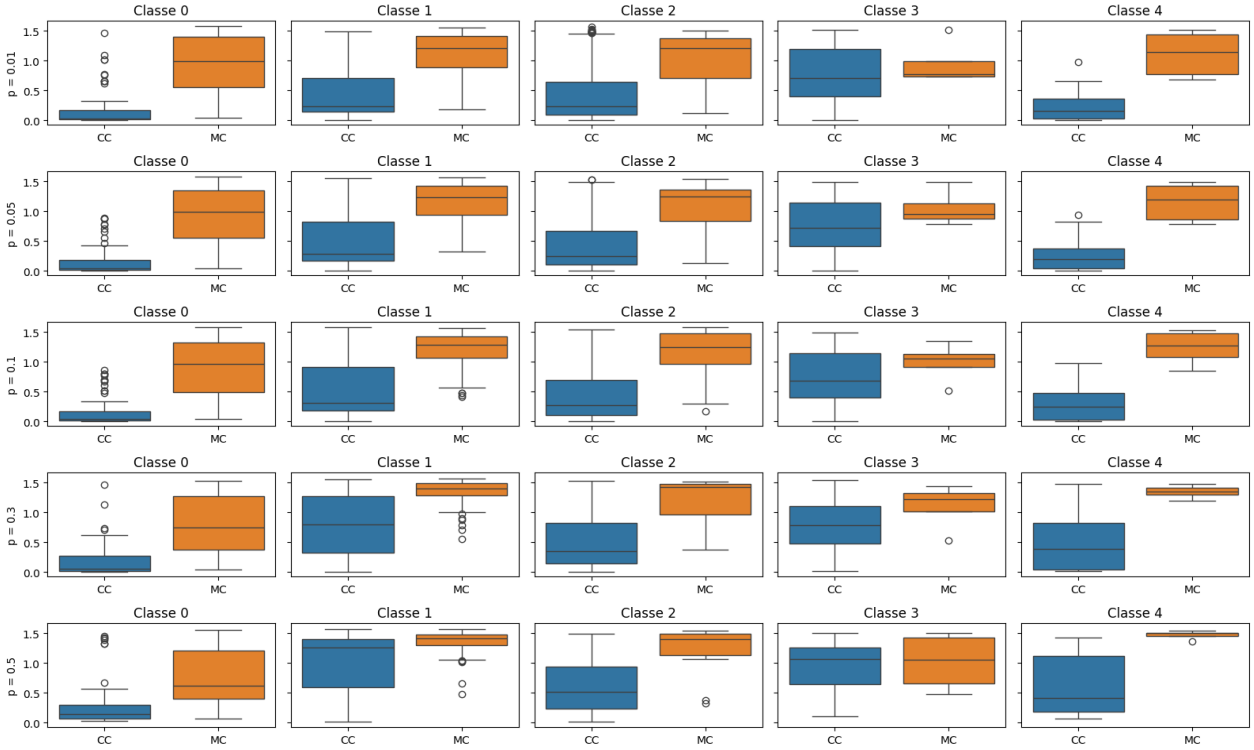

Figura 10: Boxplot entropia normalizzata per CC e MC, suddivisa per classe e per valore di ρ – Validation Set

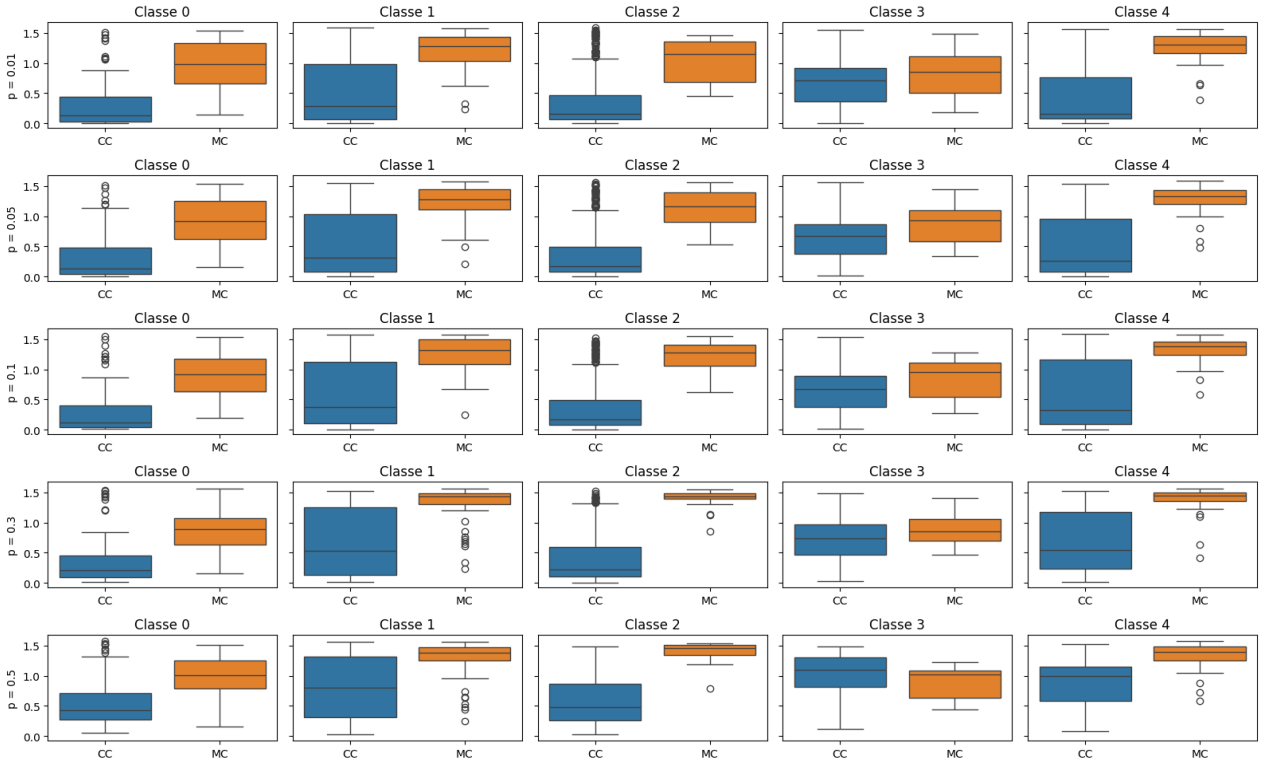

Figura 9: Boxplot entropia normalizzata per CC e MC, suddivisa per classe e per valore di ρ - Test Set

L'analisi delle distribuzioni di entropia mostra che un valore di dropout p=0.1 consente di ottenere una più chiara separazione tra CC e MC.

3.4.4 Determinazione della Soglia di Incertezza sul Validation Set

Dopo aver identificato il valore ottimale di **dropout probability (p=0.1)** per la quantificazione dell'incertezza mediante **Monte Carlo Dropout (MCD)**, il passo successivo è la determinazione di una **soglia di incertezza** che permetta di discriminare tra predizioni affidabili e non affidabili.

L'obiettivo è individuare un valore soglia dell'**entropia normalizzata** oltre il quale una predizione viene considerata troppo incerta per essere ritenuta affidabile e viene quindi scartata. Questo permette di migliorare le performance di classificazione se i campioni ad alta incertezza corrispondono in buona parte ai misclassificati dal modello.

Per valutare l'incertezza associata a ciascuna predizione, è stata calcolata l'**entropia normalizzata** delle probabilità softmax ottenute con il MCD. L'entropia rappresenta una misura della dispersione della distribuzione predittiva:

- **Bassa entropia** → Il modello assegna una probabilità elevata a una classe specifica, indicando una predizione con alta fiducia.
- **Alta entropia** → Le probabilità delle classi sono distribuite più uniformemente, segnalando una maggiore incertezza nella classificazione.

L’analisi per la determinazione della soglia di incertezza sul validation set è stata condotta sia globalmente (su tutte le classi) sia classe per classe, distinguendo tra campioni **correttamente classificati (CC)** e **misclassificati (MC)**. I risultati sono stati rappresentati graficamente attraverso **boxplot**, che evidenziano la distribuzione dell’entropia per CC e MC permettono di identificare un possibile valore soglia per discriminare le predizioni affidabili da quelle incerte.

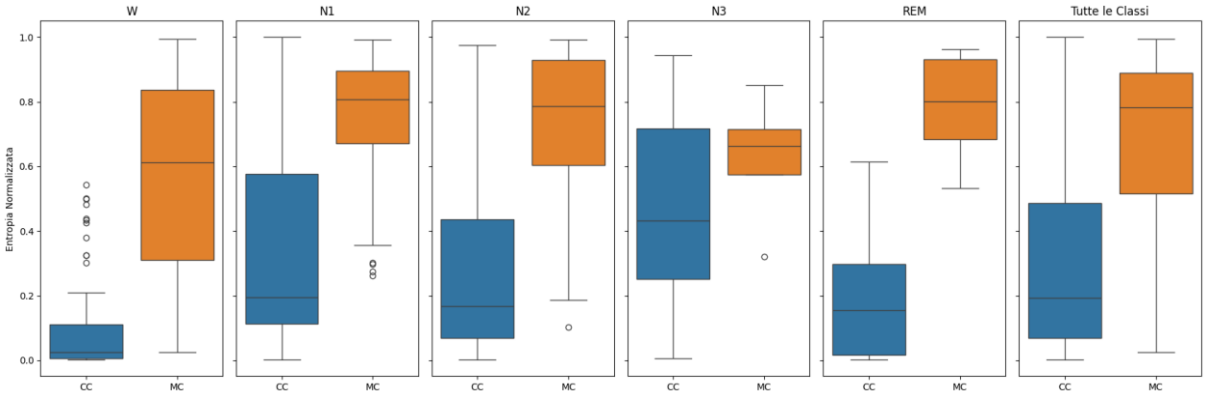

Figura 11: Boxplot entropia normalizzata su tutte le classi e sulle singole classi dividendo in CC e MC - Validation Set.

Dall’analisi dei boxplot, emerge che i campioni **misclassificati (MC)** tendono ad avere valori di entropia significativamente più elevati rispetto ai **correttamente classificati (CC)**. Questa osservazione suggerisce che l’approccio di stabilire una soglia di entropia per filtrare i campioni più incerti e migliorare l’affidabilità delle predizioni possa essere applicato in modo efficace nel tentativo di migliorare le performance di classificazione del modello.

Sono stati considerati due metodi per determinare la soglia ottimale:

1. **Criterio visivo** sui boxplot
    - Osservando la separazione tra CC e MC nel boxplot complessivo di tutte le classi, è stata identificata una prima soglia di **0.5** sull'entropia normalizzata per discriminare le predizioni affidabili da quelle incerte.
    - Tuttavia, un'analisi dettagliata sulle singole classi ha mostrato che la fase **N1** presenta una maggiore variabilità nell'incertezza, richiedendo una soglia più elevata per garantire una separazione più efficace.
2. Compromesso tra **aumento delle performance** e **numero di campioni rimossi**

   L'applicazione della soglia di incertezza deve garantire un compromesso tra il miglioramento delle performance e il mantenimento di un numero sufficiente di campioni in ciascuna classe. Dopo l'eliminazione dei campioni con entropia superiore alla soglia, il numero di campioni rimanenti in ciascuna classe può risultare sbilanciato. In particolare, l'analisi ha evidenziato che la classe N1, già poco rappresentata nel dataset di partenza, subisce una forte riduzione di campioni affidabili se si applica la stessa soglia determinata per le altre classi e potrebbe compromettere la robustezza del modello. La fase N1 presenta una maggiore variabilità nell'incertezza, richiedendo una soglia più elevata per garantire una separazione più efficace tra CC e MC e per avere un compromesso opportuno tra miglioramento delle performance e mantenimento di una numerosità adeguata post rimozione dei campioni incerti.

Per la determinazione della soglia di incertezza sulle classi W, N2, N3 e REM si è mostrato sufficiente applicare il criterio visivo che ha portato a stabilire 0.5 come soglia che permette sia di migliorare le performance sia di mantenere una numerosità adeguata dei campioni per ogni classe nel dataset dopo la rimozione dei campioni incerti.

Per evitare una perdita eccessiva di campioni nella fase N1 e per valutare l'effetto della soglia sulle prestazioni del modello, è stato realizzato un grafico che mostra la relazione tra il numero di campioni rimossi e l'aumento di accuracy. Sono state testate 50 soglie equispaziate comprese tra 0.25 e 1.

Per ogni soglia, è stata misurata la variazione percentuale dell'accuracy e il numero di campioni eliminati. La variazione percentuale di accuracy è stata calcolata tra la condizione di partenza (nessun campione rimosso) e quella post rimozione dei campioni sopra una determinata soglia. La ripetizione di questo calcolo per ogni soglia testata ha portato al seguente grafico.

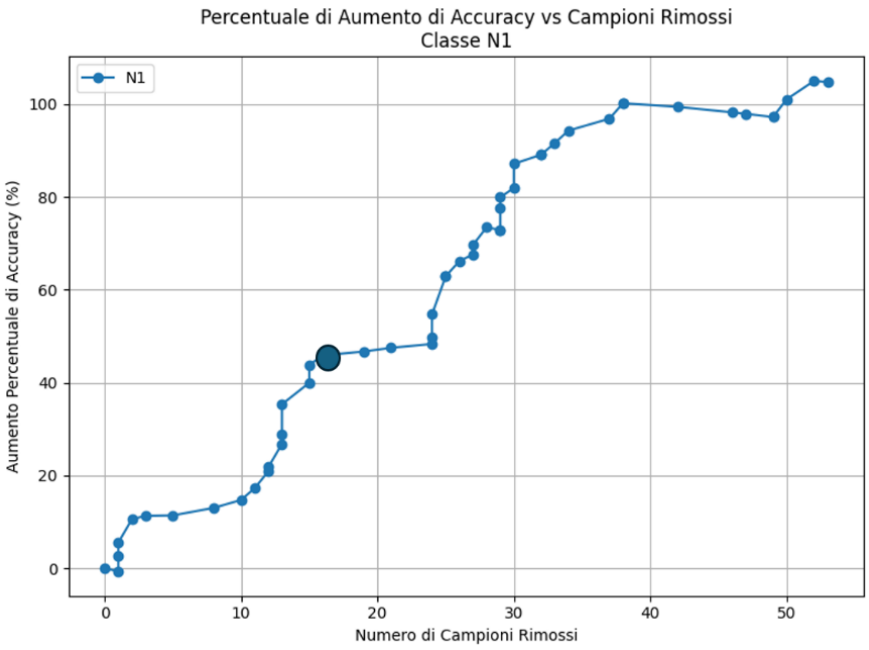

Figura 12: Andamento aumento percentuale di accuratezza in funzione del numero di campioni rimossi (valutando 50 soglie tra 0.25 e 1) per la classe N1. Il cerchio blu mostra la soglia di incertezza di 0.78.

