REPORT — Valutazione di ARTEMIS, TARNet e MITNet sul Dataset NEWS

1. Introduzione

In questo progetto è stata valutata e adattata l'architettura ARTEMIS, originariamente progettata per il trattamento binario sul dataset IHDP, al dataset NEWS.
L'obiettivo principale è stimare l'effetto causale individuale (ITE) e confrontare le prestazioni di: TARNet, MITNet, ARTEMIS mantenendo lo stesso setup sperimentale per garantire un confronto equo e metodologicamente corretto.

Il confronto è stato condotto utilizzando:

- stesso dataset
- stesso split train/test
- stessi seed
- stesse metriche di valutazione

2. Descrizione del Dataset NEWS

È stato utilizzato il dataset NEWS nella sua versione semi-sintetica con trattamento binario.

2.1 Caratteristiche principali

- Numero di campioni: N = 5000
- Numero di feature: d = 3477
- Variabile di trattamento: binaria ($T \in \{0,1\}$)
- Outcome: continuo
- Potenziali outcome disponibili: $\mu_0(x)$ e $\mu_1(x)$

2.2 Struttura dei file

Il dataset è fornito in formato sparso (.csv.x), in cui ogni riga rappresenta una tripletta: (indice_riga, indice_colonna, valore)
La matrice delle feature è stata ricostruita in formato denso utilizzando la rappresentazione COO (Coordinate Format) tramite la libreria `scipy.sparse`.
Il file .csv.y contiene:

- Colonna 0 → trattamento osservato T
- Colonna 1 → outcome fattuale YF

- Colonna 3 → $\mu_0(x)$
- Colonna 4 → $\mu_1(x)$

2.3 Natura semi-sintetica

Il dataset NEWS è semi-sintetico:
Il trattamento osservato è reale, i potenziali outcome per tutti i trattamenti sono generati artificialmente. Questo consente di calcolare direttamente metriche causali che normalmente non sarebbero osservabili in dati reali come pehe e errore sull'ate.

2.4 Complessità del Dataset

L'elevata dimensionalità delle covariate (d = 3477) rende il problema particolarmente complesso, in quanto aumenta il rischio di overfitting e richiede una rappresentazione latente informativa per stimare correttamente l'eterogeneità dell'effetto causale.
In contesti ad alta dimensionalità, modelli puramente supervisionati possono avere difficoltà a separare l'informazione rilevante per il trattamento da quella legata all'outcome, rendendo cruciale la qualità della rappresentazione latente appresa.

3. Preprocessing e Setup Sperimentale

3.1 Ricostruzione della matrice

La matrice sparsa è stata ricostruita come $X \in R$ ^(5000×3477) mediante costruzione di una `coo_matrix` e successiva conversione in formato denso.

3.2 Split Train/Test

È stato utilizzato uno split:

- 80% training
- 20% test
- Campionamento stratificato rispetto a T
- random_state = 42

Gli indici dello split sono stati salvati per garantire completa riproducibilità.
Questo garantisce:

1. Bilanciamento tra trattati e controlli
2. Confronto coerente tra modelli
3. Riproducibilità degli esperimenti

4. Modifiche Implementative rispetto alla versione originale di ARTEMIS

L'implementazione originale di ARTEMIS era progettata per il dataset IHDP, con caricamento automatico dei dati e struttura specifica per trattamento binario.
Per adattare il modello al dataset NEWS sono state effettuate le seguenti modifiche.

4.1 Sostituzione del Loader IHDP con Loader NEWS

La versione originale utilizzava un loader dedicato per IHDP, che scaricava automaticamente le simulazioni e restituiva i tensori già pronti per il training.
Nel presente lavoro:
- Il loader IHDP è stato rimosso.
- È stato implementato un caricamento manuale del dataset NEWS.
- Sono stati letti i file:
- .csv.x per le covariate
- .csv.y per trattamento e outcome

Il file .csv.x è fornito in formato sparse (COO).
È stata quindi ricostruita la matrice delle feature in formato denso utilizzando scipy.sparse.coo_matrix.

