GH+ Labs

AI-enabled obstetric ultrasound (OBUS)

September 2025

Executive Summary

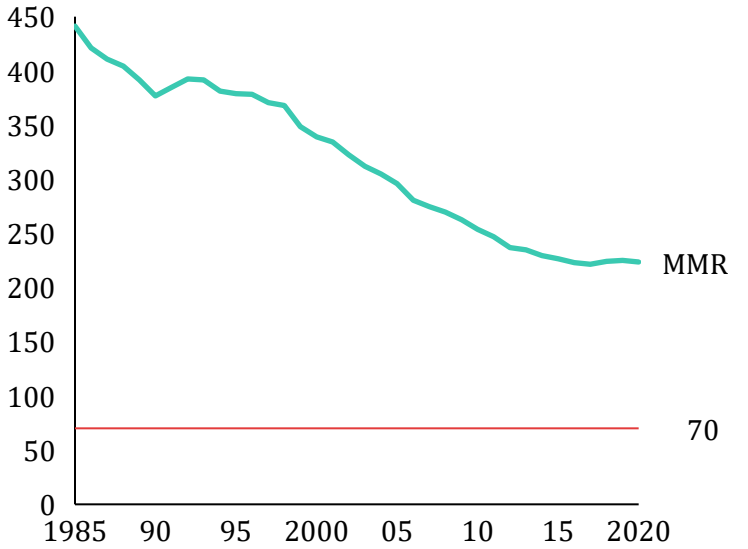
- GHL, in partnership with the Gates Foundation division for Maternal, Newborn, Child Nutrition & Health (MNCNH), has developed prototype AI models for automated interpretation of obstetric ultrasound blind sweep videos.
- Four obstetric AI models have been developed for the first phase, namely, prediction of gestational age (GA), fetal presentation (FP), fetal weight (EFW), and multiple gestation (TWIN). All of these meet internal product performance requirements.
- Code for data curation, as well as training and evaluation of these AI models along with model weights trained on FAMLI data, have been shared with GF partners.
- Critical next steps are for these partners to fine-tune these models with data from their own devices, integrate them into their device software, conduct clinical trials leading to regulatory approval, and scale up the products for deployment.
- Next steps could focus on the development of a few more obstetric features that were identified in consultation with key opinion leaders and care providers in the target markets.

Problem Statement - The burden of maternal remains unacceptably high. Neonatal mortality (death during the first 30 days of life) is still a challenge in LMICs.

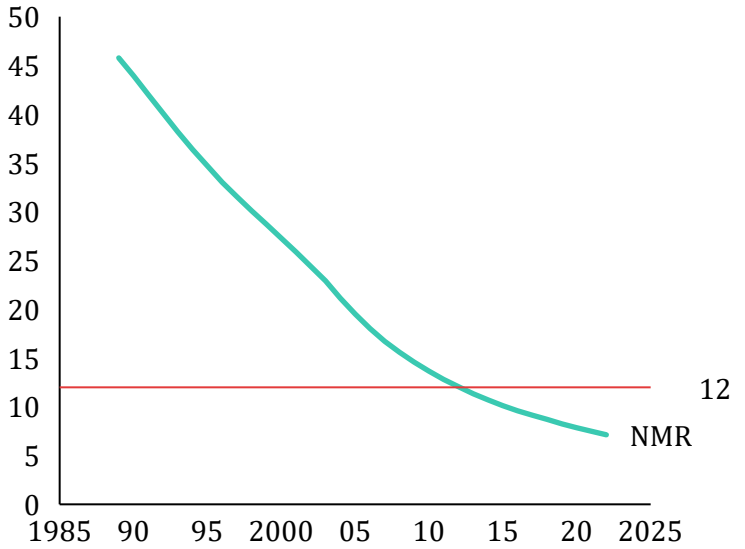
The global landscape has evolved significantly since 2000. While mortality rates have declined between 2000-2015 due to scientific advancements, political commitment, and economic growth, growth has stalled after 2015 driven by **funding constraints, global fragmentation, expanding vulnerable populations, and health system barriers**

Sustainable Development Goal (SDG)

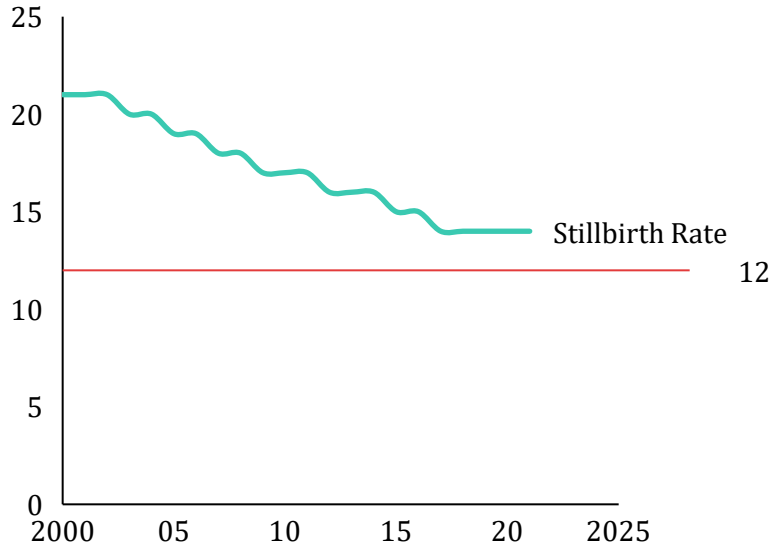
Maternal Mortality Rate[1], maternal deaths per 100,000 live births



Neonatal Mortality Rate[1], neonatal deaths per 1,000 live births

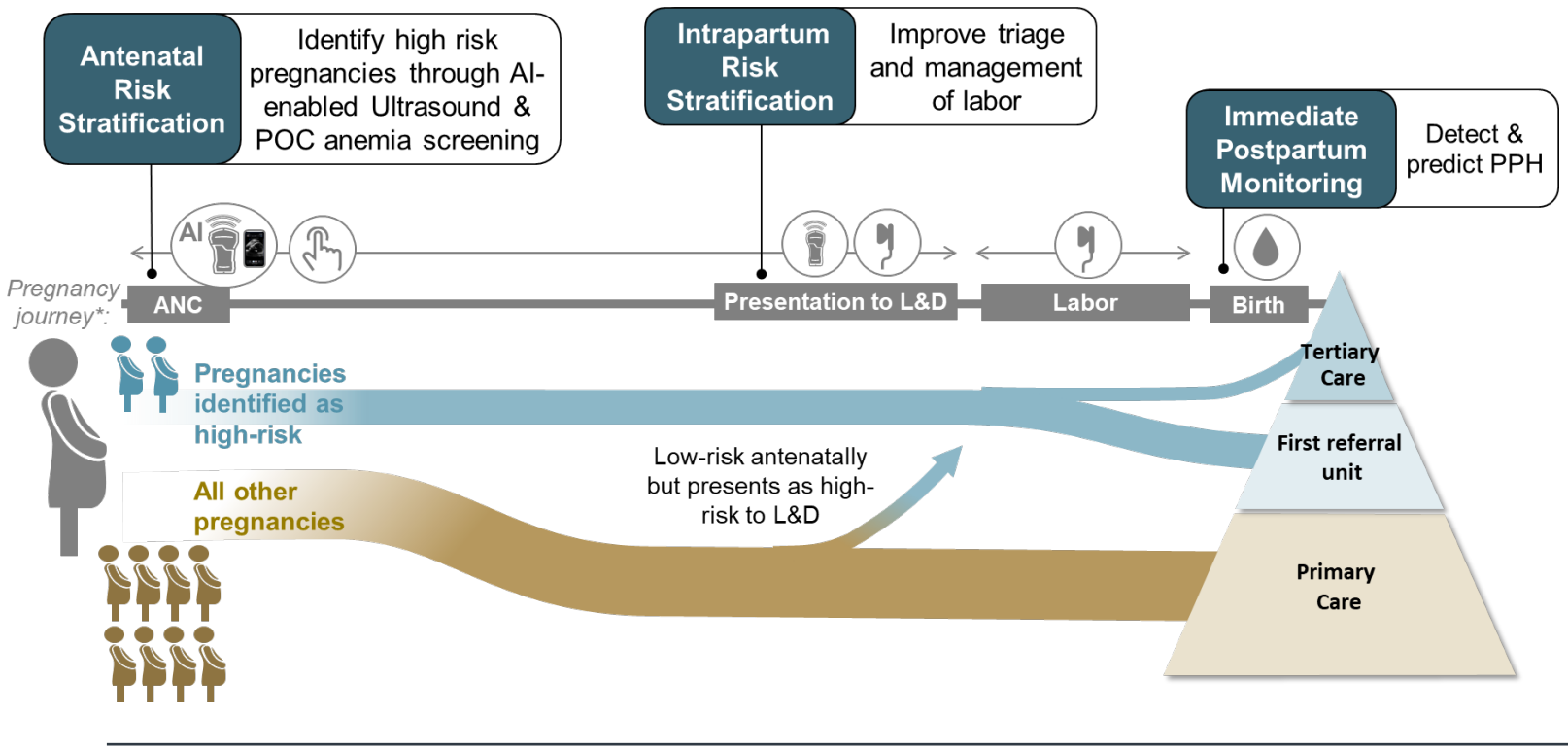


Stillbirth Rate[2], maternal deaths per 100,000 live births



1. WHO
2. UNICEF

GATES FOUNDATION MNCNH DEVICES & AI STRATEGY FOCUSED ON RISK STRATIFICATION



AI-enabled Point-of-Care Ultrasound (POCUS)

Intrapartum Sensors

Non-invasive Hemoglobin anemia screening device

PPH detection technologies

AI Ultrasound is key driver for impact in LMICs with 1.7M deaths and 145M DALYs averted by 2040

The main challenges in LMICs to overcome is the affordability for a ultrasound imaging tool and the lack of specially-trained healthcare workers