Il grafico mostra un range di soglie tra 0.78 e 0.8 in cui l'accuracy rimane quasi invariata. Di conseguenza, scegliamo un valore di **0.78** per ridurre il numero di campioni rimossi mantenendo buone performance.

Dalla combinazione di criterio visivo per la determinazione della soglia di incertezza e del criterio basato sull'aumento percentuale dell'accuratezza in funzione del numero di campioni rimossi, le soglie ottimali risultano:

- **0.50** per le fasi **W, N2, N3, REM**
- **0.78** per la fase **N1**

3.4.5 Verifica della Soglia di Incertezza sul Test Set

Dopo aver determinato la soglia di incertezza ottimale sulla base dell'analisi del validation set, è fondamentale verificare la sua applicabilità anche sul test set per valutare la capacità di generalizzazione del metodo.

L'obiettivo è applicare la soglia di entropia normalizzata determinata precedentemente e analizzare l'impatto sulla qualità delle predizioni, ricalcolando le metriche di valutazione e le confusion matrix prima e dopo la rimozione dei campioni con alta incertezza.

Procedura di verifica sul test set

1. **Applicazione del modello sul test set**
    - Il modello viene applicato a tutte le epoche del test set per ottenere le predizioni baseline (senza Monte Carlo Dropout).
    - Queste predizioni vengono utilizzate per identificare i correttamente classificati (CC) e i misclassificati (MC) in base al confronto con le etichette reali.

2. **Applicazione di Monte Carlo Dropout (MCD) per stimare l'incertezza**
    - Il modello viene eseguito nuovamente con MCD attivo per ciascun input, generando 10 predizioni softmax per ogni epoca del test set.
    - L'incertezza viene calcolata come entropia normalizzata delle predizioni ottenute con MCD.
3. **Applicazione della soglia di incertezza**
    - Viene utilizzata la soglia ottimale stabilita sul validation set:
        - 0.5 per le fasi W, N2, N3, REM
        - 0.78 per la fase N1 (più soggetta a incertezza)
    - I campioni con entropia superiore alla soglia (non affidabili) vengono rimossi.
4. Calcolo delle **confusion matrix** e delle **metriche di valutazione** (Precision, Recall, F1-score**)** prima e dopo la rimozione dei campioni incerti e la loro variazione percentuale

Di seguito si riportano le metriche di valutazione, i loro incrementi percentuali e le confusion matrix relativi a validation set e test set prima e dopo l'applicazione della soglia di incertezza e la rimozione dei campioni non affidabili.

Validation set

Classe	*Prec Pre*	*Prec Post*	*Recall Pre*	*Recall Post*	*F1-Score Pre*	*F1-Score Post*
W	1.00	1.00	0.86	0.98	0.92	0.99
N1	0.47	0.63	0.94	0.96	0.62	0.76
N2	0.96	0.99	0.82	0.89	0.88	0.93
N3	0.68	0.76	0.91	0.98	0.78	0.86
REM	0.90	0.98	0.69	0.67	0.78	0.80

Tabella 10: Metriche Valutazione (Precision, Recall e F1-Score) Pre e Post rimozione campioni incerti tramite applicazione soglia di incertezza - Validation Set

Classe	*Aumento % Precision*	*Aumento % Recall*	*Aumento % F1-Score*
W	0.00	14.72	7.42
N1	35.36	2.71	22.43
N2	3.03	8.01	5.65
N3	11.63	7.77	9.94
REM	9.28	-2.28	2.42

Tabella 11: Aumento percentuale (%) Precision, Recall e F1-Score Pre e Post rimozione campioni incerti tramite applicazione soglia di incertezza - Validation Set

Test set

Classe	*Prec Pre*	*Prec Post*	*Recall Pre*	*Recall Post*	*F1-Score Pre*	*F1-Score Post*
W	0.98	0.98	0.86	0.98	0.92	0.98
N1	0.37	0.62	0.53	0.73	0.44	0.67
N2	0.98	0.99	0.86	0.95	0.92	0.97
N3	0.74	0.85	0.97	0.99	0.84	0.92
REM	0.74	0.91	0.84	0.91	0.79	0.91

Tabella 12: Tabella 8: Metriche Valutazione (Precision, Recall e F1-Score) Pre e Post rimozione campioni incerti tramite applicazione soglia di incertezza - Test Set

Classe	*Aumento % Precision*	*Aumento % Recall*	*Aumento % F1-Score*
W	-0.06	14.69	7.32
N1	66.46	37.69	53.25
N2	0.72	9.98	5.45
N3	14.64	2.05	8.81
REM	22.69	9.06	15.88

Tabella 13: Tabella 9: Aumento percentuale (%) Precision, Recall e F1-Score Pre e Post rimozione campioni incerti tramite applicazione soglia di incertezza - Test Set

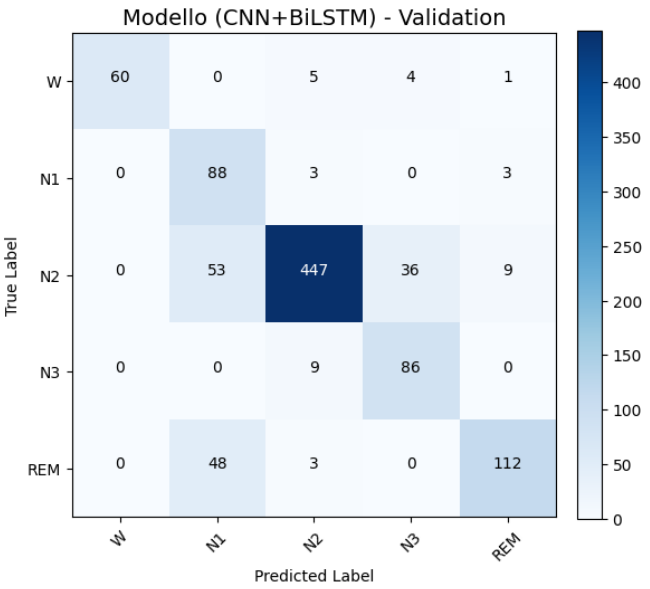

Figura 14: Confusion Matrix - CNN+BiLSTM: prima della rimozione dei campioni non affidabili - Validation Set

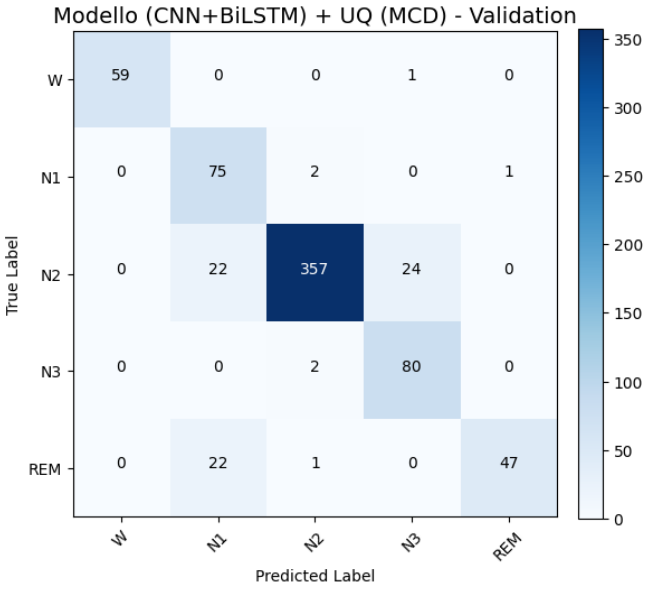

Figura 13: Confusion Matrix - CNN+BiLSTM + UQ (MCD): dopo la rimozione dei campioni non affidabili per applicazione della soglia di incertezza- Validation Set

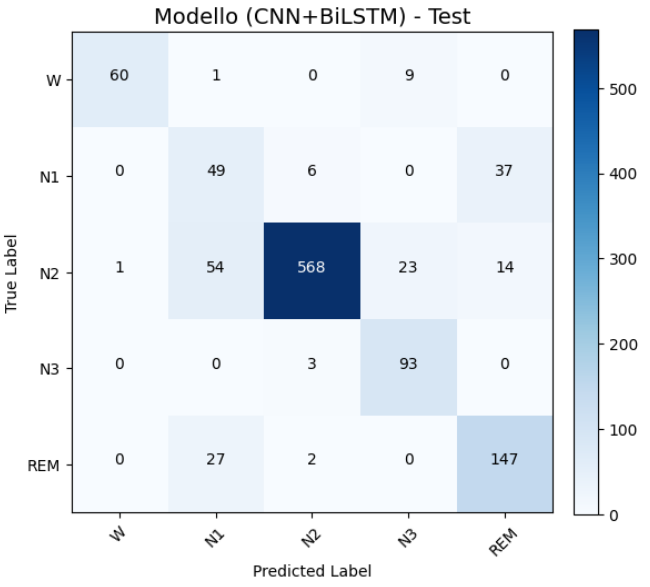

Figura 16: Confusion Matrix - CNN+BiLSTM: prima della rimozione dei campioni non affidabili - Test Set

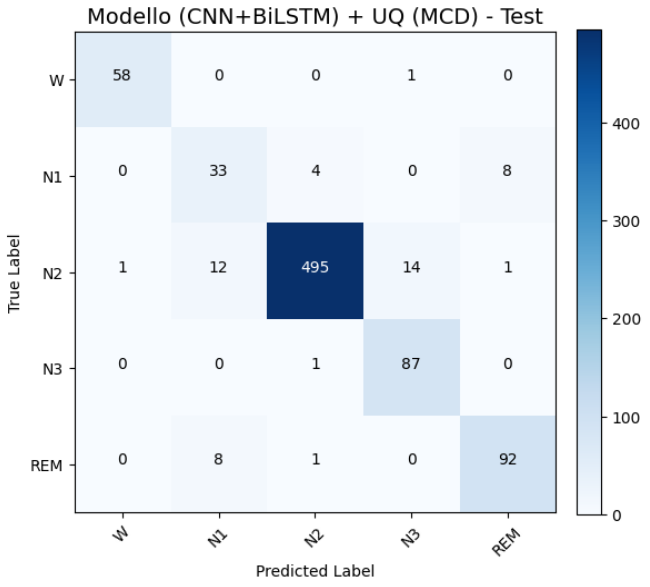

Figura 15: Confusion Matrix - CNN+BiLSTM + UQ (MCD): dopo la rimozione dei campioni non affidabili per applicazione della soglia di incertezza- Test Set

Classe	Validation Set			Test Set		
	Totale	Affidabili	Non affidabili (rimossi)	Totale	Affidabili	Non affidabili (rimossi)
W	70	60	10	70	59	11
N1	94	78	16	92	45	47
N2	545	403	142	660	523	137
N3	95	82	13	96	88	8
REM	163	70	93	176	101	75

Tabella 14: Numero di campioni totali del set di dati, affidabili (sottosoglia di incertezza) e non affidabili (rimossi, sopra soglia di incertezza) per ogni classe – Validation e Test Set

L'applicazione della soglia di incertezza determinata sul validation set ha confermato la sua efficacia anche sul test set.

- Il modello risulta più preciso dopo la rimozione delle predizioni con alta incertezza.
- Il compromesso sulla soglia della fase N1 (portata a 0.78 anziché 0.5) ha permesso di mantenere una numerosità sufficiente della classe, evitando un eccessivo squilibrio.
- Il confronto tra le confusion matrix prima e dopo la rimozione mostra una riduzione degli errori di classificazione, con un impatto positivo sulle performance del metodo.

L'applicazione della soglia di incertezza ha portato a un **miglioramento dell'accuratezza** complessiva del modello, riducendo l'impatto dei campioni misclassificati con predizioni instabili e migliorando la qualità delle classificazioni.

3.5 Artificial Intelligence Explainability (XAI)

L'Explainable AI (XAI) si riferisce a una serie di tecniche sviluppate per rendere i modelli di deep learning più interpretabili e comprensibili. Queste tecniche mirano a fornire informazioni su come i modelli di intelligenza artificiale effettuano previsioni o decisioni, consentendo di evidenziare le regioni del segnale che influenzano maggiormente la decisione del modello.

3.5.1 Metodi di Explainability XAI

I metodi di visualizzazione sono tra le strategie più utilizzate per interpretare il comportamento di una rete neurale profonda; essi evidenziano tramite mappe a colori (heatmap) le regioni dell'input che contribuiscono maggiormente alla previsione. I metodi di spiegabilità di visualizzazione si dividono in metodi di retropropagazione e metodi di perturbazione. Tra i primi, gli approcci attenzionati in questo lavoro di tesi sono i metodi basati sul gradiente (heatmap stimate con GradCAM e GradCAM++). Tra i metodi di perturbazione viene analizzata la stima delle heatmap tramite ScoreCAM.

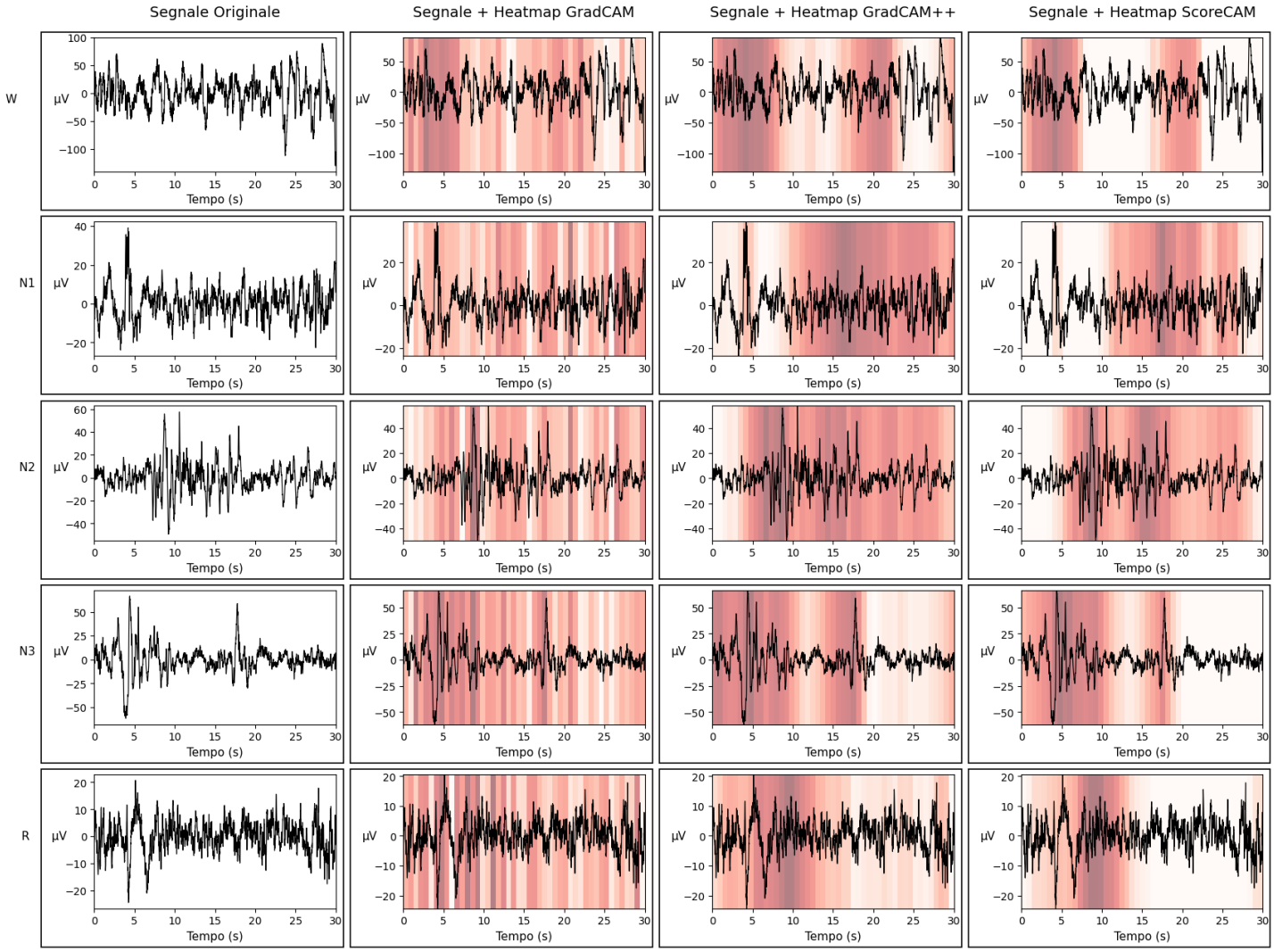

Figura 17: Epoca esemplificativa di ogni fase: segnale originale e segnale con heatmap sovrapposta per ogni metodo XAI (GradCAM, GradCAM++, ScoreCAM)

3.5.1.1 GradCAM

GradCAM (Gradient-weighted Class Activation Mapping) [29] è una tecnica che utilizza le informazioni di gradiente per generare una heatmap di attivazione che evidenzia le regioni del segnale EEG più rilevanti per la classificazione.

