Federated Learning For Heterogeneous Electronic Health Records Utilising Augmented Temporal Graph Attention Networks

Soheila Molaei
University of Oxford

Anshul Thakur
University of Oxford

Ghazaleh Niknam
University of Tehran

Andrew Soltan
University of Oxford

Hadi Zare
University of Tehran

David Clifton
University of Oxford

Abstract

The proliferation of decentralised electronic healthcare records (EHRs) across medical institutions requires innovative federated learning strategies for collaborative data analysis and global model training, prioritising data privacy. A prevalent issue during decentralised model training is the *data-view discrepancies* across medical institutions that arises from differences or availability of healthcare services, such as blood test panels. The prevailing way to handle this issue is to select a common subset of features across institutions to make data-views consistent. This approach, however, constrains some institutions to shed some critical features that may play a significant role in improving the model performance. This paper introduces a federated learning framework that relies on augmented graph attention networks to address data-view heterogeneity. The proposed framework utilises an *alignment augmentation* layer over self-attention mechanisms to weigh the importance of neighbouring nodes when updating a node's embedding irrespective of the data-views. Furthermore, our framework adeptly addresses both the temporal nuances and structural intricacies of EHR datasets. This dual capability not only offers deeper insights but also effectively encapsulates EHR graphs' time-evolving nature. Using diverse real-world datasets, we show that the proposed framework significantly outperforms conventional FL methodology for dealing with heterogeneous data-views.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

1 INTRODUCTION

Electronic Health Record (EHR) datasets can play a pivotal role in advancing medical research and optimising patient care (Yang et al., 2023; Alpert et al., 2023). However, they are often decentralised across various medical institutions due to privacy, regulatory, and commercial considerations (Thakur et al., 2021). This decentralisation makes global data analysis as well as model training impractical (Rieke et al., 2020; Wang et al., 2023). Federated Learning (FL) offers a promising solution by enabling extraction of insights from data located in multiple places, all while preserving the privacy and autonomy of the original data holders. Instead of aggregating data centrally, FL allows data to remain at its source, transmitting only model updates. Such an approach not only ensures the safety of sensitive data but also adheres to stringent privacy regulations (Cho et al., 2022; McMahan et al., 2017; Kairouz et al., 2021).

A significant challenge with FL in this context is inconsistency of the data-views across different medical institutions. This heterogeneity in data-views results from differences in the availability of medical services. Some institutions may offer a comprehensive range of medical services, while others may have limited resources or specialised focuses. For example, arterial blood gas (ABG) test panels may not be available or well documented for a particular cohort of interest across all the National Health Service (NHS) trusts in the UK. This behaviour is further amplified in low and middle-income countries (LMICs) where medical services availability changes dramatically with geographical locations. While current FL methods overcome this

heterogeneity by selecting a subset of common features across all institutions to harmonise data-views, this often forces the omission of crucial features at some institutions, risking the loss of invaluable insights (Nguyen et al., 2022; Soltan et al., 2023; Rieke et al., 2020).

To address these challenges, this paper introduces the Augmented Graph Attention Networks (AGAT) that are based on similarity-wise graphs and can be incorporated in any FL framework to allow a more effective analysis of the EHR datasets across institutes or clients. The proposed architecture introduces an innovative alignment augmentation layer into existing self-attention mechanisms, which have been tailored to underscore the relevance of neighbouring nodes in a graph during the embedding update phase (Veličković et al., 2018). A unique aspect of this alignment augmentation layer is its adeptness in aligning feature dimensions, ensuring unbiased and consistent updates to a graph node's embedding across diverse data-views. By doing so, we achieve a holistic representation that captures the intrinsic relationships and dependencies among graph nodes. This refined method promises improved performance and adaptability across varied data scenarios. Moreover, in the realm of federated learning, the traditional approach of model aggregation as depicted in *federated averaging* (McMahan et al., 2017) often doesn't account for the varying significance of contributions from individual clients. This one-size-fits-all approach may lead to sub-optimal aggregation outcomes. Consequently, we have developed an innovative attention mechanism at the server level, which skillfully weighs the significance of individual client contributions during aggregation.

Furthermore, recognising the dynamic and temporal characteristics of data from various institutions is vital. Each data point marks a distinct episode in a patient's health journey, with effective capture of this temporal information being imperative. Such insights can trigger early interventions and improve patient outcomes by offering a deeper understanding into disease progression and treatment efficiency (Atif et al., 2023; Mateos and Rajawat, 2013). With the goal of creating more reliable and sophisticated healthcare solutions, we have enhanced our architecture to mirror the evolving nature of graphs across diverse clients. A dynamic graph, evolving over time, can be depicted using a series of static graphs or snapshots (Skarding et al., 2021). Many dynamic graph representation methods utilise Recurrent Neural Networks (RNNs) (Manessi et al., 2020; Chen et al., 2019; Niknam et al., 2023; Li et al., 2019), but RNNs require vast amounts of training data and face scalability issues with increasing time steps (Sankar et al., 2020). Here, we incorporate a time-sensitive embedding, enriching current node features with a temporal dimension that reflects data from the previous time step. This modification ensures a dynamic representation of data transitions over time. The key contributions we bring to this work include:

- We unveil the Augmented Graph Attention Networks (AGAT) tailored for EHR dataset analysis in federated learning settings, built on similarity-wise graphs.
- Introduced within AGAT, the Alignment Augmentation Layer aligns feature dimensions, ensuring consistent embeddings across different data-views.
- We adopt an attention mechanism at the server level, emphasising individual client contributions during aggregation.
- Recognising EHR's temporal nature, we embed a time-sensitive layer, reflecting data transitions over time.

2 EARLIER STUDIES

Federated learning across heterogeneous data-views is a scarcely explored problem. As discussed above, earlier studies have simply selected a subset of common features across all clients to make data-views consistent and enable federated training of a global model. In a prominent example, Soltan *et al.* trained global models for triaging COVID patients across four NHS trusts, where two of the trusts lacked arterial blood gas (ABG) test panels, by selecting common features to align data-views (Soltan et al., 2023).