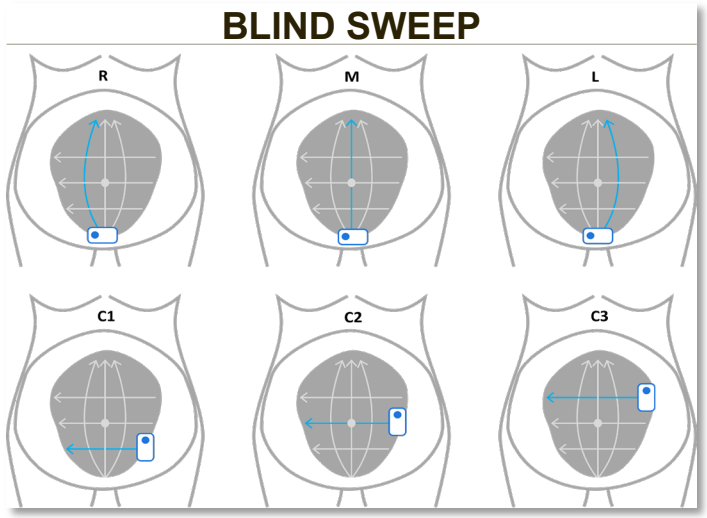
Kenya - only nurses and midwives available in most L2/L3 primary healthcare centers; currently only 25% mothers received at least 1 scan during pregnancy

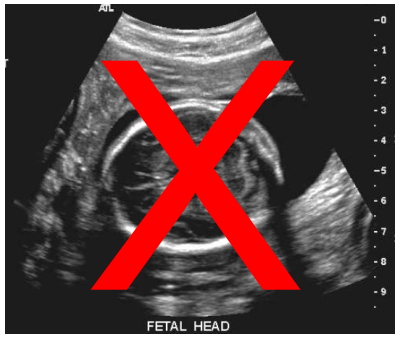
Ghana – just like this rough road, many factors prevent pregnant mothers in LMICs from travelling and seeking essential cares in upper-level facilities

Malawi – Only one U/S system in Nsanje District (Hospital) with 800K people (2022). Only a few radiologists available in the whole country.

Gates Foundation addresses these challenges by investing in Low-cost AI-powered Ultrasound innovations and enabling non-expert users

Enable Low Skilled Workers (nurses and midwives)





Blind Sweep

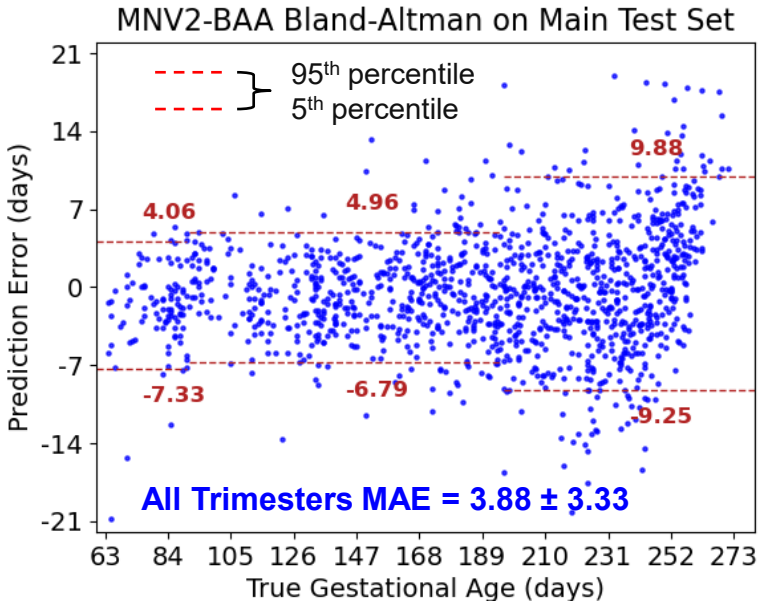
- Minimum training
- No image acquisition guidance
- No traditional U/S imaging display / interpretation

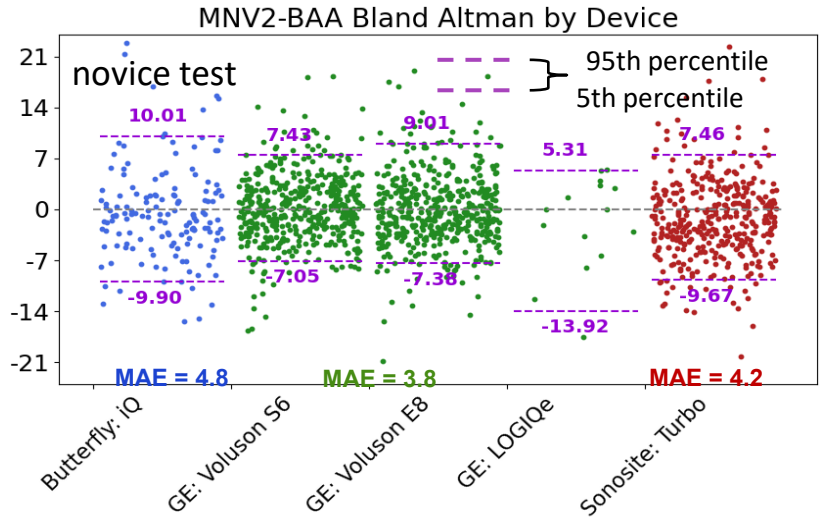
Offer Affordable and Accessible U/S Devices

Portable Device

- Handheld U/S with mobile/tablet computer
- AI on edge - no Cloud / mobile/tablet compatible
- FDA / CE mark approval to support in-country scale up

GH Labs played a key translational role by developing open AIOB algorithms and jump-starting partner development work





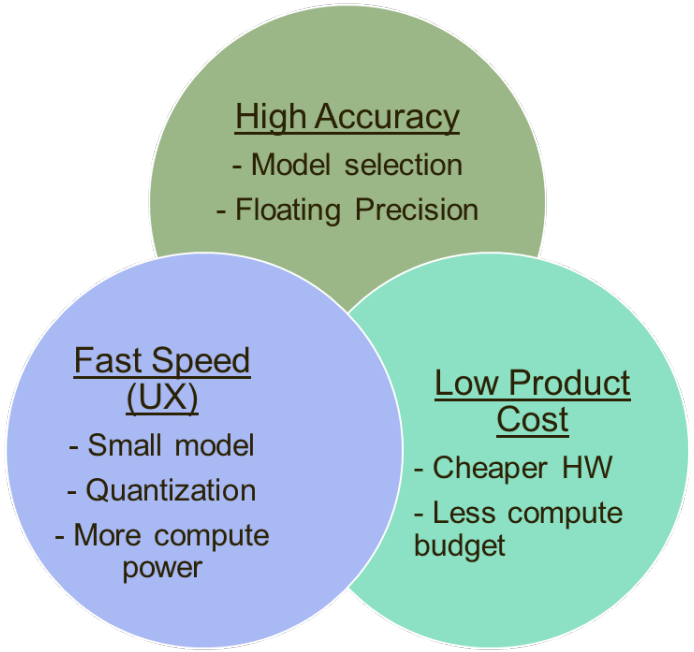
AI Blind Sweep Algorithms

1. Gestational age
2. Fetal Presentation
3. Estimated Fetal Weight
4. Multiple Gestation
5. Amniotic Fluid Volume
6. Placenta Location
7. Preeclampsia

GH+Labs

GHL developed 4 AI blind sweep algorithms (GA, FP, EFW, MG) with performance comparable to current U/S standard care