Dato un modello CNN-BiLSTM, GradCAM viene applicato all'ultimo layer BiLSTM. I gradienti dell'output della classe predetta rispetto alle feature map dell'ultimo layer vengono calcolati e mediati per ottenere una mappa di attivazione ponderata.

Questo processo genera una heatmap con le dimensioni dello strato BiLSTM sul quale GradCAM viene applicato; essa viene poi normalizzata e ridimensionata per una visualizzazione coerente con l'input.

La formula per il calcolo dell'importanza della feature map è:

$$\alpha_k^c = \frac{1}{Z}\sum_{i,j}\frac{\partial y^c}{\partial A_k^{i,j}}$$

Dove:

- y^c è il punteggio della classe di interesse c.
- $A_k^{i,j}$ rappresenta l'attivazione nella posizione (i, j) della feature map A_k.
- Z è il numero totale di elementi nella feature map.

La heatmap finale GradCAM viene ottenuta tramite combinazione lineare delle feature map pesate dai coefficienti α_k^c:

$$GradCAM = ReLU\left(\sum_k \alpha_k^c A_k\right)$$

Implementazione nel metodo proposto

In questo lavoro, GradCAM è stato implementato eseguendo i seguenti passaggi:

1. Estrazione delle attivazioni: viene selezionato il layer BiLSTM del modello.
2. Calcolo dei gradienti: si calcola il gradiente della classe predetta rispetto all'attivazione del layer selezionato.
3. Ponderazione delle feature map: i gradienti vengono mediati e utilizzati per pesare le attivazioni.
4. Generazione della heatmap: la combinazione delle attivazioni pesate produce la heatmap finale.
5. Interpolazione e normalizzazione: la heatmap viene normalizzata e riportata alle dimensioni del segnale EEG per permettere una visualizzazione sovrapposta di epoca di segnale con la rispettiva heatmap stimata.

L'implementazione è stata realizzata con TensorFlow e sfrutta GradientTape per calcolare i gradienti. I dettagli implementativi a livello di codice sono riportati nella sezione Appendici B.

3.5.1.2 GradCAM++

Grad-CAM++ è una evoluzione di GradCAM che migliora la localizzazione delle regioni più rilevanti utilizzando pesi più raffinati per la stima delle feature map [31]. Questo metodo utilizza le derivate di secondo e terzo ordine dei gradienti per ottenere i pesi.

La formula utilizzata per calcolare l'importanza delle feature map è la stessa di GradCAM ma con l'utilizzo delle derivate di ordine superiore.

Implementazione nel metodo proposto

L'implementazione segue una pipeline simile a quella di GradCAM:

1. Estrazione delle attivazioni del BiLSTM.
2. Calcolo delle derivate di ordine superiore.
3. Determinazione dei coefficienti di importanza.
4. Combinazione pesata delle feature map.
5. Interpolazione e normalizzazione della heatmap.

I dettagli implementativi a livello di codice sono riportati nella sezione Appendici B.

3.5.1.3 ScoreCAM

ScoreCAM [32] è un'alternativa ai metodi basati sui gradienti come GradCAM e GradCAM++. A differenza di questi ultimi, ScoreCAM non utilizza i gradienti ma sfrutta direttamente le attivazioni del layer selezionato per generare una heatmap più robusta e meno sensibile al rumore dei gradienti.

ScoreCAM colma il divario tra metodi basati sulla perturbazione e metodi basati su CAM e il suo principio di funzionamento si basa sulla perturbazione del segnale di input utilizzando le attivazioni delle feature map e sull'osservazione della variazione della previsione del modello per la classe di interesse.

L'idea è la seguente:

1. Per ogni feature map A_k del layer selezionato:
    - La feature map viene sovracampionata alle dimensioni dell'input e normalizzata.
    - Si genera una versione mascherata dell'epoca di segnale EEG moltiplicando il segnale originale per i punteggi di attivazione A_k e ottenendo così l'input perturbato.
    - Si applica il modello all'input perturbato e si ottiene la predizione.
    - Si assegna un peso alla feature map in base alla variazione della predizione.
2. Si combinano linearmente le feature map pesate per ottenere la heatmap finale.

La heatmap è data da:

$$ScoreCAM = \sum_k S(A_k)A_k$$

Dove $S(A_k)$ è il punteggio assegnato alla feature map in base alla variazione della predizione.

Implementazione nel metodo proposto

L'implementazione segue questi passaggi:

1. Estrazione delle attivazioni dal layer BiLSTM.
2. Creazione delle versioni mascherate (input perturbato) del segnale EEG.
3. Esecuzione del modello sulle versioni perturbate e calcolo dei pesi ScoreCAM.
4. Combinazione delle feature map pesate per ottenere la heatmap finale.
5. Interpolazione e normalizzazione della heatmap.

I dettagli implementativi a livello di codice sono riportati nella sezione Appendici B.

3.5.2 Analisi XAI tramite SUE e correlazione

Generazione delle Heatmap Baseline e con Monte Carlo Dropout

L'analisi delle heatmap generate con i metodi di Explainable AI (XAI) appena descritti consente di valutare la robustezza e la coerenza delle regioni di attivazione del modello durante la classificazione delle fasi del sonno. Per condurre questa analisi si sono applicati **GradCAM**, **GradCAM++** e **ScoreCAM** sia in configurazione **baseline** sia con **Monte Carlo Dropout (MCD)** attivo, al fine di valutare la stabilità delle heatmap e stimare l'incertezza dell'interpretabilità del modello.

Durante l'inferenza con MCD, per ogni campione sono state generate **10 heatmap**, consentendo di analizzare la ripetibilità delle attivazioni della rete neurale e di quantificare la sovrapposizione spaziale delle regioni più rilevanti per la classificazione.

Definizione di SUE e Correlazione

Per confrontare e valutare la stabilità delle heatmap ottenute con i diversi metodi XAI, si sono considerate due metriche principali:

- **Spatial Uncertainty Estimator (SUE)** [41]**:** metrica che integra aspetti XAI con principi di UQ e misura la sovrapposizione spaziale tra le heatmap baseline e quelle generate con MCD, offrendo una stima della ripetibilità delle previsioni del modello, quantificando il grado di sovrapposizione tra le caratteristiche più influenti identificate tramite heatmap generate con diversi metodi XAI.
- **Coefficiente di correlazione di Pearson:** quantifica il grado di similarità tra la heatmap baseline e quelle ottenute con MCD, fornendo un'indicazione della coerenza tra le predizioni con dropout attivo e senza.

Formula per il calcolo della correlazione di Pearson:

$$r = \frac{\sum(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 * (Y_i - \bar{Y})^2}}$$

Dove:

- X e Y rappresentano le intensità delle heatmap in baseline e con MCD.
- $\bar{X}$ e $\bar{Y}$ sono le medie delle due distribuzioni.
- r assume valori tra -1 e 1; valori vicini a 1 indicano forte correlazione positiva.

Formula per il calcolo di SUE:

$$SUE = \frac{\sum Heatmap_{BL} \cap Heatmap_{MCD}}{\sum Heatmap_{BL} \cup Heatmap_{MCD}}$$

Dove:

- ∩ rappresenta l'intersezione tra la heatmap baseline e quella con MCD.
- ∪ rappresenta l'unione delle due heatmap.
- Un valore SUE di 1 implica una stretta somiglianza tra le caratteristiche rilevanti identificate nella heatmap baseline e quelle nelle heatmap MCD, mentre una dissimilarità nelle regioni altamente rilevanti delle heatmap porta a un valore SUE che si avvicina a zero. L'analisi porta a risultati efficacemente utilizzabili se i valori SUE sono più alti per i CC rispetto ai MC, poiché permette di stabilire una correlazione positiva tra i valori di SUE e l'accuratezza della classificazione. Valori SUE più alti corrispondono a una classificazione più accurata, mentre valori SUE più bassi indicano un numero maggiore di classificazioni errate.
- Un valore di SUE vicino a 1 indica alta sovrapposizione e quindi alta ripetibilità delle heatmap.

Analisi delle Distribuzioni di SUE e Correlazione

L'analisi delle heatmap è stata eseguita seguendo questi passaggi:

1. Stima delle heatmap baseline con GradCAM, GradCAM++ e ScoreCAM per validation set.
2. Generazione delle heatmap con Monte Carlo Dropout (MCD) con GradCAM, GradCAM++ e ScoreCAM con 10 iterazioni per ogni input per validation set.
3. Calcolo di SUE medio per ogni campione tra la heatmap baseline e le 10 heatmap con MCD.
4. Calcolo della correlazione di Pearson per ogni campione tra la heatmap baseline e la media delle 10 heatmap con MCD.
5. Visualizzazione dei risultati tramite boxplot per CC e MC e per ciascun metodo XAI per velidation set.

Si sono esaminati i valori di SUE e correlazione separatamente per corretti classificati (CC) e misclassificati (MC), valutando le differenze tra i metodi GradCAM, GradCAM++ e ScoreCAM.

Boxplot di SUE e Correlazione di Pearson tra Heatmap Baseline e MCD

Heatmap stimate con **GradCAM**

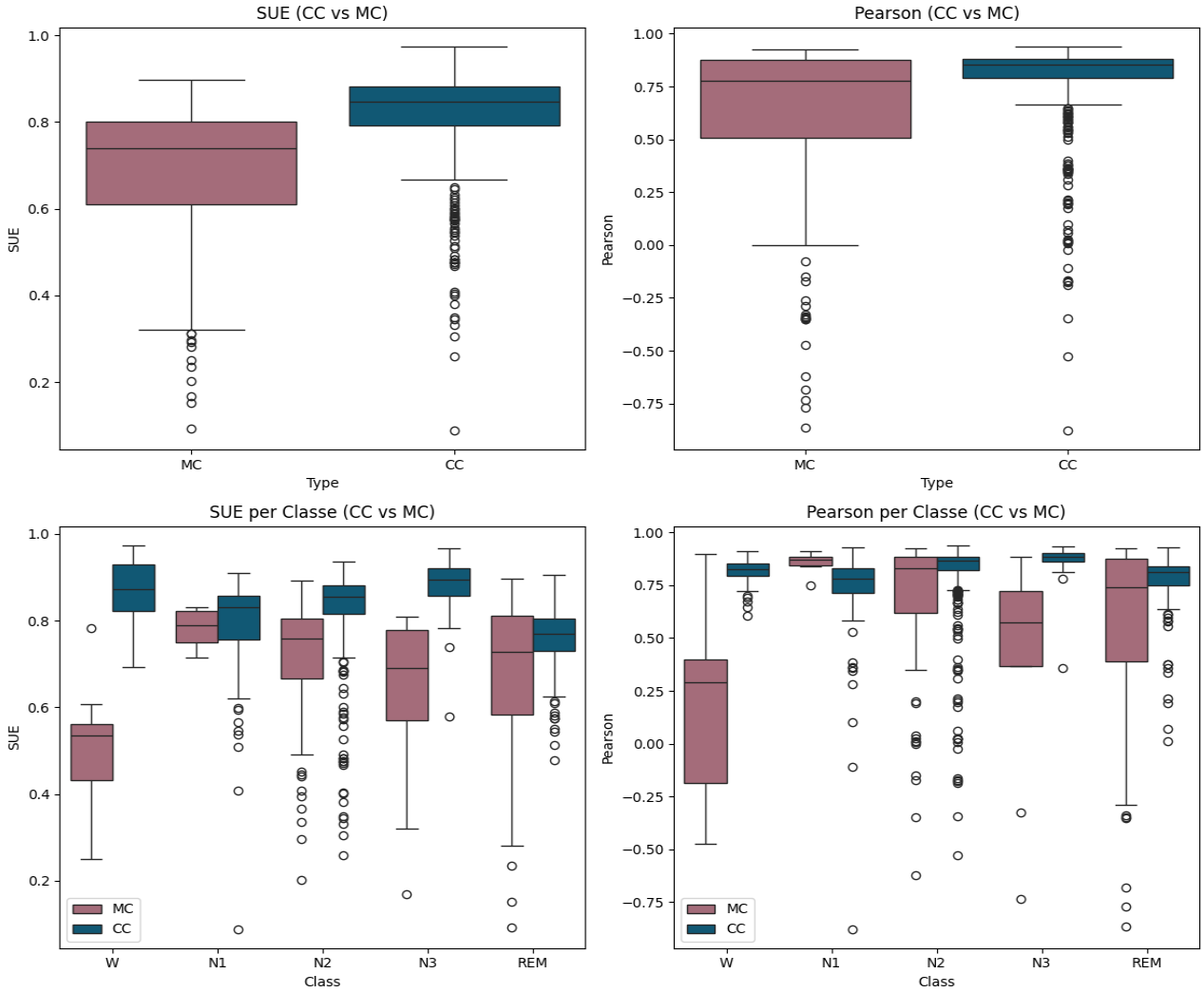

Figura 18: Boxplot di SUE e Correlazione di Pearson tra Heatmap stimate con GradCAM Baseline e MCD

