Computer Vision for Fetal Ultrasounds

Trevor Andrus
Computer Science Department
Brigham Young University
Provo, USA
Andrus.tn@gmail.com

Nathan Fastabend
College of Physical and Mathematical Sciences
Brigham Young University
Provo, USA
illiad@byu.edu

Braxton Owens
dept. of Computer Science
Brigham Young University
Provo, USA
cbo27@byu.edu

***Abstract*—**

***Index Terms*—Computer Vision, Fetal Gender, Ultrasounds**

I. Background

When it comes to prenatal care, obtaining quality ultrasounds and properly understanding those ultrasounds is critical to achieving proper diagnoses and providing premium patient care. Mistakes when interpreting an ultrasound image can result in health conditions being missed, misidentified, or falsely identified, resulting in the correct care being delayed if it is administered at all. Due to the risks associated with inaccurate results with regards to ultrasounds, this paper explores the option of using computer vision techniques to classify fetal gender, and to investigate possible future routes for building on this technology in the future.

II. Previous Work

As there is a high penalty for HIPAA violations and many healthcare providers follow the prevailing paradigm of delaying the implementation of new technologies until rigorous testing is performed, or until mandated to do so by governing bodies, very little research has been done with regards to this topic, and some of the research, like that performed by a team at Fellowship.AI, had to be done on publicly available data rather than on information provided by healthcare professionals [4]. This paper builds on the techniques for data collection used in their paper, applied to a better curated dataset provided by a local healthcare provider, to achieve a different application.

III. Data

The data for this exploration was provided by a local healthcare provider, hereafter referred to as LHP. More accurately, access to the raw data was provided, and the process for the collection, sanitation, and preparation necessary to make the data usable is detailed below. The raw data includes annotated ultrasound images dated as far back as February 1st, 2021, with corresponding hospital birth records for any patients who delivered while still remaining in the service area of LHP. In the event of fetal demise, the patient moving to a different healthcare provider, or the hospital who oversaw the delivery not reporting the information, no ground truth data was available.

A. Collection

The available ultrasound images were made available through a web portal which associated the images with a given patient, appointment type, and appointment time. The images were available as either jpg or dcm. Due to the extra patient information included in the dcm format, the dataset was exclusively composed of the jpg version in order to better preserve patient privacy. As the primary objective of this exploration was to predict fetal gender, the images selected were those annotated as depicting the gender, and indicated by the sonographer as containing the relevant information. This resulted in the majority of the images being taken from the 20 week Fetal Anatomy Survey, though they were not exclusively taken from this appointment type.

Unfortunately, the vast majority of the images saved by LHP were the annotated versions of the images, and not the unannotated images. These images would need to be adjusted in order to be useful in the later exploration, as detailed in the Preparation section below. Also, the images all contained a header with several pieces of patient identifying information which would need to be redacted, as detailed in the Sanitization section below.

After the images were collected, access was granted to LHP's patient records, including the hospital delivery reports and delivery summaries which provided information on the subsequent birth inasmuch as LHP had access. This information included the gender of the newborn, as well as the birth weight, estimated gestational age, and birth date. This additional information was collected to be used in future studies given the assumption that predicting gender from the collected images would prove to be an easy task. Not all of these data points were available for every image, due to the complications listed above. The available data was compiled in a spreadsheet, correlating an image number with the corresponding ground truth information.

In total, 303 images were collected, 270 having corresponding birth gender information, and 250 having a complete set of ground truth information. More data was available for collection, but the time requirements to collect the information, at $\tilde{5}$ minutes per record, prevented more data from being collected within the allotted timeframe of the project. It is estimated that over twenty thousand valid images are available from LHP, if a more efficient method of collecting them is devised in a future work.

