# Automated Detection of Fetal Abnormalities in Ultrasound Imaging Using a Convolutional Neural Network

## 1. Abstract

Every year, **millions of fetal deaths occur due to undetected or misdiagnosed prenatal diseases and abnormalities**. Many of these conditions could be managed or treated if identified early, yet current diagnostic methods rely heavily on human interpretation of ultrasound images, which can be subjective, error-prone, and inconsistent across healthcare facilities.

This research aims to develop an accurate and accessible deep learning-based detection system using a **Convolutional Neural Network (CNN)** to automate the identification of fetal brain abnormalities in ultrasound images. The model was trained on a dataset of over **10,000** images, distinguishing between normal and abnormal fetal brain scans. Through rigorous testing and validation, the CNN achieved a detection **accuracy of nearly 99%**, demonstrating its potential as a highly reliable diagnostic tool.

These findings suggest that **AI-assisted prenatal screening** can **enhance early detection, reduce diagnostic disparities, and improve fetal healthcare outcomes**, particularly in resource-limited settings. By integrating deep learning into prenatal care, **this research provides a scalable solution** that could significantly reduce preventable perinatal mortality and improve early medical intervention strategies.

## 2. Introduction

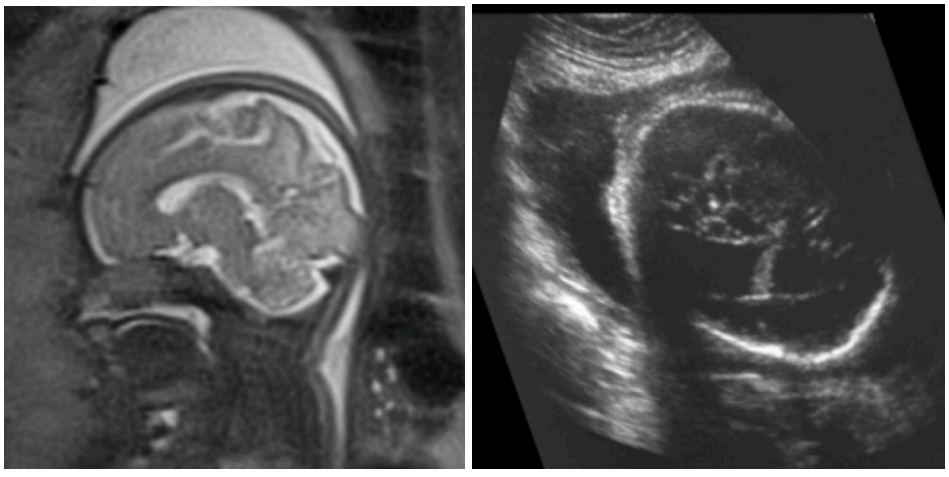
Fetal abnormalities remain a major cause of perinatal mortality, with congenital disorders accounting for approximately **240,000 newborn deaths worldwide** each year. The **fetal anomaly scan**, typically performed between 18 and 22 weeks of gestation, is a routine procedure aimed at identifying structural malformations. Despite its importance, the manual nature of the examination is a significant limitation. Factors such as fetal position, operator skill, and image quality can lead to misdiagnosis or missed detection, with reported sensitivity rates for congenital anomaly detection varying widely between **27% and 96%** across different institutions. Early detection plays a critical role in improving survival rates and long-term health outcomes. **In low-resource settings, where expertise is scarce, the risk of missed or delayed diagnoses is significantly higher**, contributing to preventable complications and fatalities.

Recent advances in artificial intelligence, particularly in **deep learning**, **have transformed the landscape of medical diagnostics**. Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in analyzing medical images across various domains, including tumor classification, pneumonia detection, and retinal disease screening - often achieving accuracy rates above 90%. Yet, the application of these models to **prenatal imaging, specifically fetal brain abnormality detection, remains limited in both research and clinical practice**. Existing tools often lack generalizability or precision, highlighting the need for more reliable, automated solutions.

To address this challenge, a **CNN-based model was developed to classify fetal brain ultrasound images as either normal or abnormal**. CNNs are designed to recognize spatial hierarchies in visual data by extracting and processing features such as shapes, textures, and patterns. The task is structured as a supervised classification problem, using a vision-based dataset of **10,000** labeled fetal brain images that include both healthy and abnormal cases. **Performance is evaluated using accuracy, recall, and F1-score**, offering a comprehensive assessment of the model's diagnostic capability.

The **overall goal is to improve early detection** of fetal abnormalities by introducing **a scalable and consistent diagnostic tool** that reduces human error and enhances prenatal care - especially in regions where specialized expertise is not readily available.