Heatmap stimate con **GradCAM++**

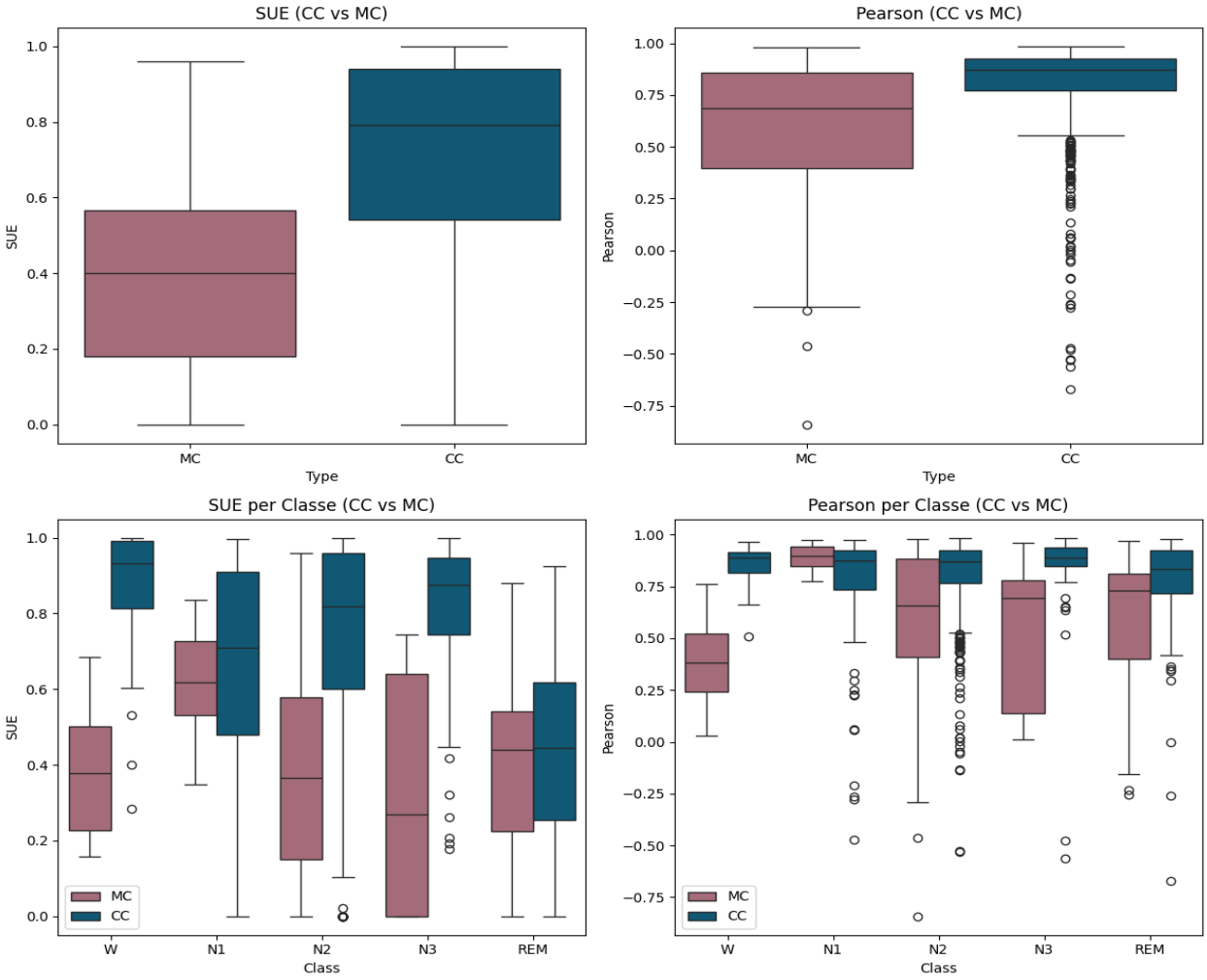

Figura 19: Boxplot di SUE e Correlazione di Pearson tra Heatmap stimate con GradCAM++ Baseline e MCD

Heatmap stimate con **ScoreCAM**

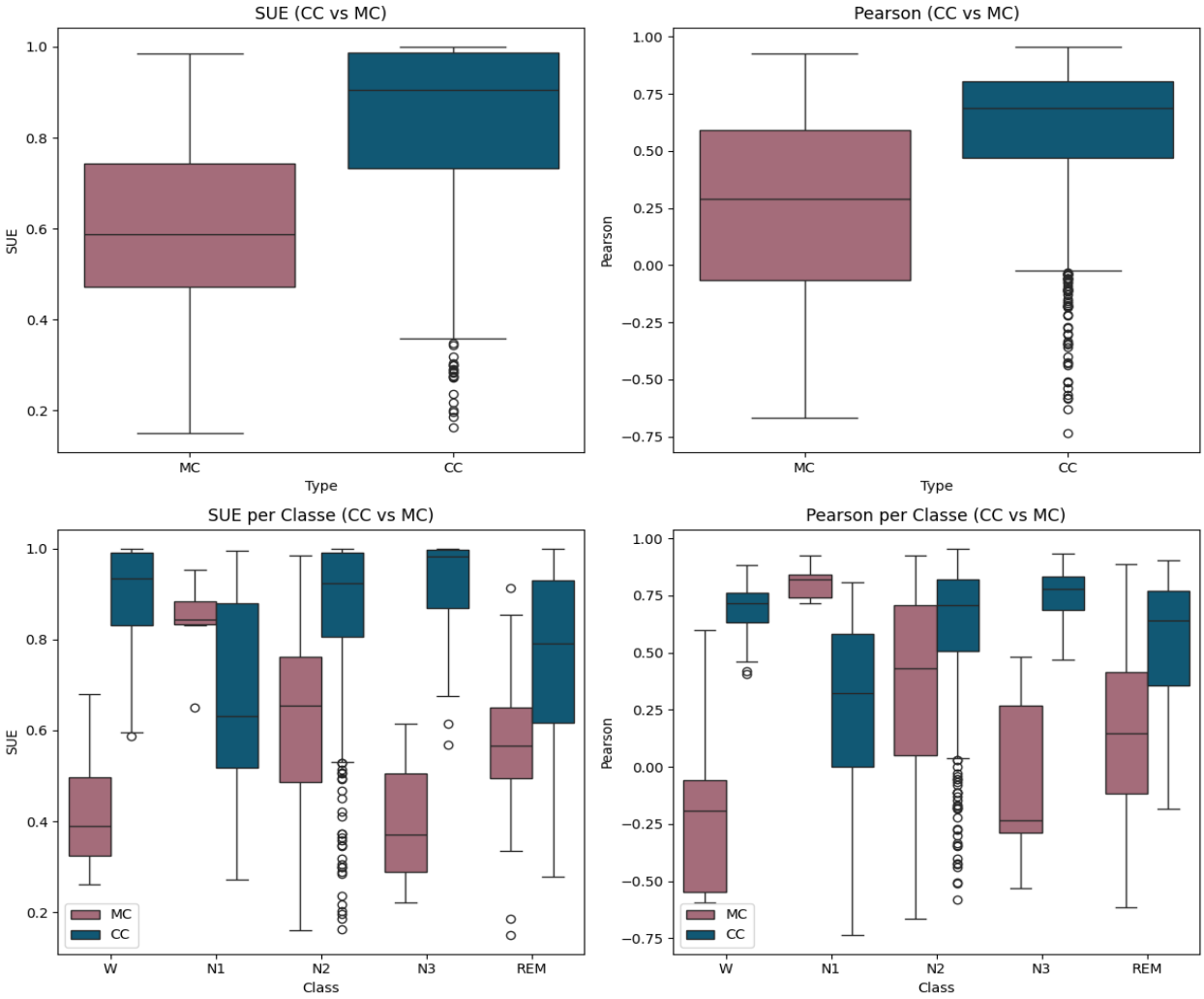

Figura 20: Boxplot di SUE e Correlazione di Pearson tra Heatmap stimate con ScoreCAM Baseline e MCD

Dai boxplot ottenuti, emerge che:

- In generale, **SUE** e correlazione sono **più alti per i CC rispetto agli MC**, indicando che nei campioni correttamente classificati le heatmap baseline e con MCD sono più simili e c'è una più elevata ripetibilità della localizzazione delle caratteristiche salienti per la classificazione.
- **ScoreCAM** si è dimostrato il **metodo più efficace** nella distinzione tra CC e MC, mostrando valori SUE più elevati nei CC e una maggiore separazione delle distribuzioni rispetto a GradCAM e GradCAM++.
- Si può osservare come la difficoltà del modello nel classificare la fase del sonno N1 venga mostrata con una vicinanza più marcata dei boxplot di SUE e correlazione relativi a CC e MC, se non addirittura con una loro sovrapposizione completa o inversione del comportamento rispetto a quello atteso.

Si è inoltre esaminata la **distribuzione di SUE e della correlazione tra le heatmap con MCD e baseline** per il metodo XAI ScoreCAM, rappresentando i risultati tramite curve di densità per CC e MC, con l'intento di indagare la possibilità di stabilire una soglia che permetta di migliorare le performance globali del metodo con l'affiancamento dei concetti di XAI e UQ.

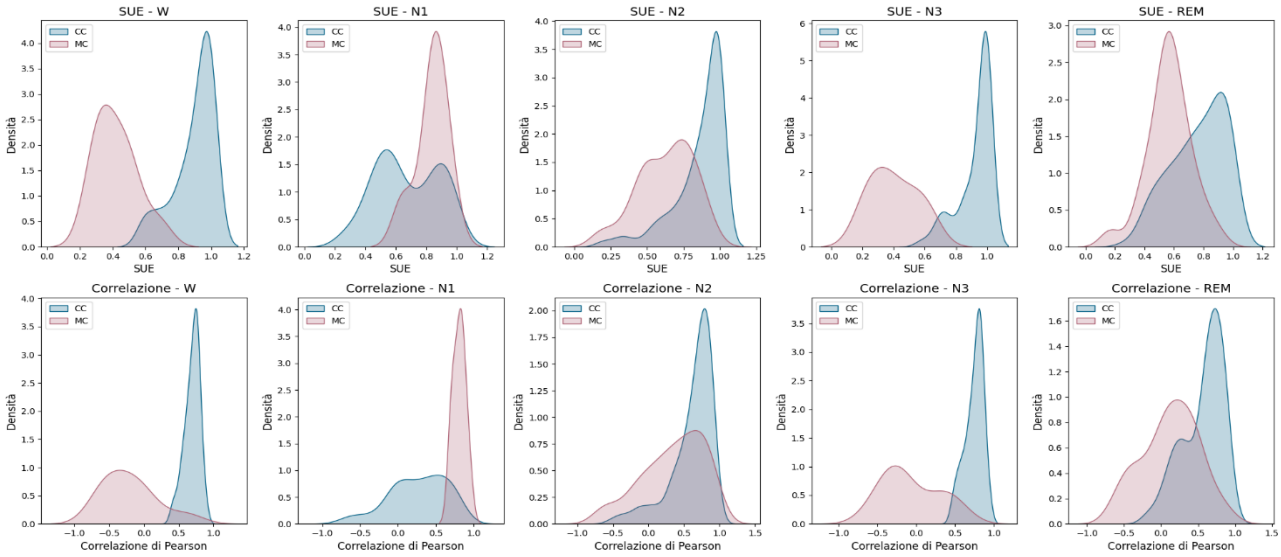

Figura 21: Distribuzioni di SUE e Correlazione di Pearson tra heatmap BL e con MCD per ScoreCAM sul Validation Set

Le distribuzioni mostrano un comportamento diverso delle due categorie, soprattutto nel caso del SUE. Tuttavia, la **sovrapposizione** tra le distribuzioni di CC e MC rende difficile estrarre una informazione quantitativa al fine di stabilire una soglia netta per discriminare in modo efficace CC e MC.

Questa analisi ha fornito informazioni sulla robustezza e sulla stabilità delle heatmap generate con diversi metodi XAI. I risultati indicano che ScoreCAM offre una maggiore capacità di discriminazione tra CC e MC rispetto a GradCAM e GradCAM++. Tuttavia, l'elevata sovrapposizione nelle distribuzioni suggerisce che, almeno in questa configurazione, le metriche SUE e correlazione delle heatmap BL e MCD per discriminare CC e MC **non è sufficiente** per migliorare direttamente le performance del modello.

3.5.3 Analisi XAI tramite CO Score

Si può quantificare la qualità dei metodi di Intelligenza Artificiale Spiegabile (XAI) basati sulla stima di heatmap, valutandone l'efficacia nel migliorare la probabilità di predizione delle classi corrette. Per svincolarsi dal concetto di localizzazione come metrica per quantificare le prestazioni di XAI, dal momento che esiste una differenza tra rilevanza computazionale e rilevanza umana (ciò che gli algoritmi considerano saliente potrebbe non essere significativo per un osservatore umano), si introduce il concetto di **Spiegazione Aumentativa AX**. Questo metodo combina l'input con la relativa heatmap stimata con un metodo XAI per ottenere una probabilità più alta di predire la classe corretta. In questo modo le heatmap non si limitano a indicare le aree di attenzione della rete, ma possono anche essere strumenti utili per migliorare la fiducia predittiva.

Per valutare quantitativamente l'efficacia del processo AX, viene introdotto il **CO (Confidance Optimization) Score** [38]. Il CO Score misura la differenza ponderata tra i valori di output grezzi prima e dopo l'applicazione delle mappe di calore. Un punteggio positivo indica un aumento della probabilità di classificazione corretta, mentre un punteggio negativo ne indica una diminuzione. Un CO Score pari a zero indica che la heatmap non ha aggiunto informazioni utili.

CO Score misura quanto una heatmap migliora la probabilità della classe corretta quando viene utilizzata per modificare l'input originale, fornendo un'indicazione quantitativa della qualità del metodo di Explainable AI (XAI).

L’analisi si articola in tre fasi principali:

1. **Determinazione del metodo XAI migliore tra (GradCAM, GradCAM++ e ScoreCAM)**: il CO Score viene calcolato utilizzando i target reali per valutare quale metodo XAI discrimina meglio tra campioni correttamente classificati (CC) e misclassificati (MC).
2. **Determinazione della soglia di CO Score per migliorare le performance del modello**: in questa fase, il CO Score viene ricalcolato utilizzando le predizioni del modello invece dei target reali, per identificare una soglia di confidenza che permetta di rimuovere le predizioni non affidabili. Questa scelta è determinata dalle condizioni reali di applicazione in ambito clinico di modelli non supervisionati, cioè in mancanza dell’informazione della classe target.
3. **Verifica della soglia sul test set** per valutare la generalizzabilità della strategia.