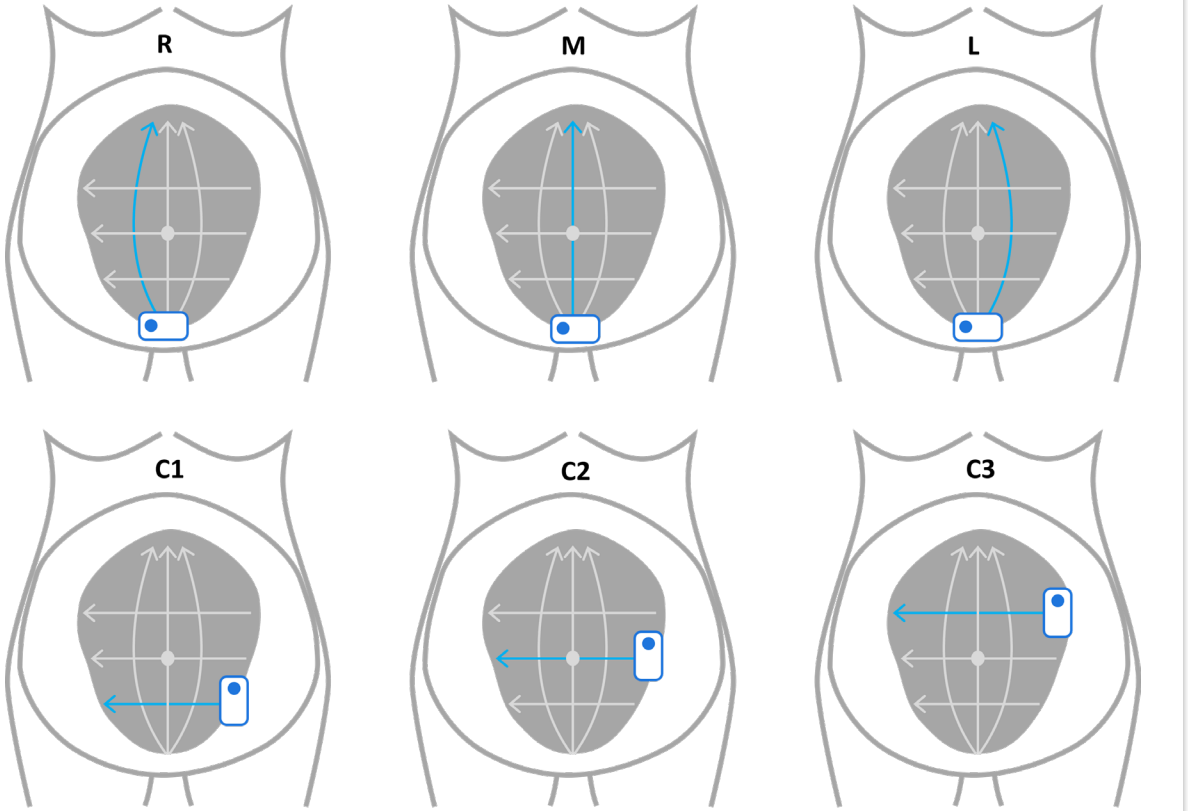
Development Partners / Grantees



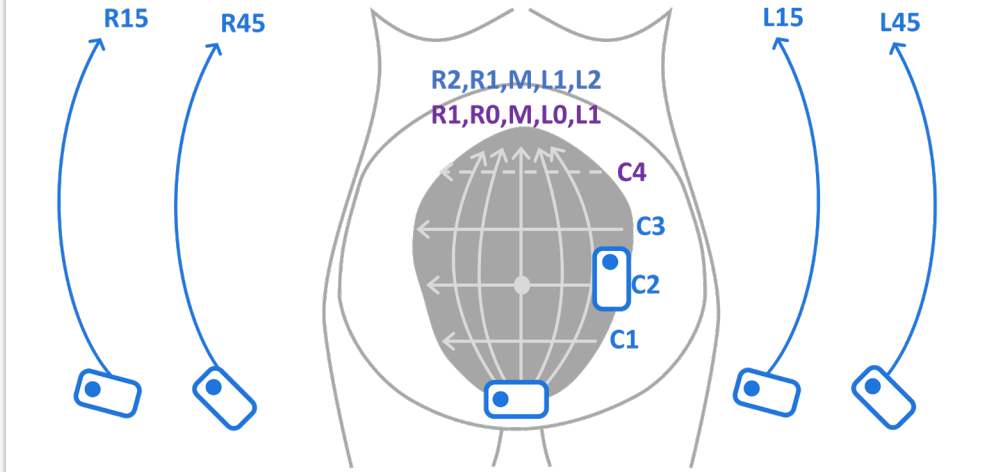
Product development trade-offs

Blind sweeps + AI enable ultrasound use by minimally trained users

BLIND SWEEP TYPES – NOMINAL



BLIND SWEEPS – ACTUAL



UNSPECIFIED SWEEP TYPES	
ASSBS	Assumed blind sweep
NaN or Unknown	Unknown sweep type

OTHER SWEEP TYPES	
GS, YS, CRL, BPD, HC, AC, FL, HL, TCD	Biometric scans
CA1, CA2, CA3, CA4, FHR	Fetal heart scans
LUQ,RUQ, LLQ, RLQ, EQ, 3Dxxx	3D volume scans
LOV, ROV	Ovaries
BREECH, CEPHALIC, OBLIQUE, TRV	Fetal presentation
DVP, Q1, Q2, Q3, Q4	Amniotic scans
xxxPLAC, CERVIX, PREVIA, FUNDUS	Maternal anatomy

FAMLI Datasets – curated and uncurated

Dataset	OBUS features	Description	Limitations	videos	valid videos	patients	exams
GA_NEJME	GA, FP	Curated subset of FAMLI2_enrolled	Uncomplicated singleton pregnancies; blind sweeps only; GA labels only—FP labels from FAMLI2_enrolled.	119,386	119,386	4,521	12,300
FAMLI2_enrolled	TWIN, EFW, DVP	Collected 2018-2022	Twins nominally excluded—but yet present. Required major curation effort for TWIN, EFW, DVP modeling.	406,164	382,456	6,205	21,835
FAMLI2	TWIN, EFW	Continuation 2022-2023	Same inclusion criteria as FAMLI2_enrolled. Curation needed.	26,418	24,633	321	1,817
FAMLI3	TWIN, EFW	Collected 2023-2025 expanded inclusion criteria	Twins explicitly included. Curation needed.	143,411	102,105	1,233	5,873

Various deep learning architectures can be used for video understanding

Category	Architecture	Phase 1 OBUS Features				Notes
		GA	FP	EFW	TWIN	
modular spatial encoding→ temporal aggregation →classifier or regressor	2D CNN — Attention — FC	✓	×	✓	×	Correlation between frames not important for GA, EFW. Current best candidate for DVP.
	2D CNN — MIL — FC				✓	Focuses attention on positive samples. Current best candidate for TWIN.
	2D CNN — Conv LSTM — FC		✓		×	Temporal correlation between frames via LSTM important for FP and potentially other features.
	2D Vision Transformer — LSTM — FC					Have started OB ultrasound foundation model experiments with this architecture.
unitary spatial encoding, temporal aggregation performed by one module	(2+1)D CNN — FC					Tried for DVP without much success.
	3D CNN — FC					Works almost as well for DVP as Attention.
	3D Vision Transformer — FC	×				Computationally intensive for video because it scales quadratically.
	3D Mamba — FC		×			Computationally feasible because it scales linearly.

Legend: × experiments, ✓ final model

Deep learning architectures for video analysis; strengths & deficiencies

How it works:

- 2D CNN computes frame embeddings
- Each frame's attention weight is based on non-linear function of its own frame embedding
- Video embedding is attention-weighted mean (or other function) of frame embeddings
- Final output is fully-connected projection of video embedding

Strengths:

- Inductive bias for images
- Mechanism to emphasize predictive frames individually
- Computationally efficient

Deficiencies:

- No mechanism to leverage interframe correlations
- Temporal patterns are ignored (frame reshuffle has no effect)

How it works:

- 2D CNN computes frame embeddings
- Each frame's attention weight is based on non-linear function of ALL frame embeddings
- Video embedding is attention-weighted mean (or other function) of frame embeddings
- Final output is fully-connected projection of video embedding

Strengths:

- Inductive bias for images
- Mechanism to emphasize predictive frames based on interframe correlations and dependencies
- Computationally efficient

Deficiencies:

- Temporal patterns are ignored (frame reshuffle has no effect)

How it works:

- 2D CNN computes frame embeddings
- At each time step (frame), LSTM decides which features to forget, which features to receive from current time step, and which features to output to next time step
- Video embedding is last temporal output of LSTM
- Final output is fully-connected projection of video embedding

Strengths:

- Inductive bias for images
- Leverages temporal patterns
- Computationally efficient

Deficiencies:

- No frame-level importance score
- Long-range dependencies can be lost (won't matter for 10-30 second video clips)

How it works:

- Video is considered a 3D volume (x, y, t) and divided into 3D patches (cubes), which are treated as tokens
- Patches are input to multiple layers of transformer encoders (each consisting of self-attention + skip connections + fully-connected layer)
- Video embedding is the global average (x, y, t) of the patch embeddings from the last encoder layer
- Final output is fully-connected projection of video embedding

Strengths:

- Powerful, non-local, spatial and temporal pattern analysis

Deficiencies:

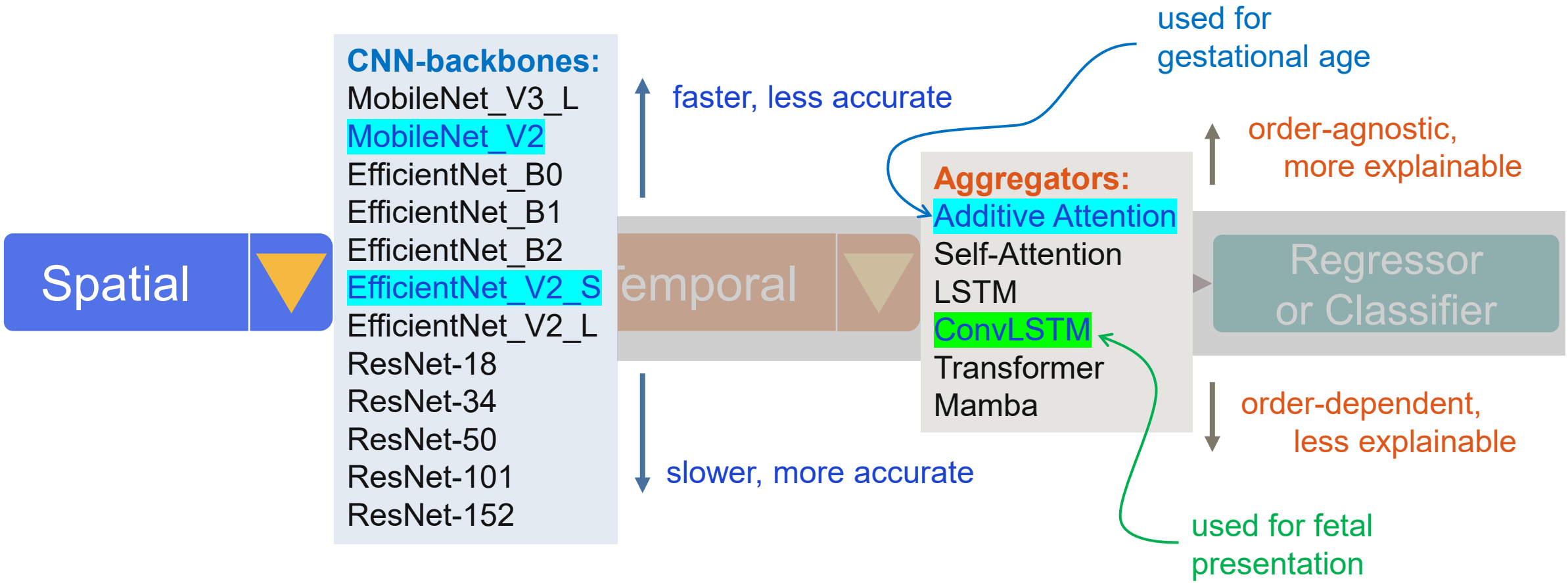
- Scales more than linearly with image resolution
- Requires lots of training data
- High computational and memory requirements

