Enhancing Sleep Apnea Detection with CO-CNN: A Novel Co-Teaching Framework Application

Alessandra D'Anna[1*] and Laura Ferretti[1*†]

[1*]Università Politecnica delle Marche, Laurea Magistrale Ingegneria Informatica e dell'Automazione.

*Corresponding author(s). E-mail(s): S1120037@studenti.univpm.it; S1119324@studenti.univpm.it;
†Gli autori hanno contribuito in maniera equa al progetto.

Abstract

La sindrome delle apnee del sonno (SAS), secondo i più recenti studi, colpisce una porzione compresa tra l'1% e il 4% della popolazione adulta. È caratterizzata da episodi di respirazione interrotta per almeno 10 secondi, dovuti a parziale o totale occlusione delle vie aeree durante il sonno. Attualmente, l'identificazione della SAS si basa principalmente sulla polisonnografia (PSG), un esame che presenta alcune limitazioni, tra cui l'invasività. Per affrontare queste sfide, proponiamo un sistema di supporto alla diagnosi che sfrutta dati raccolti tramite microfono tracheale, combinato con un approccio di deep learning per rilevare gli eventi apneici durante il sonno. Il nostro modello (CO-CRNN), addestrato utilizzando il framework co-teaching, utilizza una combinazione di layer convoluzionali, una unità ricorrente bidirezionale gated (GRU) e un layer denso. Addestrato su un training set di 20 pazienti e testato su 4 pazienti, il nostro approccio ha raggiunto un tasso di recall del 0.8431%, di precision 0.7932% e F1-score 0.8174%. Ha dimostrato buone prestazioni nel filtraggio delle label rumorose, evidenziando l'efficacia del paradigma di addestramento di co-teaching nel supportare l'identificazione della SAS in presenza di rumorosità nelle annotazioni del dataset.

Keywords: Sleep Apnea Syndrome, Polisonnografia, Co-teaching, CRNN

1 Introduzione

Le apnee notturne costituiscono un disturbo del sonno caratterizzato da ripetute interruzioni involontarie della respirazione, con una durata variabile da un minimo di

10 secondi a diversi minuti. Tali episodi determinano una riduzione dell'ossigeno nel sangue e nel cervello, con conseguente aumento di anidride carbonica.
Come riportato in [1], la Sindrome delle Apnee del Sonno (SAS) colpisce tra l'1% e il 4% della popolazione adulta, sebbene tali dati siano verosimilmente sottostimati. Studi epidemiologici, infatti, evidenziano una presenza significativamente superiore, rivelando un ampio spettro di casi non diagnosticati. La mancata diagnosi tempestiva può comportare un aggravamento delle condizioni cliniche del paziente, poiché la SAS è associata a un elevato rischio di ipertensione, patologie cardiovascolari, sonnolenza diurna, incidenti stradali, infortuni domestici e lavorativi, nonché a una riduzione delle capacità cognitive e lavorative, con conseguente deterioramento della qualità della vita. In considerazione di queste problematiche, possiamo affermare che la SAS rappresenta un rilevante problema socio-sanitario. È pertanto importante sviluppare strategie efficaci per ridurre le mancate diagnosi e mitigare le conseguenze negative sulla salute degli individui.
L'esame utilizzato dall'attuale pratica clinica per la valutazione della SAS è la polisonnografia. Quest'ultima viene effettuata con l'ausilio del polisonnigrafo, strumento collegato al paziente mediante sensori ed elettrodi, posizionati in diverse parti del corpo. I parametri fisiologici monitorati includono l'attività cerebrale, i livelli di ossigeno nel sangue, il battito cardiaco, la respirazione, il russamento, i movimenti degli arti e, talvolta, la pressione arteriosa. La natura della polisonnografia richiede che l'esame sia eseguito in un ambiente controllato, all'interno di strutture ospedaliere sotto la supervisione di personale medico specializzato. La necessità di un monitoraggio prolungato e l'utilizzo di sensori invasivi può causare disagi al paziente e, a volte, rendere difficile sostenere correttamente l'esame diagnostico.

