TabAttention: Learning Attention Conditionally on Tabular Data

Michal K. Grzeszczyk[1(✉)], Szymon Płotka[1,2,3], Beata Rebizant[4],
Katarzyna Kosińska-Kaczyńska[4], Michał Lipa[5],
Robert Brawura-Biskupski-Samaha[4], Przemysław Korzeniowski[1],
Tomasz Trzciński[6,7,8], and Arkadiusz Sitek[9]

[1] Sano Centre for Computational Medicine, Cracow, Poland
m.grzeszczyk@sanoscience.org
[2] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[3] Amsterdam University Medical Center, Amsterdam, The Netherlands
[4] The Medical Centre of Postgraduate Education, Warsaw, Poland
[5] Medical University of Warsaw, Warsaw, Poland
[6] Warsaw University of Technology, Warsaw, Poland
[7] IDEAS NCBR, Warsaw, Poland
[8] Tooploox, Wroclaw, Poland
[9] Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Abstract. Medical data analysis often combines both imaging and tabular data processing using machine learning algorithms. While previous studies have investigated the impact of attention mechanisms on deep learning models, few have explored integrating attention modules and tabular data. In this paper, we introduce TabAttention, a novel module that enhances the performance of Convolutional Neural Networks (CNNs) with an attention mechanism that is trained conditionally on tabular data. Specifically, we extend the Convolutional Block Attention Module to 3D by adding a Temporal Attention Module that uses multi-head self-attention to learn attention maps. Furthermore, we enhance all attention modules by integrating tabular data embeddings. Our approach is demonstrated on the fetal birth weight (FBW) estimation task, using 92 fetal abdominal ultrasound video scans and fetal biometry measurements. Our results indicate that TabAttention outperforms clinicians and existing methods that rely on tabular and/or imaging data for FBW prediction. This novel approach has the potential to improve computer-aided diagnosis in various clinical workflows where imaging and tabular data are combined. We provide a source code for integrating TabAttention in CNNs at https://github.com/SanoScience/Tab-Attention.

**Keywords:** Attention · Fetal Ultrasound · Tabular Data

1 Introduction

Many clinical procedures involve collecting data samples in the form of imaging and tabular data. New deep learning (DL) architectures fusing image and non-image data are being developed to extract knowledge from both sources of information and improve predictive capabilities [9]. While concatenation of tabular

and imaging features in final layers is widely used [8,11], this approach limits the interaction between them. To facilitate better knowledge transfer between these modalities more advanced techniques have been proposed. Duanmu *et al.* [5] presented the Interactive network in which tabular features are passed through a separate branch and channel-wise multiplied with imaging features at different stages of Convolutional Neural Network (CNN). Pölsterl *et al.* [16] proposed a Dynamic Affine Feature Map Transform (DAFT) to shift and scale feature maps conditionally on tabular data. In [6], Guan *et al.* presented a method for transforming tabular data and processing them together with 3D feature maps via VisText self-attention module. The importance of the attention mechanism on DL models' performance has been extensively studied [23]. Convolutional Block Attention Module (CBAM) [25] has been shown to improve the performance of DL models on high dimensional data [24,26]. Despite these advances, few studies have explored the potential of incorporating attention maps with imaging and tabular data simultaneously.

We develop such a solution and as an example of application, we use fetal birth weight (FBW) prediction from ultrasound (US) data. It is a challenging task requiring clinicians to collect US videos of fetal body parts and fetal biometry measurements. Currently, abdominal circumference (AC), head circumference (HC), biparietal diameter (BPD), and femur length (FL) are used to estimate FBW with heuristic formulae [7]. The predicted weight is the indicator of perinatal health prognosis or complications in pregnancy and has an impact on the method of delivery (vaginal or Cesarean) [17]. Unfortunately, the current approach to FBW estimation is often imprecise and can lead to a mean absolute percentage error (MAPE) of 10%, even if performed by experienced sonographers [20]. An ensemble of Machine Learning algorithms was proposed by Lu *et al.* [12] for solving this task. CNNs are applied for fetal biometry measurements estimation from US standard planes [1] or US videos [15]. Tao *et al.* [21] approach this problem with a recurrent network utilizing temporal features of fetal weight changes over weeks concatenated with fetal parameters. Płotka *et al.* [14] developed BabyNet, a hybrid CNN with Transformer layers to estimate FBW directly from US videos. Recent studies show that there is a strong correlation between the image features of the abdominal plane and the estimated fetal weight, indicating that it can serve as a dependable indicator for evaluating fetal growth [3]. We utilize the US videos of the abdomen (imaging data) and biometry measurements with other numerical values (tabular data) during our experiments.