B. Sanitation

Due to HIPAA compliance standards, which forbids the dissemination of any patient identifying information outside of authorized use cases, all of the collected data needed to be properly sanitized prior to use in training or to be shared among team members. The steps taken to properly sanitize the data are as follows: First, all team members were either employees of LHP or signed BAAs before being given access to the information. Second, all images and project data were stored in a Google Drive folder managed by LHP, and LHP holds a BAA agreement with Google to maintain data security. Third, prior to uploading the images to the secure folder, each image was renamed to correspond to a uniquely numbered ground truth record, and the header containing patient identifying information was cropped out. Fourth, all references to the patients' names were replaced with unique identifying numbers, all raw dates had the years removed, and raw dates were ultimately replaced with calculated relative dates.

After these sanitization efforts, no combination of the remaining information could conclusively be connected to any person, nor could the patient be identified by the information present without additional information, satisfying HIPAA compliance requirements.

C. Preparation

Some of the details of the preparation of the dataset have already been discussed, such as the creation of a spreadsheet to associate the downloaded images with their ground truth values. As some of the collected images did not have all of the corresponding ground truth information, some initial models were trained exclusively to predict fetal gender, and for those models the 270 images with just the gender were separated into two class folders for use in training and testing, with the remaining 33 being discarded. The results of this testing are discussed in greater detail below.

For the more advanced models attempting to predict the additional ground truth information, the 250 images with complete information available were packaged with the spreadsheet containing the relevant sanitized information as a Dataset object.

In both of these cases, the images needed one final adjustment, in the form of removing the annotations, before they could be used to train the models. This is due to not wanting the model to simply learn to recognize the words "boy", "girl", "XX", or "XY", in the images instead of the relevant features. Further, using the annotated arrows to find the relevant features was not desired, as not all images had arrows, and those arrows are not natural features of the ultrasound scan.

Luckily, all of the annotations produced by LHP were a particular shade of yellow not present anywhere else in the image. This allowed for a preprocessor to detect the yellow pixels and remove them from the image. Initially these pixels were replaced with an average color from the surrounding area, or black, but this resulted in a region that was often still legible. The second solution to remove the annotation was to use a thorough, albeit inefficient, blurring algorithm individually replacing each detected pixel and all the pixels around it, orthogonally and diagonally, with the average color of the non-detected pixels in a 5x5 area around the detected pixel. This resulted in an effective technique, as demonstrated in the figures below.