Modular deep learning architecture

Permits design choices to meet use case, performance requirements

Modular deep learning architecture

Permits design choices to meet use case, performance requirements



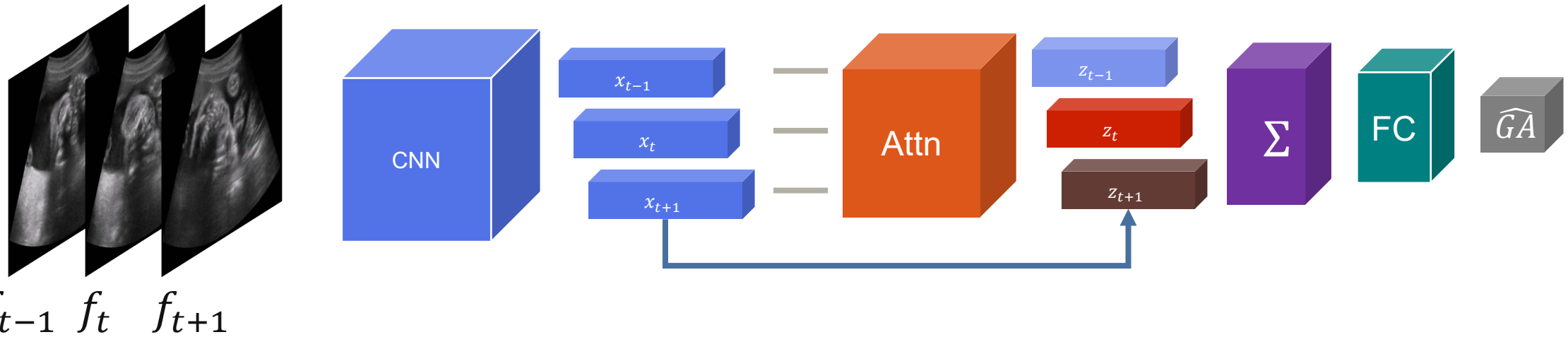
GA model uses additive attention

Model Architecture

2D CNN – Additive Attention – FC

Key Details

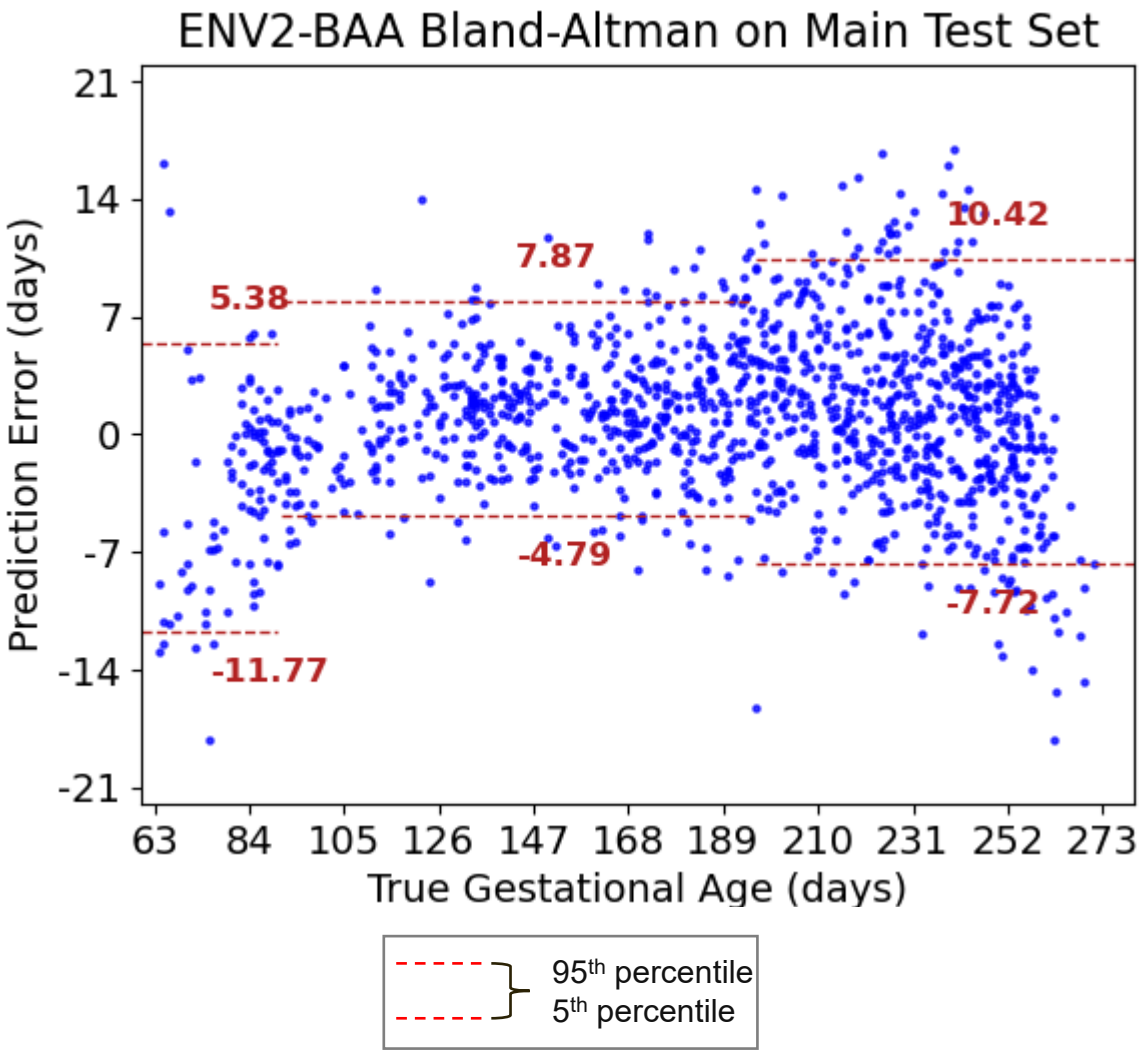
- Frame-weighting via attention mechanism
- Interframe dependencies ignored
  - Works for GA estimation
  - May not work for targets dependent on temporal correlations

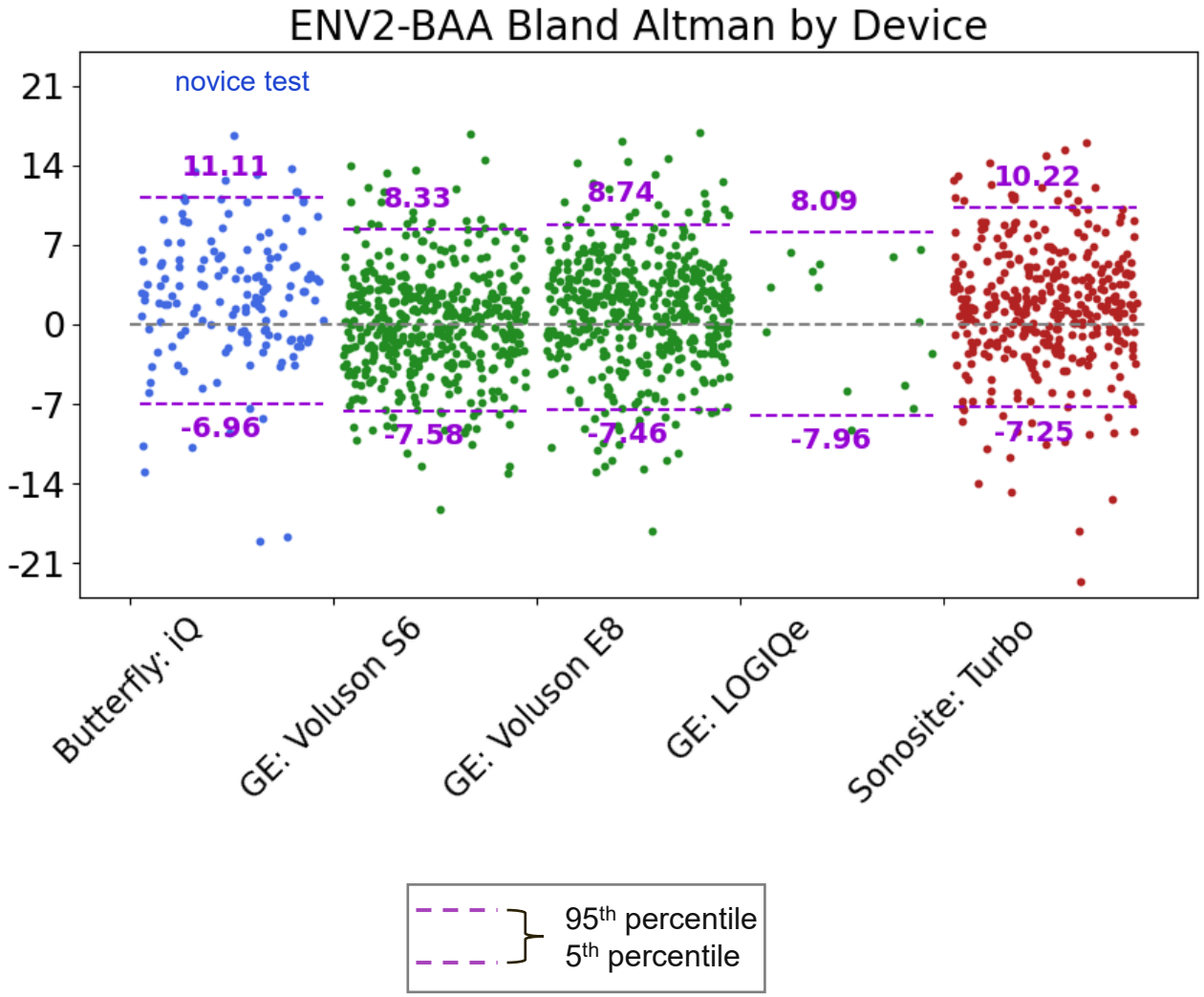


frames	CNN feature extractor	frame embeddings	temporal attention	attention-modulated embeddings	frame mean	regressor	GA estimate