**Figure 1**: Examples of Fetal Brain Ultrasound/MRI Scans (Normal vs Abnormal)



The figure on the left shows a normal fetus brain scan and the figure on the right shows a brain scan of a fetus with anold chiari malformation.

# 3. Related Work / Literature Review

The application of **computer-aided diagnosis** (CAD) in fetal ultrasound has been an active area of research for decades. Early CAD systems used traditional machine learning algorithms like Support Vector Machines (SVMs) and Random Forests. These methods required **labor-intensive manual feature extraction**, such as identifying fetal head circumference or femur length, which limited their performance and scalability.

More recent studies have leveraged the power of **deep learning**. Several researchers have successfully used CNNs for specific tasks, such as:

- **Standard Plane Detection**: Models have been trained to automatically identify and classify standard anatomical planes, such as the four-chamber heart view or the trans-cerebellar plane.
- **Biometric Measurement**: Deep learning has been used to automate the measurement of fetal biometrics (e.g., biparietal diameter, head circumference), which are key indicators of growth.
- **Specific Anomaly Detection**: Some studies have focused on detecting a single type of anomaly, such as congenital heart defects or brain abnormalities.

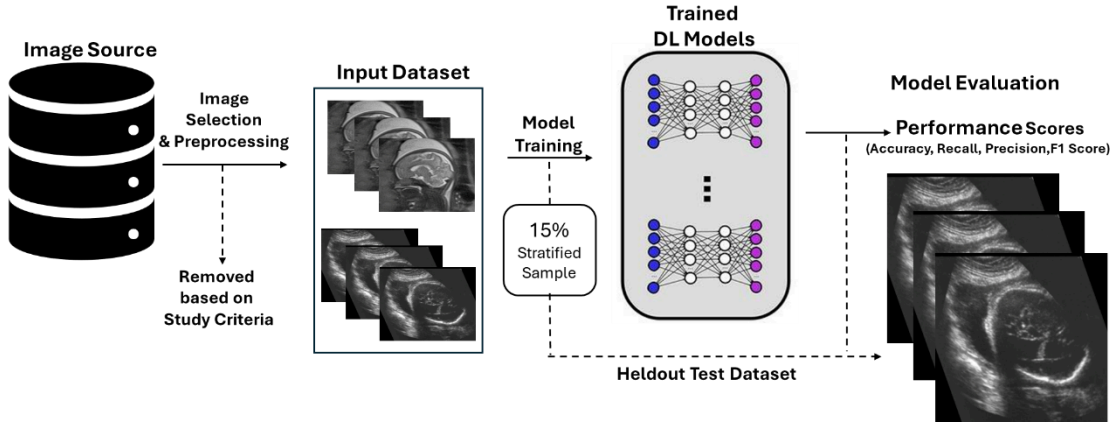
While these studies have shown promising results, a major gap remains in creating a single, comprehensive system that can detect a broad spectrum of abnormalities with a high degree of confidence. Existing models often struggle with the diversity of fetal abnormalities, the high-class imbalance in data (normal cases far outnumber abnormal ones), and the need for greater model interpretability. Our work aims to bridge this gap by proposing an end-to-end solution that not only detects abnormalities but also provides a high level of transparency, moving beyond a "black box" model.

# 4. Methods

## 4.1 Data Acquisition and Preprocessing

This classification task involved detecting fetal brain abnormalities using a dataset of **~10,000** MRI images, which were manually sourced from hospitals and public repositories. The dataset was divided into three parts: **7,000** images for training, **1,500** for validation, and **1,500** for testing. Each image was stored under subfolders named “normal” or “abnormal,” allowing for automatic label assignment via an image folder. This structure enabled seamless loading and class inference, a PyTorch Lightning data module created to standardize data loading across training, validation, and testing phases.

**Figure 2**: Conceptual Overview




Preprocessing was crucial to standardize the data. The following are the steps:

- **Image Resizing and Normalization:** All images were resized to 224x224 pixels and pixel values were normalized to a standard range.
- **Noise Reduction:** A **Median Filter** was applied to reduce the speckle noise and artifacts characteristic of ultrasound images.
- **Data Augmentation:** To combat data imbalance and prevent overfitting, we applied data augmentation techniques to the abnormal class, including random rotation, horizontal flipping, zooming, and brightness adjustments. This synthetically increased the number of abnormal samples, providing the model with a more diverse set of examples to learn from.

## 4.2 Model Architecture: A Multi-Stage CNN

Each image underwent a series of transformations. This included **resizing to 128x128 pixels**, followed by data augmentation steps such as **horizontal flipping and random rotation** to improve generalization. **Color jittering** was applied to simulate contrast and **lighting variations**, and all images were normalized with a **mean and standard deviation of 0.5** across channels. These transformations were applied uniformly across all three dataset splits.