Apart from this, as discussed by Aviv *et al.* (Shamsian et al., 2021), learning personalised or client-specific federated models using hypernetworks also presents an ideal solution for data heterogeneity. This FL framework trains a global hypernetwork at the server that outputs weights for personalised models at each client where client's model architectures is dictated by its respective data-view. Despite its potential advantages, hypernetworks are often difficult to train and information sharing across clients is only passive or weak.

Comparison against the proposed framework: As previously mentioned, the proposed framework maintains all features across clients without the need for cumbersome preprocessing associated with basic feature alignment. Unlike hypernetwork-based FL, the proposed framework can yield both global and personalised models as well as improved information sharing among clients.

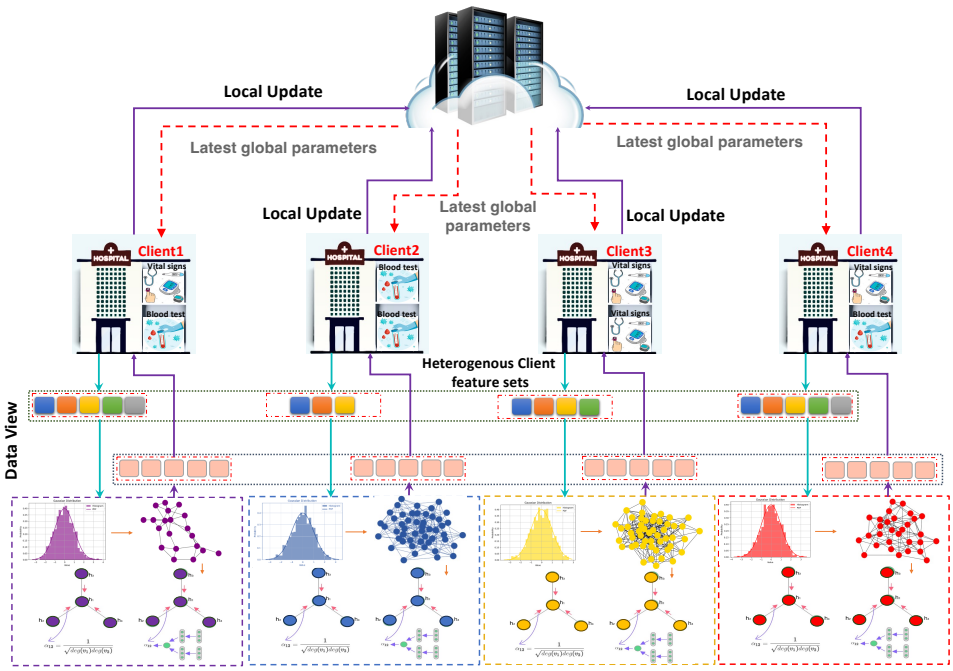


Figure 1: Illustration of the proposed federated learning framework designed to handle heterogeneous data views across multiple clients in a healthcare setting. A central server initialises global models graph attention networks and coordinates the training process. Each client employs data augmentation mechanism to harmonise different data-views and exploits Graph Attention Layers to perform effective local training as well as to generate gradient updates for the global model. The server aggregates these gradients and updates the global model.

3 PROPOSED METHOD

The proposed framework integrates federated averaging with AGAT and attention aggregation to process EHR datasets. Figure 1 depicts a systematic illustration of the proposed framework. We also introduce a dynamic variant of the proposed framework to capture temporal dynamics of the EHR datasets with distinct data-views.

3.1 Notation and Problem definition

The proposed FL framework considers K clients that are assumed to participate in each training round. Each client can have the EHR dataset with distinct data-views. A centralised server initialises the global model and coordinates the training process across these clients.

The proposed framework can be applied in two settings: static and dynamic. In static setting, the EHRs are conceptualised as a static graph $G = (P, E)$, where P is the set of nodes representing individual patients and E is the set of edges connecting them. Each node has associated features, denoted as $F_{P_i} = \{f_{(1,P_i)}, f_{(2,P_i)}, \ldots, f_{(M,P_i)}\}$, that include a diverse array of medical and demographic details such as age, medical history, and current conditions. The dimensions of these features could vary for different clients. This static setting can be seen as analogous to working with *tabular data*.

In dynamic setting, we introduce a dynamic graph G_d, which is comprised of a series of static graph snapshots. Specifically, $G_d = \{G^{(1)}, G^{(2)}, \ldots, G^{(T)}\}$ for T time steps. At each time step t, the sets of nodes and edges are represented by $P^{(t)}$ and $E^{(t)}$, respectively. These sets are allowed to evolve over time to accommodate potential changes in either the node or edge sets. Although the dimensionality of the features associated with each node may differ across clients, it is assumed to remain constant across various snapshots for each individual client.

3.2 Server-side Processing

The server decides the architecture of global augmented graph attention (AGAT) network and ini-

tialises the parameters $\boldsymbol{\theta}$. At this stage, it assumes that each client has a fixed or consistent data-view with n features. During each round of training, the server invokes the ith client to perform local training and provide gradients $\nabla_{\boldsymbol{\theta}_i}$. The gradients provided by each client are aggregated using attention scores to update the global model $\boldsymbol{\theta}$ as:

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \eta \sum_{i=1}^{K} \beta_i \nabla_{\boldsymbol{\theta}_i}, \tag{1}$$

Here β_i is the attention score for ith client and η is the learning rate to update global model $\boldsymbol{\theta}$.

3.3 Client-side Training

During each round of training, the ith client receives the global parameters from the server, and it initialises its local model using these parameters i.e. $\boldsymbol{\theta}_i = \boldsymbol{\theta}$. The client performs local training using the local dataset for multiple epochs (irrespective of the data-view) to obtain updated local model parameters $\boldsymbol{\theta}'_i$. Then, the gradient update for global model can be computed as:

$$\nabla_{\boldsymbol{\theta}_i} = \boldsymbol{\theta}_i - \boldsymbol{\theta}'_i. \tag{2}$$

This gradient update is sent to the server for updating global model parameters.

3.3.1 Augmented Graph Attention Network

Augmentation Layer: We introduce a layer before the attention layers (of graph attention networks) to align the data-views or feature dimension across all clients. To achieve this, we generate an "additional" feature representation for each patient that is augmented or concatenated with the original features to obtain the cohesive n-dimensional representation as required by the design of the global model.