Definizione e Calcolo del CO Score

Il **CO Score** è una metrica che misura il cambiamento nella probabilità predetta dal modello quando il segnale originale viene combinato con la heatmap generata da un metodo XAI. Il principio di base è che, se la heatmap fornisce informazioni utili, allora la probabilità della classe corretta dovrebbe aumentare dopo la combinazione della heatmap con l’input del modello. In questo lavoro si è deciso di effettuare questa combinazione tramite moltiplicazione tra heatmap e segnale in ingresso, cioè pensando maggiormente le parti di segnale evidenziate come più influenti dal metodo XAI nella determinazione della classe da parte del modello. Il CO Score viene calcolato come la differenza tra la probabilità della classe di interesse prima e dopo l’applicazione della heatmap al segnale di input.

La formula generale del CO Score è:

$$CO_{Score}(x, h) = k * [P(x * h) - P(x)]$$

Dove:

- x è il segnale originale (epoca di 30 s di segnale EEG).
- h è la heatmap generata con un metodo XAI.
- x∗h è il segnale modificato, ottenuto moltiplicando segnale e heatmap.
- P(x) è la softmax delle predizioni originali del modello.
- P(x∗h) è la softmax delle predizioni ottenute con il segnale pesato con l'heatmap.
- k è un vettore di pesi definito in due modi diversi a seconda della fase dell'analisi; nella prima fase di selezione del miglior metodo XAI la definizione di k è basata sui target reali, nella seconda fase di determinazione della soglia ottimale di CO Score la definizione di k è basata sulle predizioni del modello.
    - $k_j = 1$ se j è la classe corretta (per la prima fase di analisi) o la classe predetta (per la seconda fase).
    - $k_i = -\frac{1}{C-1}$ per tutte le altre classi (C=5, numero totale di classi).

Un punteggio positivo indica un aumento della probabilità di classificazione corretta, mentre un punteggio negativo ne indica una diminuzione. Un CO Score pari a zero indica che la heatmap non ha aggiunto informazioni utili. CO Score più elevati corrispondono a miglioramenti nella probabilità predittiva. Evidenziando differenze di CO Score tra dati classificati correttamente e quelli classificati erroneamente, il CO Score può fungere da indicatore di correttezza predittiva.

3.5.3.1 Determinazione del Metodo XAI Migliore

Per identificare quale metodo XAI è più efficace, si calcola il CO Score utilizzando i **target reali**. Il processo seguito è il seguente:

1. **Predizione standard**: si applica il modello alle epoche di segnale x del validation set per ottenere la softmax originale P(x).
2. **Generazione heatmap**: si calcola la heatmap h con GradCAM, GradCAM++ e ScoreCAM.
3. **Modifica del segnale**: il segnale x viene moltiplicato per la heatmap h per enfatizzare le regioni ritenute più rilevanti.
4. **Nuova predizione**: si applica il modello al segnale modificato $x*h$, ottenendo la softmax $P(x*h)$.
5. **Calcolo del CO Score**: si calcola la differenza $P(x*h)-P(x)$ e la si moltiplica per k, basato sui target reali, per ogni metodo XAI.
6. Analisi della **distribuzione del CO Score** per **corretti classificati (CC)** e **misclassificati (MC) per ogni metodo XAI**: si osserva se il CO Score permette di distinguere i due gruppi. Se un metodo XAI è efficace, il CO Score deve essere significativamente più alto nei CC rispetto ai MC.

Di seguito si osservano le distribuzioni del CO Score calcolato usando le Heatmap stimate con GradCAM, GradCAM++ e ScoreCAM.

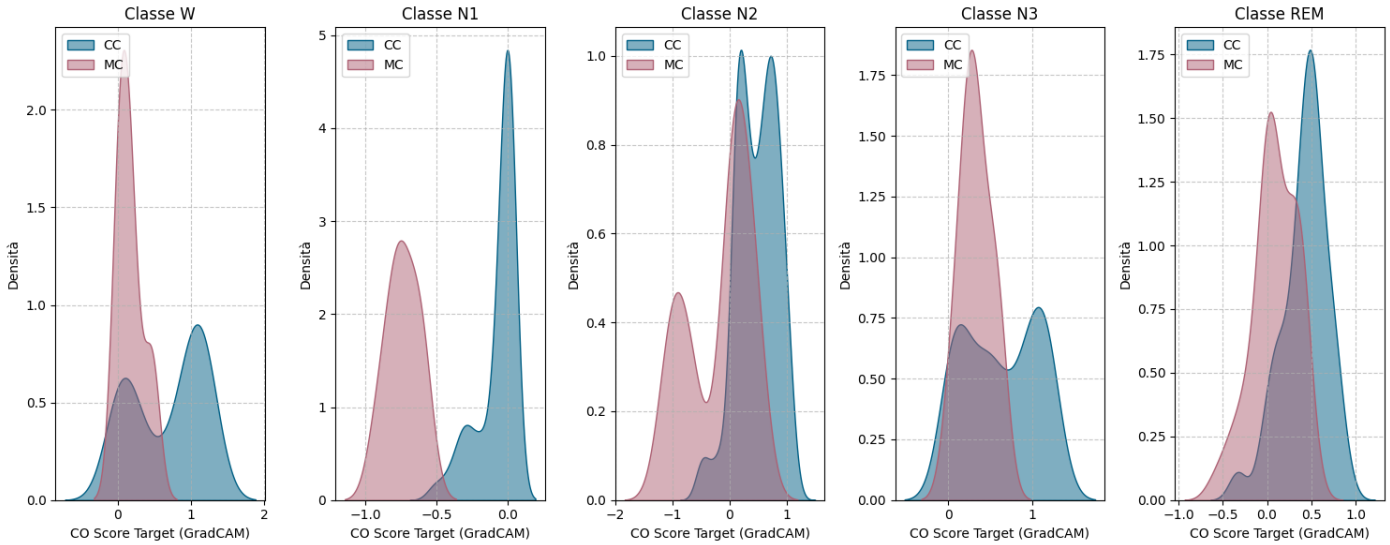

Figura 22: Distribuzione COScore calcolato sui target - Heatmap stimate con GradCAM – Validation Set

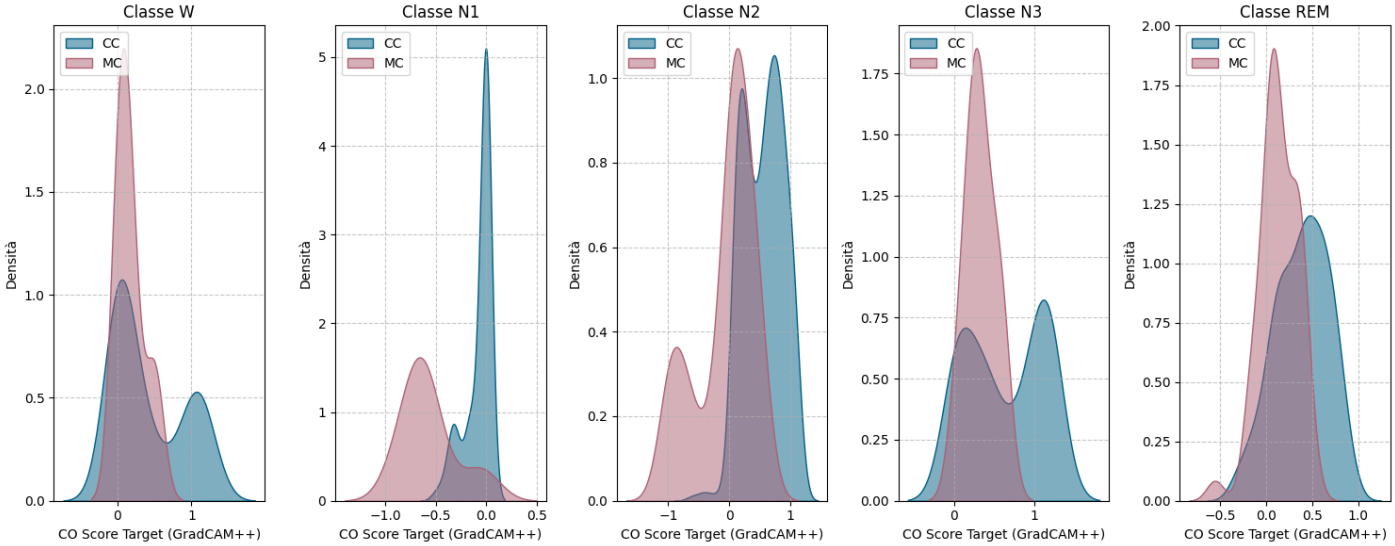

Figura 23: Distribuzione COScore calcolato sui target - Heatmap stimate con GradCAM++ – Validation Set

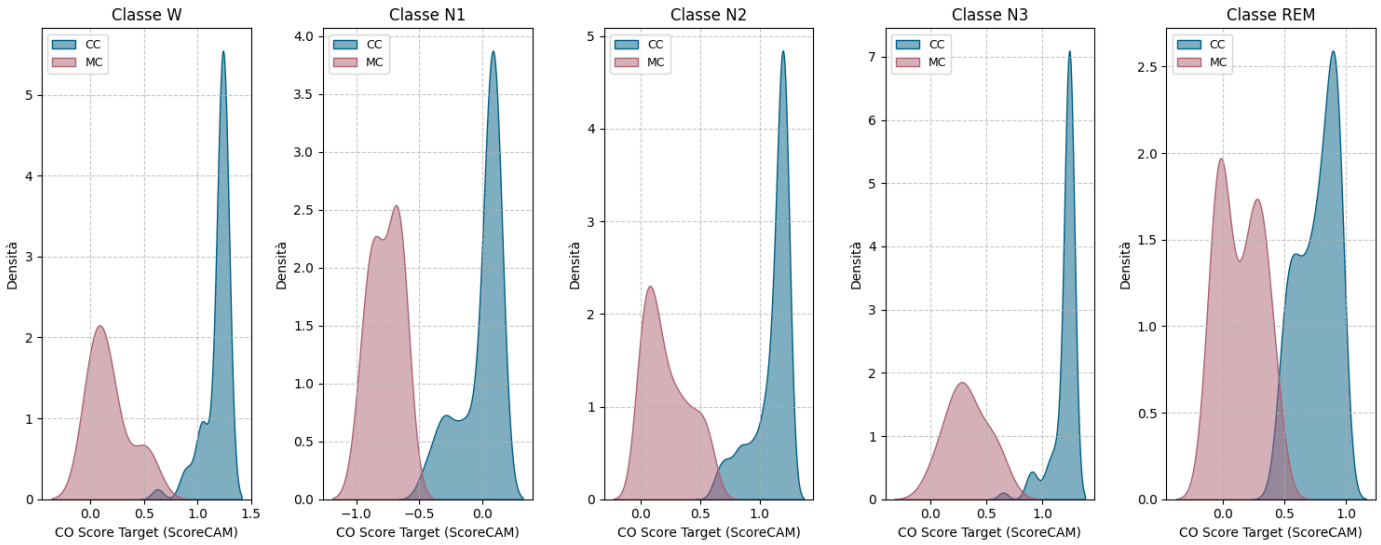

Figura 24: Distribuzione COScore calcolato sui target - Heatmap stimate con ScoreCAM – Validation Set

L'analisi delle distribuzioni dei CO Score per CC e MC per i 3 metodi XAI utilizzati mostra che **ScoreCAM è il metodo più efficace**, in quanto genera heatmap che portano a un incremento della probabilità della classe corretta molto più evidente nei CC rispetto agli MC. Questo supporta l'utilizzo delle heatmap ScoreCAM come strumenti efficaci non solo per l'interpretabilità, ma anche per il miglioramento delle prestazioni del modello.

3.5.3.2 Determinazione della Soglia del CO Score sul Validation Set

Una volta selezionato ScoreCAM come miglior metodo XAI, si utilizza il CO Score per migliorare la qualità delle predizioni rimuovendo i campioni con predizioni inaffidabili.

In questa fase, invece dei **target reali**, si utilizzano le **predizioni baseline** del modello per la definizione dei pesi nella combinazione tra segnale in ingresso e heatmap stimata con ScoreCAM.

I passaggi seguiti in questa fase dell'analisi sono i seguenti:

1. **Predizione standard**: si applica il modello alle epoche di segnale x del validation set, ottenendo la softmax P(x).
2. **Generazione heatmap**: si calcola la heatmap h utilizzando ScoreCAM.
3. **Modifica del segnale**: il segnale x viene moltiplicato per la heatmap h.
4. **Nuova predizione**: si calcola la softmax P(x∗h).
5. **Calcolo del CO Score**: si usa la formula con il vettore k costruito sulla base della classe predetta dal modello.
6. **Analisi della distribuzione e determinazione della soglia**: si osservano le distribuzioni del CO Score per CC e MC per determinare una soglia ottimale di discriminazione. **La determinazione della soglia è stata effettuata tramite criterio visivo:**

- **W, N2, N3:** soglia 1.0
- **REM, N1: soglia** 0.3

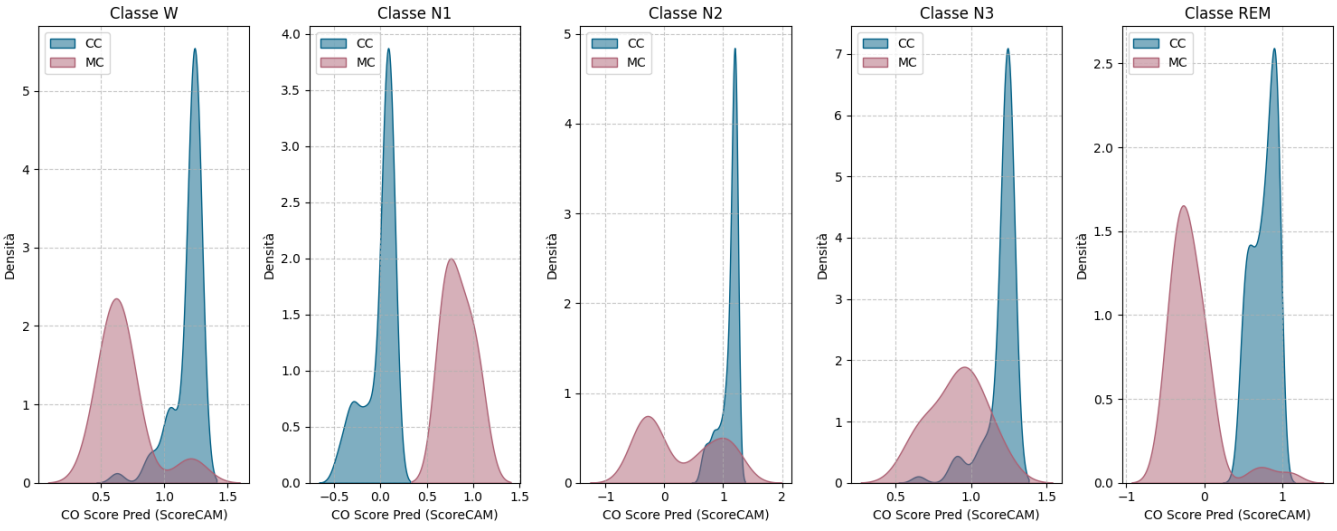

Figura 25: Distribuzione COScore calcolato sulle predizioni- Heatmap stimate con ScoreCAM – Validation Set

Evidenziando differenze di CO Score tra dati classificati correttamente e quelli classificati erroneamente, il CO Score può fungere da indicatore di correttezza predittiva e la ScoreCAM può essere utilizzata come strumenti efficaci non solo per l'interpretabilità, ma anche per il miglioramento delle prestazioni del modello.