In this work, we introduce TabAttention, a novel module designed to enhance the performance of CNNs by incorporating tabular data. TabAttention extends the CBAM to the temporal dimension by adding a Temporal Attention Module (TAM) that leverages Multi-Head Self-Attention (MHSA) [23]. Our method utilizes pooled information from imaging feature maps and tabular data (represented as tabular embeddings) to generate attention maps through Channel Attention Module (CAM), Spatial Attention Module (SAM), and TAM. By incorporating tabular data, TabAttention enables the network to better identify *what*, *where*, and *when* to focus on, thereby improving performance. We evaluate our method on the task of estimating FBW from abdominal US
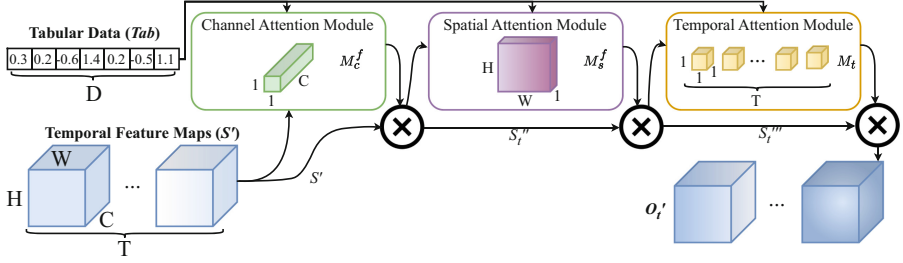
Fig. 1. The overview of proposed TabAttention module inspired by CBAM [25]. We add a Temporal Attention Module to the existing architecture to extend the method to 3D data processing. In our setting, channel, spatial and temporal attention maps are trained conditionally on tabular data. $\otimes$ indicates element-wise multiplication.

videos and demonstrate that TabAttention is at least on par with existing methods, including those based on tabular and/or imaging data, as well as clinicians. The main contributions of our work are: 1) the introduction of TabAttention, a module for conditional attention learning with tabular data, 2) the extension of CBAM to the temporal dimension via the TAM module, and 3) the validation of our method on the FBW estimation task, where we demonstrate that it is competitive with state-of-the-art methods.

2   Method

In this section, we introduce the fundamental components of the TabAttention module. We detail the development of CBAM augmented with a Temporal Attention Module. Then, we elaborate on how TabAttention leverages tabular embeddings to modulate the creation of attention maps and outline how the module can be seamlessly incorporated into the residual block of ResNet.

Figure 1 presents the overview of the TabAttention module. Given US video sequence $S \in \mathbb{R}^{T_0 \times 1 \times H_0 \times W_0}$ of height H_0, width W_0 and frame number T_0 as the input, 3D CNN produces intermediate temporal feature maps $S' \in \mathbb{R}^{T \times C \times H \times W}$ where C is the number of channels. In our setting, the CBAM block generates T 1D channel attention maps $M_c \in \mathbb{R}^{T \times C \times 1 \times 1}$ and T 2D spatial attention maps $M_s \in \mathbb{R}^{T \times 1 \times H \times W}$. We create attention maps separately for every temporal feature map as the information of *what* is meaningful and *where* it is important to focus on might change along the temporal dimension. To account for the temporal changes and focus on *when* is the informative part we add TAM which infers temporal attention map $M_t \in \mathbb{R}^{T \times 1 \times 1 \times 1}$. Intermediate temporal feature maps S' are refined with attention maps in the following way:

$$S'' = M_c(S') \otimes S' \qquad S''' = M_s(S'') \otimes S'' \qquad \boldsymbol{O'} = M_t(S''') \otimes S''' \quad (1)$$