Il file .csv.y è stato interpretato come:
- Colonna 0 → Trattamento osservato T
- Colonna 1 → Outcome fattuale $Y_{(F)}$
- Colonna 3 → $\mu_0(x)$
- Colonna 4 → $\mu_1(x)$

Sono stati quindi costruiti manualmente gli array necessari al training:
X, T, $Y_{(F)}$, μ_0, μ_1

4.2 Setup sperimentale e split riproducibile

È stato implementato uno split:
- 80% training
- 20% test
- Stratificato rispetto alla variabile di trattamento
- random_state = 42

Gli indici di split sono stati salvati su Drive per garantire riproducibilità e confronto equo tra i modelli.

4.3 Estensione strutturale dell'OutcomeHead

La versione originale di ARTEMIS prevedeva due teste separate:
- head0 → μ_0(x)
- head1 → μ_1(x)

Questa struttura era rigida e limitata al trattamento binario.
È stata implementata una versione generalizzata multi-head che restituisce:
$\hat{\mu}(x) \in \mathbb{R}^{BxK}$
dove:
- B è la dimensione del batch
- K è il numero di trattamenti

Nel caso corrente K = 2, ma la struttura è ora estendibile a K > 2.
La loss fattuale è stata modificata da:
torch.where(T, μ_1, μ_0)
a una selezione indicizzata:
$Y_{(F)} = \hat{\mu}(x)[range(B), T]$
Questo rende il codice compatibile con trattamento multi-classe.

4.4 Mantenimento della componente contrastiva

La loss contrastiva originale è stata mantenuta invariata. Essa opera sugli embedding latenti z prodotti dall'encoder e forza:
- Avvicinamento tra embedding con effetti causali simili
- Allontanamento tra embedding con effetti differenti

Questo meccanismo favorisce la strutturazione dello spazio latente in funzione dell'eterogeneità dell'effetto causale.

4.5 Mantenimento della regolarizzazione tramite Mutual Information

Sono state mantenute le due reti MINE per la stima della Mutual Information:
MI(z, T)
MI(z, Y)

Questa componente:
- Penalizza dipendenza eccessiva tra rappresentazione latente e trattamento
- Preserva informazione utile rispetto all'outcome

La struttura originale della loss complessiva è stata mantenuta:
$\mathcal{L} = \mathcal{L}_{(}\text{sup}_{)} + \alpha\ \mathcal{L}_{(}\text{extra}_{)}$
dove la componente extra include:
- Loss contrastiva
- Termini di Mutual Information

4.6 Training multi-seed

Per garantire robustezza dei risultati, ogni modello è stato addestrato con 5 inizializzazioni differenti (seed 0–4). Sono state riportate:
- Media
- Deviazione standard
- Test statistici paired (t-test e Wilcoxon)

5. TARNet

5.1 Architettura

TARNet utilizza:
- Encoder MLP condiviso
- Due teste di outcome:
  - $\mu_0(x)$
  - $\mu_1(x)$

- Loss supervisionata (MSE/SmoothL1)

Non utilizza:

- Loss contrastiva
- Regolarizzazione tramite Mutual Information

5.2 Risultati — TARNet su NEWS (5 seed)

-PEHE = 3.3861 ± 0.1570
-Errore ATE = 0.1916 ± 0.1499

5.3 Osservazioni

1. PEHE elevato
2. Varianza significativa nell'ATE
3. Nessun meccanismo di regolarizzazione della rappresentazione

TARNet non utilizza meccanismi di regolarizzazione basati su Mutual Information né struttura contrastiva, il che può spiegare la performance inferiore in presenza di covariate ad alta dimensionalità.

6. MITNet

6.1 Architettura

MITNet utilizza:

- Encoder MLP
- Due teste di outcome (μ_0, μ_1)
- Loss supervisionata MSE

In questa implementazione:
-Non è stata integrata loss contrastiva
-Nella presente implementazione non è stata integrata una penalizzazione esplicita della Mutual Information, rendendo il modello equivalente a una variante supervisionata standard

6.2 Risultati — MITNet su NEWS (5 seed)

-PEHE = 3.7083 ± 0.1619
-Errore ATE = 0.2551 ± 0.0493

6.3 Osservazioni

- Prestazioni inferiori rispetto a TARNet
- Nessun miglioramento significativo rispetto alla baseline
- Difficoltà nella stima dell'effetto individuale su feature ad alta dimensionalità

7. ARTEMIS

7.1 Architettura

ARTEMIS integra:

- Encoder MLP
- Multi-head outcome (estendibile a K trattamenti)
- Loss supervisionata
- Loss contrastiva sugli embedding
- Regolarizzazione tramite Mutual Information (MINE networks):
  - MI(z, T)
  - MI(z, Y)

Questa combinazione mira a separare informazione legata al trattamento da quella legata all'outcome, migliorando la qualità della rappresentazione latente e la capacità del modello di catturare eterogeneità dell'effetto causale.