Applicazione della soglia e miglioramento delle performance

Dopo aver definito le soglie, i campioni del validation set con CO Score inferiore alla soglia vengono scartati. Le confusion matrix e le metriche di performance (precision, recall, F1-score) sono state ricalcolate prima e dopo la rimozione dei campioni.

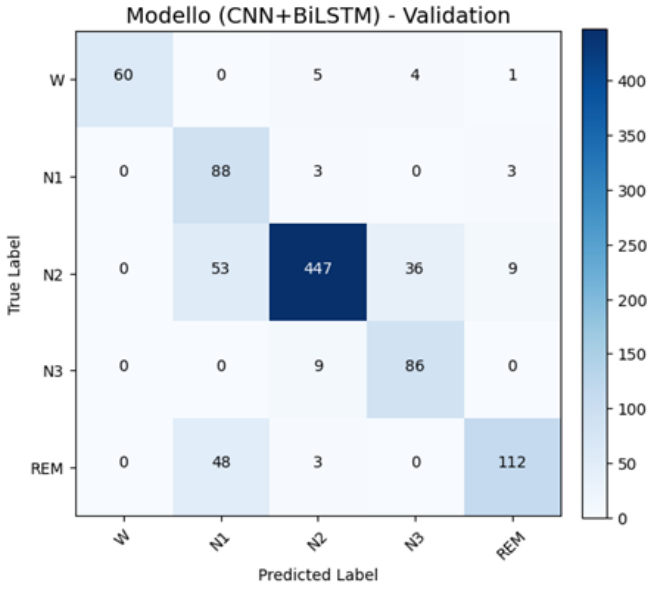

Figura 27: Confusion Matrix - CNN+BiLSTM: prima della rimozione dei campioni sopra soglia - Validation Set

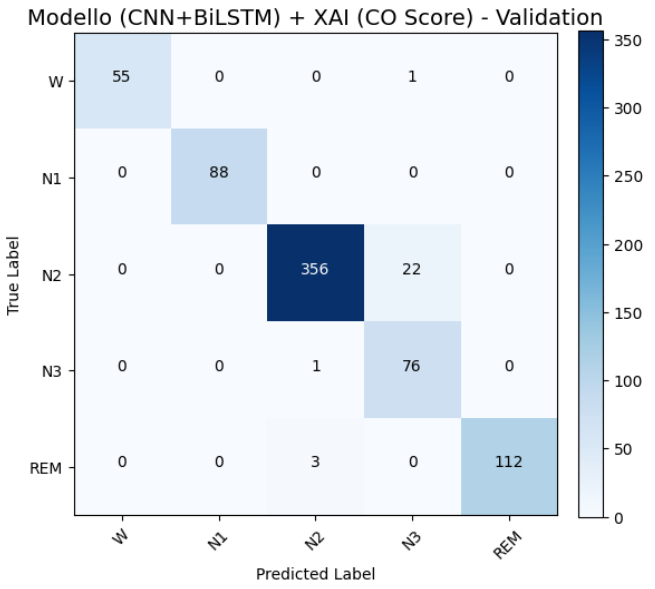

Figura 26: Confusion Matrix - CNN+BiLSTM + XAI (CO Score): dopo la rimozione dei campioni non affidabili per applicazione della soglia sul CO Score - Validation Set

Classe	*Prec Pre*	*Prec Post*	*Recall Pre*	*Recall Post*	*F1-Score Pre*	*F1-Score Post*
W	1.00	1.00	0.86	0.98	0.92	0.99
N1	0.47	1.00	0.94	1.00	0.62	1.00
N2	0.96	0.99	0.82	0.94	0.88	0.96
N3	0.68	0.77	0.91	0.99	0.78	0.86
REM	0.90	1.00	0.69	0.97	0.78	0.99

Tabella 15: Metriche Valutazione (Precision, Recall e F1-Score) Pre e Post rimozione campioni incerti tramite applicazione soglia sul CO Score - Validation Set

Classe	*Aumento % Precision*	*Aumento % Recall*	*Aumento % F1-Score*
W	0.00	14.58	7.36
N1	114.77	6.82	60.80
N2	3.31	14.83	9.21
N3	12.47	9.03	10.97
REM	11.61	41.74	26.87

Tabella 16: Aumento percentuale (%) Precision, Recall e F1-Score Pre e Post rimozione campioni incerti tramite applicazione soglia sul CO Score - Validation Set

- **Miglioramento delle performance**: si osserva un incremento della precisione e un miglioramento generale delle metriche, senza una perdita eccessiva di campioni correttamente classificati.
- **Confronto con UQ**: rispetto alla sogliatura basata sull'entropia normalizzata (UQ), il CO Score ha il vantaggio di rimuovere selettivamente solo i MC, preservando un maggior numero di CC.

3.5.3.3 Verifica della soglia sul Test Set

Per valutare la generalizzabilità della strategia, la soglia determinata sul validation set è stata applicata al test set.

1. **Applicazione della soglia CO Score** stabilita sul validation set al test set.
2. Ricalcolo delle **confusion matrix** e delle **metriche di performance**.

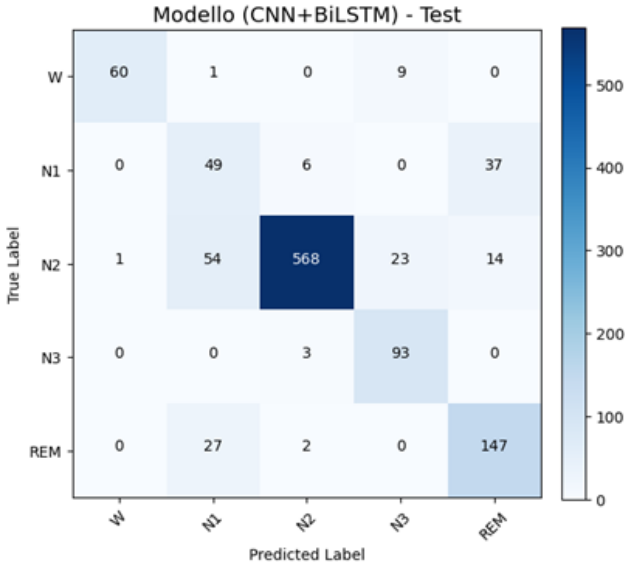

Figura 29: Confusion Matrix - CNN+BiLSTM: prima della rimozione dei campioni sopra soglia - Test Set

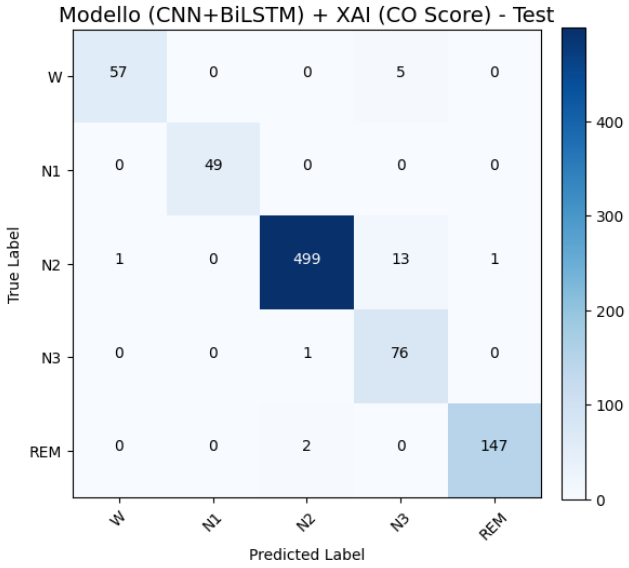

Figura 28: Confusion Matrix - CNN+BiLSTM + XAI (CO Score): dopo la rimozione dei campioni non affidabili per applicazione della soglia sul CO Score - Test Set

Classe	*Prec Pre*	*Prec Post*	*Recall Pre*	*Recall Post*	*F1-Score Pre*	*F1-Score Post*
W	0.98	0.98	0.86	0.92	0.92	0.95
N1	0.37	1.00	0.53	1.00	0.44	1.00
N2	0.98	0.99	0.86	0.97	0.92	0.98
N3	0.74	0.81	0.97	0.99	0.84	0.89
REM	0.74	0.99	0.84	0.99	0.79	0.99

Tabella 17: Metriche Valutazione (Precision, Recall e F1-Score) Pre e Post rimozione campioni incerti tramite applicazione soglia sul CO Score - Test Set

Classe	*Aumento % Precision*	*Aumento % Recall*	*Aumento % F1-Score*
W	-0.09	7.26	3.71
N1	167.35	87.76	127.55
N2	1.33	12.81	7.13
N3	8.67	1.89	5.62
REM	33.78	18.12	25.93

Tabella 18: Aumento percentuale (%) Precision, Recall e F1-Score Pre e Post rimozione campioni incerti tramite applicazione soglia sul CO Score - Test Set

3. Analisi della **numerosità delle classi** post-rimozione.

Classe	Validation Set			Test Set		
	Totale	Affidabili	Non affidabili (rimossi)	Totale	Affidabili	Non affidabili (rimossi)
W	70	56	14	70	62	8
N1	94	88	6	92	49	43
N2	545	378	167	660	514	146
N3	95	77	18	96	77	19
REM	163	115	48	176	149	27

Tabella 19: Numero di campioni totali del set di dati, affidabili (sottosoglia di incertezza) e non affidabili (rimossi, sopra soglia di incertezza) per ogni classe – Validation e Test Set

Si osserva come con l'approccio XAI ScoreCAM e l'applicazione della soglia su CO Score si ha una rimozione ridotta di campioni appartenenti alle fasi N1 e REM che sono le fasi più problematiche e soprattutto quelle in cui vengono rimossi più campioni con l'approccio di filtraggio sulla base della soglia definita con la quantificazione dell'incertezza UQ.

Fase del sonno	UQ		XAI	
	Val (%)	Test (%)	Val (%)	Test (%)
W	14.29	15.71	20.00	11.43
N1	17.02	51.09	6.38	46.74
N2	26.06	20.76	30.64	22.12
N3	13.68	8.33	18.95	19.79
REM	57.06	42.61	29.45	15.34

Tabella 20: Percentuale campioni rimossi con l'applicazione degli approcci UQ e XAI

I risultati confermano che ScoreCAM + CO Score è una strategia efficace per migliorare le performance senza ridurre eccessivamente il numero di campioni disponibili. Anche sul test set il confronto con UQ porta ad osservare che, rispetto alla sogliatura basata sull'entropia normalizzata (UQ), il CO Score ha il vantaggio di rimuovere principalmente i MC, preservando un maggior numero di CC.

L'analisi tramite CO Score ha dimostrato che:

1. Il CO Score è una valida metrica empirica per valutare il contenuto informativo delle heatmap e per determinare quantitativamente l'incremento della probabilità di predizione delle classi corrette di un certo metodo XAI.
2. ScoreCAM è il miglior metodo XAI per la nostra applicazione, mostrando la maggiore capacità di distinguere CC e MC.
3. L'uso del CO Score permette di filtrare predizioni incerte in modo efficace, migliorando la precisione del modello rimuovendo principalmente gli errori e preservando i CC.
4. Il metodo è più robusto rispetto alla quantificazione dell'incertezza basata sull'entropia, poiché preserva meglio le classi meno rappresentate come N1, che è anche la più difficile da classificare.

L'applicazione di ScoreCAM con CO Score rappresenta quindi un approccio innovativo per migliorare l'affidabilità delle predizioni, le prestazioni e la trasparenza del modello.

4. Risultati

4.1 Presentazione dei Risultati

L'obiettivo principale di questa tesi è esplorare il ruolo dell'Explainable Artificial Intelligence (XAI) e della quantificazione dell'incertezza (Uncertainty Quantification, UQ) per migliorare l'affidabilità e l'interpretabilità dei modelli di deep learning applicati alla classificazione delle fasi del sonno. Per valutare l'efficacia dell'approccio proposto, sono state analizzate le performance del modello sul test set, attraverso l'uso di metriche di valutazione, come accuracy, precision, recall, F1-score e confusion matrix. La validazione dei metodi XAI e UQ è stata condotta per determinare l'impatto di ciascuna tecnica sulle prestazioni del modello.

4.2 Quantificazione dell'Incertezza (UQ)

L'uso del Monte Carlo Dropout (MCD) ha permesso di stimare l'incertezza associata alle predizioni del modello. L'entropia normalizzata è stata utilizzata per quantificare la dispersione delle probabilità predittive e identificare i campioni con alta incertezza. L'analisi dei boxplot (*Fig. 11*) ha evidenziato che i campioni misclassificati (MC) tendono ad avere valori di entropia più elevati rispetto ai correttamente classificati (CC), rendendo possibile la determinazione di una soglia di incertezza per la rimozione dei campioni meno affidabili.

Di seguito si riportano le metriche di valutazione, i loro incrementi percentuali e le confusion matrix relativi al test set prima e dopo l’applicazione della soglia di incertezza e la rimozione dei campioni non affidabili.