GA model with EfficientNet_V2_S backbone

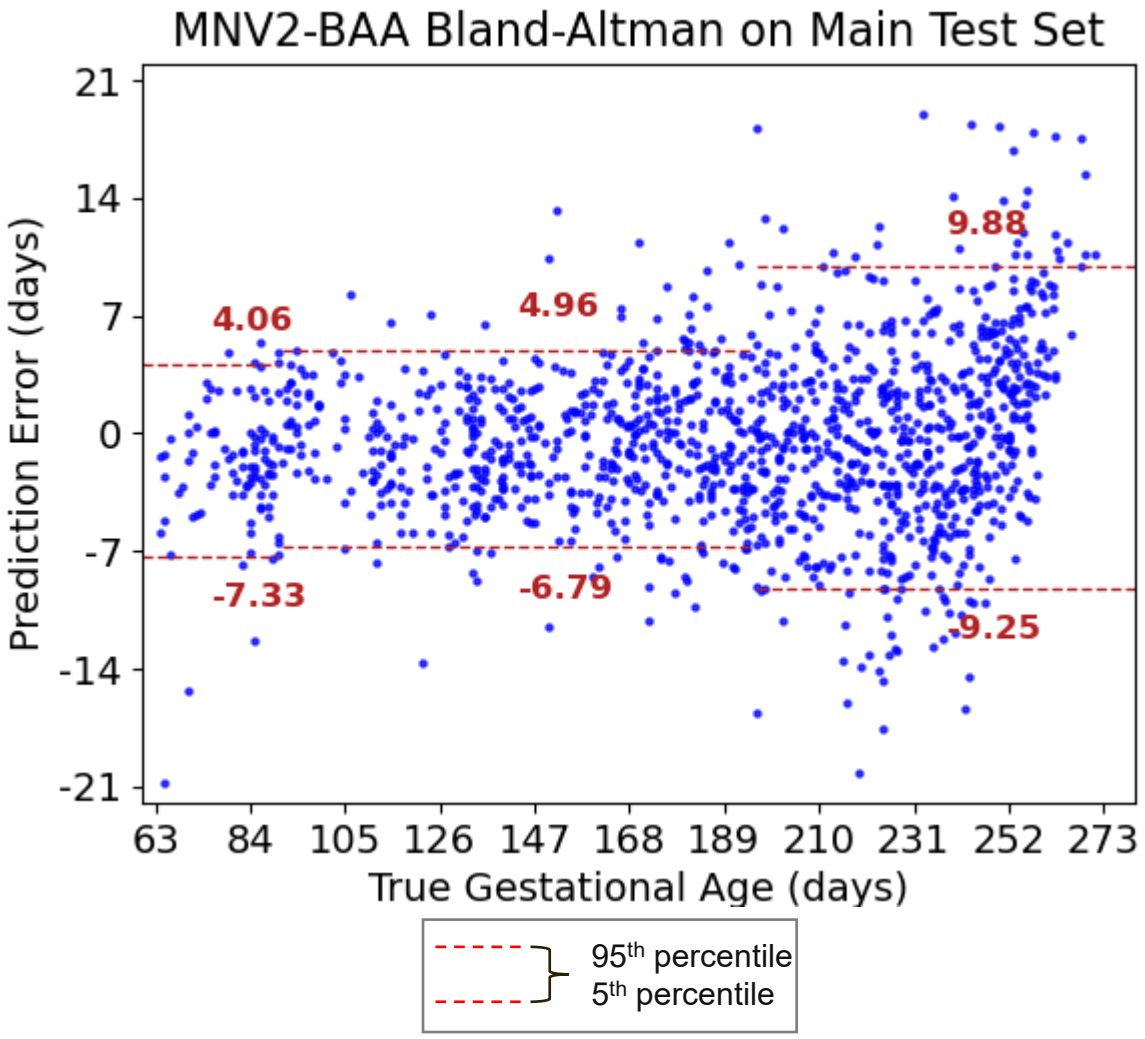
All Trimesters MAE = 3.94 ± 3.28

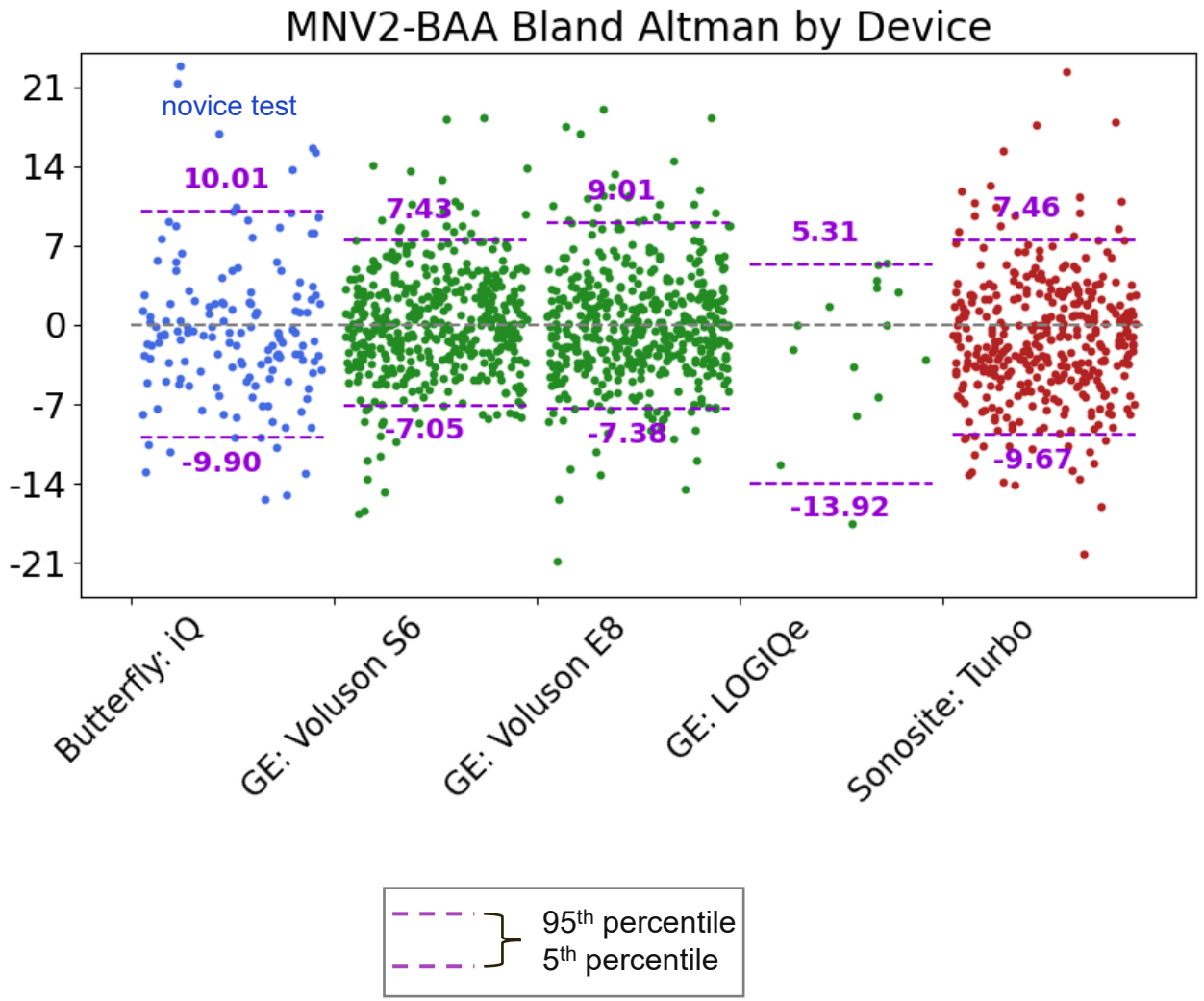




GA model with MobileNet_V2 backbone

All Trimesters MAE = 3.88 ± 3.33



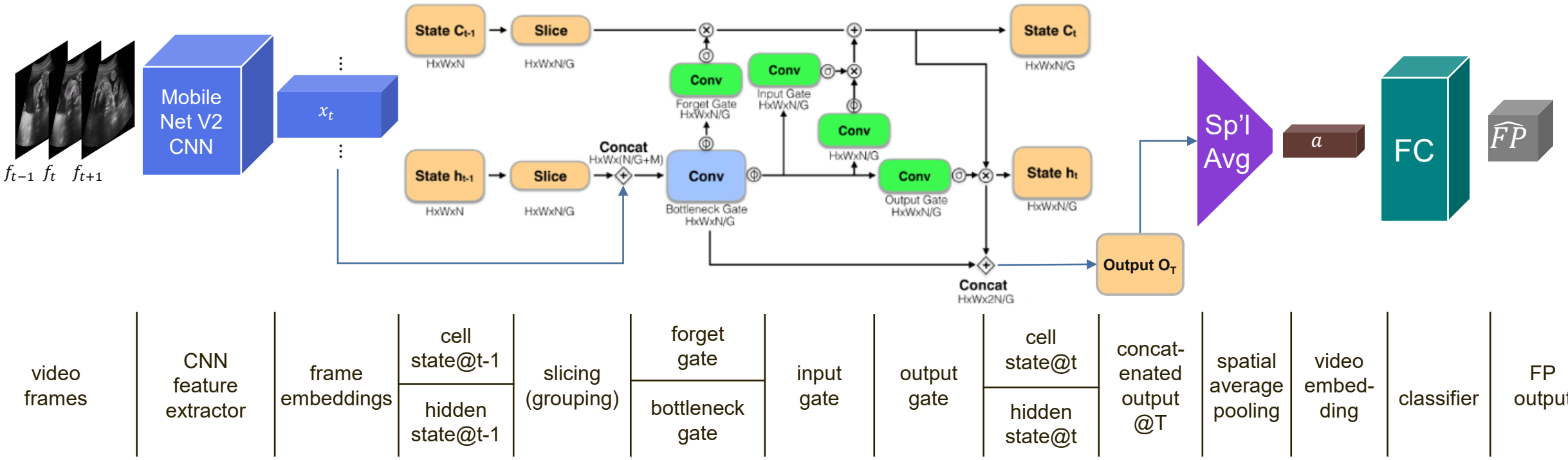


FP model uses grouped convolutional LSTM (ConvLSTM)