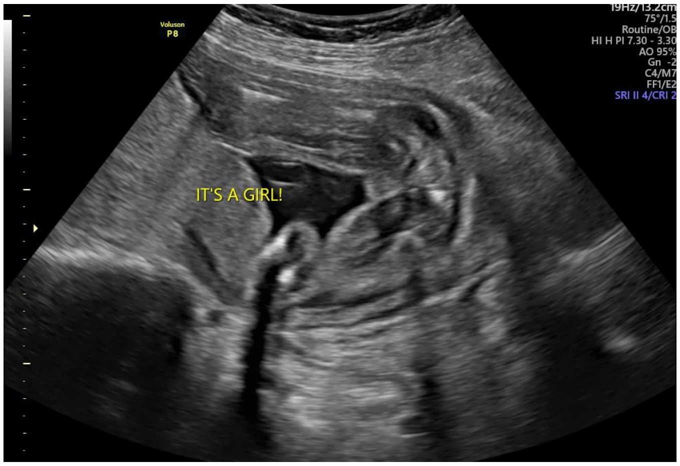

Fig. 1. One of the images with the header removed

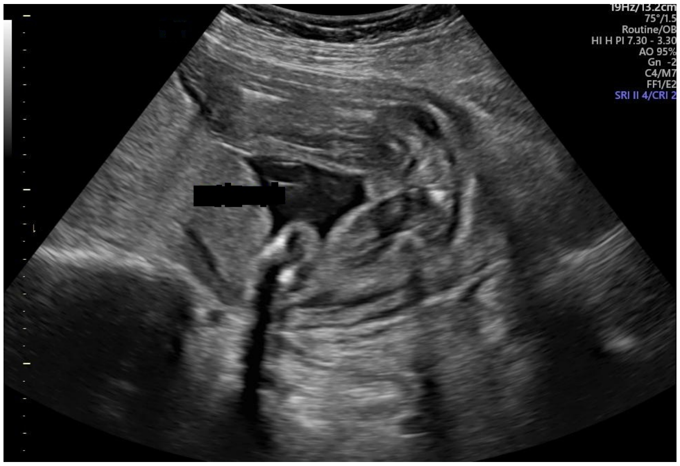

Fig. 2. The same image using the masking technique

As shown in the figures, the blurring technique is not perfect. The figures above show an average case for the quality, and the difference in quality most often depends on where the annotation is placed. For the images where the annotation was in the empty space, the blurred text was undetectable. For the images where the annotation covered a complicated area, the blurred text was more obvious. Regardless, the use of the blurring technique did not appear to have an adverse effect on the accuracy of the model.

D. Augmentation

Given the difficulty of collecting the data, as described in detail above, a sufficiently large dataset was not achievable

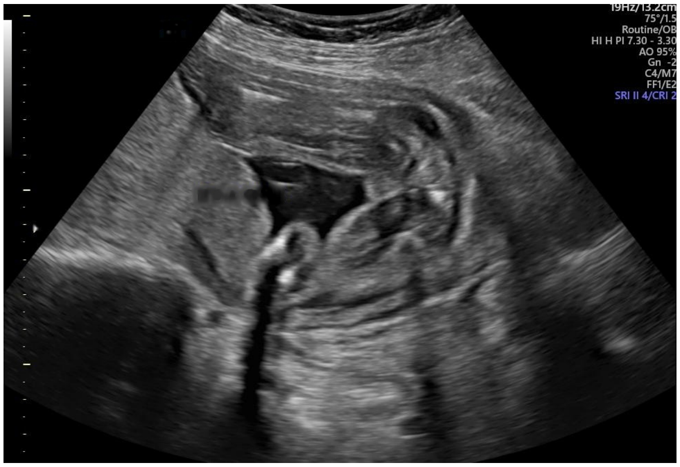

Fig. 3. The same image using the blurring technique

within the timeframe of the project. As such, some data augmentation efforts were employed to artificially increase the size of the dataset for use in training and testing, with varied results. Given the important features of the images were rotation and flip independent, and while some sonographers had a preferred orientation when creating the images during the ultrasound appointment, there were instances of a wide variety of orientations present within the dataset. As such, one of the first data augmentation procedures to be applied was four 90 degree rotations on each image after cropping them to a square. The desired feature was not always centered within the square, as it was not desired that the model learn that the feature would always be present in the center, but it was always within the cropped area. These rotations provided a fourfold increase in the number of usable images. Additionally, adding a horizontal flip before doing the rotations allowed for that increase to be doubled. Increasing the number of usable images to 2000 from the initial 250.

Additionally, random rotations were considered and briefly tested, but they were not sufficiently successful in initial testing and were discarded. Future iterations on this project may consider revisiting the effectiveness of random rotations, as well as exploring the use of regression models to fill in the gaps for images that did not have complete corresponding ground truth information, as another option for data augmentation.

IV. Alternative Data

Due to the time constraints on the project, and the difficulty of obtaining the desired dataset, a premade alternative ultrasound dataset, concerned with detecting and diagnosing tumors in ultrasounds, was used in an effort to test potential architectures in preparation for when the fetal dataset would be ready for use. This was done with the expectation that an architecture that performed well on the tumor dataset would perform similarly on the fetal dataset. The results in the subsequent sections will thus make reference to this alternative dataset, and the models explored will often compare the results on the two datasets in order to determine performance. This tumor dataset, (hereafter referred to as tumor data), consists of 780 labeled ultrasound images split into 3 classes: malignant, benign, and normal. There was moderate skew in the class distributions in the tumor data (while we saw an even split between male and female in the final fetal data). The before mentioned augmentation methods (90 degree rotations, and random rotation sampling) were also applied to this dataset to increase the sample size. However, as will be shown in subsequent sections, augmentation actually decreased the accuracy of our classification models on both datasets.