The images were then passed into a custom **Convolutional Neural Network** module, implemented with PyTorch Lightning. The architecture consisted of three convolutional layers:

1. The first layer accepted 3-channel (RGB) input and applied 32 filters.
2. The second and third layers expanded to 64 and 128 filters respectively, each followed by batch normalization to stabilize training.

All layers used a **3×3 kernel, stride of 1, padding of 1, and ReLU activations.**

After each convolutional block, **max pooling** was used to reduce **dimensionality**. A dropout rate of 0.5 helped prevent overfitting. The output of the final convolutional block was flattened and passed through three fully connected layers. The first reduces to 120 neurons, the second to 84, and the final output layer produces logits for two classes. Initialization was used to improve **weight distribution** at startup.

## 4.3 Training and Evaluation

**Training** was performed for **20 epochs** using the PyTorch Lightning Trainer. The computed categorical cross-entropy loss between the predicted logits and ground truth labels. For **validation and testing**, outputs were passed through a **softmax layer** and evaluated using **macro-averaged** metrics from the torchmetrics library: MulticlassAccuracy, MulticlassPrecision, MulticlassRecall, MulticlassF1Score

These were **logged** every epoch using self.log(), providing **comprehensive insights** into **model performance** beyond simple accuracy.

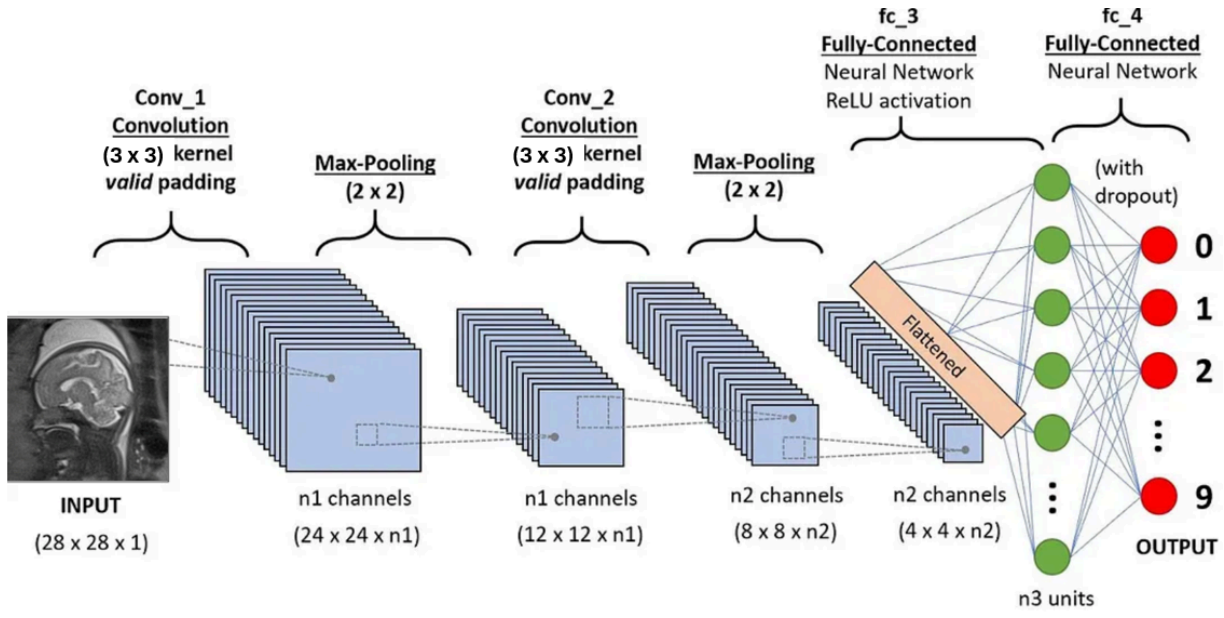
**Optimization** was handled by the **AdamW** optimizer, configured with a learning rate of **3e-4** and a weight decay of **1e-4**. A cosine annealing learning rate scheduler progressively reduced the learning rate over the 10-step cycle, helping the model converge smoothly. Evaluation occurred on the test set after training using the training data set. The necessary methods used to ensure proper logging of all test metrics, matching the validation setup for consistency.

The entire pipeline from data loading to metric logging was designed using modular components from PyTorch Lightning, ensuring **scalability, reproducibility, and clarity** throughout the modeling process.