7.2 Risultati — ARTEMIS su NEWS (5 seed)

-PEHE = 1.5693 ± 0.2342
-Errore ATE = 0.1479 ± 0.0890

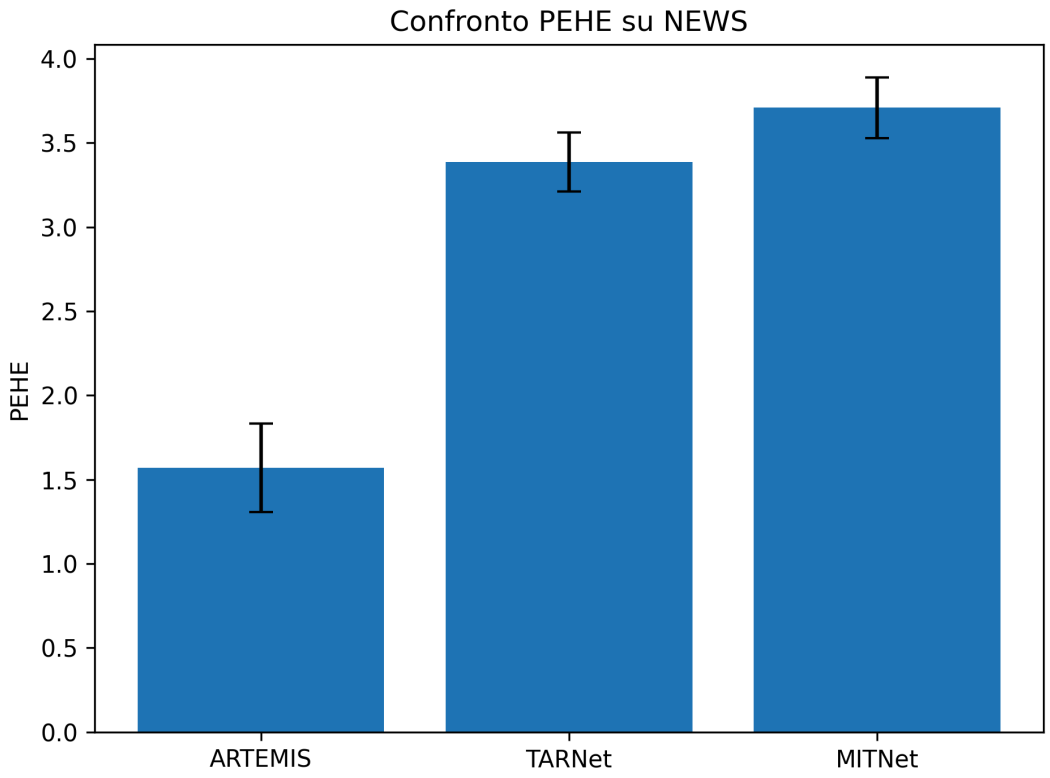
8. Confronto Finale tra Modelli

MODELLO	PEHE	ATE error
TARNet	3.3861 ± 0.1570	0.1916 ± 0.1499
MITNet	3.7083 ± 0.1619	0.2551 ± 0.0493
ARTEMIS	1.5693 ± 0.2342	0.1479 ± 0.0890

Si osserva una riduzione del PEHE di circa il 53% rispetto a TARNet e del 57% rispetto a MITNet.