V. Feature Extraction

The final step of preparing the data for classification was feature extraction. Feature extraction transforms raw input images into a more compact (and sometimes more informative) representation of input data. Feature extraction methods are commonly used in image classification to reduce dimensionality, improve generalizations, and reduce computation time for large algorithms. While one could hypothetically pass full-length arrays into classifiers, feature extracted representations of images have been shown to improve classification accuracy in some applications. As such, two of the most common feature extraction techniques were applied to our data before classification:

A. HOG Transform

The HOG transform (Histogram of Oriented Gradients) was first introduced conceptually in 1986, but wasn't popularized in classification until 2005 when Navneet Dalal and Bill Triggs presented supplementary work on the concept at CVPR. [1] HOG transforms are motivated by the fact that object appearances within an image can be described by the intensity of their gradients. In application, the descriptor splits an image into cells, and computes the histograms of gradient directions for each cell. The compilation of all the cell's histograms constitutes the output of the HOG transform. HOG is particularly useful for capturing the shape and texture information of objects, and is renowned for its efficacy in classification tasks.

B. SIFT Descriptors

Invented by David Lowe in 1999, [2] SIFT descriptors (Scale Invariant Feature transform) are very similar to the HOG transform. They also compute the 'edginess' of pixels through gradient orientation, but instead of computing in blocks over a whole image, they calculate points of interest on which to focus. This keypoint detection is one of the main features that separate the two methods. Because of this focus on points of interest, SIFT descriptors often excel in object recognition tasks, rather than classifying entire images.

More detailed analyses of the specific computations and strengths of each of these methods is surely useful, but outside

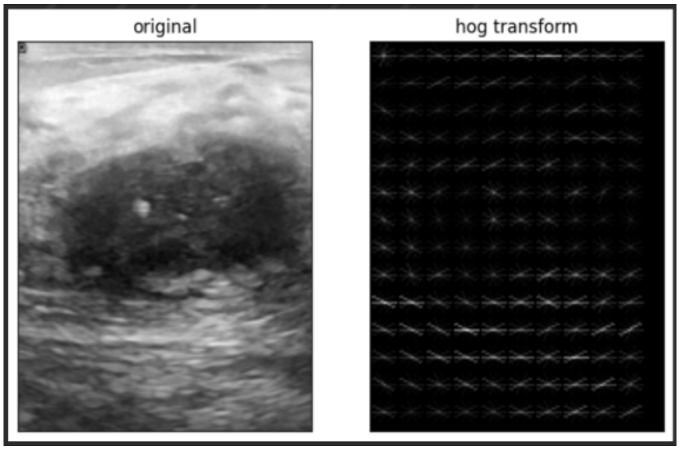

Fig. 4. Example of a HOG transform on a sample tumor image

the scope of this project. We find it sufficient to say that each of these methods serves to extract features from input images, and the efficacy of each is worth testing in our specific classification problem.

VI. Classification

Like was described in the previous alternative data section, the majority of our architecture exploration was first performed on tumor data, and then transferred to fetal data once data collection was complete. As such, we will report classification accuracies for both datasets, to give insight into the methods of model choice. A note here - we decided that given our application, accuracy would be our most applicable evaluation metric. We toyed with the idea of using precision, recall, and F1, but in determining fetal gender, accuracy would provide us with the most effective measure of success (As there is no reason to minimize type 1 or type 2 error).