For a specific client C_k with an associated dataset D_k having m-dimensional features, we estimate the parameters $\mu_{(k)}$ and $\sigma^2_{(k)}$ for the normal distribution of the features of nodes in D_k. New $(n-m)$ features for node P_i are then generated using the equation:

$$S^{\text{new}}_{(P_i,k)} \sim \mathcal{N}(\mu_{(k)}, \sigma^2_{(k)}) \tag{3}$$

Subsequently, we construct the new n-dimensional feature set for node P_i by concatenating these newly generated features with the existing ones, as formulated below,

$$F^{\text{new}}_{(P_i,k)} = S^{\text{new}}_{(P_i,k)} \,||\, F_{(P_i,k)} \tag{4}$$

Constructing Graph: As discussed earlier, each patient forms a node of the graph. We leverage the K-Nearest Neighbour (K-NN) to identify the neighbourhood ($\mathcal{N}$) of each node where nodes are represented by n-dimensional features (Dong et al., 2011). Euclidean distance is used as a metric for identifying the nearest neighbours. Two nodes P_i and P_j with feature representations $F^{\text{new}}_{P_i}$ and $F^{\text{new}}_{P_j}$ are connected by an edge in graph $G = (P, E)$ if $P_i \in \mathcal{N}_{P_j}$ (where $\mathcal{N}_{P_j}$ represents the neighbourhood of P_j) or vice-versa.

Attention Layers: After the augmentation layer, AGAT network consists of attention layers. The input comprises a graph with N nodes corresponding to N patients where each node is represented by new features $\{F^{new}_{P_1}, F^{new}_{P_2}, ..., F^{new}_{P_N}\}, F^{new}_{P_i} \in \mathbb{R}^n$. Each attention layer generates a new set of patient or node features as its output $\mathbf{H} = \{H_{P_1}, H_{P_2}, ..., H_{P_N}\}, H_{P_i} \in \mathbb{R}^{n'}$, which then serves as the input for the subsequent layer. The forward propagation for each patient or node P_i in layer l is given by,

$$H^{(l+1)}_{p_i} = \textsc{Sigma}\left(\sum_{u \in \mathcal{N}_{P_i}} \alpha_{P_u} \mathbf{W}^{(l)} H^{(l)}_u\right), \tag{5}$$

where $\mathcal{N}_{P_i}$ is the set of neighbours of P_i, $\mathbf{W}^{(l)}$ is the weight matrix for layer l, Sigma is the activation function, and α_{P_u} is the attention score calculated as,

$$\alpha_{P_u} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{a}^\top [\mathbf{W}H_{P_i} || \mathbf{W}H_u]\right)\right)}{\sum_{v \in \mathcal{N}_{P_i}} \exp\left(\text{LeakyReLU}\left(\vec{a}^\top [\mathbf{W}H_{P_i} || \mathbf{W}H_v]\right)\right)} \tag{6}$$

Here, the attention mechanism is a single-layer feed-forward neural network, parameterised by a weight vector $\vec{a} \in \mathbb{R}^{2n}$, and applying the LeakyReLU non-linearity. Where $.^\top$ represents transposition and $||$ is the concatenation operation. After attention layers, the node or patient's embedding is mapped to predict outputs using fully-connected layers.

3.4 Dynamic Graph Learning

Graph attention mechanisms are often designed to work in static settings where the graph structure is fixed over time. When the graph is dynamic, time becomes another crucial aspect to consider (Veličković et al., 2018; Oskarsson et al., 2023). In the context of discrete dynamic graphs, changes can occur at specific time steps $t = \{1, 2, \ldots, T\}$, where edges and nodes may be added or removed (Skarding et al., 2021; Zhang et al., 2023). To adapt our framework for a discrete dynamic graph, we extend the architecture to take into account the dynamic nature of the graph as shown in Figure 2. To this end, we add a time-dependent embedding to our modelling. A time-dependent vector is added to the existing node features, providing a glimpse of what the graph looks like at the previous time step. The idea is to use node features at the

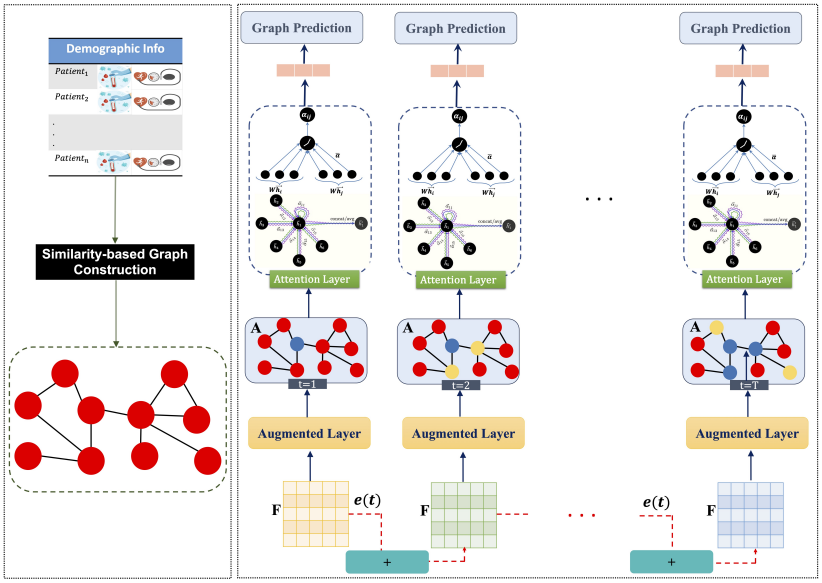


Figure 2: Schematic representation of the Temporal Augmented Graph Attention Network (AGAT) within a federated learning framework, emphasising the integration of dynamic graph learning, attention aggregation, and the processing of Electronic Health Records (EHR) datasets.

previous time step to generate a temporal representation that can be added to the current node features. For any ith node P_i, we create a temporal embedding $e_{P_i}(t)$ at time-step t by feeding the node's features $F_{P_i}(t-1)$ from the previous time step $t-1$ through a feed-forward network or multi-layer perceptron as shown in Equation 7.

$$e_{P_i}(t) = \text{MLP}(F_{P_i}(t-1)) \tag{7}$$

The temporal embedding $e_{P_i}(t)$ is then added to the existing node features F_{P_i} for the current time step t to generate updated node features $F'(t)$.

$$F'_{P_i}(t) = F_{P_i} + e_{P_i}(t) \tag{8}$$

Finally, we compute the attention scores using the updated node features.

