Weight Space Correlation Analysis: Quantifying Feature Utilization in Deep Learning Models

Chun Kit Wong[1,2] CKWO@DTU.DK
[1] *Technical University of Denmark, Kongens Lyngby, Denmark*
[2] *Pioneer Centre for AI, Copenhagen, Denmark*

Paraskevas Pegios[1,2] PPAR@DTU.DK
Nina Weng[1] NINWE@DTU.DK
Emilie Pi Fogtmann Sejer[3,4] EMILIE.PI.FOGTMANN.SEJER.01@REGIONH.DK
[3] *University of Copenhagen, Copenhagen, Denmark*
[4] *CAMES Rigshospitalet, Copenhagen, Denmark*

Martin Grønnebæk Tolsgaard[3,4] MARTIN.GROENNEBAEK.TOLSGAARD@REGIONH.DK
Anders Nymark Christensen[1] ANYM@DTU.DK
Aasa Feragen[1,2] AFHAR@DTU.DK

Editors: Under Review for MIDL 2026

Abstract

Deep learning models in medical imaging are susceptible to shortcut learning, relying on confounding metadata (e.g., scanner model) that is often encoded in image embeddings. The crucial question is whether the model actively utilizes this encoded information for its final prediction. We introduce Weight Space Correlation Analysis, an interpretable methodology that quantifies feature utilization by measuring the alignment between the classification heads of a primary clinical task and auxiliary metadata tasks. We first validate our method by successfully detecting artificially induced shortcut learning. We then apply it to probe the feature utilization of an SA-SonoNet model trained for Spontaneous Preterm Birth (sPTB) prediction. Our analysis confirmed that while the embeddings contain substantial metadata, the sPTB classifier's weight vectors were highly correlated with clinically relevant factors (e.g., birth weight) but decoupled from clinically irrelevant acquisition factors (e.g. scanner). Our methodology provides a tool to verify model trustworthiness, demonstrating that, in the absence of induced bias, the clinical model selectively utilizes features related to the genuine clinical signal.

Keywords: Shortcut learning, feature utilization, obstetric ultrasound.

1. Introduction

Deep learning models have achieved impressive performance across numerous medical imaging tasks, often matching or exceeding human expert capabilities. However, their reliance on vast, complex datasets introduces significant concerns regarding model interpretability, trustworthiness, and, critically, generalization. One of the primary threat to model reliability is shortcut learning, where a model learns a simple, non-causal predictive rule that performs well on training data but fails when deployed in new environments. In medical imaging, this often manifests as a model relying on confounding factors, which are variables correlated with both the image features and the clinical outcome, but without a direct causal link to the underlying anatomy or pathology. These confounders can include machine settings, acquisition protocols, scanner models, institution-specific artifacts, or patient demographics. When a model relies on such features, it is effectively taking a ”shortcut” to prediction, bypassing the true clinical signal (Geirhos et al., 2020).

There is an ever-present possibility that metadata attributes, patient demographics, or acquisition-related confounding information are subtly encoded within a medical image's visual features. Deep learning models are powerful enough to identify these minute, global, or textural subtleties that are extremely difficult for the human eye to notice. A stark illustration of this capability was provided by Gichoya et al. (2022), who demonstrated that deep learning classifiers could predict a patient's self-reported race from chest X-ray images with high accuracy. Such a finding confirms that models can extract and utilize non-clinical demographic information inadvertently embedded in the image, raising serious questions about fairness and generalization across different populations and healthcare settings (Obermeyer et al., 2019).

However, while the work of Gichoya et al. (2022) confirms that confounding information is encoded in the image input, this does not automatically mean the model has made use of that information in its inference process. The mere presence of a feature in the embedding space is a prerequisite for shortcut learning, but not proof of its use. Glocker et al. (2023) addressed this gap by proposing a method to visualize the distribution of embeddings relative to metadata attributes. Their approach allowed researchers to see if the image embeddings naturally separated based on a specific confounder, such as an age group. While this successfully allows one to visualize the distribution of the images in the embedding space, it still does not directly address the fundamental question: does the model's final classification layer, i.e. the decision boundary, actively leverage the axis of variance related to that confounder? It remains possible that the decision boundary is orthogonal to the confounder's axis, indicating the information is present in the embedding but functionally ignored by the classifier.

To definitively address the question of information utilization, we introduce a methodology that focuses on analyzing and comparing the attention of the neurons in the classification head. Our method, which we termed Weight Space Correlation Analysis, distinguishes itself by comparing the weight vectors (i.e. the ”attention”) of the primary task against the weight vectors of auxiliary metadata tasks. By quantifying the alignment between these decision boundaries using correlation analysis, we can directly determine whether the features used for a clinical prediction are the same features used to identify a confounder.

We validate this methodology by successfully detecting artificially introduced shortcut learning in a controlled environment. We then apply the validated method to probe the feature utilization of another model, SA-SonoNet (Pegios et al., 2023), trained for the prediction of Spontaneous Preterm Birth (sPTB). Our results show that while the sPTB model's embeddings contain information about the metadata, the model selectively utilizes clinically relevant features, such as those related to fetal weight, while successfully avoiding shortcuts based on acquisition factors like scanner models.

The remainder of this paper is organized as follows: section 2 details the datasets and the proposed Weight Space Correlation methodology; section 3 presents our experiments and results; and section 4 concludes the paper.