A. *Classical Machine Learning*

While the world of computer vision is often dominated by the complexity of neural algorithms, we wanted to explore the capabilities of classical machine learning algorithms on this vision task. We created a program to test data on 10 popular machine learning classifiers: SVM (Support Vector Machine), SGD (Stochastic Gradient Descent), RF (Random Forest), KNN (K-Nearest Neighbors), NB (Naive Bayes), DT (Decision Tree), LOGREG (Logistic Regression), MLP (Multilayer Perceptron), and VOTE (Voting Classifier -Ensemble of 5 previous classifiers). In an attempt to be as thorough as possible in testing these classical methods, we trained them on a variety of data, augmentation, and feature extraction combinations. For both the tumor and fetal dataset, each of these algorithms was tested with non-feature extracted data, HOG transformed data, and SIFT descriptor data - each in augmented and non-augmented form. The accuracy scores from the best augmentation-feature extraction pairs are shown below:

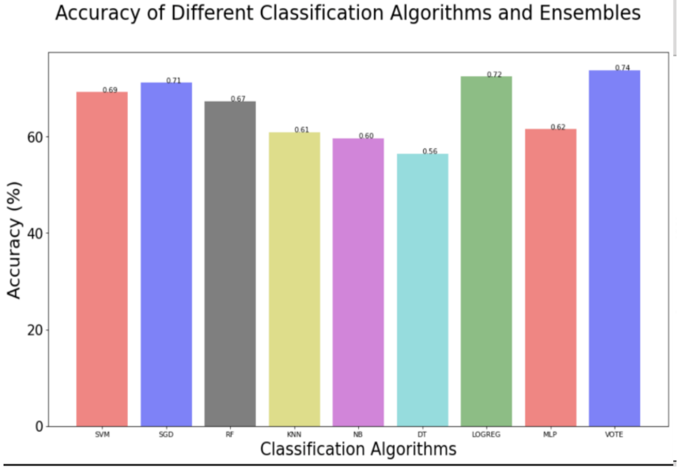

Fig. 5. Tumor Model Accuracies

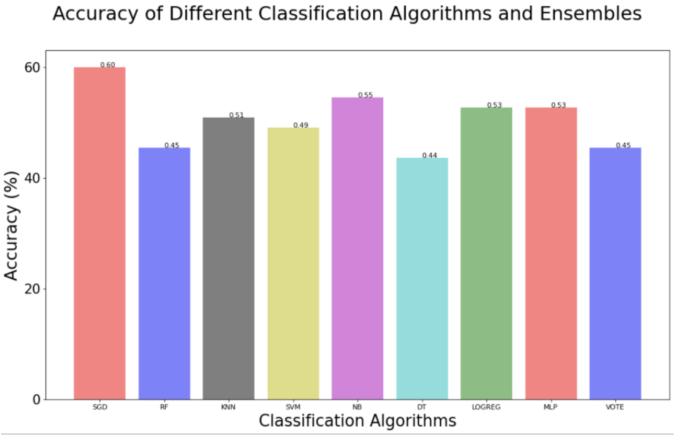

Fig. 6. Fetal Model Accuracies

After trying all possible augmentation-feature extraction combinations, we found that the non-augmented, HOG transformed data returned the highest model accuracy. Like we had originally hoped, the combinations that worked best for the tumor data also worked best for the fetal data. Interestingly, we found that in both datasets, data augmentation actually reduced accuracy. We're not sure if this is due to our augmentation methods, or if the small dataset causes issues with augmentation. We also interestingly found a difference in the specific best-performing model between the two datasets. The format of this paper renders the model labels in figures 5 and 6 basically unreadable, however, in figure 5, we see that the maximum accuracy is reported by the Voting ensemble method (far right bar) with an accuracy of 74% across the 3 classes of the tumor data. However, in the fetal data (figure 6) we see that our maximum accuracy is reported by the Stochastic Gradient Descent algorithm (far left bar) for an accuracy of 60% across the two classes.