Here $\boldsymbol{O'}$ denotes the output of the module and $\otimes$ is an element-wise multiplication during which attention maps are broadcasted along all unitary dimensions.
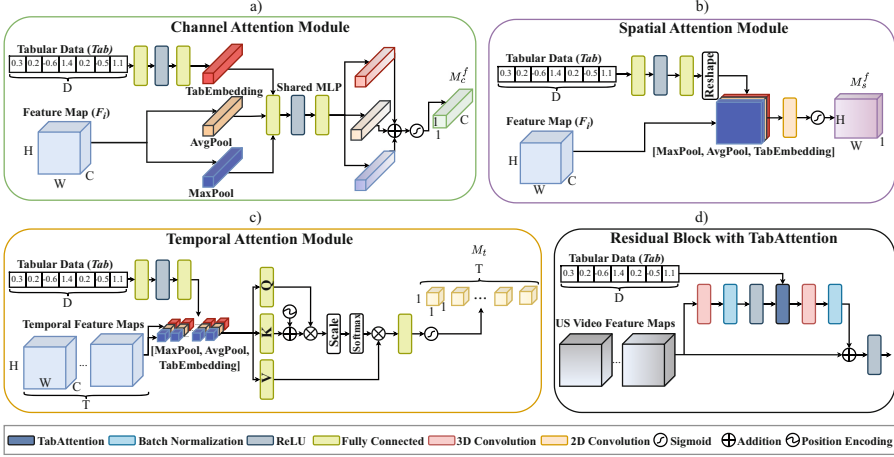
Fig. 2. Details of TabAttention components: CAM (a), SAM (b) and TAM (c) with tabular embeddings. In TAM, only one attention head is visualised. TabAttention is integrated with residual block as presented in (d). $\otimes$ indicates matrix multiplication.

In general, attention maps are computed based on information aggregated by average- and max-pooling along specified dimensions which are then passed through shared layers for refinement (Fig. 2). Then, these refined descriptors are passed through the sigmoid function to create final attention maps. To account for the tabular information during attention maps computing, we embed the input tabular data $Tab \in \mathbb{R}^D$, where D is the number of numerical features, with two linear layers and Rectified Linear Unit (ReLU) activation in between. The tabular data is embedded to the size of pooled feature maps. The embedding is passed through shared layers in the same way as pooled feature maps. Therefore, the attention maps are computed conditionally on tabular data. Thus, the output of TabAttention $\boldsymbol{O'_t}$ is computed as follows:

$$S''_t = M_c(S', Tab) \otimes S' \qquad S'''_t = M_s(S''_t, Tab) \otimes S''_t \qquad \boldsymbol{O'_t} = M_t(S'''_t, Tab) \otimes S'''_t \tag{2}$$

Channel Attention Module. We follow the design of the original CBAM [25]. We split temporal feature maps into T feature maps F_i where $i \in 1, ..., T$ so that each of them is passed through CAM separately. To compute the channel attention (M_c), we aggregate the spatial information through average- and max-pooling to produce descriptors ($F^c_{avg_i}, F^c_{max_i} \in \mathbb{R}^{C \times 1 \times 1}$). We pass the tabular data through a multi-layer perceptron (MLP_{emb_c}) with one hidden layer (of size $\mathbb{R}^{\frac{C}{z}}$, where z is the reduction ratio set to 16) and ReLU activation to embed it into the same dimension as spatial descriptors. Then, both descriptors, with tabular embedding are passed through the shared network which is MLP with a hidden activation size of $\mathbb{R}^{\frac{C}{z}}$ and one ReLU activation. After the MLP is applied, the output vectors are element-wise summed to produce the attention map. We

concatenate attention maps of all feature maps to produce M_c:

$$M_c(S', Tab) = [M_c^f(F_i, Tab)]_{i=1,...,T}$$
$$M_c^f(F_i, Tab) = \sigma(MLP(F_{max_i}^c) + MLP(F_{avg_i}^c) + MLP(MLP_{emb_c}(Tab)))$$
(3)

