DistCare: Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis

CSE 6250 Final Project

- Team C3: Bo Fan & Surya Garikipati

- GitHub Link:

https://github.com/BoFanDY/BD4H-C3-project

- Presentation slide and presentation video link (google drive):

https://drive.google.com/drive/folders/1MFlUNiHoK9KlXRnepsJrIzSrW6GdzZOJ?usp=share_link

video Link

Data Preprocessing

Teacher Model

- Pre-trained on the PhysioNet dataset.

Student Model

- Trained on the Tongji Hospital (TJH) dataset, leveraging knowledge distillation from the teacher model.

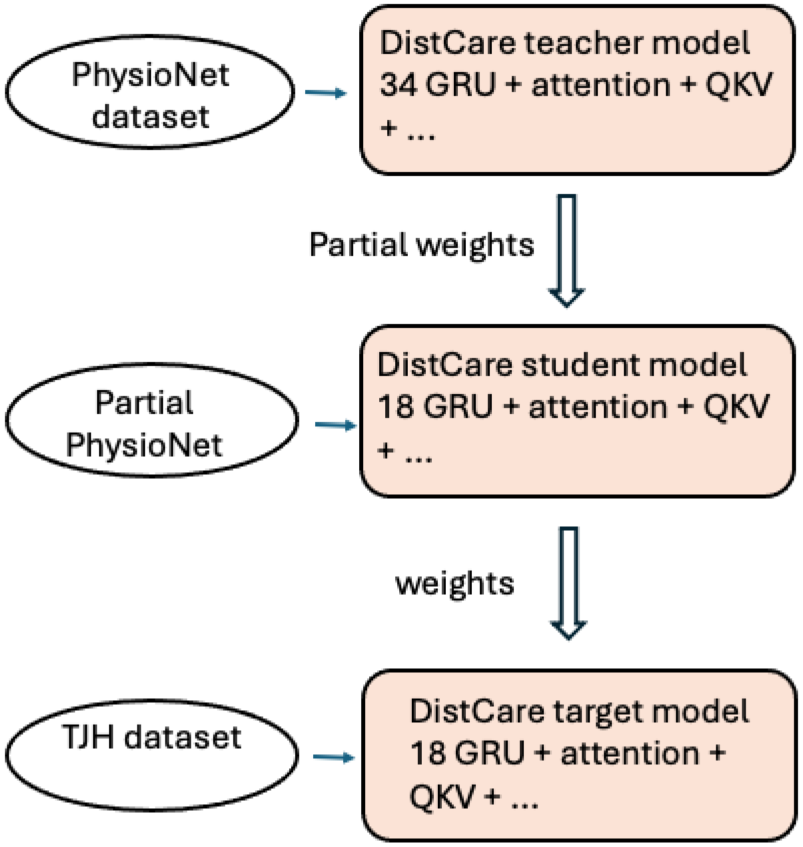
Target (Length of Stay - LOS)

- If outcome = 0: t(stay)= min(35,t)
- Otherwise: t(stay) = 70 - min(35,t)

Preprocessing Steps

- **Data Cleaning:** Forward fill to handle missing patient IDs.
- **Feature Selection:** Used numerical features for modeling.
- **Imputation:** Applied KNN imputation (k =5) to fill missing numerical data.
- **Aggregation:** Aggregated data by patient ID for a single record per patient.

DistCare Architecture



PhysioNet dataset

DistCare teacher model
34 GRU + attention + QKV
+ ...

Partial weights

Partial PhysioNet

DistCare student model
18 GRU + attention + QKV
+ ...

weights

TJH dataset

DistCare target model
18 GRU + attention +
QKV + ...

- DistCare transfer learning models contains 3 models: teacher, student, target.

- Teacher model was trained on PhysioNet dataset. Its architecture contains
  - Positional encoding layers
  - 34 GRU layers
  - Attention layer
  - QKV layer for transfer learning loss
  - Multi-headed attention layers
  - Sublayer connection layer
  - Position wise feed forward layers
  - Fully connection layers with dropout and tanh activation function
  - Output Fully connection layer

- Student model has less complex architecture (18), borrow weights from teacher model and tuned by partial PhysioNet dataset

- Target model has same architecture with student model but fined tuned on TJH dataset.

GRU Transfer Learning Architecture

Teacher Model

- 4 GRU layers (hidden size: 256) with attention mechanism.
- Two fully connected layers with batch normalization.
- Dropout: 40%.
- Pretrained on PhysioNet, fine-tuned on Tongji Hospital (TJH) dataset.

Student Model

- 2 GRU layers (hidden size: 128) without attention.
- Two fully connected layers with batch normalization.
- Dropout: 20%.
- Trained using knowledge distillation.

Knowledge Distillation

- Hard Loss: MSE between student predictions and true labels.
- Soft Loss: KL Divergence between teacher and student outputs
- Total Loss = $\alpha \cdot$ Hard Loss + $(1-\alpha) \cdot$ Soft Loss, $\alpha=0.5$

Reduces complexity while maintaining performance.