B. *Grid Search and Ensemble Methods*

Impressed by the efficacy of the voting ensemble method on the tumor data, we also explored other ensemble methods - namely ADA boosted decision trees, Gradient boosting machine, ensemble stacking, and bagging methods. Unfortunately, we didn't see any significant increases in accuracy from these methods. As a final attempt to increase accuracy, we also employed a parameter grid search to tune the hyperparameters of our estimators. Again, unfortunately, despite testing multiple types of loss functions, penalty functions, and numbers of iterations, our accuracy never significantly improved over the base classifiers.

C. *Neural Methods*

After exhausting the utility of our classical machine learning methods, we turned to neural methods to further improve our classification accuracy. Specifically, we employed an implementation of VGG-19. VGG-19 is an open source neural classifier out of the Visual Graphics Group at Oxford university. [3] The Model is a deep convolutional neural network consisting of 16 convolutional layers, 5 max pooling layers, 3 fully connected layers and 1 softmax layer in the following configuration:

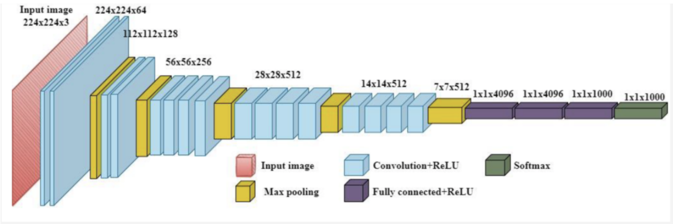

Fig. 7. VGG architecture

Using this out of the box CNN, we were able to see significant improvements in the accuracies of both tumor and fetal data. As can be seen in figures 8 and 9, the VGG classification accuracy is a head above all of the classical methods (including ensemble methods).

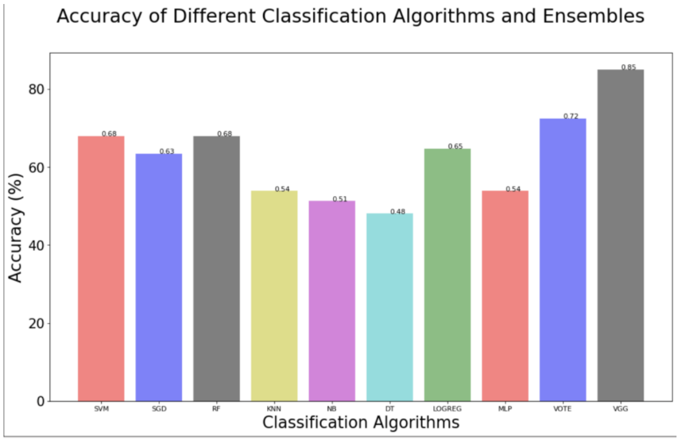

Fig. 8. Tumor classification with VGG

For Tumor data, VGG-19 reported a classification accuracy of 85% (Far right bar) - over 10% higher than any accuracy we were able to achieve with classical methods.

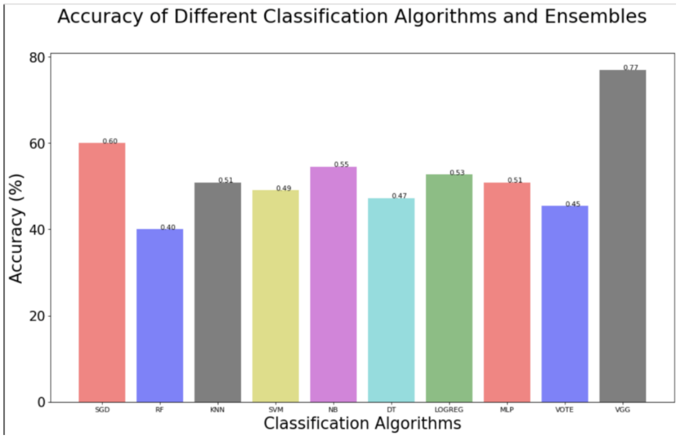

Fig. 9. Fetal Classification with VGG

Similarly with the fetal data, we see that VGG (far right bar) far outperforms all classical methods with a classification accuracy of 77% - A full 17% higher than our best accuracy with classical methods.