1.1 Stato dell'arte

Per affrontare le problematiche derivanti dall'attuale pratica clinica, numerosi ricercatori hanno condotto studi volti a supportare la diagnosi della SAS mediante l'impiego di sistemi di machine learning o deep learning e l'ausilio di microfoni per l'acquisizione dei dati.
Un primo approccio, proposto da Korompili et al. [2], prevede l'utilizzo di microfoni ambientali e tracheali per acquisire registrazioni audio del ciclo del sonno dei pazienti. Queste registrazioni vengono successivamente elaborate da un algoritmo di Voice Activity Detection (VAD) per stimare la probabilità di presenza della respirazione. Tuttavia, come riportato dagli stessi autori, l'utilizzo di questo algoritmo non ha prodotto risultati ottimali, soprattutto in presenza di rumore ambientale.
Lo studio di Wang et al. [3] prevede anch'esso l'uso di registrazioni audio ambientali, processate mediante l'estrazione dei MEL Spectrograms. Questi dati vengono poi utilizzati per l'addestramento di una rete neurale convoluzionale (OSAnet) per la rilevazione degli eventi apneici. Anche in [4], viene proposta una classificazione di eventi apneici e stato sonno-veglia a mezzo di registrazioni tracheali. Gli autori dell'articolo, Nakano et al., propongono un approccio basato su due differenti deep neural network (DNN), ognuna addestrata per un compito specifico: la rilevazione delle apnee e la rilevazione dello stato sonno-veglia.

Nel recente studio di Singtothong et al. [5], la localizzazione degli eventi di apnea si basa su diverse caratteristiche, tra cui lo studio dei rumori notturni, l'ossigenazione del sangue e il battito cardiaco. L'implementazione proposta utilizza tre differenti modelli, ognuno finalizzato ad un compito specifico. In questo approccio, i dati di input vengono forniti come spettrogrammi MEL, dimostrando l'efficacia di una strategia multi-modale nel migliorare le prestazioni di rilevamento. Tuttavia, un limite significativo del loro approccio è che un evento di apnea viene classificato come tale solo se supera i 30 secondi di durata, a causa dell'approccio multimodale proposto, rendendo impossibile la rilevazione di apnee più brevi, che possono comunque essere clinicamente rilevanti.
Infine, lo studio di Lillini et al. [6] propone un framework che coinvolge una Rete Neurale Convoluzionale Ricorrente (S-CRNN), addestrata secondo il paradigma siamese, per l'estrazione delle caratteristiche, e una successiva fase decisionale basata su un algoritmo di clustering non supervisionato (K-means) per discriminare tra eventi di apnea e non-apnea all'interno delle registrazioni audio.
I lavori sopra citati utilizzano dataset privati o il dataset open source proposto da [7], il quale contiene numerose label rumorose che possono influenzare i risultati dell'addestramento.
A seguito dell'analisi dello stato dell'arte, proponiamo un metodo basato sull'uso di una rete neurale convoluzionale ricorrente (CO-CRNN), addestrata secondo il framework Co-Teaching [8], per l'identificazione degli eventi di apnea. Il dataset utilizzato è stato estratto dal dataset messo a disposizione da [7]. La scelta nell'utilizzo del framework sopra citato è guidata dalla natura rumorosa delle label presenti nel dataset.

2 Materiali e Metodi

In figura è possibile visualizzare lo schema del flusso di lavoro (processamento dati, training e testing), fasi principali del lavoro proposto.