Classe	*Prec Pre*	*Prec Post*	*Recall Pre*	*Recall Post*	*F1-Score Pre*	*F1-Score Post*
W	0.98	0.98	0.86	0.98	0.92	0.98
N1	0.37	0.62	0.53	0.73	0.44	0.67
N2	0.98	0.99	0.86	0.95	0.92	0.97
N3	0.74	0.85	0.97	0.99	0.84	0.92
REM	0.74	0.91	0.84	0.91	0.79	0.91

Tabella 21: Risultati. Metriche Valutazione (Precision, Recall e F1-Score) Pre e Post rimozione campioni incerti tramite applicazione soglia di incertezza - Test Set

Classe	*Aumento % Precision*	*Aumento % Recall*	*Aumento % F1-Score*
W	-0.06	14.69	7.32
N1	66.46	37.69	53.25
N2	0.72	9.98	5.45
N3	14.64	2.05	8.81
REM	22.69	9.06	15.88

Tabella 22: Risultati. Aumento percentuale (%) Precision, Recall e F1-Score Pre e Post rimozione campioni incerti tramite applicazione soglia di incertezza - Test Set

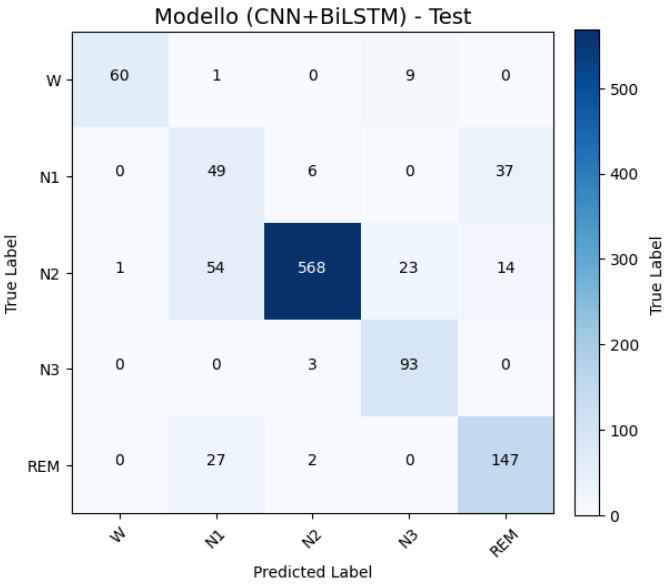

Figura 31: Risultati. Confusion Matrix - CNN+BiLSTM: prima della rimozione dei campioni non affidabili - Test Set

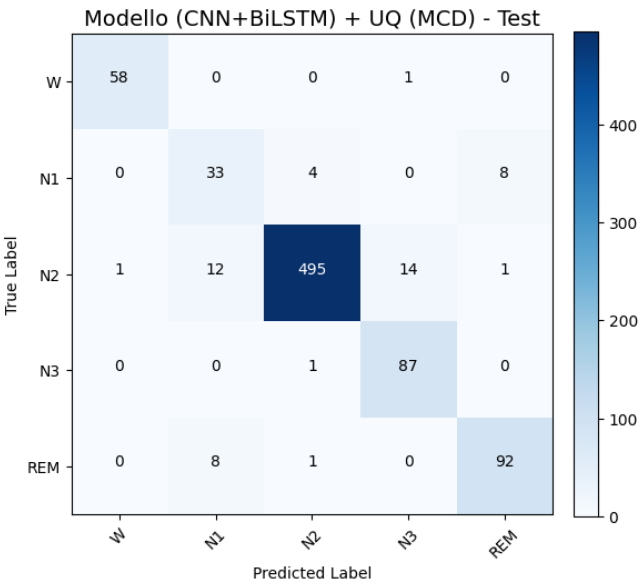

Figura 30: Risultati. Confusion Matrix - CNN+BiLSTM + UQ (MCD): dopo la rimozione dei campioni non affidabili per applicazione della soglia di incertezza- Test Set

Dopo l'applicazione della soglia di incertezza:

- Miglioramento di precision, recall e F1-score, in particolare per la fase N1 (fase la cui classificazione in letteratura risulta più problematica)
- Riduzione delle predizioni errate. Il confronto tra le confusion matrix prima e dopo la rimozione mostra una riduzione degli errori di classificazione, con un impatto positivo sulle performance del metodo.

L'applicazione della soglia di incertezza ha portato a un **miglioramento dell'accuratezza** complessiva del modello, riducendo l'impatto dei campioni misclassificati con predizioni instabili e migliorando la qualità delle classificazioni.

4.3 Explainable Artificial Intelligence (XAI)

Sono stati utilizzati tre metodi di explainability per generare heatmap e interpretare le decisioni del modello: GradCAM, GradCAM++ e ScoreCAM.

Il **CO Score** ha permesso di misurare il cambiamento nella probabilità predetta dal modello quando il segnale originale viene combinato con la heatmap generata da un metodo XAI.

Il CO Score è stato utilizzato per determinare quale metodo XAI generasse heatmap più efficaci per migliorare le predizioni del modello. I risultati mostrano che **ScoreCAM** ha ottenuto distribuzioni di CO Score con una migliore discriminatività tra CC e MC (*Fig. 24*).

Evidenziando differenze di CO Score tra dati classificati correttamente e quelli classificati erroneamente, come mostrato dall'analisi delle distribuzioni del CO Score per i diversi metodi XAI, esso può fungere da indicatore di correttezza predittiva ed è stato quindi utilizzato per definire una soglia e filtrare predizioni poco affidabili, migliorando ulteriormente le performance del modello.

Le confusion matrix e le metriche di performance (precision, recall, F1-score) sono state ricalcolate prima e dopo la rimozione dei campioni sul test set.

Classe	*Prec Pre*	*Prec Post*	*Recall Pre*	*Recall Post*	*F1-Score Pre*	*F1-Score Post*
W	0.98	0.98	0.86	0.92	0.92	0.95
N1	0.37	1.00	0.53	1.00	0.44	1.00
N2	0.98	0.99	0.86	0.97	0.92	0.98
N3	0.74	0.81	0.97	0.99	0.84	0.89
REM	0.74	0.99	0.84	0.99	0.79	0.99

Tabella 23: Risultati. Metriche Valutazione (Precision, Recall e F1-Score) Pre e Post rimozione campioni incerti tramite applicazione soglia sul CO Score - Test Set

Classe	*Aumento % Precision*	*Aumento % Recall*	*Aumento % F1-Score*
W	-0.09	7.26	3.71
N1	167.35	87.76	127.55
N2	1.33	12.81	7.13
N3	8.67	1.89	5.62
REM	33.78	18.12	25.93

Tabella 24: Risultati. Aumento percentuale (%) Precision, Recall e F1-Score Pre e Post rimozione campioni incerti tramite applicazione soglia sul CO Score - Test Set

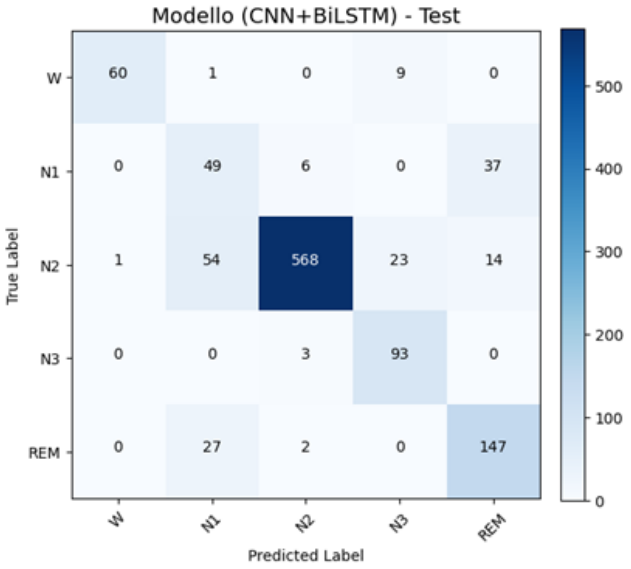

Figura 33: Risultati. Confusion Matrix - CNN+BiLSTM: prima della rimozione dei campioni sopra soglia - Test Set

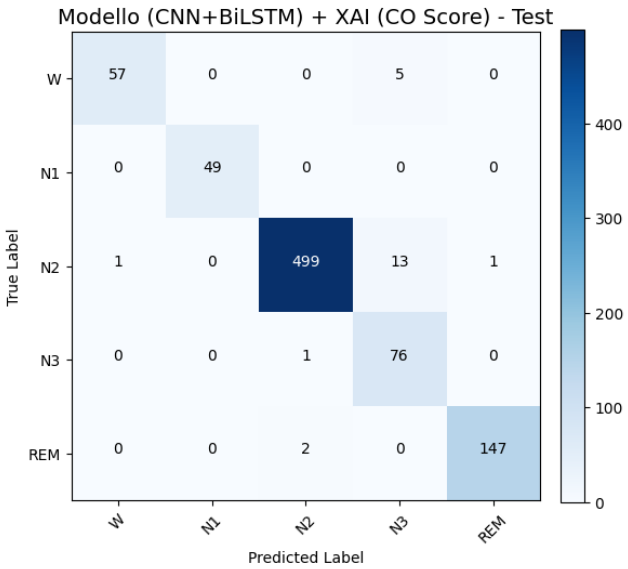

Figura 32: Confusion Matrix - CNN+BiLSTM + XAI (CO Score): dopo la rimozione dei campioni non affidabili per applicazione della soglia sul CO Score - Test Set

Si osserva un **miglioramento delle performance generale**, senza una perdita eccessiva di campioni correttamente classificati. L'uso del CO Score permette di filtrare predizioni incerte in modo efficace, migliorando la precisione del modello rimuovendo principalmente gli errori e preservando i CC.

L’analisi tramite CO Score ha dimostrato che:

- Il CO Score è una valida metrica empirica per valutare il contenuto informativo delle heatmap e per determinare quantitativamente l’incremento della probabilità di predizione delle classi corrette di un certo metodo XAI.
- ScoreCAM è il miglior metodo XAI per la nostra applicazione, mostrando la maggiore capacità di distinguere CC e MC.

4.4 Confronto tra UQ e XAI

L'analisi comparativa tra UQ e XAI ha evidenziato che entrambe le tecniche hanno portato a miglioramenti nelle performance, ma con approcci diversi:

- UQ (Monte Carlo Dropout + Entropia Normalizzata) ha permesso di filtrare le predizioni ad alta incertezza, migliorando l'affidabilità generale.
- XAI (ScoreCAM + CO Score) ha permesso di migliorare la comprensione del modello e di rimuovere solo le predizioni effettivamente errate, preservando più campioni.

Analisi sulla numerosità delle classi post-rimozione ha portato ad osservare come con l’approccio XAI ScoreCAM e l’applicazione della soglia su CO Score si ha una rimozione ridotta di campioni appartenenti alle fasi N1 e REM che sono le fasi più problematiche e soprattutto quelle in cui vengono rimossi più campioni con l’approccio di filtraggio sulla base della soglia definita con la quantificazione dell’incertezza UQ.

Fase del sonno	UQ		XAI	
	Val (%)	Test (%)	Val (%)	Test (%)
W	14.29	15.71	20.00	11.43
N1	17.02	51.09	6.38	46.74
N2	26.06	20.76	30.64	22.12
N3	13.68	8.33	18.95	19.79
REM	57.06	42.61	29.45	15.34

Figura 34: Risultati. Percentuale campioni rimossi con l'applicazione degli approcci UQ e XAI

I risultati confermano che ScoreCAM + CO Score è una strategia efficace per migliorare le performance senza ridurre eccessivamente il numero di campioni disponibili; il CO Score ha il vantaggio di rimuovere principalmente i MC, preservando un maggior numero di CC.

L'applicazione di ScoreCAM con CO Score rappresenta quindi un approccio innovativo per migliorare l'affidabilità delle predizioni, le prestazioni e la trasparenza del modello.

4.5 Confronto Stato dell'Arte – Metodi Proposti

Le performance sono state confrontate con lo stato dell'arte attraverso una tabella comparativa. Si è osservato che il modello CNN+BiLSTM con tecniche di UQ e XAI ha raggiunto prestazioni superiori rispetto a quelle disponibili in letteratura.

Metodo	Precision					Recall					F1-score					PCC					Acc Tot
	W	N1	N2	N3	REM	W	N1	N2	N3	REM	W	N1	N2	N3	REM	W	N1	N2	N3	REM	
1. [17]	0.81	0.55	0.78	0.76	0.71	0.86	0.38	0.86	0.65	0.69	0.83	0.45	0.82	0.70	0.70						0.75
2. [18]	0.86	0.45	0.88	0.88	0.77	0.89	0.26	0.89	0.89	0.82											
3. [45]	0.86	0.52	0.77	0.86	0.80	0.87	0.31	0.84	0.81	0.82	0.86	0.39	0.81	0.84	0.82	0.73	0.28	0.49	0.76	0.68	
4. [7]																0.95	0.61	0.89	0.92	0.84	0.87
Proposto	0.98	0.38	0.98	0.74	0.74	0.86	0.53	0.86	0.97	0.84	0.92	0.44	0.92	0.84	0.79	0.97	0.30	0.85	0.73	0.65	0.84
Proposto + UQ	0.98	0.62	0.99	0.85	0.91	0.98	0.73	0.95	0.99	0.91	0.98	0.67	0.97	0.92	0.91	0.98	0.71	0.96	0.99	0.92	0.94
Proposto + XAI	0.98	1.00	0.99	0.81	0.99	0.92	1.00	0.97	0.99	0.99	0.95	1.00	0.98	0.90	0.99	0.93	1.00	0.97	0.98	0.97	0.96

Tabella 25: Risultati. Confronto Perfromance Stato dell'Arte - Metodi Proposti (XAI e UQ)

Metodo	Modello	Dataset
1. [17]	CNN-RNN	DS-1 Charité DS-2 Umich
2. [18]	AT-BiLSTM	DRM-SUB PSEE
3. [45]	CNN+LSTM	Bruxism Sleep apnea
4. [7]	Cascade of 2 LSTM+RNN	Sleep-EDF-12 Physionet
Proposto	CNN+BiLSTM	Sleep-EDF-12 Physionet

Tabella 26: Specifiche Metodi Stato dell'Arte (Modelli di DL e Dataset)

5. Limiti e Lavori Futuri

Sintesi del Lavoro

In questa tesi è stato sviluppato un approccio innovativo per la classificazione delle fasi del sonno basato sull'integrazione di tecniche XAI e UQ in un modello di deep learning CNN+BiLSTM. I risultati hanno dimostrato che la quantificazione dell'incertezza migliora l'affidabilità delle predizioni, mentre le tecniche XAI forniscono strumenti per l'interpretabilità e per il miglioramento delle performance.

Punti di Forza

- Applicazione di tecniche XAI e UQ a segnali fisiologici, settore ancora scarsamente esplorato rispetto all'applicazione di queste tecniche ad immagini mediche.
- Miglioramento delle performance rispetto allo stato dell'arte, soprattutto della fase N1, grazie alla rimozione dei campioni ad alta incertezza.
- Applicazione e confronto tra diversi metodi XAI.
- Utilizzo del CO Score per migliorare la discriminazione tra CC e MC, con un approccio innovativo che ha permesso di migliorare la selezione dei campioni affidabili senza eliminare troppi dati.

Sviluppi Futuri

- Applicazione a diversi dataset EEG per validare la generalizzabilità del metodo.
- Integrazione in ambienti clinici per testare l'efficacia pratica del modello in scenari di applicazione reale.
- Sperimentazione di altri metodi di quantificazione dell'incertezza per migliorare la robustezza del modello.
- Applicazione del framework sviluppato ad altri task di classificazione EEG, come la rilevazione di disturbi neurologici.
- Integrazione con sistemi di supporto decisionale clinico per rendere l'approccio utilizzabile nella pratica medica.

6. Conclusioni

In questa tesi è stato sviluppato un approccio innovativo per la classificazione delle fasi del sonno basato sull'integrazione di tecniche XAI e UQ in un modello di deep learning CNN+BiLSTM. I risultati hanno dimostrato che la quantificazione dell'incertezza migliora l'affidabilità delle predizioni e le tecniche XAI forniscono strumenti per l'interpretabilità utili per il miglioramento delle performance del modello.