2. Materials and Methods

2.1. Clinical Datasets

In this study, we utilize two distinct, private clinical ultrasound datasets to evaluate the interplay between feature encodability and shortcut learning. Both datasets are accompanied by a rich set of demographic and acquisition metadata, including ultrasound scanner manufacturer, hospital site ID, and maternal ethnicity.

- **The Fetal Dataset**: This dataset focuses on anatomical classification. It comprises 2D ultrasound images of the four standard fetal planes: the fetal head, abdomen, femur, and thorax. The primary task is a multi-class classification problem where the model must identify the anatomical plane present in the image.

- **The Cervix Dataset**: This dataset focuses on a prognostic task related to preterm birth. It consists of transvaginal cervical ultrasound images, balanced equally between two classes: term birth and preterm birth (defined as delivery before 37 weeks of gestation). The primary task is the binary classification of the image into these prognostic outcomes.

2.1.1. Preprocessing of Metadata Attributes

To unify the prediction tasks under a single supervised learning framework, we reformulated the prediction of continuous metadata variables as a multi-class classification problem. Continuous attributes were discretized via binning, transforming scalar values into distinct categorical labels. Specifically, the range of each continuous variable was partitioned into k intervals. Any value falling within a specific interval was assigned the class label corresponding to that bin. This discretization step mitigates the impact of outliers and allows for the application of classification metrics across all target variables.

2.2. Network Architectures and Training Regimes

To systematically disentangle the encoding of spurious correlations from their functional utilization, we employ a ResNet50 backbone across two distinct architectural configurations. These variations allow us to contrast a model trained solely for the clinical task against one explicitly forced to encode metadata features.

1. **Baseline:** We train a standard ResNet50 where the input is solely the ultrasound image x_{img} and the output is the primary clinical classification y_{class} (Anatomy for the Fetal Dataset; Birth outcome for the Cervix Dataset).

2. **Multi-Task Learning:** We utilize the image-only input but extend the architecture with auxiliary output heads. The model is trained to simultaneously predict the primary class y_{class} and the metadata attributes (e.g., scanner ID) y_{meta}. This encourages the shared backbone to learn features relevant to both the clinical target and the potential confounders.

2.3. Probing for Encoded Information

To establish a baseline for ”encodability”, i.e. the degree to which metadata is present in the latent space regardless of its utility, we adopt the linear probing methodology described by Gichoya et al. (2022); Glocker et al. (2023).

For the **Baseline** models, we first complete the training for the primary clinical task. We then freeze the parameters of the backbone encoder, treating it as a fixed feature extractor. We discard the primary classification head and attach new, randomly initialized fully connected heads corresponding to the metadata attributes. These probing heads are then trained to predict the metadata (e.g., Hospital ID) using only the frozen embeddings. High performance on this probing task indicates that the model has encoded information about the metadata, even if it was not explicitly trained to do so.

Meanwhile, the **Multi-Task** models do not require fine-tuning, since the multiple distinct heads connected to the shared embedding base were already trained jointly.

2.4. Quantifying Shortcut Reliance via Weight Space Correlation

While the probing method described in section 2.3 confirms the *presence* of confounding information, it does not quantify the *extent* to which the model relies on this information for its primary prediction. To bridge this gap, we introduce a method to quantify reliance by analyzing the correlation of the decision weights in a reduced dimensionality space.

We conceptualize the weights of the final fully connected (FC) layer not merely as regression coefficients, but as attention vectors acting upon the latent embedding. If the weight vector for the primary task (W_{task}) is highly correlated with the weight vector for a metadata attribute (W_{meta}), it suggests the model attends to similar features for both predictions, implying a reliance on that specific shortcut.

To compute this robustly, we address the high dimensionality and potential sparsity of the ResNet50 latent space (2048 dimensions) using the following pipeline:

1. **Manifold Estimation:** We compute the embeddings for the entire training set X_{train} using the frozen backbone.

2. **Dimensionality Reduction:** We apply Principal Component Analysis (PCA) to these embeddings to identify the active data manifold. We construct a projection matrix P that retains the top principal components explaining 99% of the variance in the dataset, while enforcing a minimum floor of 50 components to ensure sufficient representational capacity is preserved even in lower-rank scenarios.

3. **Weight Projection:** We project the weights of the prediction heads (both the primary task and the probing heads) into this PCA-reduced space. Let W_{fc} represent the weights of a fully connected head; the transformed weights are calculated as:

$$W'_{fc} = W_{fc} \cdot P^T \tag{1}$$

   This step ensures that the correlation is calculated based on the directions of variance that actually exist in the data, rather than the less informative orthogonal dimensions.

4. **Correlation Analysis:** Finally, we compute the pairwise correlation (cosine similarity) between every pair of projected weight vectors.

A high correlation in this projected space serves as a quantitative proxy for shortcut learning: it implies that the decision boundary for the clinical task aligns closely with the decision boundary for the confounder.

3. Experiments and Results

3.1. Establishing the Embedding of Metadata in Images

The foundational question addressed in our experiments is whether the metadata attributes, both clinical and acquisition-related, are implicitly embedded within the visual features of the medical images themselves. To test this, we used baseline models to predict metadata factors directly from images in both fetal and cervix datasets (see section 2.1). We trained separate deep learning classification models for each metadata factor in each dataset. The performance of these baseline models, which simply predict a single metadata factor from the raw image input, is documented under Appendix A.