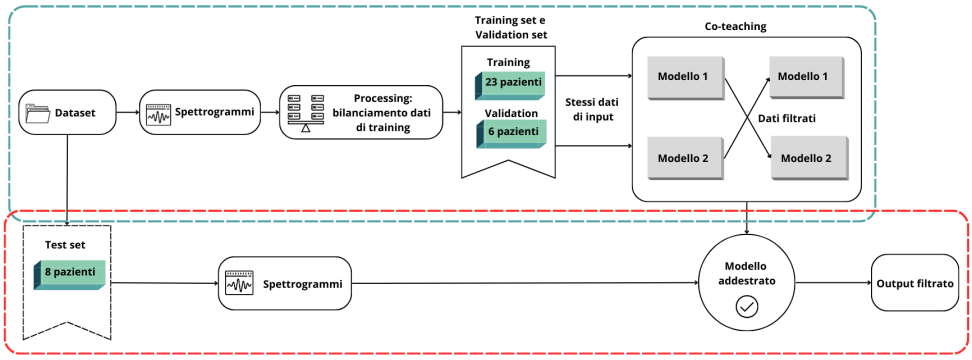

Fig. 1 Schema del flusso di lavoro per l'addestramento e il testing del modello CO-CRNN con approccio di co-teaching.

2.1 Dataset

Il dataset originale [7], da cui deriva il sub-set utilizzato nel nostro studio, è costituito da 193 pazienti in cura presso l'ospedale generale Sismanoglio di Atene. Tutti i pazienti in esso inclusi presentano forme di apnee notturne valutate come gravi o moderate.
Le registrazioni del ciclo di sonno sono state raccolte mediante un microfono tracheale (Clockaudio CTH100) e un microfono ambientale (Behringer ECM8000), durante l'esecuzione dell'esame polisonnografico. Entrambi i segnali sonori sono stati campionati a 48 kHz e memorizzati in formato audio con estensione .wav a 24 bit.
Come affermato in [7], le fasi del sonno e gli eventi di apnea sono stati valutati ed etichettati da specialisti dell'Unità del Sonno. Tutte le osservazioni sono state riportate e rese disponibili nei file di annotazione (.rml). Per semplicità di gestione, le registrazioni vengono fornite in blocchi da un'ora ciascuno.
Nel nostro studio abbiamo utilizzato un subset di 27 pazienti selezionati casualmente tra i 193 presenti nel dataset originale.
In particolare, di questi:

- 20 (corrispondente a circa il 65%) sono stati utilizzati per il training
- 3 (corrispondente a circa il 15%) per la validation
- 4 (corrispondente a circa il 20%) per il testing

Le annotazioni degli eventi di apnea necessarie per il nostro scopo, sono state estratte dai file .rml e riportate in un file .csv. I dati relativi all'inizio e alla durata di ogni apnea sono stati utilizzati per la realizzazione di una maschera binaria di lunghezza pari a quella della registrazione audio. I valori alti (1) nella maschera rappresentano le istanze di apnea mentre quelli bassi (0) le porzioni di non apnea. Un esempio è riportato nelle figure 2 e 3.

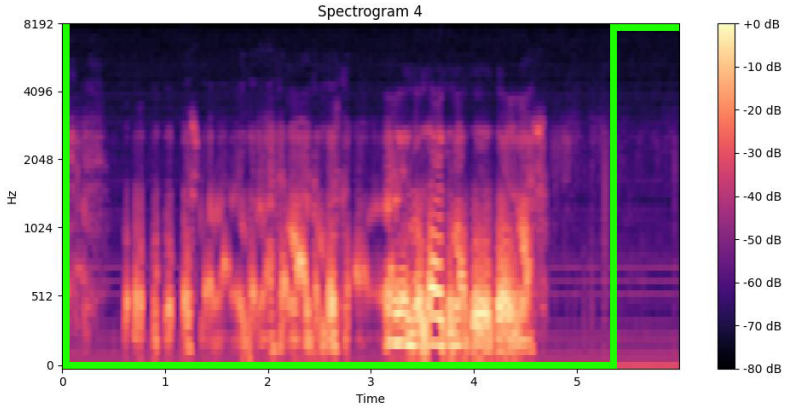

Fig. 2 Rappresentazione esplicativa della sovrapposizione della maschera binaria ad uno spettrogramma a prevalenza *non apnea*