This stands as a testament to the power of neural methods in computer vision applications - CNN's and transformers are far ahead of classical machine learning methods in these types of problems.

Given more time, we would have liked to experiment further with neural methods - perhaps fine tuning VGG to work specifically with ultrasound data, or experimenting with custom methods for classification. We do employ a custom CNN in subsequent sections, but were not able to see the kind of accuracy we saw from the out-of-the-box VGG implementation.

VII. Predicting Fetal Weight and Birth age

In this project, we aimed to develop a convolutional neural network (CNN) capable of predicting baby weight and birth age. Our initial CNN model, implemented in Pytorch, was relatively simple, designed to capture fundamental patterns in the dataset. However, it soon became evident that the model was primarily predicting mean values for both baby weight and birth age. This issue indicated that the model was not learning effectively from the data and was unable to capture the nuances necessary for accurate predictions. There are plenty of reasons why the model was not able to learn the appropriate patterns. So as a first attempt, we decided to implement a more complicated model using VGG19, a CNN architecture known for its depth and complexity. VGG19's architecture, with its multiple convolutional layers, was expected to extract more detailed features from the data, thus enabling more nuanced predictions. VGG19 is characterized by its depth, consisting of 19 layers (16 convolutional layers, 3 fully connected layers). It uses small (3x3) convolution filters throughout the model, which helps in capturing fine-grained patterns in the data. We adapted the VGG19 model to suit our specific task. This involved modifying the final layers of the network to output predictions for baby weight and birth age. We also adjusted the input layer to accommodate the format and dimensions of our dataset.

Training the VGG19 model required significant computational resources due to its complexity. We used a combination of advanced optimizers and regularization techniques to manage overfitting. The training process was closely monitored to adjust parameters as needed for optimal learning. Because of the desire to make all of the research reproducible, our code was written in Google Colab notebooks. About 15GB of GPU RAM was available for us to use. Because of this, When training a deep and complex neural network like VGG19, especially on datasets with large input sizes or high-resolution images, memory constraints can become a significant challenge. To address this, utilizing techniques like automatic mixed precision (autocast) and gradient scaling (Scaler) can be crucial. Autocast is a feature in deep learning frameworks that automatically chooses the precision (either 32-bit or 16-bit) for different operations in the neural network. By performing certain operations in lower precision (like 16-bit floating points), it significantly reduces the memory usage and speeds up computation, without a substantial loss in model accuracy. This is particularly useful for training large models like VGG19, where memory overhead can be a limiting factor. Alongside, the Scaler, often used in conjunction with autocast, helps in managing and scaling the gradients during backpropagation. This ensures that the use of mixed precision does not lead to underflow or overflow in the gradients, which can hinder the training process or lead to NaN errors. Implementing autocast and Scaler together allows for efficient memory usage and maintains the stability of the training process, enabling the training of VGG19 on more constrained hardware without running out of memory.

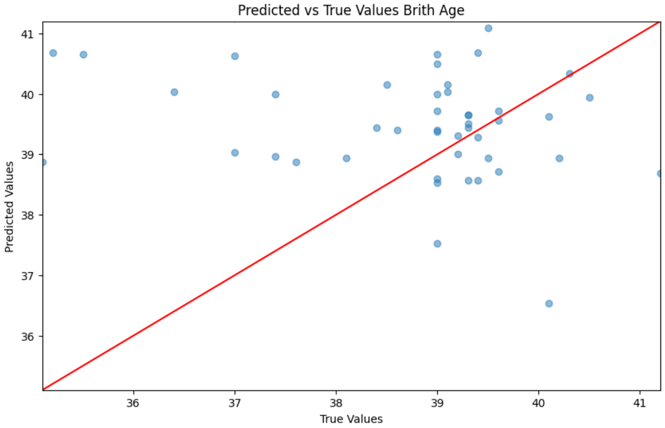

Fig. 10. Scatter plot displaying the relationship between predicted and true values of birth age with a line of best fit, indicating the degree of prediction accuracy