Spatial Attention Module. After splitting the temporal feature maps, we average- and max-pool them along channel dimension to produce feature descriptors $(F_{avg_i}^s, F_{max_i}^s \in \mathbb{R}^{1 \times H \times W})$. We pass the tabular data through MLP_{emb_s} with one hidden layer of size $\mathbb{R}^{\frac{H \times W}{2}}$ and ReLU activation to embed it into the same dimension as spatial descriptors. We reshape this embedding to the size of feature descriptors and concatenate it with them. We pass the following representation through a 2D convolution layer and the sigmoid activation:

$$M_s(S'', Tab) = [M_s^f(F_i, Tab)]_{i=1,...,T}$$
$$M_s^f(F_i, Tab) = \sigma(Conv([F_{max_i}^s, F_{avg_i}^s, Reshape(MLP_{emb_s}(Tab))]))$$
(4)

Temporal Attention Module. We create temporal descriptors by average- and max-pooling temporal feature maps along all non-temporal dimensions $(F_{avg_i}^t, F_{max_i}^t \in \mathbb{R}^{T \times 1 \times 1 \times 1})$. We embed tabular data with MLP_{emb_t} with one hidden layer of size $\mathbb{R}^{\frac{T}{2}}$ into the same dimension. We concatenate created vectors and treat them as the embedding of the US sequence which we pass to the MHSA layer (with 2 heads). We create the query (Q), key (K) and value (V) with linear layers and an output size of d (4). We add relative positional encodings [19] r to K. After passing through MHSA, we squash the refined representation with one MLP layer and sigmoid function to create a temporal attention map M_t:

$$MHSA(S_{emb}) = MLP\left(\left[softmax\left(\frac{Q_j(K_j + r)^T}{\sqrt{d}}\right)V_j\right]_{j=1,2}\right)$$
$$M_t(S''', Tab) = \sigma(MHSA([F_{max}^t, F_{avg}^t, MLP_{emb_t}(Tab)]))$$
(5)

TabAttention can be integrated within any 3D CNN (or 2D CNN in case TAM is omitted). As illustrated in Fig. 2, we add TabAttention between the first ReLU and the second convolution in the residual block to integrate our module with 3D ResNet-18.

3 Experiments and Results

This section describes the dataset used and provides implementation details of our proposed method. We benchmark the performance of TabAttention against several state-of-the-art methods and compare them to results obtained by clinicians. Additionally, we conduct an ablation study to demonstrate the significance of each key component utilized in our approach.

Dataset. This study was approved by the Ethics Committee of the Medical University of Warsaw (Reference KB.195/2021) and informed consent was obtained
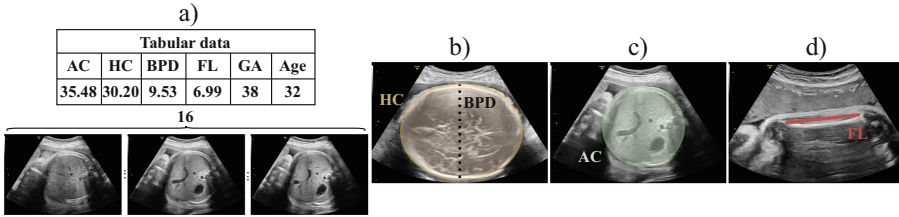
Fig. 3. Exemplary input to our method (a) of tabular data and frames of abdominal scans. Fetal US scans of the head (b), abdomen (c), and femur (d) were used to obtain AC, HC, BPD and FL utilized in tabular data.