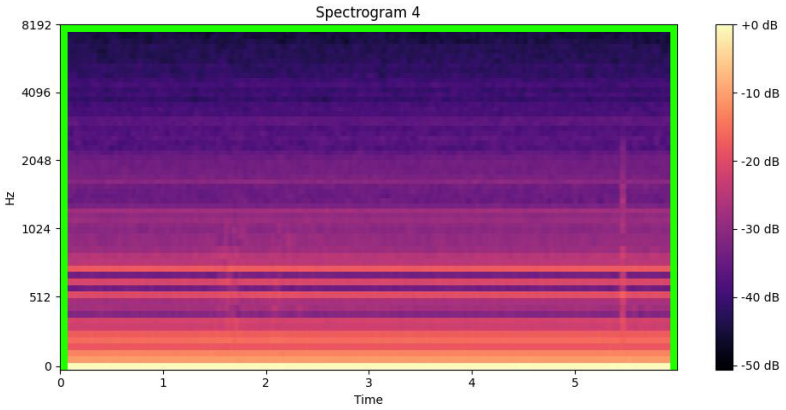

Fig. 3 Rappresentazione esplicativa della sovrapposizione della maschera binaria ad uno spettrogramma a prevalenza *apnea*

Successivamente, i file audio sono stati suddivisi in chunk di 6 secondi e per ogni chunk è stato prodotto il corrispettivo spettrogramma. Tramite l'utilizzo della maschera binaria, ogni spettrogramma è stato analizzato e, solamente i chunk contenenti interamente fenomeni di non apnea o di apnea, sono stati salvati in apposite cartelle, nominate "apnea" e "non apnea".
Per i pazienti utilizzati nella fase di training si è resa necessaria un'altra fase di processamento dei dati. Infatti, a causa dello sbilanciamento delle istanze di apnea e non apnea all'interno del dataset, è stato necessario applicare tecniche di data augmentation per ottenere un dataset bilanciato. In particolare, i dati della classe minoritaria sono stati sottoposti a resampling fino a raggiungere la quantità di dati presenti nella classe maggioritaria. Il resample è stato effettuato prestando attenzione alle eventuali problematiche che una riproduzione di dati nel dataset può causare e per questo, mediante operazioni di shuffle di dati, si è cercato di evitare che dati duplicati comparissero negli stessi batch.

2.2 Metodo proposto

L'approccio proposto prevede l'utilizzo di una rete neurale convoluzionale ricorrente addestrata secondo il paradigma del co-teaching (CO-CRNN) per il riconoscimento dei fenomeni di apnea. Tale modello è costituito da una combinazione di blocchi convoluzionali per l'estrazione delle caratteristiche, di un livello ricorrente per la modellazione delle dipendenze temporali e da un ultimo strato composto da layer convoluzionale fully connected combinato con una funzione di attivazione sigmoide.
In dettaglio, la rete conta quattro blocchi convoluzionali sequenziali, ciascuno dei quali comprende una convoluzione bidimensionale, batch normalization, una funzione di attivazione ReLU e un'operazione di max pooling. Di seguito riportiamo le specifiche tecniche di ogni strato convoluzionale:

- *conv1*: Convoluzione con 96 kernel di dimensioni (5, 5), seguita da un'operazione di pooling con finestra di dimensioni (4, 2).

- *conv2*: Convoluzione con 128 kernel di dimensioni (5, 5), seguita da un'operazione di pooling con finestra di dimensioni (4, 2).
- *conv3*: Convoluzione con 128 kernel di dimensioni (5, 5), seguita da un'operazione di pooling con finestra di dimensioni (2, 2).
- *conv4*: Convoluzione con 128 kernel di dimensioni (3, 3), seguita da un'operazione di pooling con finestra di dimensioni (2, 2).