4 EXPERIMENTATION

4.1 Datasets Used

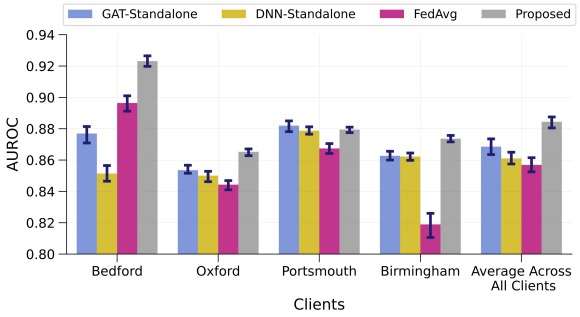
The proposed framework is evaluated on the following healthcare datasets:

- **CURIAL Datasets:** The CURIAL datasets (Soltan et al., 2021, 2022) consist of anonymised electronic health record (EHR) data (including demographic information, blood tests, and vital signs) from emergency departments (EDs) across four independent United Kingdom (UK) National Health Service (NHS) Trusts i.e. Bedford, Oxford, Portsmouth and Birmingham Trusts. These datasets are used for the binary classification task of diagnosing COVID-19. Among four trusts, Bedford and Portsmouth do not contain blood gas panel tests resulting in data-view heterogeneity. As a result, patients at Bedford and Portsmouth are represented by 21-dimensional feature vectors, while other Trusts use 28-dimensional vectors to represent patients.

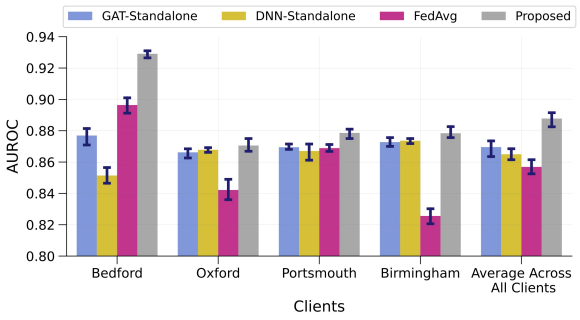
- **PhysioNet 2012 Dataset** (Silva et al., 2012): This dataset deals with in-hospital mortality prediction based on the first 48 hours of ICU stay. Each ICU stay is represented by a time-series with 48 time steps (separated by 1 hour). Each time-step is represented by a 44-dimensional feature representation. This dataset contains 8000 time-series examples that are randomly divided into 5 "hypothetical" clients for enabling federated learning.

- **eICU Collaborative Research Database:** eICU (Pollard et al., 2018; Tang et al., 2020) is a large multi-centre dataset containing 164,333 patients' ICU stays as time-series samples. We use a pre-processed version (Tang et al., 2020) of this dataset for predicting *shock* in the next 4 hours. As a result, each ICU stay is represented by a time-series containing 4 time-steps where each time-step is represented by the 375-dimensional feature (mainly vital signs and demographic features).

  For federated settings, we select 15 centres or hospitals containing the maximum number of examples.

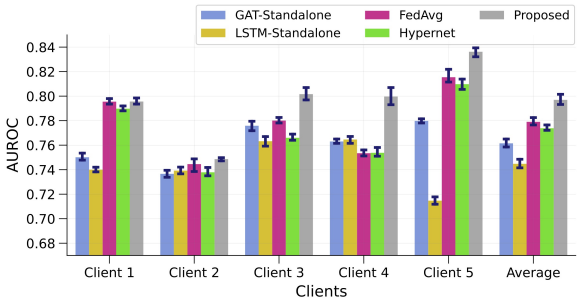


(a) Standard scenario

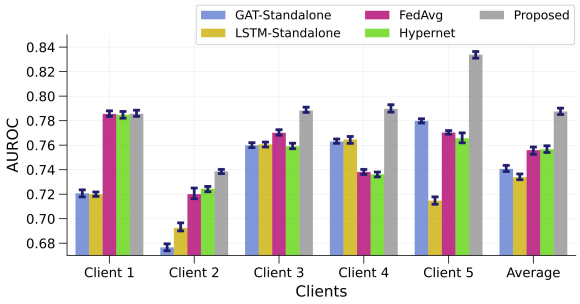


(b) Heterogeneous data-views

Figure 3: Performance of the proposed framework against different baselines on *CURIAL* datasets.



(a) Standard scenario



(b) Heterogeneous data-views

Figure 4: Performance of the proposed framework against different baselines on *PhysioNet 2012 challenge* datasets.

4.2 Designed Experiments

We compare performance of the proposed federated learning framework against the standard FedAvg (McMahan et al., 2017), hypernetwork-based FL (Shamsian et al., 2021) and standalone or client-specific models (only trained on a client's data). For baselines, we use three-layered DNN for CURIAL and LSTM-based binary classifiers for both PhysioNet and eICU datasets. These models are deployed as standalone baselines as well as employed in FedAvg and hypernet-based FL for federated baselines.

We design two different scenarios to evaluate the proposed framework and baselines:

- *Consistent data-views:* This scenario is characterised by consistent data-views across all clients. In case of CURIAL, we manually aligned data-views across NHS trusts by selecting common subset of features across all trusts.
- *Heterogeneous data-views:* In these scenarios, the data-views across clients are forced to be different. In CURIAL, there is an inherent heterogeneity between data-views across different Trusts. For PhysioNet and eICU, we randomly remove a significant number of features from clients to induce heterogeneity.

  As FedAvg can only work on consistent data-views, we manually aligned features to make data-views consistent before FedAvg.

Train-test distribution & performance evaluation: 60%, 15% and 25% of examples at each client are used for training, validation and testing. The testing is performed at each client after personalising the global model (after clients' local training). The area under ROC curve (AUROC) is used as a performance metric for all experiments.