The results consistently demonstrate that the visual features extracted by the model contain substantial information regarding the metadata across both domains. In the fetal dataset, models achieved strong predictive accuracy for the primary task, as well as for acquisition-related factors like scanner, pixel spacing, and hospital ID. Similar performance was also observed among models predicting auxiliary factors from the cervix images.

These findings confirm a consistent observation: for both datasets, some of the metadata factors are implicitly embedded within the image's texture, geometry, and presentation. This suggests that any model trained on these images may inadvertently encode features related to these attributes alongside the primary clinical task, necessitating the subsequent investigation into their utilization (see section 3.2).

3.2. Utilization of Clinically Irrelevant Factors in Classification

Section 3.1 confirmed that metadata attributes are visually embedded in the images. This section addresses our second core research question: Does the primary classifier (fetal plane identification) actively utilize these clinically irrelevant, but embedded, factors in its decision-making process? To answer this, we employed the fine-tuning methodology described in section 2.3. We first trained a base model for the primary task of Fetal Standard Plane classification. We then fine-tuned the final layers of this pre-trained model to perform secondary classification tasks for the clinically irrelevant metadata factors (e.g., scanner, hospital ID). The performance of the fine-tuned model on the metadata prediction tasks, shown in table 1, confirms the continued presence of this information in the embeddings of the primary classifier. The model achieved a relatively high AUROC when fine-tuned to predict certain metadata variables (e.g., scanner). This result reinforces the finding that the image embeddings, generated by a model focused solely on plane classification, still contain sufficient features to distinguish between different acquisition parameters.

While the embeddings hold the information, the critical step is, however, determining if that information is being used. As demonstrated in fig. 1(a), weight space correlation analysis (see section 2.4) on the weight vectors of the final classification heads suggests that the correlation between the weight vectors for the fetal plane classes and the weight vectors for the scanner classes was consistently low. This indicates that although the necessary infor-

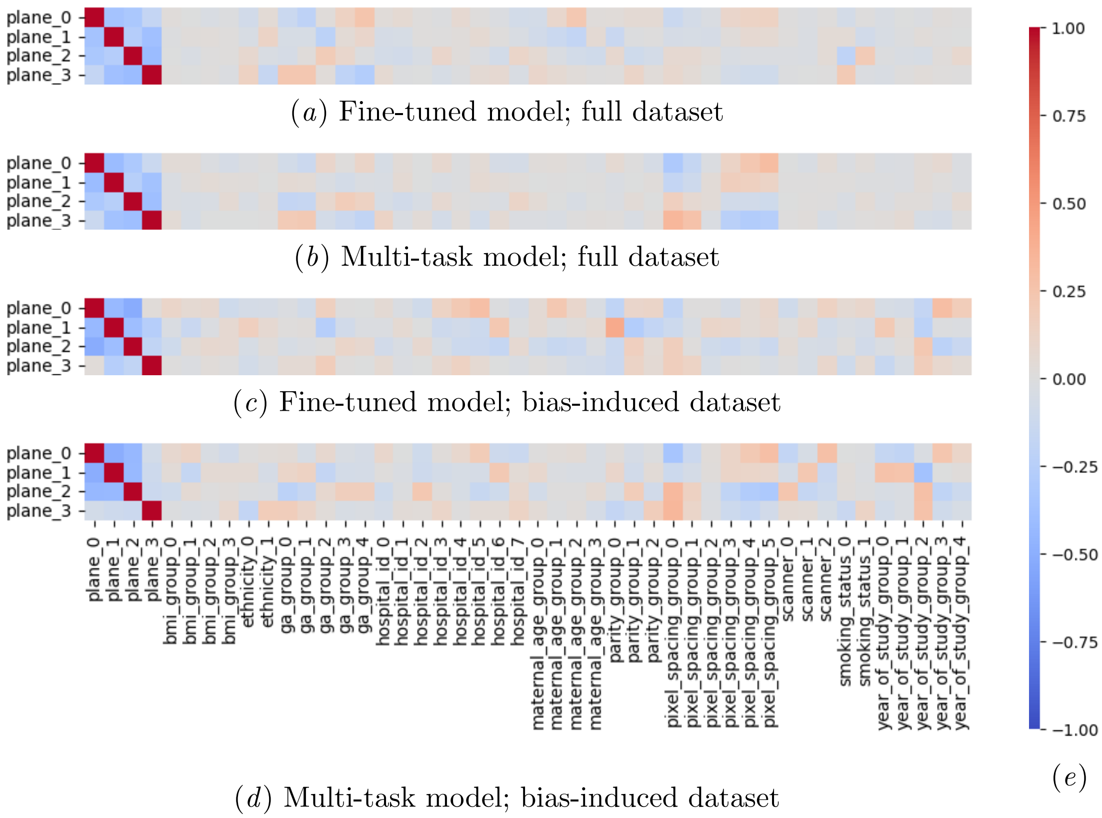
Target	**Accuracy**	**Precision**	**Recall**	**F1**	**AUROC**
Fine-tuned model					
Plane	95.8 ± 0.6	95.4 ± 0.6	95.4 ± 0.6	95.4 ± 0.6	99.6 ± 0.1
Scanner	76.8 ± 1.4	75.8 ± 1.2	75.2 ± 1.9	75.3 ± 1.7	91.4 ± 0.9
Pixel spacing	56.8 ± 1.1	53.2 ± 1.3	54.1 ± 1.3	52.9 ± 1.3	88.2 ± 0.3
GA	50.5 ± 1.7	45.0 ± 1.4	45.7 ± 1.1	44.4 ± 1.2	80.0 ± 1.0
Hospital ID	49.8 ± 3.4	36.1 ± 2.1	36.9 ± 2.2	35.0 ± 2.2	77.0 ± 1.1
Year of study	46.4 ± 3.2	40.7 ± 0.8	45.3 ± 3.1	39.4 ± 0.8	75.6 ± 1.9
BMI	36.5 ± 1.6	35.8 ± 1.5	36.5 ± 1.4	35.2 ± 1.5	64.1 ± 1.1
Ethnicity	85.9 ± 5.8	50.4 ± 2.0	52.5 ± 3.8	50.0 ± 1.6	58.8 ± 5.8
Parity	49.7 ± 3.2	35.8 ± 0.6	35.9 ± 0.9	33.9 ± 0.3	55.0 ± 1.0
Smoking status	77.2 ± 7.0	50.4 ± 1.7	50.7 ± 2.3	49.5 ± 1.9	50.8 ± 3.9
Maternal age	25.6 ± 1.8	25.1 ± 1.5	25.7 ± 2.0	24.6 ± 1.4	50.8 ± 1.7
Multitask model					
Plane	95.7 ± 0.1	95.4 ± 0.1	95.3 ± 0.1	95.3 ± 0.1	99.6 ± 0.1
Scanner	96.9 ± 0.5	96.9 ± 0.6	96.5 ± 0.6	96.7 ± 0.6	99.7 ± 0.1
Pixel spacing	69.4 ± 1.3	67.9 ± 0.8	67.2 ± 1.3	67.0 ± 1.3	94.0 ± 0.4
GA	59.5 ± 1.4	52.8 ± 1.2	51.5 ± 1.0	51.6 ± 1.1	86.8 ± 0.4
Hospital ID	73.7 ± 0.8	55.6 ± 1.8	54.9 ± 1.8	54.5 ± 1.8	91.9 ± 0.7
Year of study	68.7 ± 1.2	51.4 ± 3.0	51.7 ± 0.7	49.9 ± 1.0	90.5 ± 0.2
BMI	39.8 ± 1.9	43.4 ± 0.8	39.9 ± 2.0	40.1 ± 1.9	68.6 ± 0.9
Ethnicity	94.3 ± 0.4	47.6 ± 0.0	49.5 ± 0.2	48.5 ± 0.1	42.0 ± 2.4
Parity	70.2 ± 2.3	33.7 ± 3.0	34.3 ± 0.4	32.1 ± 0.7	64.0 ± 1.6
Smoking status	84.5 ± 0.5	52.1 ± 4.8	50.4 ± 0.8	48.2 ± 1.4	51.2 ± 2.9
Maternal age	28.3 ± 1.4	27.5 ± 3.2	27.2 ± 2.2	25.0 ± 2.9	53.3 ± 1.7