L'obiettivo principale di questi blocchi convoluzionali è l'estrazione di caratteristiche dagli spettrogrammi MEL in input, al fine di identificare pattern utili per il riconoscimento dei fenomeni di apnea.
Successivamente, il layer GRU viene alimentato con l'output del backbone CNN. Esso agisce per la caratterizzazione temporale degli eventi [9]. In questo lavoro utilizziamo una GRU con una dimensione di input di 128 canali. L'ultimo strato della rete è un layer denso (fully connected) che, combinato con una funzione di attivazione sigmoide, è responsabile della classificazione finale degli eventi di apnea.
Per incrementare la robustezza del modello durante *l'addestramento* e affrontare la presenza di label rumorose, abbiamo implementato il framework di co-teaching proposto da [8]. Questo approccio è particolarmente utile per il filtraggio delle label rumorose nel set di addestramento e coinvolge l'uso di due reti gemelle che cooperano durante il processo di apprendimento.
L'idea principale del co-teaching è di addestrare due modelli paralleli che si aiutano reciprocamente a filtrare i dati rumorosi. In particolare, vengono inizializzati due modelli con seed diversi. Durante ogni iterazione dell'addestramento, entrambi i modelli ricevono gli stessi dati in input. Ogni batch di dati viene elaborato separatamente dalle due reti, che calcolano la perdita (loss) per ogni singolo dato nel batch. Forniamo una rappresentazione grafica di tale processo in Figura 4 (estratta dal paper [8]).

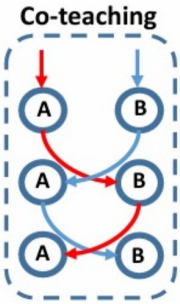

Fig. 4 Rappresentazione del framework di co-teaching

Successivamente, i valori di loss calcolati vengono ordinati in ordine decrescente, e i dati con i valori di loss più elevati vengono scartati. Questo processo permette di ottenere due mini-batch di dati filtrati, uno elaborato dalla prima rete e l'altro dalla seconda. Ciascuna rete utilizza il mini-batch filtrato dall'altra per ricalcolare la loss e aggiornare i propri pesi. Questo scambio crea un processo di insegnamento reciproco che migliora la capacità delle reti, apprendendo dai dati meno rumorosi.

Il numero di dati scartati aumenta gradualmente ad ogni epoca di addestramento. Questo incremento è regolato dal tasso di forget rate e dal numero di epoche entro cui deve essere raggiunto il massimo valore di filtraggio. La gradualità di questo processo consente un filtraggio accurato e proporzionato all'avanzamento dell'addestramento della rete, riducendo progressivamente l'influenza dei dati rumorosi. Per determinare il modello con performance migliori, abbiamo effettuato un controllo rigoroso sui valori di F1-score.
Il modello risultante dalla fase di training è stato valutato utilizzando un test set composto da dati di 4 pazienti sconosciuti al modello. Le metriche impiegate per valutare le performance includono: recall, precision e F1-score. Poiché il dataset utilizzato per il test non è bilanciato, l'accuratezza non è stata considerata una metrica rilevante. Gli spettrogrammi relativi alle registrazioni audio dei pazienti presenti nel test set, sono stati generati e sottoposti alla rete sequenzialmente, così da simulare il caso reale di analisi di un intero ciclo di sonno di un paziente.