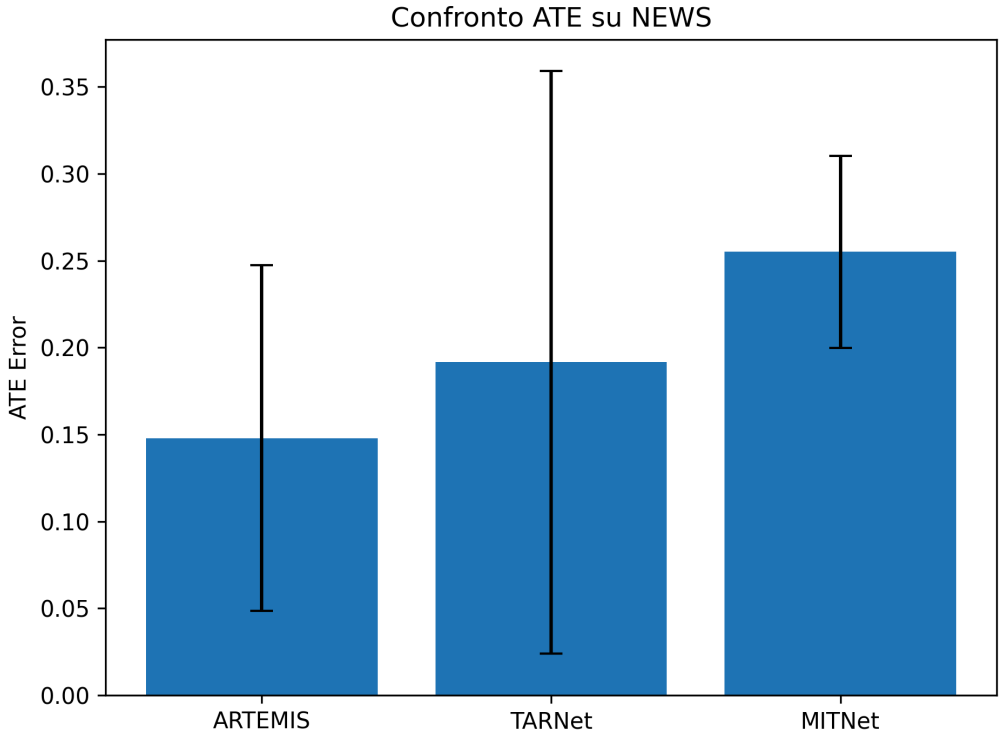
8.1 Analisi Grafica dei Risultati

- FIGURA 1: Confronto PEHE (media ± deviazione standard)



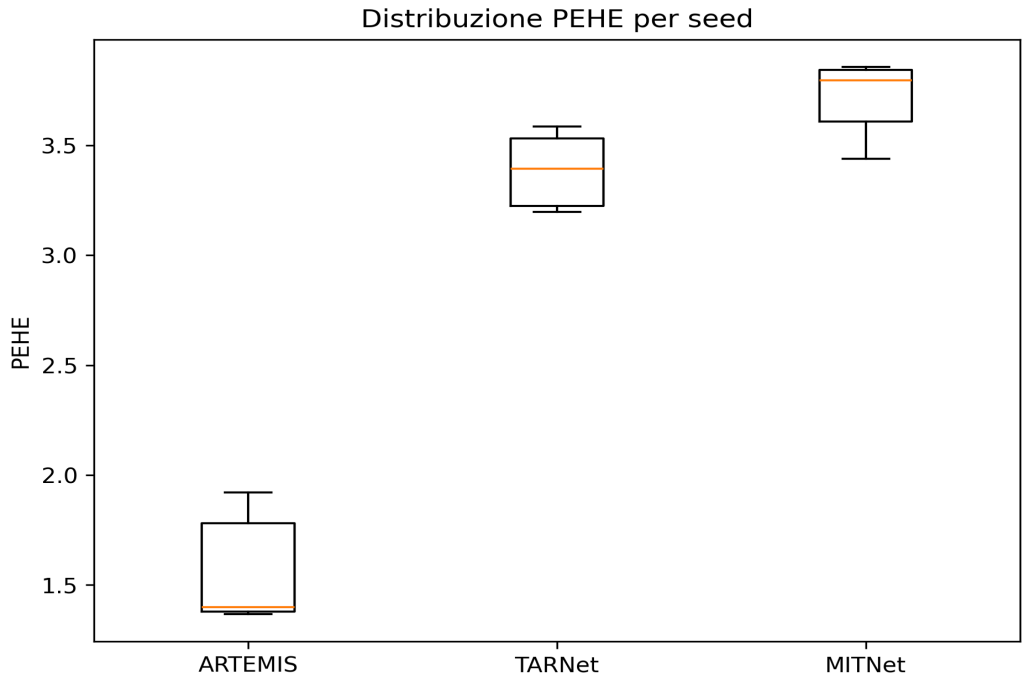
Il grafico evidenzia un netto miglioramento di ARTEMIS in termini di PEHE rispetto a TARNet e MITNe