for all subjects. The multi-site dataset was acquired using international standards approved by [18]. The dataset consists of 92 2D fetal US video scans captured in the standard abdominal plane view. These scans were collected from 92 pregnant women (31.89 ± 4.76 years), across three medical centers, and obtained as part of routine US examination done less than 24 h before delivery. This allowed us to obtain the real ground truth which was baby weight soon after birth. Five experienced sonographers (14.2 ± 4.02 years of experience) acquired the data using a single manufacturer device (General Electric) and several models (GE Voluson E6, S8, P8, E10, and S10). The abdominal fetal US videos (5–10 seconds, 13–37 frames per second) were saved in the DICOM file format. We resized the pixel spacing to $0.2\,\text{mm} \times 0.2\,\text{mm}$ for all video clips. As tabular data, we used six numerical features: AC ($34.51 \pm 2.35\,\text{cm}$), HC ($33.56 \pm 1.41\,\text{cm}$), BPD ($9.40 \pm 0.46\,\text{cm}$), FL ($7.28 \pm 0.33\,\text{cm}$), GA ($38.29 \pm 1.47\,\text{weeks}$), and mother's age. The examples of how the measurements were obtained are presented in Fig. 3. The actual birth weight of the fetus obtained right post-delivery ($3495 \pm 507\,\text{g}$) was used as the target of the prediction.

Implementation Details. We use 3D ResNet-18 as our base model. We implement all experiments with PyTorch and train networks using NVIDIA A100 80GB GPU for 250 epochs with a batch size of 16 and an initial learning rate chosen with grid search from the set of $\{1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}\}$. To minimize the Mean Squared Error loss function, we employ the Adam [10] optimizer with L2 regularization of 1×10^{-4} and cosine annealing learning rate scheduler. To evaluate the reliability of the regression algorithm, we conduct five-fold cross-validation (CV) and ensure that each patient's data is present in only one fold. To ensure similar birth weight distribution in all folds, we stratify them based on the assignment of data samples into three bins: $< 3000\,\text{g}$ g, $> 4000\,\text{g}$ g, and in-between. The input frames are of size 128×128 pixels. We follow the approach presented in [14], we set the number of input frames to 16 and average per-patient predictions of all 16 frame segments from the single video. Throughout the training process, we employ various data augmentation techniques such as rotation, random adjustments to brightness and contrast, the addition of Gaussian noise, horizontal flipping, image compression, and motion blurring for every batch. We standardize all numerical features to a mean of 0

Table 1. Five-fold cross-validation results of state-of-the-art methods utilizing imaging (Img.) and/or tabular (Tab.) data. The mean of MAE, RMSE, and MAPE across all folds are presented. The best results are bolded.

Method	Img.	Tab.	mMAE [g]	mRMSE [g]	mMAPE [%]
BabyNet [14]	✓	✗	294 ± 30	386 ± 56	8.5 ± 1.0
3D ResNet-18 [22]	✓	✗	289 ± 38	373 ± 43	8.5 ± 1.1
XGBoost	✗	✓	259 ± 23	328 ± 26	7.6 ± 0.1
Linear Regression	✗	✓	207 ± 18	260 ± 19	6.0 ± 0.1
Clinicians	✗	✓	205 ± 14	253 ± 13	5.9 ± 0.0
DAFT [16]	✓	✓	175 ± 30	244 ± 42	5.3 ± 1.0
Interactive [5]	✓	✓	172 ± 27	230 ± 44	5.2 ± 0.9
TabAttention (ours)	✓	✓	$\mathbf{170 \pm 26}$	$\mathbf{225 \pm 37}$	$\mathbf{5.0 \pm 0.8}$

and a standard deviation of 1. We use Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and MAPE to evaluate the regression performance.

Comparison with State-of-the-Art Methods. We compare TabAttention with several methods utilizing tabular data only (Linear Regression [13], XGBoost [4]), imaging data only (3D ResNet-18 [22], BabyNet [14]), both types of data (Interactive [5], DAFT [16]), and Clinicians. The predictions of Clinicians were achieved using Hadlock III [7] formula and AC, HC, BPD, FL measurements. The comparison of results from the five-fold CV is presented in Table 1. TabAttention achieves the lowest MAE, RMSE and MAPE (170 ± 26, 225 ± 37, 5.0 ± 0.8 respectively) among all tested methods. Our approach outperforms clinically utilized heuristic formulae, machine learning, and image-only DL methods (two-tailed paired t-test p-value < 0.05). Results of TabAttention are also best compared with all DL models utilizing tabular and imaging modalities, however, the difference does not reach statistical significance with a p-value around 0.11.