The number of examples at each client, hyperparameters and the details regarding the number of features available to each clients (in heterogeneous scenario) are provided in the supplementary document.

5 RESULTS & DISCUSSION

5.1 Performance on CURIAL Datasets

Figure 3 illustrates the performance of the proposed framework and comparative baselines on four different NHS clients. The analysis of this figure highlights the following:

- In case of heterogeneous data-views, the proposed framework is able to perform either better or comparable to the standalone DNN and GAT. This shows that the proposed framework is able to overcome dif-

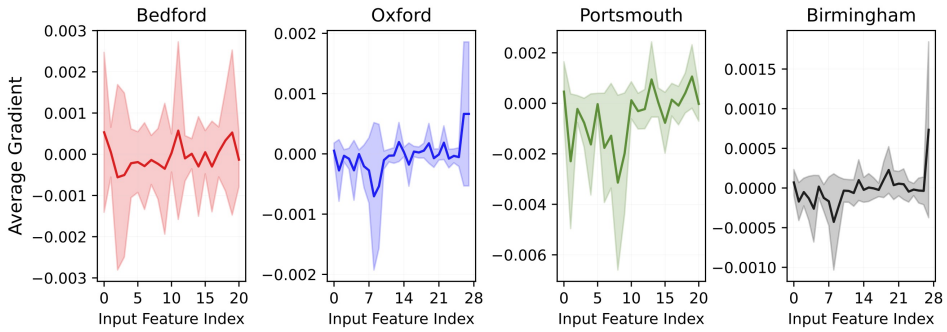


Figure 5: Average gradient of inputs with respect to trained model outputs at each client. The first 21 features are same across all clients. Oxford and Birmingham have additional 7 features.

Table 1: Performance of different comparative method in consistent data-view scenario on the eICU dataset.

Client	Methods				
	GAT-standalone	LSTM-standalone	FedAvg	Hypernet	Proposed
1	0.611±0.05	0.598±0.08	0.679±0.11	0.623±0.03	**0.681±0.15**
2	0.579±0.02	0.582±0.13	**0.614±0.19**	0.609±0.04	0.611±0.12
3	0.538±0.17	0.483±0.10	0.573±0.09	0.521±0.07	**0.588±0.06**
4	0.577±0.14	0.537±0.03	**0.692±0.20**	0.636±0.16	0.684±0.05
5	0.601±0.10	0.589±0.02	**0.702±0.04**	0.651±0.12	0.697±0.07
6	0.578±0.11	0.572±0.14	0.581±0.15	0.575±0.01	**0.611±0.20**
7	0.581±0.13	0.549±0.11	0.611±0.07	0.582±0.09	**0.675±0.18**
8	0.655±0.19	0.561±0.04	0.716±0.10	0.579±0.14	**0.728±0.13**
9	0.645±0.01	0.612±0.05	0.691±0.06	0.629±0.21	**0.885±0.03**
10	0.654±0.12	0.676±0.16	**0.762±0.13**	0.692±0.18	0.761±0.14
11	0.628±0.20	0.610±0.09	**0.648±0.02**	0.608±0.11	0.647±0.10
12	0.634±0.03	0.645±0.12	0.689±0.14	0.671±0.05	**0.697±0.08**
13	0.541±0.09	0.545±0.18	0.542±0.17	**0.548±0.16**	0.547±0.21
14	0.71±0.15	0.717±0.07	0.723±0.08	0.689±0.13	**0.756±0.19**
15	0.654±0.07	0.638±0.10	0.734±0.16	0.611±0.06	**0.788±0.09**
Average	0.613±0.109	0.595±0.093	0.670±0.109	0.620±0.097	**0.707±0.115**

ferences in data-views and induce information transfer across clients. Moreover, a maximum relative performance boost of 4.5% is observed at Bedford (that has fewer number of training samples) over the standalone-GAT highlighting the information transfer from other clients to Bedford.

- The proposed framework exhibits a noticeable improvement over both federated baselines i.e. FedAvg and Hypernet. The lesser performance of FedAvg can be attributed to the loss of features at some clients to align data-views. Moreover, better performance of the proposed framework can further be attributed to intrinsic graph structures resulting in better modelling of patients relationship. The performance of Hypernet is either comparable or lesser than standalone model confirming the lesser information transfer among clients.

- In standard or consistent data-view scenario, the performance of the proposed framework is either comparable or better that all baselines. This shows that the proposed framework can work well with both consistent and inconsistent data-views.

5.2 Performance on PhysioNet and eICU Datasets

The efficacy of our method is unequivocally evident when evaluated on the benchmark datasets of PhysioNet and eICU, as depicted in Figure 4 and detailed in Tables 1 and 2. The performance trends on these datasets are also pretty similar to CURIAL. Remarkably, the consistency in performance trends across PhysioNet, eICU, and CURIAL datasets underscores our model's robust adaptability and resilience across diverse datasets. Beyond mere comparability, our framework frequently demonstrates superiority over other baselines across all five client evaluations. This strength is further illuminated by the superior average AUROC across all 15 clients of the eICU dataset, where our method surpasses both standalone and federated baselines in scenarios encompassing heteroge-

Table 2: Performance of different comparative method in heterogeneous data-view scenario on the eICU dataset.