Despite implementing the more sophisticated VGG19 model in our project, we encountered a persistent issue where the model continued to predict mean values, indicating a lack of diversity and precision in its predictions. To combat this, we employed a range of strategies within the PyTorch framework aimed at enhancing model performance and avoiding mean prediction. The first of these strategies included data augmentation. By introducing a random rotation transform we hoped to give the model more data to learn from and to introduce more variability in the training data. We adjusting learning rates and other hyperparameters, and experimenting with different loss functions that could penalize average predictions more effectively. Additionally, we explored the use of dropout and regularization techniques to prevent overfitting. We also considered the use of advanced optimization algorithms like Adam and RMSprop to refine the learning process. Despite these comprehensive efforts, we unfortunately did not observe significant improvements in the model's ability to move away from predicting mean values. This outcome highlighted the complex nature of our task and suggested the need for further investigation into both the data and model architecture to identify and address the underlying issues more effectively.

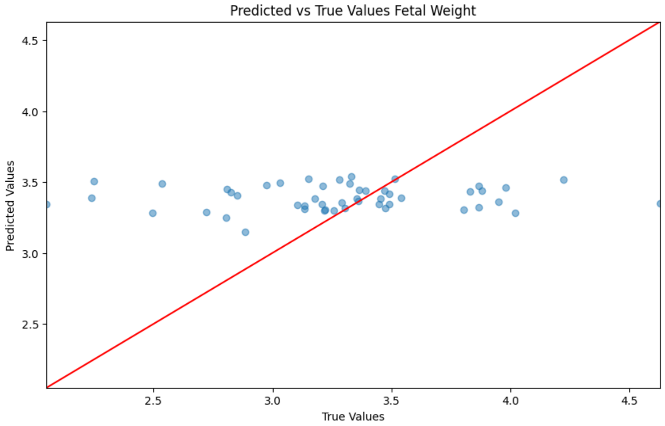

Fig. 11. Scatter plot comparison of predicted versus true fetal weight values with a linear regression line illustrating the prediction model's performance

The persisting challenges in accurately predicting weight and birth age using the VGG19 model led us to a crucial realization regarding the nature of our dataset. It became increasingly evident that the dataset might lack the essential information necessary for making these predictions. This hypothesis aligns with the fact that our training dataset predominantly consisted of sonographic images focused on the genitalia, a region that does not typically provide the critical measurements (like femur length, abdominal circumference, head circumference, etc.) used in standard fetal weight and age estimations. The absence of these key anatomical features in our dataset likely deprived the model of the vital data required for accurate predictions. This situation underscores the principle often referred to in data science as ”garbage in, garbage out,” which in a more scientific context, emphasizes the importance of relevant and comprehensive data for training machine learning models. The quality and relevance of input data are paramount in determining the effectiveness of the model's learning and prediction capabilities. Therefore, this project's outcomes suggest a need for a more carefully curated dataset, one that includes images or measurements directly correlated with fetal weight and age, to train the model more effectively. This refinement in data selection is crucial for ensuring that the input data encompasses the necessary information for the model to learn and make precise predictions.

REFERENCES

[1] N. Dalal and B. Triggs, ”Histograms of oriented gradients for human detection,” 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), San Diego, CA, USA, 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

[2] D. G. Lowe, ”Object recognition from local scale-invariant features,” Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 1999, pp. 1150-1157 vol.2, doi: 10.1109/ICCV.1999.790410.

[3] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

[4] M. Fuller et al. (2021), “A Deep Learning Approach for Masking Fetal Gender in Ultrasound Images,” arXiv:2109.06790 [cs.CV].

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.