Results

DistCare Model (Reproduced)

- Our reproduced model struggled to match the original paper’s results due to missing details in preprocessing and post-processing.

DistCare Model (Paper)

- Indicates superior performance likely due to refined preprocessing steps unavailable in the paper.

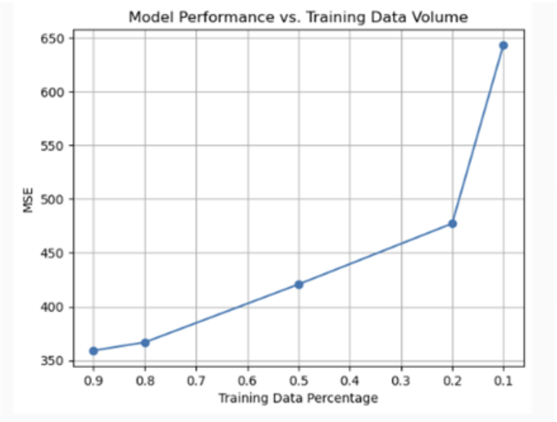
GRU Transfer Learning Model

- Demonstrates strong performance, highlighting the value of transfer learning and knowledge distillation.

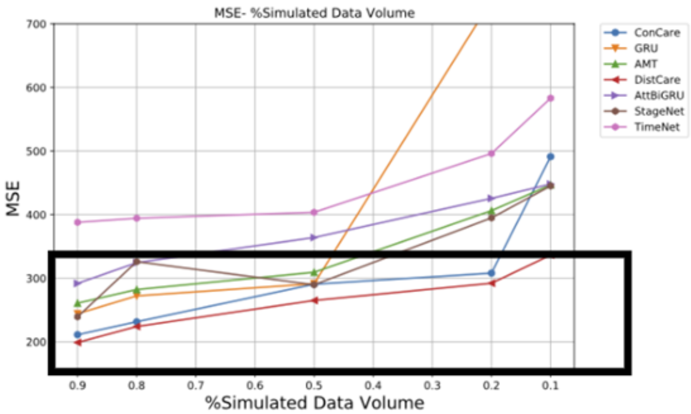
	MSE	MAE
DistCare model results we reproduced	305.8588	16.4427
DistCare model results shown in the paper	198.9287	9.7518
GRU transfer learning model	218.8638	6.6734
MLP	152.2419	6.4463

Results

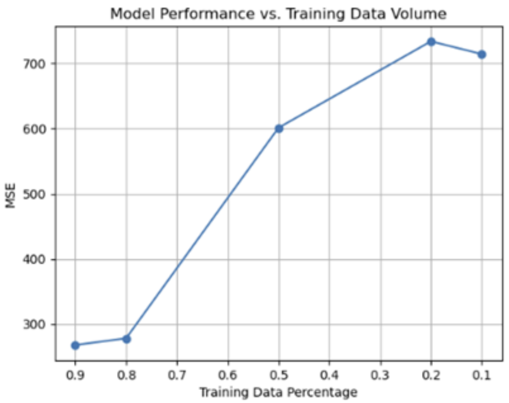
DistCare Model (Reproduced)



DistCare Model (From Paper)



GRU Transfer Learning



- The reproduced DistCare model achieves an MSE of 305.86, higher than the original paper's 198.93, likely due to missing preprocessing details.

- GRU achieves an MSE of 218.86 at 90% training data, demonstrating competitive results but still lagging behind the original DistCare model.

- Both models show increasing MSE as training data decreases, with sharp performance deterioration below 20%, consistent with the original paper.

- GRU transfer learning highlights the effectiveness of pre-trained weights, offering a simpler yet efficient alternative to complex architectures.

Results

GRU Transfer Learning Model

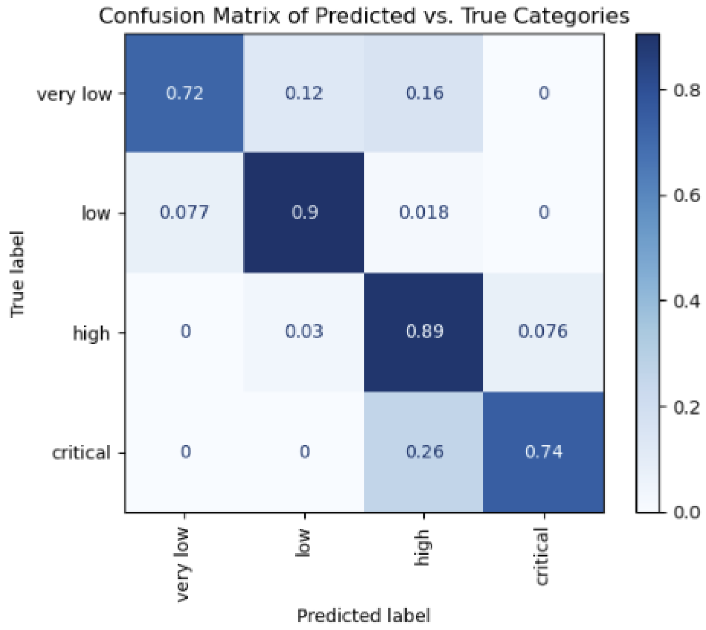
- Performs well in "low" and "high" categories but struggles to classify "critical" cases accurately, with 26% misclassified as "high."