Table 1: Test performance of ResNet50 classifier model when fine-tuned to predict the other targets. Below Test performance of ResNet50 classifier model trained to predict various target values in a multitask setting. Baseline performance is shown in Appendix A

mation about the scanner is present in the preceding embedding layer, the model's decision boundary for the primary plane classification task is largely orthogonal to the directionality required to classify the scanner. In other words, the embedded, clinically irrelevant information is not being actively utilized by the classifier for its primary prediction.

3.2.1. Stress Test with Multi-Task Learning

To stress-test this finding, we trained a multi-task learning model designed to encourage the simultaneous encoding of all factors. This model was trained to predict the fetal plane and all metadata factors concurrently, thereby explicitly maximizing the metadata information content within the shared embedding space, as shown in table table 1. Upon performing the



(a) Fine-tuned model; full dataset

(b) Multi-task model; full dataset

(c) Fine-tuned model; bias-induced dataset

(d) Multi-task model; bias-induced dataset

(e)

Figure 1: Correlation matrix between weight vectors from classification head of the primary task versus that of each metadata attributes, extracted from fine-tuned or multi-task classifier models, trained using the entire fetal dataset or with induced bias. Full matrix is available in the Appendix at fig. 3.

same weight space correlation analysis on this model, the correlation between the plane and scanner weight vectors remained low. This further reinforces the initial conclusion: even when the model is explicitly forced to encode metadata information into the embeddings, the weight vectors for the primary task (plane classification) remain largely decoupled from the weight vectors of the irrelevant factors (scanner classification).

3.3. Validation of Shortcut Learning Detection via Induced Bias

Section 3.2 suggested that the model did not utilize available scanner information for plane classification. This section aims to validate our hypothesis: If a model does adopt a shortcut learning strategy, this behavior will be detectable via the weight space correlation analysis. To induce shortcut learning, we intentionally biased the training set by introducing a strong correlation between the primary classification target (fetal plane) and a clinically irrelevant factor (scanner). We performed data culling on the fetal dataset, discarding images such that the majority of images for each standard plane were acquired by a distinct, single scanner model. The final biased dataset composition is documented in table 2, illustrating the high degree of induced correlation between the two factors. This manipulation forces the classifier to potentially adopt a shortcut learning approach, where classifying the scanner becomes an efficient proxy for classifying the fetal plane.