Ablation Study. We conduct ablation experiments to validate the effectiveness of key components of our proposed method (Table 2). We employ 3D ResNet-18 as the baseline model. The integration of TAM or CBAM with attention maps learned conditionally on tabular data into the 3D ResNet-18 architecture improves the predictive performance of the network. Subsequently, the incorporation of full TabAttention further enhances its capabilities.

Table 2. Five-fold cross-validation results of ablation study with key components of TabAttention. The first row is the result of the baseline method. The next rows refer to modules of TabAttention with or without tabular embeddings (Tab.) and the last one is full TabAttention.

Method	Img.	Tab.	mMAE [g]	mRMSE [g]	mMAPE [%]
3D ResNet-18 [22]	✓	✗	289 ± 38	373 ± 43	8.5 ± 1.1
+ TAM	✓	✗	288 ± 43	389 ± 65	8.4 ± 1.2
+ CBAM + Tab	✓	✓	271 ± 51	371 ± 99	7.7 ± 1.3
+ TAM + Tab	✓	✓	180 ± 32	237 ± 45	5.5 ± 0.1
+ **TabAttention (ours)**	✓	✓	$\mathbf{170 \pm 26}$	$\mathbf{225 \pm 37}$	$\mathbf{5.0 \pm 0.8}$

4   Discussion and Conclusions

In this work, we present a novel method, TabAttention, that can effectively compete with current state-of-the-art image and/or tabular-based approaches in estimating FBW. We found that it outperformed Clinicians achieving mMAPE of 5.0% vs. 5.9% (p-value < 0.05). A key advantage of our approach is that it does not require any additional effort from clinicians since the necessary data is already collected as part of standard procedures. This makes TabAttention an alternative to the heuristic formulas that are currently used in clinical practice. We should note that while TabAttention achieved the lowest metrics among the DL models we evaluated, the differences between our approach and other DL methods using tabular data were not statistically significant, partly due to the small performance change. This small difference in the performance is likely caused by the fact that the tabular features used in TabAttention are mainly derived from the same modality (i.e. US scans), so they do not carry additional information, but instead can be considered as refined features already present in the scans. To develop TabAttention, we used tabular data as a hint for the network to learn attention maps and gain additional knowledge about essential aspects presented in the scans. This approach significantly improved the performance of baseline methods and demonstrated its practical applicability.

Accurate estimation of FBW is crucial in determining the appropriate delivery method, whether vaginal or Cesarean. Low birth weight (less than $2500\,g$ g) is a major risk factor for neonatal death, while macrosomia (greater than $4000\,g$) can lead to delivery traumas and maternal complications, such as birth canal injuries, as reported by Benacerraf *et al.* [2]. Thus, precise prediction of FBW is vital for very low and high weights. Notably, in this respect, our method is robust to outliers with high or low FBW since there is no correlation between true FBW and absolute prediction error (Pearson correlation coefficient of -0.029).

This study has limitations. Firstly, a relatively small study cohort was used, which may affect the accuracy and generalization of the results. To address this, future work will include a larger sample size by using additional datasets. Secondly, our dataset is limited to only Caucasian women and may not be rep-

resentative of other ethnicities. It is important to investigate the performance of our method with datasets from different ethnic groups and US devices to obtain more robust and generalizable results. Lastly, our method relies on fetal biometry measurements that are subject to inter- and intra-observer variabilities. This variability could potentially affect the network's performance and influence the measurements' quality. Future studies should consider strategies to reduce measurement variabilities, such as standardized protocols or automated measurements, to improve the accuracy of the method.