Client	Methods				
	GAT-standalone	LSTM-standalone	FedAvg	Hypernet	Proposed
1	0.571±0.015	0.363±0.018	**0.639±0.019**	0.429±0.017	0.634±0.012
2	0.511±0.011	0.561±0.014	0.564±0.016	**0.569±0.012**	0.517±0.019
3	0.521±0.016	0.451±0.013	**0.557±0.017**	0.487±0.015	0.548±0.011
4	0.557±0.012	0.457±0.019	0.54±0.018	0.511±0.016	**0.695±0.014**
5	0.576±0.017	0.381±0.015	0.68±0.013	0.531±0.014	**0.696±0.019**
6	0.564±0.019	0.552±0.012	0.559±0.015	0.565±0.013	**0.607±0.016**
7	0.572±0.012	0.544±0.018	0.599±0.017	0.552±0.011	**0.695±0.012**
8	0.651±0.014	0.501±0.015	0.706±0.019	0.489±0.016	**0.698±0.018**
9	0.637±0.018	0.439±0.016	0.681±0.012	0.561±0.017	**0.865±0.013**
10	0.587±0.015	0.573±0.011	0.529±0.013	0.582±0.019	**0.779±0.016**
11	0.608±0.017	0.310±0.012	0.628±0.015	0.532±0.014	**0.647±0.018**
12	0.521±0.019	0.555±0.013	0.469±0.016	**0.565±0.012**	0.529±0.011
13	0.532±0.014	0.530±0.016	0.531±0.019	**0.548±0.017**	0.539±0.015
14	0.697±0.012	0.717±0.018	0.577±0.014	0.679±0.016	**0.776±0.013**
15	0.592±0.018	0.596±0.011	0.41±0.012	0.576±0.019	**0.781±0.017**
Average	0.579 ± 0.015	0.502 ± 0.015	0.587 ± 0.016	0.545 ± 0.015	**0.685 ± 0.015**

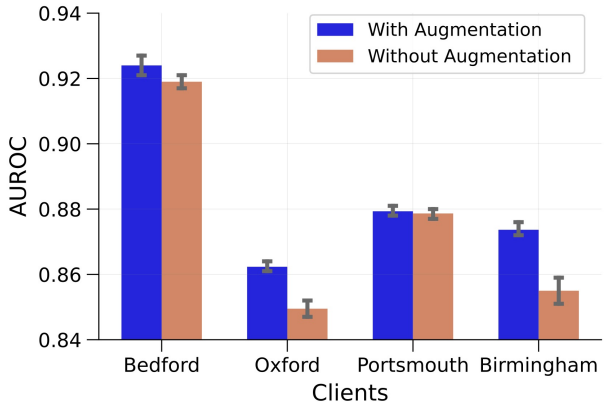


Figure 6: Impact of augmentation mechanism on the performance of the proposed framework on CURIAL datasets

neous and standard data views. Such results validate the model's adeptness at handling the intricate nuances of heterogeneous time-series data, affirming its capability to offer sophisticated and effective temporal modelling in the ever-evolving domain of Electronic Health Records analytics.

5.3 Impact of Augmentation Mechanism in Proposed Framework

The augmentation mechanism alleviates requirement of feature alignment in cases where clients have heterogeneous data-views. Apart from minimising pre-federated learning processing requirements, this mechanism also allows clients to fully utilise the available features. To analyse the impact of augmentation mechanism on performance, we removed augmentation layer from the proposed framework and aligned features manually across four CURIAL clients. Figure 6 illustrates the relative drop in performance of the proposed framework on removing augmentation layer. Since the number of features (i.e. 21) at Bedford and Portsmouth remain same after manual feature alignment, we do not witness any noticeable performance drop. However, at Oxford and Birmingham, we needed to drop some features (from 28 to 21) for data-view alignment. As a result, we witness a significant performance drop at both sites. This highlights that the augmentation mechanism allows the proposed framework to exploit all available features at clients resulting in better performance.

To further prove this fact, we compute gradient-based input feature importance (i.e. gradient of input with respect to model outputs) in each personalised or client-specific model. Figure 5 illustrates average gradient-based feature importance in each model. The first 21 features at each client are same, whereas Oxford and Birmingham has 7 more features (indexed between 21 and 28). The analysis of Figure 5 shows that at Oxford and Birmingham, additional features (that are absent at other clients) are more important for COVID-19 prediction. Again, this shows that rather than discarding additional features for data-view alignment, the proposed framework can exploit them for improving predictive modelling.

6 CONCLUSION

This paper introduced augmented temporal graph attention networks for handling heterogeneous data-views in federated learning within the context of

healthcare informatics. An exhaustive comparative analysis was conducted to show that the proposed framework can overcome data-views heterogeneity and exhibits either better or comparable performance against centralised and federated baselines. The successful evaluation of the proposed framework opens up new avenues for federated learning in healthcare that were earlier remained distant due to massive preprocessing requirements for data-view alignment. Future work will deal with extending the proposed framework by exploiting integrative federated methodologies to elevate performance further.

Code

The implementation of the proposed method is available at `http://github.com/AnshThakur/FL4HeterogenousEHRs`.

Acknowledgments

David Clifton was supported by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; and the InnoHK Hong Kong Centre for Centre for Cerebro-cardiovascular Engineering (COCHE).

References

Ash B Alpert, Jamie E Mehringer, Sunshine J Orta, Emile Redwood, Tresne Hernandez, Lexis Rivers, Charlie Manzano, Roman Ruddick, Spencer Adams, Catherine Cerulli, et al. Experiences of transgender people reviewing their electronic health records, a qualitative study. *Journal of General Internal Medicine*, 38(4):970–977, 2023.

Muhammad Atif, Muhammad Shafiq, and Friedrich Leisch. Applications of monitoring and tracing the evolution of clustering solutions in dynamic datasets. *Journal of Applied Statistics*, 50(4):1017–1035, 2023.

Jinyin Chen, Jian Zhang, Xuanheng Xu, Chenbo Fu, Dan Zhang, Qingpeng Zhang, and Qi Xuan. E-lstm-d: A deep learning framework for dynamic network link prediction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(6):3699–3712, 2019.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10351–10375. PMLR, 2022.

Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*, pages 577–586, 2011.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Jia Li, Zhichao Han, Hong Cheng, Jiao Su, Pengyun Wang, Jianfeng Zhang, and Lujia Pan. Predicting path failure in time-evolving graphs. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1279–1289, 2019.

Franco Manessi, Alessandro Rozza, and Mario Manzo. Dynamic graph convolutional networks. *Pattern Recognition*, 97:107000, 2020.

Gonzalo Mateos and Ketan Rajawat. Dynamic network cartography: Advances in network health monitoring. *IEEE Signal Processing Magazine*, 30 (3):129–143, 2013.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(3):1–37, 2022.