	Full			Simulated Biased		
	Voluson S	V830	Voluson E10	Voluson S	V830	Voluson E10
Abdomen	717	333	658	150	150	658
Head	1018	805	888	150	700	150
Femur	760	315	533	700	150	150
Thorax	855	602	523	700	150	150

Table 2: Composition of the full and bias-induced fetal plane dataset.

We repeated the analysis described in section 3.2 using the newly biased dataset. Specifically, we trained both the single-task plane classifier and the multi-task classifier on this biased data. The resulting weight space correlations between the plane classes and the scanner classes are shown in figs. $1(c)$ and $1(d)$. For the single-task classifiers, the correlation matrix showed a noticeable increase in value compared to the original, unbiased experiment. More crucially, the multi-task model, which is explicitly encouraged to encode all features, exhibited a much stronger correlation between the weight vectors for the fetal plane classes and the scanner classes. This significant increase in correlation confirms that the model did adopt the shortcut when the bias was present, demonstrating that the decision boundary for plane classification now aligns with the directionality required for scanner classification.

While the strong correlation observed between the plane and scanner weight vectors in the biased scenario is indicative of a dependency between the two prediction tasks, it is important to note the nature of this association. A high correlation coefficient signifies that the two tasks assign highly similar attention vectors to the shared model embeddings; they are looking at similar features in the embedding space to make their respective decisions. It does not explicitly define the direction of the shortcut. That is, the result does not definitively prove whether the model is using scanner information to predict the plane, or if the plane information is strongly predictive of the scanner due to the induced bias. It simply confirms that, under conditions of high dataset bias, the feature utilization for the two tasks becomes strongly coupled.

The findings from this experiment provide crucial validation. High weight space correlation is a reliable indicator of shortcut learning, where the model utilizes a non-causal, highly correlated factor for its prediction. The successful detection of this artificially induced shortcut proves that the weight space correlation analysis is a sensitive and effective tool for determining the active utilization of embedded, irrelevant information.

3.4. Probing a Trained Model: Analysis of SA-SonoNet Embeddings

Having validated our weight space correlation methodology in section 3.3, we now apply this technique to probe a model trained on a real-world, highly relevant clinical task. For this analysis, we utilize the SA-SonoNet model (Pegios et al., 2023), which achieved competitive performance in the challenging task of Spontaneous Preterm Birth (sPTB) prediction.

3.4.1. Architecture and Methodology Adaptation

SA-SonoNet is a shape- and spatially-aware network based on the SonoNet (Baumgartner et al., 2017) architecture, modified to predict term or preterm birth from transvaginal cervix ultrasound images. The key innovation is its multimodal input: From a given cervix image, the model first leverages a separate network to compute a segmentation map of important anatomical structures (e.g., cervical canal, boundaries, bladder). The final input to the SA-SonoNet classifier is the concatenation of the original image, the segmentation map, and the pixel spacing values, which are repeated and reshaped to image dimensions to inject spatial information.

To integrate this pre-trained SA-SonoNet model into our correlation framework, we modified its final classification layer for fine-tuning and analysis. The original SA-SonoNet model uses an average pooling layer on a 14 × 18 2D embedding feature map for its final prediction. We first flattened this 14 × 18 feature map into a 252-long 1D embedding vector. This vector was then connected to a newly initialized fully connected classification head for predicting the metadata variables. For the correlation analysis, we represented the original average pooling operation, which maps the 252-long embedding to the final sPTB prediction, by including a 252-long vector of all ones in the set of weight vectors. This step allows us to compare the feature utilization direction of the original sPTB task against the fine-tuned metadata tasks.

3.4.2. Analysis of Feature Utilization in SA-SonoNet

Target	Accuracy	Precision	Recall	F1	AUROC
Term Birth	67.5 ± 2.5	67.7 ± 2.4	67.5 ± 2.5	67.4 ± 2.6	73.9 ± 3.0
Px Spacing	70.9 ± 3.8	69.4 ± 3.1	70.3 ± 3.3	69.4 ± 3.1	95.3 ± 0.9
Py Spacing	70.5 ± 3.0	66.6 ± 2.9	68.6 ± 3.1	66.6 ± 3.1	95.2 ± 1.0
Cervical Length	52.6 ± 2.0	52.8 ± 2.3	55.6 ± 1.9	53.3 ± 2.0	80.0 ± 1.6
Scanner	56.3 ± 6.4	32.0 ± 3.1	55.7 ± 9.9	30.4 ± 5.5	77.2 ± 5.7
Birth Weight	21.4 ± 1.5	22.0 ± 1.8	23.7 ± 2.5	21.0 ± 1.9	61.2 ± 1.9
Placenta Weight	28.8 ± 1.5	27.5 ± 1.2	30.0 ± 3.7	25.0 ± 1.6	55.4 ± 2.9
BMI	27.4 ± 2.5	26.9 ± 2.2	28.4 ± 3.6	23.1 ± 2.4	53.3 ± 3.0
GA	34.6 ± 3.0	27.7 ± 3.3	27.7 ± 3.5	27.5 ± 3.4	48.4 ± 2.0

Table 3: Test performance of SA-SonoNet model trained for pre-term birth prediction and subsequently fine-tuned to predict the other targets.