DistCare Model (Paper)

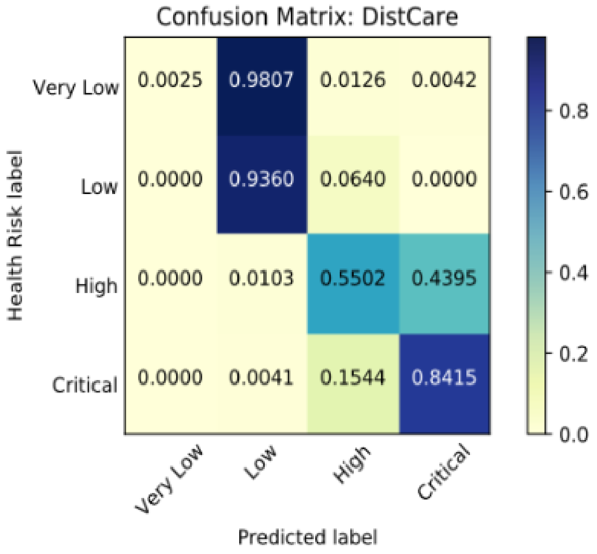
- Shows superior performance, especially in "critical" cases with 84% accuracy, demonstrating better precision across all categories.

The GRU model is competitive for simpler cases but falls short in more challenging predictions, highlighting the advantage of DistCare's robust design.

GRU Transfer Learning



DistCare Model (From Paper)



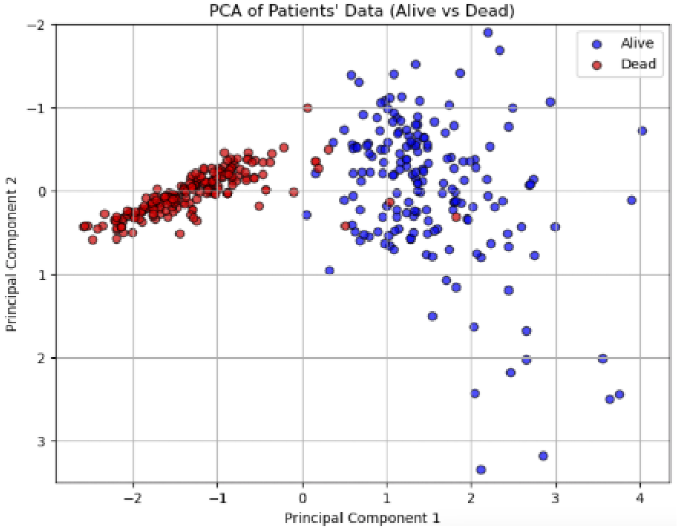
Results

- The reproduced DistCare model and the original model from the paper show similar clustering patterns, effectively separating "Alive" and "Dead" patients with minor visual differences.

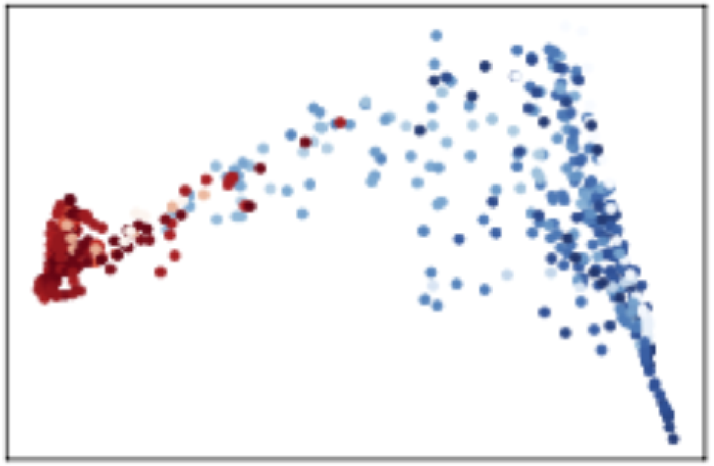
- Our post processed data reveals the capacity to capture the pattern between dead and alive patients in the features

- The small discrepancy may due to the potential difference in the data processing.

DistCare Model (Reproduced)



DistCare Model (From Paper)



Difficulties and Challenges

Incomplete Details in the Paper

- The data processing section in the paper lacked sufficient detail, making it difficult to replicate the exact pre-processing procedures.
- This required assumptions and adjustments during preprocessing to align with the model's requirements.

Hyperparameter Tuning

- Fine-tuning the numerous hyperparameters, such as learning rate, dropout, etc was time-consuming.
- Achieving the right balance for both teacher and student models required iterative experimentation.

Conclusion

- The GRU transfer learning model demonstrates competitive performance, validating the effectiveness of knowledge distillation. However, the reproduced DistCare model falls short of the original paper's results due to incomplete preprocessing details.

Thank You!