Ghazaleh Niknam, Soheila Molaei, Hadi Zare, Shirui Pan, Mahdi Jalili, Tingting Zhu, and David Clifton. Dyvgrnn: Dynamic mixture variational graph recurrent neural networks. *Neural Networks*, 165:596–610, 2023.

Joel Oskarsson, Per Sidén, and Fredrik Lindsten. Temporal graph neural networks for irregular data. In *International Conference on Artificial Intelligence and Statistics*, pages 4515–4531. PMLR, 2023.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman,

Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1): 119, 2020.

Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, and Hao Yang. Dysat: Deep neural representation learning on dynamic graphs via self-attention networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 519–527, 2020.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pages 9489–9502. PMLR, 2021.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology*, pages 245–248. IEEE, 2012.

Joakim Skarding, Bogdan Gabrys, and Katarzyna Musial. Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *IEEE Access*, 9:79143–79168, 2021.

Andrew AS Soltan, Samaneh Kouchaki, Tingting Zhu, Dani Kiyasseh, Thomas Taylor, Zaamin B Hussain, Tim Peto, Andrew J Brent, David W Eyre, and David A Clifton. Rapid triage for covid-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. *The Lancet Digital Health*, 3(2):e78–e87, 2021.

Andrew AS Soltan, Jenny Yang, Ravi Pattanshetty, Alex Novak, Yang Yang, Omid Rohanian, Sally Beer, Marina A Soltan, David R Thickett, Rory Fairhead, et al. Real-world evaluation of rapid and laboratory-free covid-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening. *The Lancet Digital Health*, 4(4):e266–e278, 2022.

Andrew AS Soltan, Anshul Thakur, Jenny Yang, Anoop Chauhan, Leon G D'Cruz, Phillip Dickson, Marina A Soltan, David R Thickett, David W Eyre, Tingting Zhu, et al. Scalable federated learning for emergency care using low cost microcomputing: Real-world, privacy preserving development and evaluation of a covid-19 screening test in uk hospitals. *medRxiv*, pages 2023–05, 2023.

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934, 2020.

Anshul Thakur, Pulkit Sharma, and David A Clifton. Dynamic neural graphs based federated reptile for semi-supervised multi-tasking in healthcare applications. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1761–1772, 2021.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, pages 1–18, 2018.

Wenshuo Wang, Xu Li, Xiuqin Qiu, Xiang Zhang, Jindong Zhao, and Vladimir Brusic. A privacy preserving framework for federated learning in smart healthcare systems. *Information Processing & Management*, 60(1):103167, 2023.

Siyue Yang, Paul Varghese, Ellen Stephenson, Karen Tu, and Jessica Gronsbell. Machine learning approaches for electronic health records phenotyping: a methodical review. *Journal of the American Medical Informatics Association*, 30(2):367–381, 2023.

Kaike Zhang, Qi Cao, Gaolin Fang, Bingbing Xu, Hongjian Zou, Huawei Shen, and Xueqi Cheng. Dyted: Disentangled representation learning for discrete-time dynamic graph. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3309–3320, 2023.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Not Applicable
   (b) Complete proofs of all theoretical results. Not Applicable
   (c) Clear explanations of any assumptions. Not Applicable

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. Yes

(b) The license information of the assets, if applicable. Not Applicable

(c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable

(d) Information about consent from data providers/curators. Not Applicable

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. Not Applicable

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

A Dataset Details

Tables A1, A2 and A3 documents the number of examples at each client in CURIAL, PhysioNet and the eICU dataset.

Table A1: Number of examples at each NHS trust in CURIAL dataset.

	Bedford	Oxford	Portsmouth	Birmingham
# Examples	1,865	161,955	38,717	95,236
# Positive Examples	210	2,791	2,005	790
# Features	21	28	21	28

Table A2: Number of examples at each "hypothetical" site in PhysioNet dataset.

	Client 1	Client 2	Client 3	Client 4	Client 5
# Examples	800	800	800	800	800
# Positive Examples	105	102	129	106	112
# Features	20	44	32	44	40

Table A3: Hospital IDs used as clients and the number of examples at each client in the eICU dataset.

	Client 1	Client 2	Client 3	Client 4	Client 5	Client 6	Client 7	Client 8	Client 9	Client 10	Client 11	Client 12	Client 13	Client 14	Client 15
Hospital ID	388	301	280	390	272	152	244	226	310	79	154	318	146	444	252
# Examples	898	913	933	1012	1014	1028	1033	1057	1135	1149	1182	1191	1192	1197	2709
# Positive Examples	103	141	70	83	44	96	44	56	60	116	104	133	92	128	148
# Features	100	100	100	216	216	216	375	375	375	375	100	100	100	375	375

B Model Architectures

DNN Baseline For CURIAL:

Dense layer with 128 nodes → ReLU Activation → Dropout with 0.25 rate → Dense layer with 1 node → Sigmoid Activation

LSTM Baseline For PhysioNet and eICU:

LSTM with 128 nodes → ReLU Activation → Dropout with 0.25 rate → Dense layer with 1 node → Sigmoid Activation

Graph Attention Network (GAT):

GATConv (128 nodes, 1 attention head) → ReLU Activation → Dropout with 0.25 rate → GATConv (1 node, 1 attention head) → Sigmoid Activation

Here GATConv stands for graph attention convolutional layer.

Temporal Augmented Graph Attention Network (Dynamic GAT) For Time-series: The framework architecture used for Dynamic GAT is illustrated in Figure A1.

Hypernetwork for Hypernet-based FL Baseline:

Hypernetwork is fully-connected DNN that is used to generated weights for DNN and LSTM models. Hypernet used by Shamsian et al. (2021) is also used in this work. Their implementation is available at `https://github.com/AvivSham/pFedHN`.