The fine-tuned model's performance on various auxiliary metadata tasks and the resulting weight space correlations are presented in table 3 and the associated correlation matrix fig. 2. The analysis reveals several expected correlations, confirming that the model utilizes clinically significant factors, as well as some desirable decoupling from irrelevant

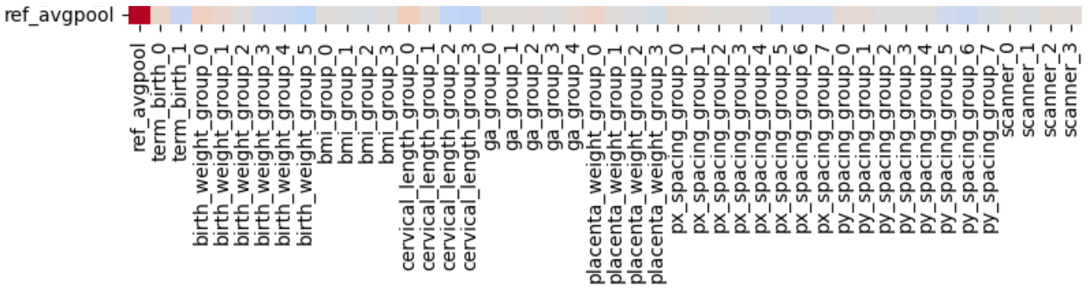


Figure 2: Correlation between weight vectors from classification head of the targets of the fine-tuned SA-SonoNet model. A reference line is added to represent the average pool layer in the original model. Full covariance matrix is available in the Appendix as fig. 4

factors. Birth weight exhibited a strong correlation, consistent with the expected link between prematurity and low birthweight, indicating that the model leverages features tied to clinical outcome. Cervical length showed a moderate correlation, aligning with its role as the clinical gold standard for sPTB risk. Pixel spacing also demonstrated a moderate correlation, reflecting its known influence on model performance and confirming that this imaging parameter contributes to the prediction. In contrast, scanner type showed only a weak correlation, suggesting the model does not rely on acquisition hardware–related shortcuts.

4. Conclusion

In this work, we introduced Weight Space Correlation Analysis, a simple and interpretable methodology designed to move beyond simply identifying the presence of confounding information to definitively quantifying its utilization by a deep learning classifier. We validated our method by successfully detecting artificially induced shortcut learning in a controlled environment. Applying this method to the SA-SonoNet model, we confirmed that while clinically irrelevant factors are indeed encoded in the image embeddings, the model's decision boundary for Spontaneous Preterm Birth prediction is selectively aligned with clinically meaningful metadata and, crucially, decoupled from confounding acquisition factors like scanner model. These findings provide a necessary level of trustworthiness in complex medical imaging models by confirming that the classifier is learning features relevant to the clinical task, rather than relying on spurious correlations.

Acknowledgments

This work is funded by the Danish Pioneer Centre for AI (DNRF grant number P1), SONAI - a Danish Regions' AI Signature Project, and the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606).

References

Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE transactions on medical imaging*, 36(11):2204–2215, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *EBioMedicine*, 89, 2023.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019.

Paraskevas Pegios, Emilie Pi Fogtmann Sejer, Manxi Lin, Zahra Bashir, Morten Bo Søndergaard Svendsen, Mads Nielsen, Eike Petersen, Anders Nymark Christensen, Martin Tolsgaard, and Aasa Feragen. Leveraging shape and spatial information for spontaneous preterm birth prediction. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 57–67. Springer, 2023.