2.2.1 Protocollo Sperimentale

Nella seguente sezione presentiamo i parametri chiave utilizzati nel nostro protocollo sperimentale:

1. *learning_rate*: Il valore inizialmente impostato è di 0.01, tale valore durante le epoche subisce una graduale diminuzione.
2. *batch_size*: Abbiamo utilizzato un batch size di 64.
3. *epochs*: Il numero di epoche scelto è di 150.
4. *macchina*: Tutto il processo è stato svolto utilizzando il framework PyTorch su una NVIDIA® TITAN RTX (Turing) con 23 GB di VRAM.
5. *ottimizzatore*: È stato utilizzato l'ottimizzatore Adam.
6. *funzione di loss*: La funzione di loss utilizzata è la binary cross entropy, scelta per il problema di classificazione binaria tra le classi di apnea e non apnea.
7. *forget_rate*: Il tasso di dimenticanza è un parametro cruciale che controlla la frazione di esempi più problematici selezionati per l'ottimizzazione dei modelli. Nel nostro caso, abbiamo impostato un valore di 0.3.
8. *num_gradual*: Rappresenta il numero di epoche durante le quali il forget_rate viene gradualmente incrementato da 0 al valore massimo definito. Nel nostro caso, abbiamo impostato questo parametro a 30 epoche.
9. *Incremento istanze da filtrare*: L'incremento del numero di istanze da filtrare durante l'addestramento è determinato dalla formula riportata in Figura 5 in cui

   - *R(T)*: percentuale di istanze da non scartare nell'epoca corrente T.
   - T_k: numero di epoche entro cui si vuole raggiungere il massimo valore di filtraggio.
   - T: epoca corrente.
   - τ: massimo tasso di filtraggio (percentuale).

$$R(T) = 1 - \min\left\{\frac{T}{T_k}\tau, \tau\right\}$$

Fig. 5 Formula di incremento delle istanze da scartare

2.3 Analisi comparativa

Abbiamo effettuato un analisi comparativa con una baseline che non prevede l'implementazione del framework co-teaching. Tale baseline consiste in una rete neurale convoluzionale ricorrente (CRNN), avente la stessa struttura presentata nel paragrafo 2.2. La scelta di effettuare questo tipo di analisi deriva dalla volontà di misurare quantitativamente l'impatto che l'introduzione del framework co-teaching comporta sulle performance della rete.

3 Risultati e Discussioni

Sono stati eseguiti vari esperimenti al fine di avere una chiara indicazione delle performance del modello.
Il modello CO-CRNN è stato addestrato con un variabile numero di pazienti, in particolare 10 e 20. Queste prove ci hanno permesso di individuare l'impatto che, una più ampia dimensione del dataset di training, può avere sulle performance del modello. I modelli ottenuti sono stati testati sul medesimo test set non bilanciato utilizzando le metriche precision, recall e F1-score. Nella tabella 1 sono riportati i risultati per ogni modello:

Table 1 Risultati in test per numero di pazienti CO-CRNN

Numero di pazienti	Precision	Recall	F1-score
10	0.7412	0.8741	0.8022
20	0.7932	0.8431	0.8174

Inoltre, seguendo l'approccio di *ablation study* [10], abbiamo indagato l'impatto che la modifica della struttura del modello ha sui risultati. In particolare, abbiamo ridotto il numero di layer convoluzionali da 4 a 3, eliminando l'ultimo blocco. Questo cambiamento ha portato a risultati in fase di test che possono essere visualizzati nella tabella 2, dimostrando l'importanza dell'ultimo layer convoluzionale per la predizione di risultati migliori. I valori riportati per entrambi i modelli derivano dall'addestramento operato sul set di training costituito da 20 pazienti.

Table 2 Risultati dell'ablation study

Modello	Precision	Recall	F1-score
CO-CRNN 4 conv layer	0.7932	0.8431	0.8174
CO-CRNN 3 conv layer	0.6064	0.9201	0.7310

Le stesse metriche e lo stesso set di dati sono stati usati anche per valutare le performance della baseline al fine di aver risultati confrontabili tra le soluzioni. I risultati ottenuti evidenziano l'effettivo impatto che label rumorose posso avere sulle performance del modello.

In tabella 3 riportiamo i risultati della baseline a confronto con quelli ottenuti dalla soluzione da noi proposta.