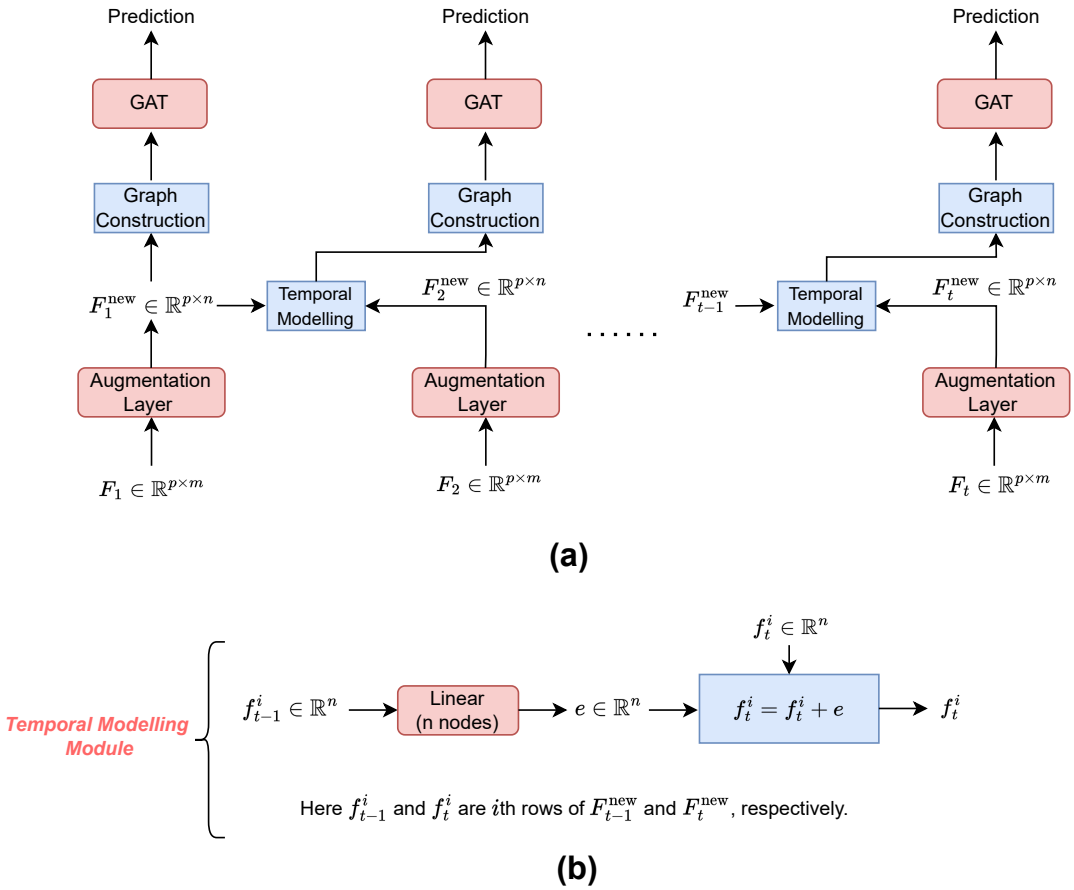


Figure A1: Illustration of dynamic GAT architecture used for time-series modelling. GAT architecture described earlier is also used here. Temporal modelling module is shared across all time-steps.

The input embedding of 32 dimensions, 2 shared hidden layers with 128 nodes and spectral norm on initialised weights is used across all datasets. The dense layers mapping hidden embedding (output of shared hidden layers) to output weights vary depending on whether weights are to be generated for DNN or LSTM.

C Parameter Setting

Centralised Baselines: Across both DNN/LSTM and GAT-based standalone baselines, we used the Adam optimiser with a fixed learning rate of 0.001. The models were trained for 300 epochs with early stopping to store best performing model configuration on validation examples. A batch-size of 64 examples is used across all datasets.

Federated Baselines and the Proposed Framework: Similar to standalone baselines, we use batch-size of 64 across all datasets. We train all federated frameworks for 300 rounds of training (client-server communication rounds). In each round, the local training is done for 2 epochs using Adam optimiser with learning rate of 0.001. For updating global model (at the server), a learning of 0.001 is used across both FedAvg and the proposed framework. In hypernet-FL, global hypernetwork is also trained using a learning rate of 0.001.

All the results are obtained with 5 different seeds, and mean along with standard deviation are presented in the main manuscript.

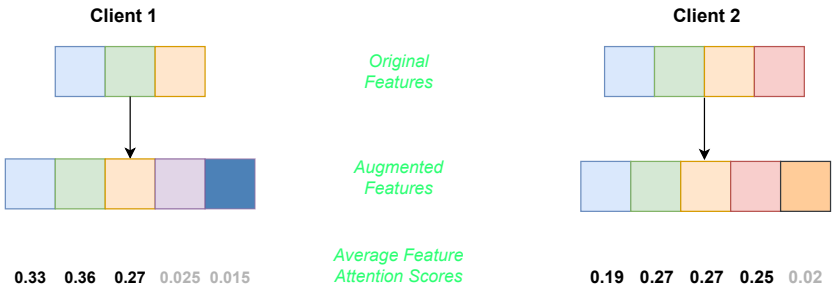


Figure A2: An illustration depicting how graph attention layers assign different attention scores to features in augmented data-views. Augmented features are expected to be given lesser attention scores than the original features.

D Navigating Data-view Heterogeneity using GAT

In federated learning with heterogeneous data-views, clients have differing dataset dimensions. For example, one client has features A and B, whereas another one has features A, B and C. To align the data-views across both these clients for enabling federated learning or global model training, we augment additional feature C' to the first client's data-view.

In this scenario, the GAT or graph attention mechanism has inherent capabilities to handle these additional features. GAT's attention mechanism is uniquely positioned to assign weights based on a feature's relevance. Specifically, it can recognise that C', being derived or related to A and B, may not hold the same intrinsic value as the original features, and thus, assign it a lower weight. This capability ensures that GAT can manage nuanced relationships between genuine and synthesised features across clients. Figure A2 graphically demonstrates this behaviour.

Similarly, *Dynamic GAT* is adept at modelling temporal changes in data relationships across clients with heterogeneous data-views. Graph structures, determined by nodes (like patients) and edges (similarities between them), might evolve over time with new data or changing relationships. Dynamic GAT's ability to handle these structural changes without needing complete retraining is a testament to its flexibility. As relationships (edges) between nodes change, the attention weights adjust accordingly, focusing on the most current and relevant information. This adaptability to shifting data relationships solidifies Dynamic GAT's position as a potent tool for federated learning scenarios, where data structures and relationships across clients might shift over time.