Appendix A. Full Results: Establishing the Embedding of Metadata in Images

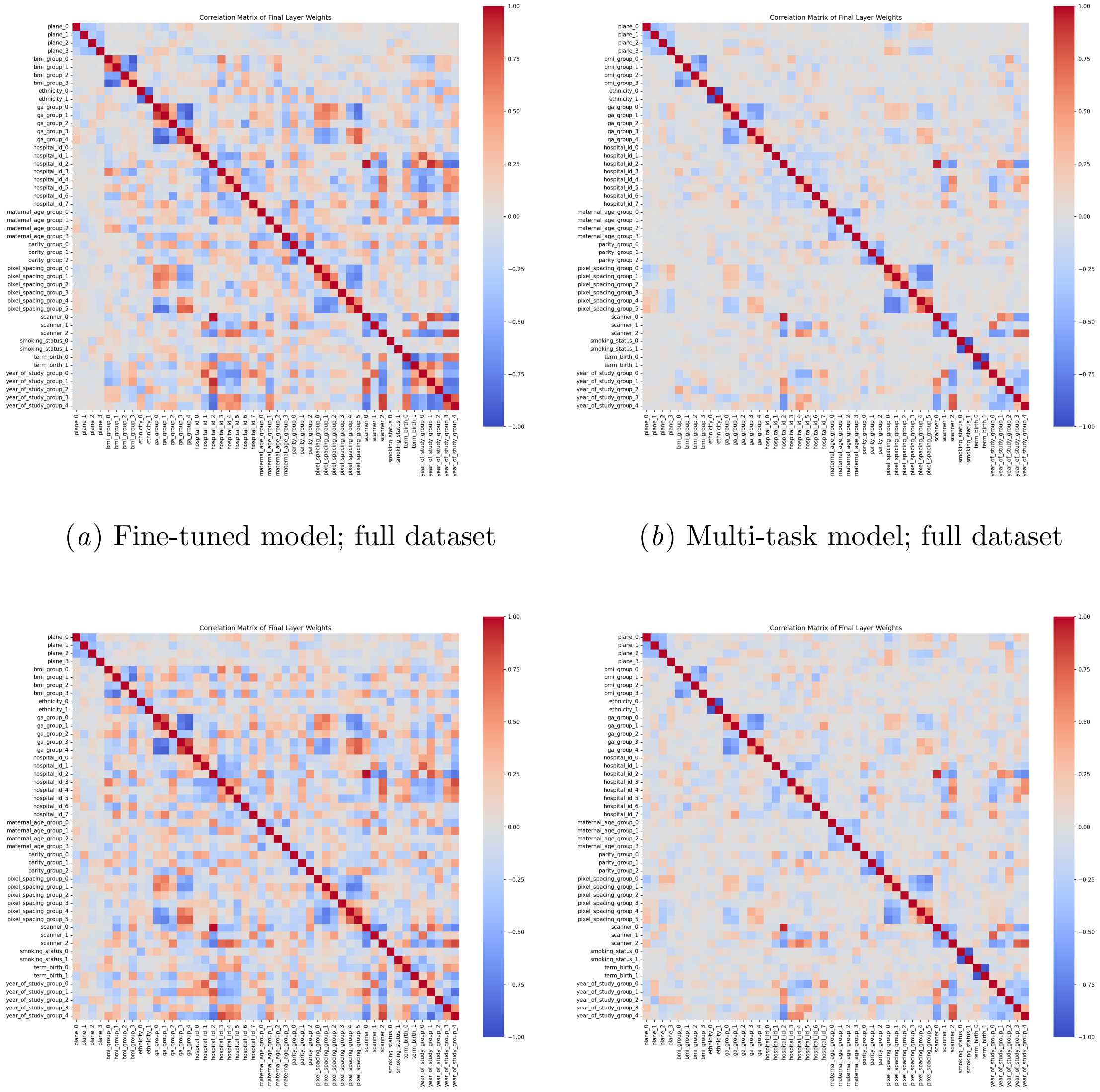
Target	Accuracy	Precision	Recall	F1	AUROC
Plane	95.8 ± 0.6	95.4 ± 0.6	95.4 ± 0.6	95.4 ± 0.6	99.6 ± 0.1
Scanner	97.6 ± 0.1	97.7 ± 0.2	97.2 ± 0.1	97.4 ± 0.1	99.8 ± 0.0
Pixel spacing	70.6 ± 0.9	69.5 ± 1.0	68.8 ± 1.0	68.5 ± 1.0	94.1 ± 0.2
Hospital ID	73.5 ± 2.0	56.7 ± 1.6	54.9 ± 2.4	54.5 ± 2.4	90.9 ± 1.4
Year of study	63.9 ± 5.5	51.4 ± 4.4	51.6 ± 3.4	49.4 ± 2.5	88.7 ± 1.2
GA	56.0 ± 1.5	50.3 ± 1.8	49.4 ± 0.4	49.3 ± 0.9	82.1 ± 0.6
BMI	37.2 ± 1.5	39.2 ± 1.3	37.1 ± 1.4	37.5 ± 1.3	65.4 ± 0.7
Parity	58.2 ± 7.5	38.0 ± 1.6	37.0 ± 1.8	35.4 ± 1.4	59.2 ± 2.1
Maternal age	26.9 ± 2.1	27.6 ± 1.6	27.5 ± 1.5	24.9 ± 2.5	52.0 ± 1.4
Smoking status	73.5 ± 8.4	51.0 ± 0.5	51.2 ± 0.9	50.0 ± 1.5	52.0 ± 1.0
Ethnicity	89.3 ± 0.8	49.9 ± 1.4	50.0 ± 1.6	49.8 ± 1.4	46.0 ± 3.9

Table 4: Test performance of ResNet50 classifier model trained to predict various target values in the fetal plane dataset. Continuous values are discretized and grouped into bins.