- FIGURA 2: Confronto ATE



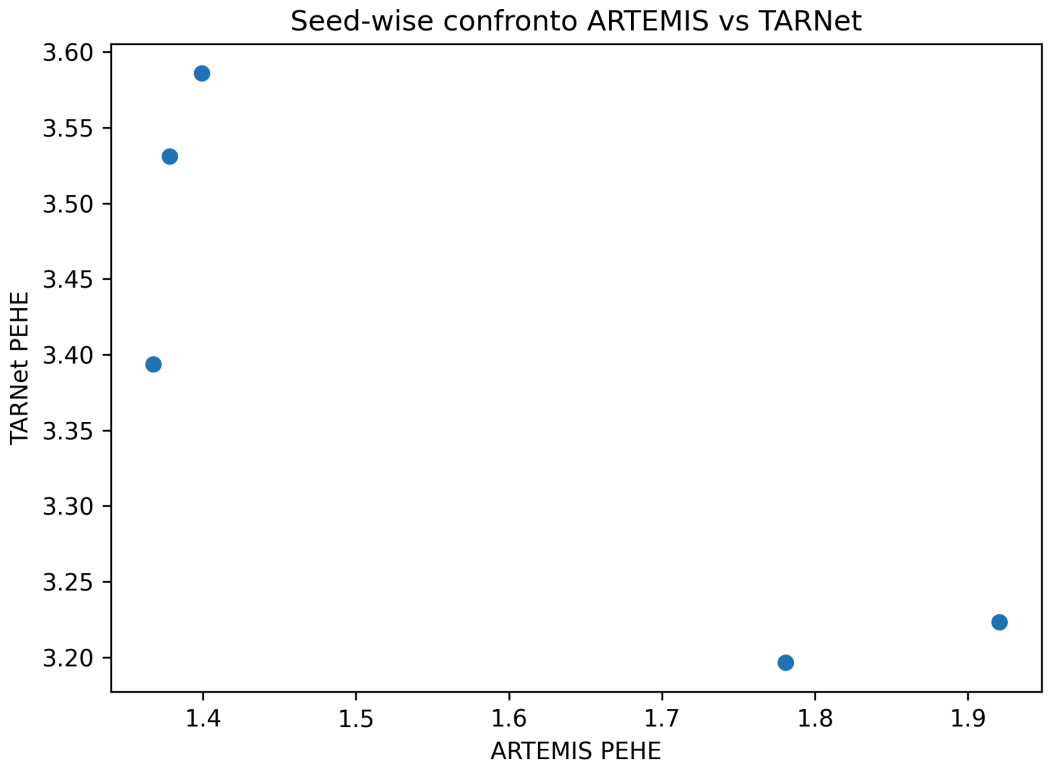
Le differenze tra i modelli risultano meno marcate sull'ATE, coerentemente con la natura aggregata della metrica.

- FIGURA 3: Distribuzione PEHE per seed



La distribuzione mostra una separazione consistente tra ARTEMIS e gli altri modelli.

- FIGURA 4: Confronto seed-wise ARTEMIS vs TARNet



Ogni punto rappresenta uno stesso seed per entrambi i modelli.
Si osserva che ARTEMIS ottiene sistematicamente valori inferiori di PEHE.

9. Analisi Comparativa

Dai risultati emerge chiaramente che:

- ☐ TARNet mostra errori elevati.
- ☐ MITNet non introduce miglioramenti sostanziali.
- ☐ ARTEMIS ottiene un miglioramento significativo in PEHE.

La combinazione di: contrastive learning e penalizzazione della Mutual Information
sembra consentire una migliore separazione delle rappresentazioni latenti e quindi una stima più accurata dell'effetto causale individuale.
Il miglioramento in termini di PEHE suggerisce che la componente contrastiva giochi un ruolo determinante nella qualità della stima dell'ITE.
Il miglioramento osservato in termini di PEHE può essere interpretato alla luce dell'elevata dimensionalità del dataset (d = 3477).

In contesti ad alta dimensionalità, modelli puramente supervisionati come TARNet e MITNet possono apprendere rappresentazioni latenti che catturano correlazioni spurie o informazioni non direttamente rilevanti per la stima dell'effetto causale individuale.

L'integrazione della loss contrastiva e della regolarizzazione tramite Mutual Information in ARTEMIS favorisce invece una strutturazione più informativa dello spazio latente, riducendo la dipendenza non desiderata tra rappresentazione e trattamento e migliorando la capacità del modello di distinguere correttamente l'eterogeneità dell'effetto.