To summarize, we have introduced TabAttention, a new module that enables the conditional learning of attention on tabular data and can be integrated with any CNN. Our method has many potential applications, including serving as a computer-aided diagnosis tool for various clinical workflows. We have demonstrated the effectiveness of TabAttention on the FBW prediction task, utilizing both US and tabular data, and have shown that it outperforms other methods, including clinically used ones. In the future, we plan to test the method in different clinical applications where imaging and tabular data are used together.

References

1. Bano, S., et al.: AutoFB: automating fetal biometry estimation from standard ultrasound planes. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 228–238. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_22

2. Benacerraf, B.R., Gelman, R., Frigoletto, F.D., Jr.: Sonographically estimated fetal weights: accuracy and limitation. Am. J. Obstet. Gynecol. **159**(5), 1118–1121 (1988)

3. Campbell, S., Wilkin, D.: Ultrasonic measurement of fetal abdomen circumference in the estimation of fetal weight. BJOG: Int. J. Obstet. Gynaecol. **82**(9), 689–697 (1975)

4. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. KDD '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785

5. Duanmu, H., et al.: Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 242–252. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_24

6. Guan, Y., et al.: Predicting esophageal fistula risks using a multimodal self-attention network. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 721–730. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_69

7. Hadlock, F.P., Harrist, R., Sharman, R.S., Deter, R.L., Park, S.K.: Estimation of fetal weight with the use of head, body, and femur measurements-a prospective study. Am. J. Obstet. Gynecol. **151**(3), 333–337 (1985)

8. Holste, G., Partridge, S.C., Rahbar, H., Biswas, D., Lee, C.I., Alessio, A.M.: End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3294–3303 (2021)

9. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digital Med. **3**(1), 136 (2020)

10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)

11. Liu, M., Zhang, J., Adeli, E., Shen, D.: Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. IEEE Trans. Biomed. Eng. **66**(5), 1195–1206 (2018)

12. Lu, Y., Zhang, X., Fu, X., Chen, F., Wong, K.K.: Ensemble machine learning for estimating fetal weight at varying gestational age. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9522–9527 (2019)

13. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

14. Płotka, S., et al.: BabyNet: residual transformer module for birth weight prediction on fetal ultrasound video. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV, pp. 350–359. Springer (2022). https://doi.org/10.1007/978-3-031-16440-8_34

15. Płotka, S., et al.: Deep learning fetal ultrasound video model match human observers in biometric measurements. Phys. Med. Biol. **67**(4), 045013 (2022)

16. Pölsterl, S., Wolf, T.N., Wachinger, C.: Combining 3D image and tabular data via the dynamic affine feature map transform. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 688–698. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_66

17. Pressman, E.K., Bienstock, J.L., Blakemore, K.J., Martin, S.A., Callan, N.A.: Prediction of birth weight by ultrasound in the third trimester. Obstet. Gynecol. **95**(4), 502–506 (2000)

18. Salomon, L., et al.: ISUOG practice guidelines: ultrasound assessment of fetal biometry and growth. Ultrasound Obstet. Gynecol. **53**(6), 715–723 (2019)

19. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)

20. Sherman, D.J., Arieli, S., Tovbin, J., Siegel, G., Caspi, E., Bukovsky, I.: A comparison of clinical and ultrasonic estimation of fetal weight. Obstet. Gynecol. **91**(2), 212–217 (1998)

21. Tao, J., Yuan, Z., Sun, L., Yu, K., Zhang, Z.: Fetal birthweight prediction with measured data by a temporal machine learning method. BMC Med. Informa. Decis. Making **21**(1), 1–10 (2021)

22. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)

23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)

24. Wang, X., Liu, D., Zhang, Y., Li, Y., Wu, S.: A spatiotemporal multi-stream learning framework based on attention mechanism for automatic modulation recognition. Digit. Signal Process. **130**, 103703 (2022)
25. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
26. Yadav, S., Rai, A.: Frequency and temporal convolutional attention for text-independent speaker recognition. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6794–6798. IEEE (2020)