Target	Accuracy	Precision	Recall	F1	AUROC
Term Birth	55.7 ± 1.8	57.5 ± 2.1	57.1 ± 1.6	55.4 ± 2.0	60.7 ± 2.3
Scanner	97.4 ± 0.8	93.1 ± 3.7	85.6 ± 2.5	88.6 ± 2.6	98.9 ± 0.2
Pixel spacing	72.7 ± 15.2	70.4 ± 16.1	70.0 ± 16.4	70.1 ± 16.3	93.7 ± 4.8
Cervical length	62.5 ± 7.8	65.2 ± 7.0	61.8 ± 9.0	62.8 ± 8.7	85.0 ± 4.9
Hospital ID	46.3 ± 14.3	39.9 ± 14.9	38.9 ± 15.3	37.9 ± 15.4	81.2 ± 9.2
GA	39.3 ± 4.5	32.6 ± 3.8	30.6 ± 3.3	30.7 ± 3.3	66.9 ± 3.2
Year of study	26.5 ± 3.7	25.6 ± 3.0	24.2 ± 3.2	24.2 ± 3.2	62.9 ± 3.7
Birth weight	21.5 ± 2.2	19.7 ± 1.6	18.1 ± 0.9	16.9 ± 1.2	54.2 ± 0.5
BMI	45.6 ± 4.8	24.3 ± 1.2	25.0 ± 0.8	23.8 ± 0.5	51.9 ± 2.0
Maternal Age	30.7 ± 3.8	24.8 ± 5.1	26.7 ± 0.9	25.1 ± 3.4	51.9 ± 0.8
Placenta weight	35.5 ± 3.2	25.2 ± 2.0	25.4 ± 2.2	24.8 ± 2.1	50.5 ± 5.4
Smoking status	85.4 ± 1.8	50.4 ± 1.6	50.6 ± 1.7	50.4 ± 1.6	50.3 ± 2.5
Ethnicity	89.8 ± 1.5	48.9 ± 1.6	49.0 ± 1.6	48.9 ± 1.6	50.3 ± 2.4
Parity group	61.5 ± 1.1	34.1 ± 3.4	32.9 ± 1.1	31.7 ± 1.7	46.4 ± 3.9

Table 5: Test performance of ResNet50 classifier model trained to predict various target values in the cervix dataset.

Appendix B. Full covariance matrix plot



(a) Fine-tuned model; full dataset

(b) Multi-task model; full dataset

(c) Fine-tuned model; bias-induced dataset

(d) Multi-task model; bias-induced dataset

Figure 3: Full correlation matrix between weight vectors from classification head of each targets, extracted from fine-tuned or multitask classifier models, trained using the entire fetal dataset or with induced bias.

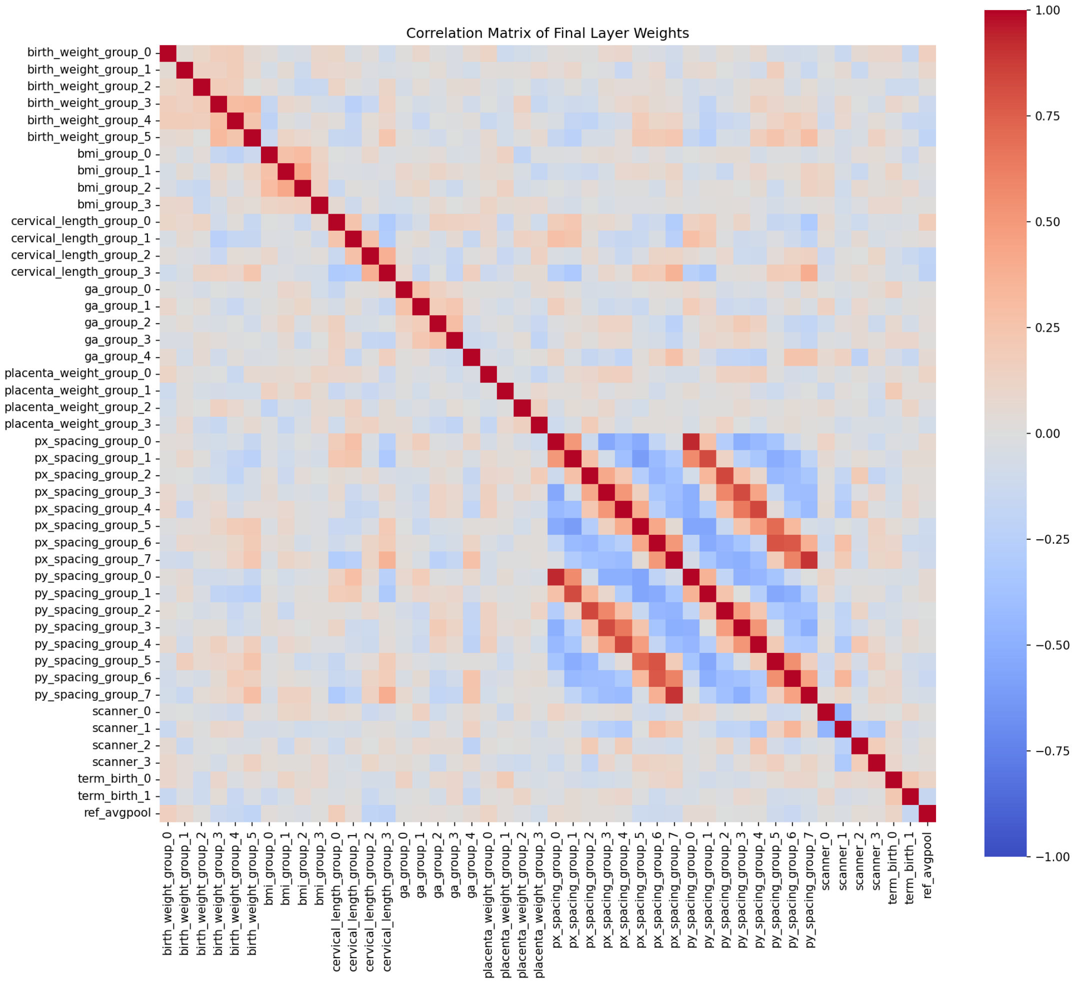


Figure 4: Full correlation between weight vectors from classification head of the targets of the fine-tuned SA-SonoNet model. A reference line is added to represent the average pool layer in the original model.