Model Architecture

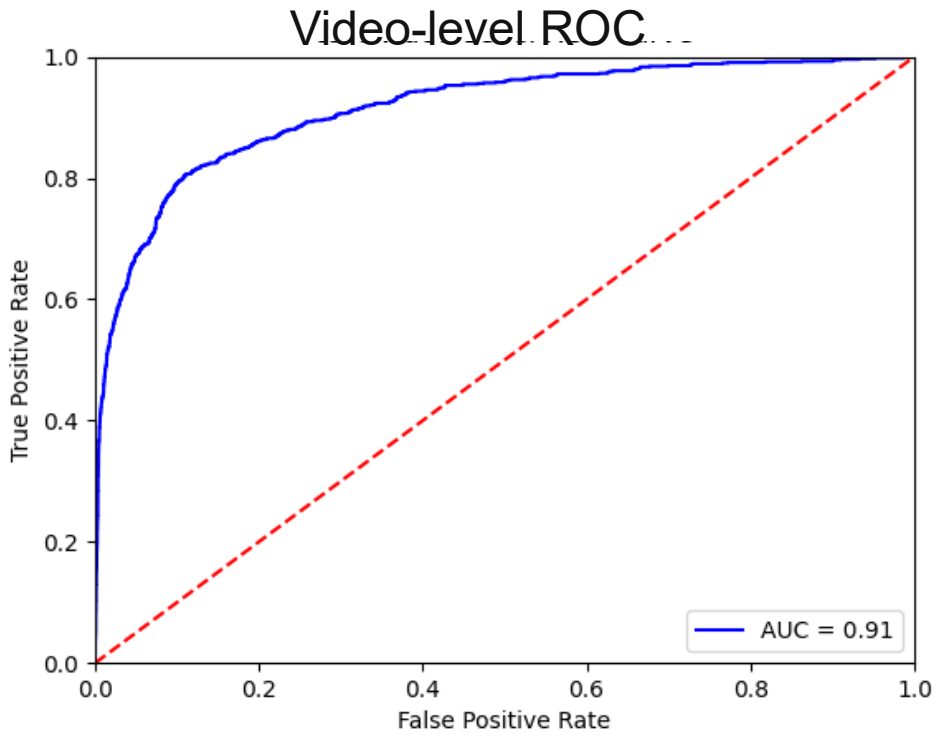
Key Details

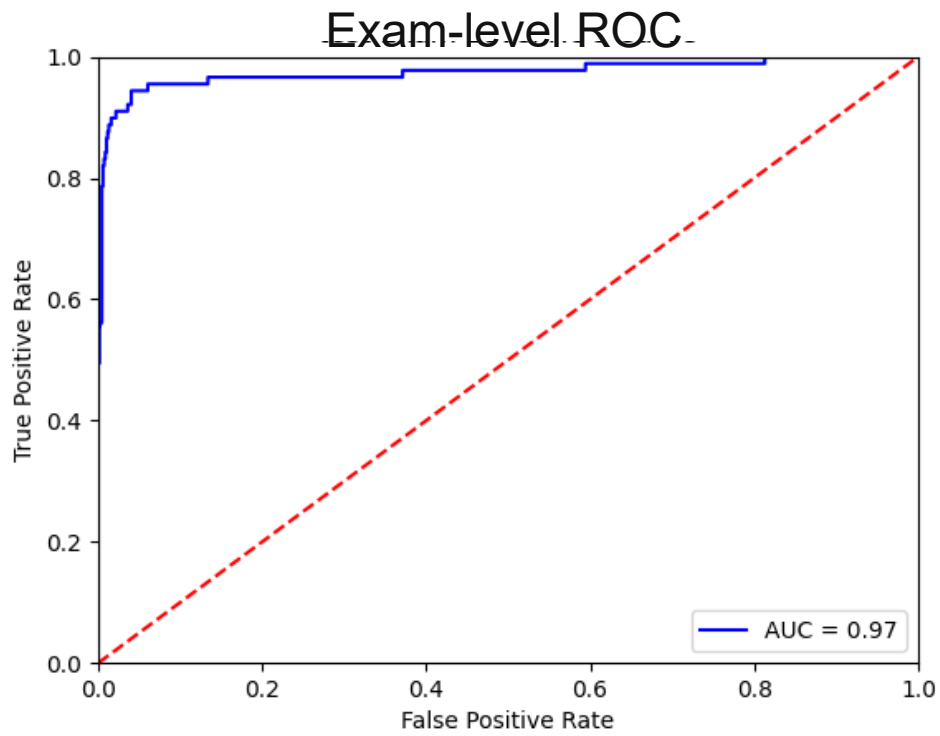
- Temporal aggregation via ConvLSTM
- Interframe dependencies modeled!
  - Necessary for FP classification



FP model with MobileNet_V2 backbone

	Exam-level results	Number of exams
Sensitivity	0.90	89 non-cephalic exams
Specificity	0.98	695 cephalic exams
AUROC	0.97	

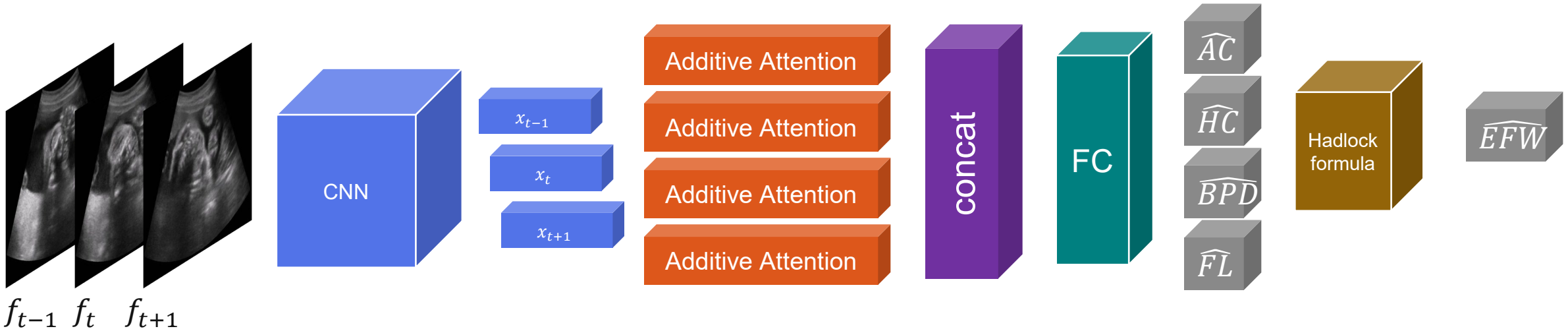




EFW model similar to GA model, with key differences

Model Architecture

2D CNN — Multiple Additive Attention — FC



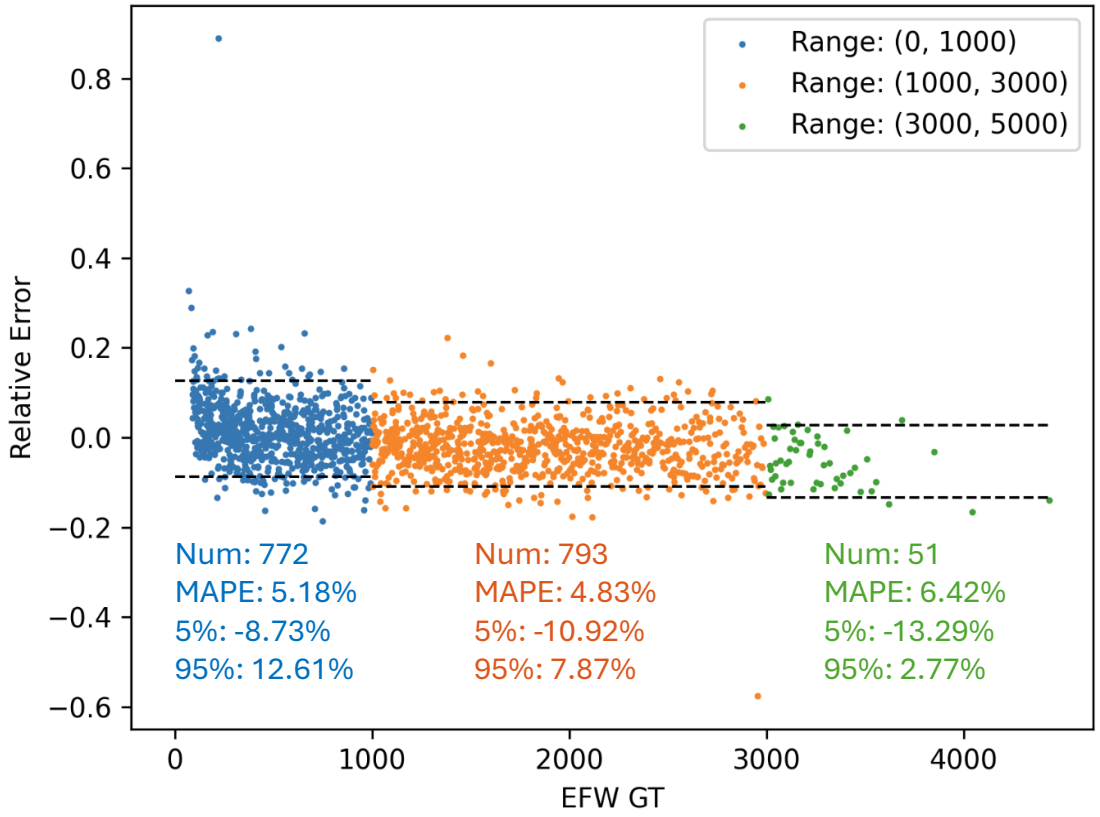
frames	CNN feature extractor	frame embeddings	multiple additive attention	concatentation of attention-modulated embeddings	multi-task regressor	biometric estimates	4-component Hadlock formula	EFW estimate