Questo spiega perché il vantaggio di ARTEMIS emerge in modo significativo sulla metrica PEHE, mentre le differenze risultano meno evidenti sull'ATE, che rappresenta una misura aggregata meno sensibile alla qualità fine della rappresentazione latente.

Per verificare la significatività statistica delle differenze tra i modelli, sono stati eseguiti:

- Paired t-test
- Wilcoxon signed-rank test

utilizzando i risultati ottenuti sugli stessi 5 seed e sugli stessi split

.

9.1 Confronto su PEHE

Confronto Mean Difference	t-test p-value	Wilcoxon p-value
ARTEMIS vs TARNet	-1.8168	0.0007 0.0625
ARTEMIS vs MITNet	-2.1391	0.0003 0.0625
TARNet vs MITNet	-0.3222	0.0309 0.0625

- I risultati del t-test indicano che ARTEMIS ottiene un miglioramento statisticamente significativo rispetto a TARNet e MITNet in termini di PEHE ($p < 0.001$).

- Il test di Wilcoxon, pur mostrando p-value leggermente superiori alla soglia classica di 0.05, conferma una tendenza coerente con il t-test. La discrepanza tra i due test può essere attribuita alla ridotta numerosità del campione (5 seed), che limita la potenza statistica del test non parametrico.

9.2 Confronto su Errore ATE

Confronto	Mean Difference	t-test p-value	Wilcoxon p-value
ARTEMIS vs TARNet	-0.0437	0.6987	1.0000
ARTEMIS vs MITNet	-0.1071	0.0787	0.1250
TARNet vs MITNet	-0.0635	0.5540	0.4375

Non emergono differenze statisticamente significative sull'ATE, Questo suggerisce che:
I modelli stimano in modo comparabile l'effetto medio e che la differenza principale riguarda la capacità di catturare l'eterogeneità individuale

9.3 Sintesi dei Risultati Statistici

I risultati indicano che ARTEMIS ottiene un miglioramento statisticamente significativo nella stima dell'effetto causale individuale (PEHE), mentre le differenze sull'ATE non risultano significative.
Questo conferma che la componente contrastiva e la regolarizzazione informativa contribuiscono principalmente alla modellazione dell'eterogeneità individuale piuttosto che alla stima dell'effetto medio.

10. Considerazioni Metodologiche

L'esperimento evidenzia che la qualità della rappresentazione latente gioca un ruolo cruciale nella stima dell'eterogeneità dell'effetto causale.
I risultati suggeriscono che meccanismi di regolarizzazione strutturale, come la loss contrastiva e la penalizzazione della Mutual Information, possono migliorare la capacità del modello di catturare differenze individuali, pur mantenendo stime comparabili sull'effetto medio.

11. Prossimi Passi

Le possibili estensioni includono:
1. Estensione completa al caso multi-trattamento (News-4)
2. Analisi della sensibilità rispetto agli iperparametri
3. Studio dell'impatto separato di:
- solo contrastive
- solo MI
- contrastive + MI

12. Limitazioni dello Studio

Nonostante i risultati ottenuti evidenzino un miglioramento significativo delle prestazioni di ARTEMIS rispetto a TARNet e MITNet, il presente studio presenta alcune limitazioni metodologiche. In primo luogo, il dataset NEWS è di natura semi-sintetica. Sebbene permetta la valutazione diretta delle metriche causali, i potenziali outcome non derivano da osservazioni reali ma da una generazione artificiale. Questo può limitare la completa generalizzabilità dei risultati a contesti reali. In secondo luogo, è stato utilizzato un unico split train/test (80/20) fisso, sebbene valutato su 5 seed differenti. Una validazione incrociata più estesa potrebbe fornire una stima ancora più robusta della variabilità del modello.

Un'ulteriore limitazione riguarda l'assenza di un'analisi approfondita della sensibilità agli iperparametri. Sebbene siano stati utilizzati valori ragionevoli e coerenti tra modelli, un'ottimizzazione sistematica (ad esempio tramite ricerca bayesiana) potrebbe influenzare le prestazioni finali. Infine, la versione di MITNet implementata non include una penalizzazione esplicita della Mutual Information, ma solo una struttura supervisionata standard. Questo potrebbe aver limitato la sua capacità di apprendere rappresentazioni latenti più informative.