Table 3 Confronto dei risultati tra baseline e soluzione proposta

Modello	Precision	Recall	F1-score
Baseline	0.6815	0.8482	0.7558
CO-CRNN	0.7932	0.8431	0.8174

Tutti i risultati precedentemente riportati sono stati raggiunti attraverso un processo continuo di miglioramento tramite un approccio a griglia, che ha coinvolto l'ottimizzazione dei parametri chiave descritti nel paragrafo precedente.
Ulteriori esperimenti sono stati svolti, sono state implementate tecniche di post processing dei risultati al fine di filtrare i risultati non considerati significativi nel contesto della SAS. Tali esperimenti però, non hanno condotto a risultati significativi.

3.0.1 Conclusione e sviluppi futuri

In questo articolo proponiamo una rete neurale convoluzionale ricorrente (CO-CRNN), addestrata mediante approccio di co-teaching, per il riconoscimento delle apnee notturne basandosi su registrazioni audio, effettuate con microfoni tracheali. I risultati ottenuti (Recall: 0.8431%, Precision: 0.7932%, F1-score: 0.8174%) evidenziano come il nostro approccio rappresenti un avanzamento rispetto alla baseline considerata, mostrando miglioramenti sulle metriche di valutazione utilizzate, ovvero precision e F1-score.
Il lavoro svolto sottolinea il potenziale del nostro approccio nel migliorare il supporto alla diagnosi delle apnee notturne, offrendo una soluzione efficace e accessibile anche in ottica di monitoraggio domestico. Il valore di questo metodo risiede non solo nelle buone performance del modello, ma anche nella sua robustezza in presenza di label rumorose, grazie al framework di co-teaching implementato.
Per avanzare nello sviluppo futuro, nell'ottica di migliorare il processo di identificazione delle apnee, potrebbe essere interessante esplorare un'integrazione tra la S-CRNN presentata in [6] e il framework di co-teaching.
Inoltre, un'ulteriore miglioramento potrebbe derivare dall'ampiamento e dal bilanciamento tra il numero di pazienti uomini e donne nel dataset.

References

[1] Maspero, C., Giannini, L., Galbiati, G., Rosso, G., Farronato, G., *et al.*: Obstructive sleep apnea syndrome: a literature review. Minerva Stomatol **64**(2), 97–109 (2015)

[2] Georgia Korompili, S.A.M.N.-A.T. Lampros Kokkalas, Potirakis, S.M.: Detecting apnea/hypopnea events time location from sound recordings for patients with severe or moderate sleep apnea syndrome (2021)

[3] Bochun Wang, H.A.Y.L.W.X.X.W..D.H. Xianwen Tang: Obstructive sleep apnea detection based on sleep sounds via deep learning (2022)

[4] Hiroshi Nakano, T.T. Tomokazu Furukawa: Tracheal sound analysis using a deep neural network to detect sleep apnea (2019)

[5] Chutinan Singtothong, T.S.: Deep-learning based sleep apnea detection using sleep sound, spo2, and pulse rate (2024)

[6] Davide Lillini, L.M.L.G. Carlo Aironi, Squartini, S.: A clinical decision support system based on deep learning for identifying sleep apneas using audio signals (2023)

[7] Korompili, G., Amfilochiou, A., Kokkalas, L., Mitilineos, S.A., Tatlas, N.-A., Kouvaras, M., Kastanakis, E., Maniou, C., Potirakis, S.M.: Psg-audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. Scientific data **8**(1), 197 (2021)

[8] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018)

[9] Shiri, F.M., Perumal, T., Mustapha, N., Mohamed, R.: A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru. arXiv preprint arXiv:2305.17473 (2023)

[10] Meyes, R., Lu, M., Puiseau, C.W., Meisen, T.: Ablation studies in artificial neural networks. arXiv preprint arXiv:1901.08644 (2019)