Key Details

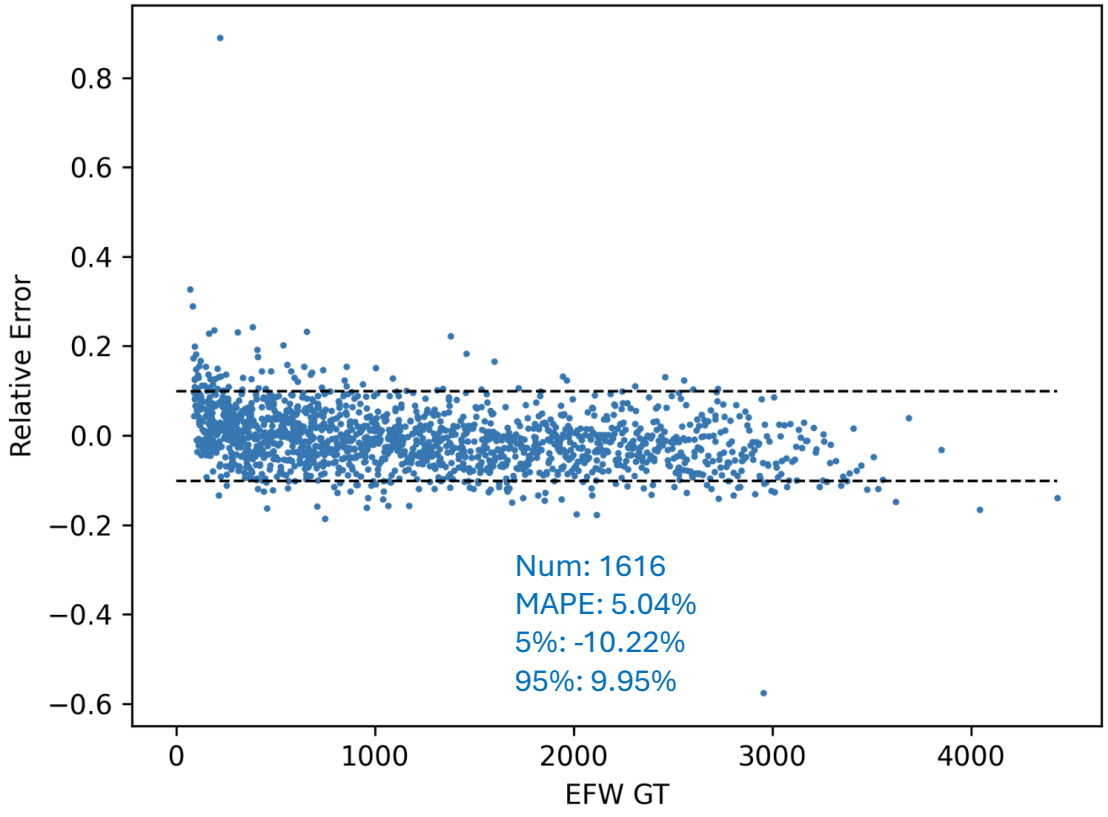
- Uses multiple additive attention modules
- Multi-task regressor predicts four biometric quantities: AC, HC, BPD, FL
- Four component Hadlock formula outputs estimated fetal weight from biometric inputs

EFW model performance

The evaluation results on data with EFW in different ranges



The evaluation results on all data



Multiple Gestation data overview

Data Sources: The Multiple Gestation dataset is constructed from three datasets: Famli2, Famli2_enrolled, and Famli3

Ground Truth Definition:

- Twin labels were present in case report forms and structured reports from data source (UNC).
- A small number of conflicting labels were referred to UNC for adjudication.
- V9 dataset contains **346 MG exams** from **105 twin patients** and **1 triplet**.

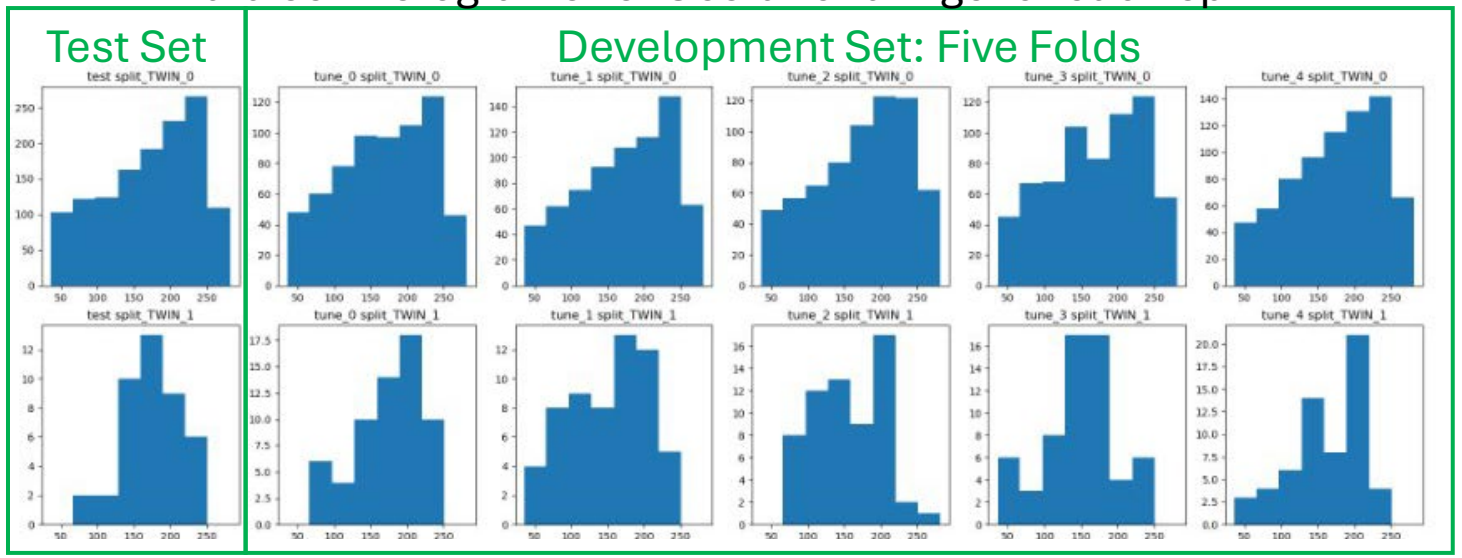
Data Splitting:

- The dataset is split into a development set (80%) and a holdout test set (20%).
- The development set was further split into five cross-validation folds.
- Sought to match the twin : singleton ratio and the overall GA distribution for the test set and each development fold.
- Ensured that exams from the same patient are assigned to the same dataset split.

Data set histograms vs. Gestational Age for each split

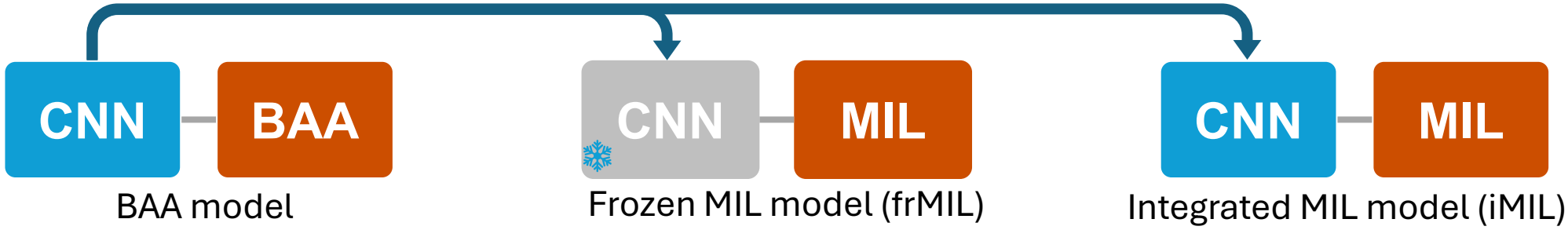
singletons →

twins →



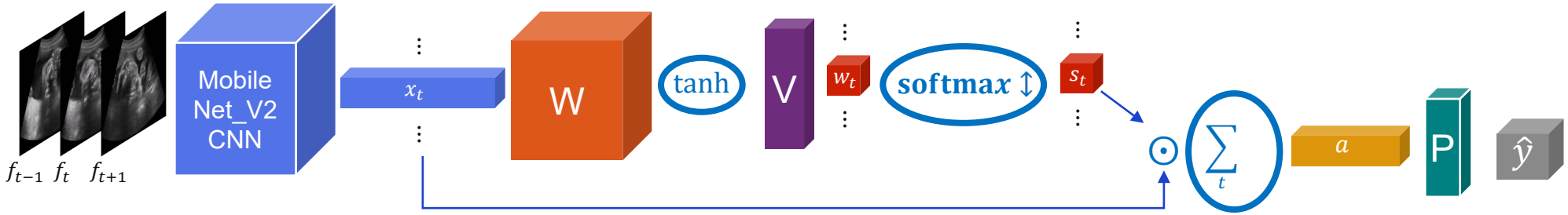
Multiple Gestation model architectures

- The models follow a **Spatial → Temporal → Classifier** pipeline
- Three twin-detection architectures have been developed:
  - **CNN → BAA (Basic Additive Attention):** This is essentially the same architecture used for the GA model. The CNN and BAA are trained concurrently at the video level, then evaluated at the exam level.
  - Two **Multiple Instance Learning (MIL)** versions:
    - **Frozen MIL (frMIL)** – a pre-trained CNN (from the BAA architecture above) is followed by an attention-based MIL network. Only the MIL weights are trained—at the exam level; the CNN weights are frozen.
    - **Integrated MIL (iMIL)** – a CNN (initialized with BAA weights) is followed by an attention-based MIL network. The CNN is initialized with the CNN from the BAA architecture, but the CNN and MIL weights are trained concurrently—again, at the exam level.