L'approccio proposto rappresenta un passo in avanti significativo verso l'uso di modelli di deep learning interpretabili e affidabili, con potenziali applicazioni in ambito clinico pratico.

7. Riferimenti bibliografici

[1] B. Goodman and S. Flaxman, “European Union Regulations on Algorithmic Decision Making and a “Right to Explanation”,” AI magazine, vol. 38, no. 3, pp. 50–57, 2017.

[2] Ronald R. Gilley, "The Role of Sleep in Cognitive Function," Sleep Medicine, 2022.

[3] Kryger, M. et al., "Principles and Practice of Sleep Medicine," Elsevier, 2016.

[4] M.J. Sateia, International classification of sleep disorders-third edition, Chest 146 (2014) 1387–1394, https://doi.org/10.1378/chest.14-0970.

[5] Wulff, K.; Gatti, S.; Wettstein, J.G.; Foster, R.G. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. Nat. Rev. Neurosci. 2010, 11, 589–599.

[6] A. Gevins, Non-invasive human neurocognitive performance capability testing method and system, U.S. Pat. (1994). https://patents.google.com/patent/ US5295491A/en (accessed February 24, 2018).

[7] Nicola Michielli, U. Rajendra Acharya, Filippo Molinari, “Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals”, Computers in Biology and Medicine, Volume 106, 2019, Pages 71-81, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2019.01.013.

[8] Rechtschaffen, A., & Kales, A., "A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects," U.S. Government Printing Office, 1968.

[9] Berry, R.; Brooks, R.; Gamaldo, C.; Harding, S.M.; Lloyd, R.M.; Quan, S.F.; Troester, M.T.; Vaughn, B.V. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, version 2.6.0; American Academy of Sleep Medicine: Darien, IL, USA, 2020.

[10] Iber, C. et al., "The AASM Manual for the Scoring of Sleep and Associated Events," American Academy of Sleep Medicine, 2007.

[11] Rechtschaffen, A. A Manual for Standardized Terminology, Techniques and Scoring System for Sleep Stages in Human Subjects; Brain Research Institute: Washington, DC, USA, 1968.

[12] Danker-Hopfe, H. et al., "Interrater Reliability for Sleep Scoring According to the Rechtschaffen & Kales and the AASM Criteria," Sleep, 2009.

[13] L. Fiorillo, A. Puiatti, M. Papandrea, P.L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, F.D. Faraci, Automated sleep scoring: a review of the latest approaches, Sleep Med. Rev. 48 (2019), 101204, https://doi.org/10.1016/j.smrv.2019.07.007.

[14] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, J.F. Payen, A convolutional neural network for sleep stage scoring from raw single-channel EEG, Biomed. Signal Process Control 42 (2018) 107–114.

[15] Biswal, S. et al., "SleepNet: Automated Sleep Staging System Using Deep Learning," IEEE Transactions on Biomedical Engineering, 2017.

[16] Li, C.; Qi, Y.; Ding, X.; Zhao, J.; Sang, T.; Lee, M. A Deep Learning Method Approach for Sleep Stage Classification with EEG Spectrogram. *Int. J. Environ. Res. Public Health* 2022, *19*, 6322. https://doi.org/10.3390/ijerph19106322

[17] ElMoaqet, H.; Eid, M.; Ryalat, M.; Penzel, T. A Deep Transfer Learning Framework for Sleep Stage Classification with Single-Channel EEG Signals. *Sensors* 2022, *22*, 8826. https://doi.org/10.3390/s22228826

[18] Fu M, Wang Y, Chen Z, Li J, Xu F, Liu X, Hou F. Deep Learning in Automatic Sleep Staging With a Single Channel Electroencephalography. Front Physiol. 2021 Mar 3;12:628502. doi: 10.3389/fphys.2021.628502. PMID: 33746774; PMCID: PMC7965953.

[19] Michielli N, Acharya UR, Molinari F. Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. Comput Biol Med. 2019 Mar;106:71-81. doi: 10.1016/j.compbiomed.2019.01.013. Epub 2019 Jan 19. PMID: 30685634.

[20] Goldberger, A.L.; Amaral, L.; Glass, L.; Hausdorff, J.M.; Ivanov, P.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; Stanley, H.E. The Sleep-EDF (Expanded) Database. 2000. Available online: https://physionet.org/content/sleep-edf/1.0.0/ (accessed on 12 December 2002).

[21] Goldberger, A.L.; Amaral, L.; Glass, J.; Hausdorff, J.M.; Ivanov, P.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; Stanley, H.E. The Sleep-EDF (Expanded) Database. 2013. Available online: https://physionet.org/content/sleep-edfx/1.0.0/ (accessed on 24 October 2013).

[22] Yen, An-Zi & Wu, Cheng-Kuang & Chen, Hsin-Hsi. (2023). Opportunities and challenges of explainable artificial intelligence in medicine. 10.1016/B978-0-323-99136-0.00009-X.

[23] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.

[24] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.

[25] R. Hamon, H. Junklewitz, I. Sanchez, Robustness and Explainability of Artificial Intelligence, Publ. Off. Eur. Union, 2020.

[26] Samek, W. & Müller, K.-R. Towards explainable artificial intelligence. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 5–22. (Springer, 2019).

[27] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, “Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022),” Computer Methods and Programs in Biomedicine, vol. 226. Elsevier Ireland Ltd, Nov. 01, 2022. doi: 10.1016/j. cmpb.2022.107161.

[28] S. Seoni, et al., Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023), Comput. Biol. Med. (2023), https://doi.org/10.1016/j.compbiomed.2023.107441.

[29] Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, 618–626 (2017).

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

[31] Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 839–847 (IEEE, 2018).

[32] Wang, H. et al. Score-cam: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 24–25 (2020).

[33] D. Alvarez-Melis and T. S. Jaakkola, “On the Robustness of Interpretability Methods,” arXiv preprint arXiv:1806.08049, 2018.

[34] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards Deep Learning Models Resistant to Adversarial Attacks,” in International Conference on Learning Representations, 2018

[35] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran and N. C. Bouaynaya, "Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks," in *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 73-84, July 2022, doi: 10.1109/MSP.2022.3142719.

[36] Shivam Kumar Sharma*, Suvadeep Maiti*, S. Mythirayee, P R Srijithesh and Raju Surampudi Bapi, Transparency in Sleep Staging: Deep Learning Method for EEG Sleep Stage Classification with Model Interpretability, IEEE Journal of Biomedical and Health Informatics, 2024

[37] Fernando Vaquerizo-Villar, Gonzalo C. Gutiérrez-Tobal, Eva Calvo, Daniel Álvarez, Leila Kheirandish-Gozal, Félix del Campo, David Gozal, Roberto Hornero, An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea, Computers in Biology and Medicine, Volume 165, 2023, 107419, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2023.107419.

[38] Tjoa, Erico, et al. Improving deep neural network classification confidence using heatmap-based eXplainable AI. *arXiv preprint arXiv:2201.00009*, 2021.

[39] E. Begoli, T. Bhattacharya, D. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, Nat. Mach. Intell. 1 (1) (2019) 20–23.

[40] Fiorillo L., Favaro P., Faraci F.D. DeepSleepNet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates IEEE Trans. Neural Syst. Rehabil. Eng., 29 (2021), pp. 2076-2085.

[41] Silvia Seoni, Filippo Molinari, U. Rajendra Acharya, Oh Shu Lih, Prabal Datta Barua, Salvador García, Massimo Salvi, Application of spatial uncertainty predictor in CNN-

BiLSTM model using coronary artery disease ECG signals, Information Sciences, Volume 665, 2024, 120383, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2024.120383.

[42] Craik, A., He, Y. & Contreras-Vidal, J. L. Deep learning for electroencephalogram (EEG) classification tasks: a review. J. Neural Eng. 16, 031001 (2019).

[43] Yulita, I.N.; Fanany, M.I.; Arymuthy, A.M. Bi-directional long short-term memory using quantized data of deep belief networks for sleep stage classification. Procedia Comput. Sci. 2017, 116, 530–538.

[44] Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 1997, 45, 2673–2681. [CrossRef].

[45] A. Leino *et al.*, "Deep Learning Enables Accurate Automatic Sleep Staging Based on Ambulatory Forehead EEG," in *IEEE Access*, vol. 10, pp. 26554-26566, 2022, doi: 10.1109/ACCESS.2022.3154899.

Appendici

A. Dettagli Implementativi Training

Allenamento rete CNN+BiLSTM

```
n_units1 = 10 #numero unità nascoste primo layer BiLSTM

n_units2 = 5 #numero unità nascoste secondo layer BiLSTM

prob = 0.3 #probabilità layer di Dropout

num_classi = 5 # Numero di classi nel tuo nuovo task di classificazione multiclasse

# Adatto modello al task multiclasse di classificazione delle fasi del sonno

model = keras.Sequential(

 [

    # Strati convoluzionali e di pooling (input atteso dal primo layer a 3000)

    layers.Conv1D(filters = 8 , kernel_size = 55, padding='same',input_shape=(3000,1)),

    layers.MaxPooling1D(pool_size=2,strides=2, padding='same'),

    layers.Conv1D(filters = 16, kernel_size = 41, padding='same'),

    layers.MaxPooling1D(pool_size=2,strides=2, padding='same'),

    layers.Conv1D(filters = 24, kernel_size = 33, padding='same'),

    layers.MaxPooling1D(pool_size=2,strides=2, padding='same'),

    layers.Dropout(prob),

    layers.Conv1D(filters = 36, kernel_size = 21, padding='same'),

    layers.MaxPooling1D(pool_size=2,strides=2, padding='same'),

    layers.Conv1D(filters = 48, kernel_size = 9, padding='same'),

    layers.MaxPooling1D(pool_size=2,strides=2, padding='same'),

    layers.Conv1D(filters = 56, kernel_size = 3, padding='same'),
```

```
    layers.MaxPooling1D(pool_size=2,strides=2, padding='same'),
    layers.Dropout(prob),
    layers.Bidirectional(LSTM(n_units1,return_sequences = True)),
    layers.Bidirectional(LSTM(n_units2,return_sequences = True)),
    layers.Dropout(prob), #Dovrebbe prevenire l'overfitting, parametro di regolarizzazione
    layers.Flatten(),
    # Strati completamente connessi
    layers.Dense(50,activation = 'relu'),
    layers.Dense(20,activation = 'relu'),
    layers.Dense(num_classi, activation='softmax')
 ]
)
learning_rate = 0.00001
optimizer = keras.optimizers.Adam(learning_rate=learning_rate)
model.compile(loss='categorical_crossentropy',optimizer=optimizer,
metrics=['accuracy'])
summary = print(model.summary())
# Bilanciamento pesi classi
import numpy as np
from sklearn.utils.class_weight import compute_class_weight
Y_train_labels = np.argmax(Y_train, axis=1)  # Y_train è in formato one-hot
class_labels = np.unique(Y_train_labels)  # Ottieni le etichette uniche
class_weights = compute_class_weight('balanced', classes=class_labels,
y=Y_train_labels)
class_weight_dict = {i: class_weights[i] for i in range(len(class_labels))}
# Per salvare checkpoint intermedi e applicare poi l'iterazione migliore
```

```
from tensorflow.keras.callbacks import ModelCheckpoint, Callback

class CustomModelCheckpoint(Callback):

  def __init__(self, filepath, save_every_n_epochs=25, monitor='val_loss', mode='min'):

# Allenamento del modello con il callback personalizzato (salvo epoca migliore ogni tot) e class weight

allenamento = model.fit(x=X_train, y=Y_train, batch_size=20, validation_data=(X_val, Y_val), epochs=500, callbacks=[custom_checkpoint], verbose=1, class_weight=class_weight_dict)
```

B. Dettagli Implementativi Metodi XAI

GradCAM

```
import numpy as np
import tensorflow as tf
def gradcam_heatmap(model, x, layer_name):
  last_conv_layer = model.get_layer(layer_name)
  grad_model = tf.keras.models.Model([model.inputs], [last_conv_layer.output, model.output])
  with tf.GradientTape() as tape:
    last_conv_layer_output, preds = grad_model(x, training=False)
    pred_index = tf.argmax(preds[0])
    class_channel = preds[:, pred_index]
    grads = tape.gradient(class_channel, last_conv_layer_output)
    pooled_grads = tf.reduce_mean(grads, axis=(0, 1, 2))
    last_conv_layer_output = last_conv_layer_output[0] * pooled_grads
    heatmap = tf.reduce_sum(last_conv_layer_output, axis=-1)
    heatmap = np.maximum(heatmap.numpy(), 0)
  return (heatmap - np.min(heatmap)) / (np.max(heatmap) - np.min(heatmap))
```

GradCAM++

```
import numpy as np
import tensorflow as tf
def gradcampp_heatmap(model, x, layer_name):
  last_conv_layer = model.get_layer(layer_name)
  grad_model = tf.keras.models.Model([model.inputs], [last_conv_layer.output, model.output])
  with tf.GradientTape() as tape:
    last_conv_layer_output, preds = grad_model(x, training=False)
    pred_index = tf.argmax(preds[0])
    class_channel = preds[:, pred_index]
    grads = tape.gradient(class_channel, last_conv_layer_output)
    grads_squared = tf.square(grads)
    grads_cubed = tf.pow(grads, 3)
    alpha = grads_squared / (2 * grads_squared + tf.reduce_sum(last_conv_layer_output * grads_cubed, axis=(0,1,2)))
    weights = tf.reduce_sum(alpha * tf.nn.relu(grads), axis=(0,1,2))
    last_conv_layer_output = last_conv_layer_output[0] * weights
    heatmap = tf.reduce_sum(last_conv_layer_output, axis=-1)
    heatmap = np.maximum(heatmap.numpy(), 0)
  return (heatmap - np.min(heatmap)) / (np.max(heatmap) - np.min(heatmap))
```

ScoreCAM

```
import numpy as np

import tensorflow as tf

def scorecam_heatmap(model, x, layer_name):

  last_conv_layer = model.get_layer(layer_name)

  grad_model = tf.keras.models.Model([model.inputs], [last_conv_layer.output, model.output])

  last_conv_layer_output = grad_model(x)[0][0].numpy()

  num_channels = last_conv_layer_output.shape[-1]

  activations = np.maximum(last_conv_layer_output, 0)

  activations = (activations - np.min(activations)) / (np.max(activations) - np.min(activations))

  scorecam_heatmap = np.zeros(activations.shape[:-1])


  for i in range(num_channels):

    upsampled_activation = tf.image.resize(tf.expand_dims(activations[..., i], axis=-1), (x.shape[1], x.shape[2]))

    masked_input = x * upsampled_activation

    preds = grad_model(masked_input)[1]

    score = preds[0, tf.argmax(preds[0])]

    scorecam_heatmap += score.numpy() * activations[..., i]

  return (scorecam_heatmap - np.min(scorecam_heatmap)) / (np.max(scorecam_heatmap) - np.min(scorecam_heatmap))
```