Multiple Gestation BAA model architecture

Additive Attention: 2D CNN – Additive Attention – FC

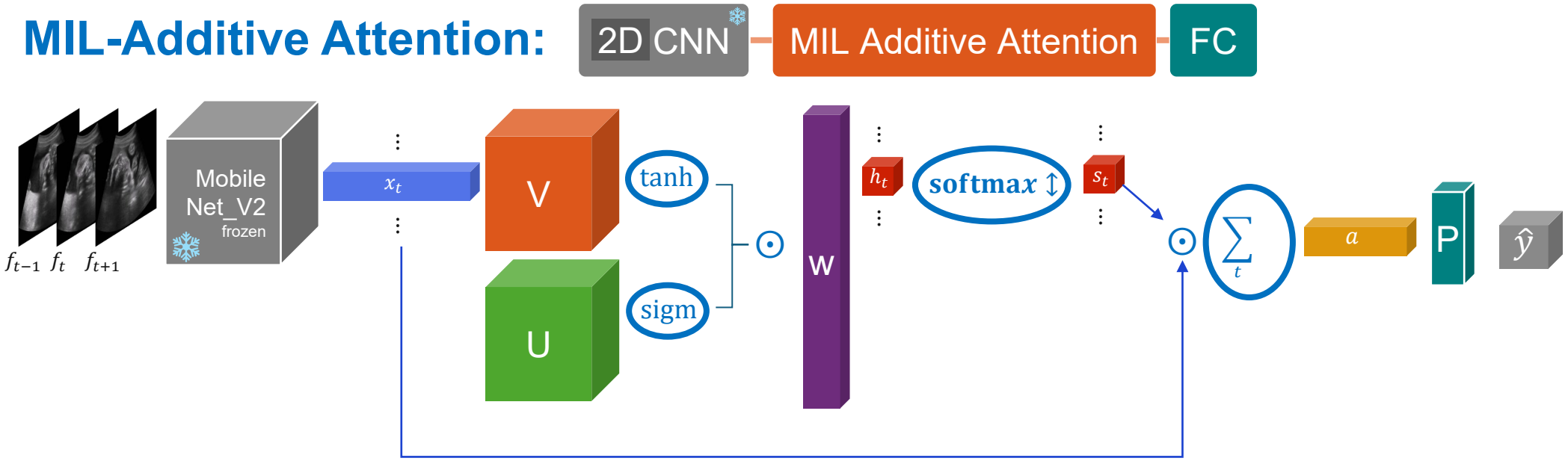


- Advantages
  - Similar architecture as used in GA and EFW models – greatly reduces development time.

- Specific hyperparameters/design choices:
  - Trained (and loss optimized) at video level: Not all videos/frames have twin evidence.
  - All frames used for each video at inference time: Not likely to miss any frames with twin evidence.
  - Augmentation applied at video level: Ensures consistency across frames.

Multiple Gestation frozen MIL architecture

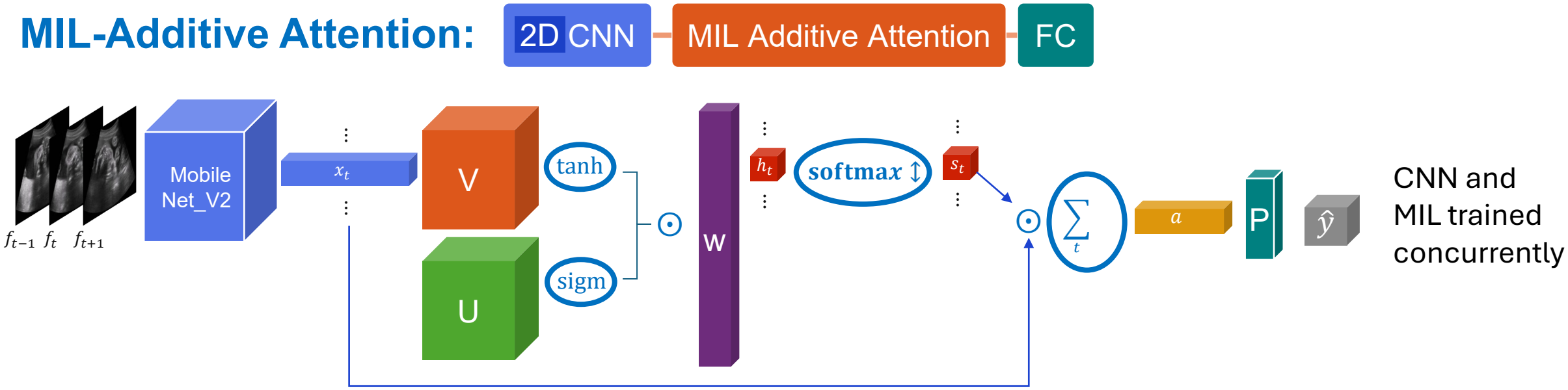
MIL-Additive Attention:



- Advantages
  - We know the CNN is extracting twin-relevant features (BAA model performed reasonably well).
  - MIL-Additive Attention is more expressive than Basic Additive Attention due to sigma channel.
  - Reduces training time – the MIL component is small relative to CNN.
- Specific hyperparameters/design choices:
  - Trained at **exam level**: Picks out predictive evidence across all frames of all exam videos.
  - 1000 frames used: Can use more frames in frozen architecture.

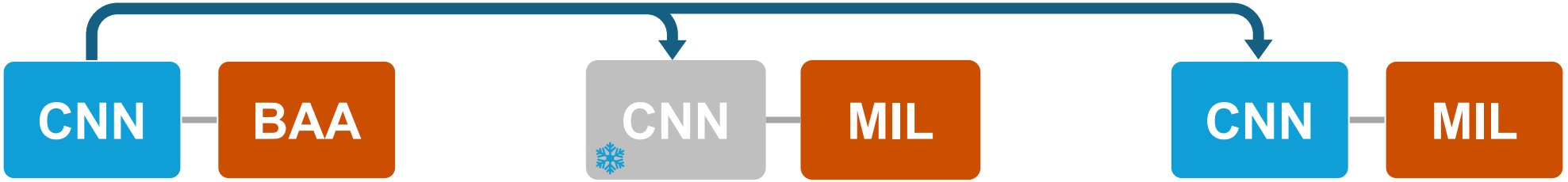
Multiple Gestation integrated MIL architecture

MIL-Additive Attention:



- Advantages over frozen MIL:
  - Fine-tuning CNN may generate frame embeddings more useful/specific for MIL.
  - Image augmentation can be used, which improves generalization. The frozen MIL approach didn't enable augmentation during MIL training.
- Specific hyperparameters/design choices:
  - Trained at **exam level**: Picks out predictive evidence across all frames of chosen videos in an exam.
  - 300 frames used: Balance between performance and speed/memory use during training.

Multiple Gestation performance summary



	BAA model	Frozen MIL model (frMIL)	Integrated MIL model (iMIL)
Single-fold Validation Result	Sensitivity 86.9% Specificity 86.4% AUC 0.938	Sensitivity 94.4% Specificity 90.0% AUC 0.965	Sensitivity 88.9% Specificity 95.3% AUC 0.964
All 5-folds Average Result	Sensitivity 80.4% Specificity 87.2% AUC 0.892	Sensitivity 86.4% Specificity 89.7% AUC 0.916	Sensitivity 85.7% Specificity 92.9% AUC 0.922
Test Result			Sensitivity 89.7% Specificity 94.6% AUC 0.951

Several iMIL models perform similarly on the test set

	Fold(s)	Threshold	Model ranking	AUC	Sens	Spec
'Best' Model	Ensemble 3+4, average scores	0.2	#1	0.951	89.7%	94.6%
Alternatives, ranked	Fold 3	0.2	#2	0.947	87.2%	94.6%
	Fold 4	0.2	#2	0.933	87.2%	95.7%
	Ensemble 0+3+4, average scores	0.2	#3	0.954	92.3%	94.8%
	Ensemble 0+3+4, majority vote	From each fold	#3	NA	84.6%	95.8%
	Fold 0	0.1	#3	0.942	82.1%	94.4%
	Fold 2	0.25	#4	0.945	87.2%	92.9%
	Ensemble 0+2+3+4, average scores	0.2	#4	0.961	92.3%	93.3%
Rejected models	Fold 1 (and any ensemble with it)	0.3	NA	0.930	87.2%	92.0%
Potential future work	Average weight models (3+4, 0+3+4, 0+2+3+4)					

- Consistency between models is reassuring.
- Ensembles have a slight advantage, but at cost of higher inference times.

References

+ Publications from UNC, Google, and Hadlock:

- Pokaprakarn, Stringer, et al., AI estimation of gestational age from blind ultrasound sweeps in low-resource settings. NEJM Evidence, 1 (5), 2022.
- Gomes, Vwalika, Lee, et al., A mobile-optimized artificial intelligence system for gestational age and fetal malpresentation assessment. Commun Med 2, 128 (2022).
- Hadlock et al, Estimation of fetal weight with the use of head, body, and femur measurements--A prospective study, Am J Obs Gyn, 151 (3), pp 333-337, 1985.

+ GitHub repository:

- GA, FP, EFW, and TWIN code repository: Global-Health-Labs/OBUS-GHL

Acknowledgements

Global Health Labs

Courosh Mehanian
Dan Shea
Olivia Zahn
Sourabh Kulhare
Matt Horning
Wei Luo
Kate McClean
Charles Delahunt

GF: MNCNH

Ari Moskowitz
Richard Zong
Manu Vatish

UNC

Jeff Stringer
Srihari Venkatesh Chari
Teeranan Pokaprakarn

RayShape

Wenlong Shi

Note: Being included in this list does not imply endorsement of this deck nor the technical work.

We also acknowledge additional GHL and other organization support staff, personnel who collected and prepared the data, and participants in the studies.