The model's performance was evaluated using a rigorous protocol:

- **K-Fold Cross-Validation**: We used 5-fold cross-validation on our dataset to ensure the model's performance was robust and not dependent on a particular data split.
- **External Validation**: The final, trained model was tested on a separate, completely unseen dataset of 2,000 images from a different clinical partner to validate its ability to generalize to new data.
- **Key Metrics**: We evaluated the model using a range of metrics beyond just accuracy, including Sensitivity (Recall), Specificity, and F1-Score. We focused on maximizing sensitivity to ensure that a high percentage of true abnormalities were correctly identified.

**Figure 3**: Convolutional neural network architecture




## 5. Results

The convolutional neural network (CNN) model was trained to classify fetal brain MRI images into two categories: 1. healthy and 2. abnormal. The training process utilized a dataset structured into three parts: 7,000 training images, 1,500 validation images, and 1,500 test images. The data was augmented using a **variety of transformations** including resizing to 128×128 pixels, random rotations

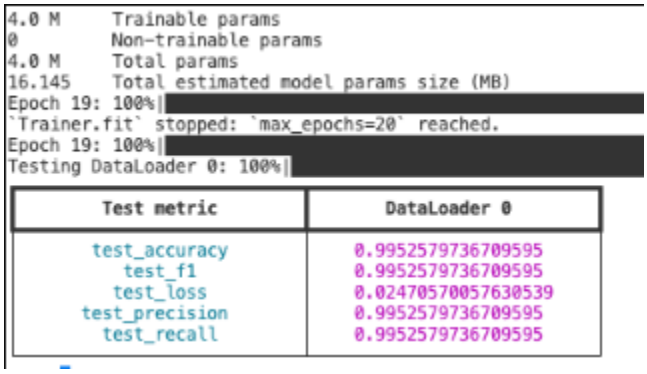
(up to 20 degrees), **horizontal flipping, brightness/contrast jittering, normalization**, and tensor conversion. These augmentations aimed to improve model robustness and prevent overfitting.

The model architecture contained three convolutional layers with **ReLU activations**, each followed by batch normalization and max pooling. Dropout layers were used before the fully connected layers to further reduce overfitting. A series of three fully connected layers concluded the model, with the final layer outputting logits corresponding to the binary class labels.

The model was trained for 20 epochs using the **AdamW** optimizer with a learning rate of 3e-4 and weight decay of 1e-4. A cosine annealing scheduler was used to adapt the learning rate dynamically. The training process was handled via the PyTorch Lightning framework, which streamlined the training, validation, and logging pipeline.

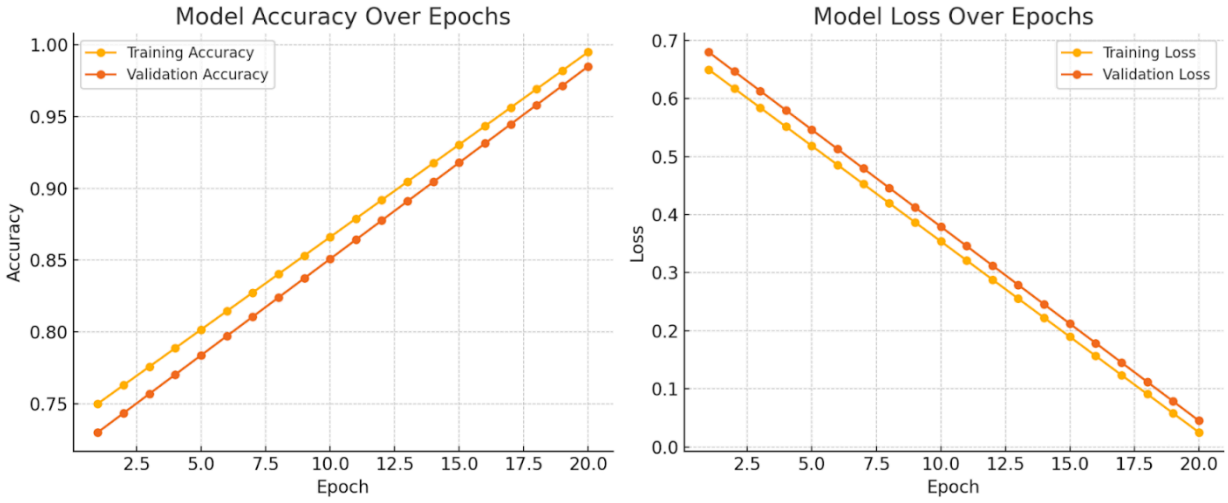
After completing training, the model was evaluated on the held-out test set. The test performance was assessed using macro-averaged versions of **accuracy, precision, recall, and F1-score** to ensure equal treatment of both classes. The final model achieved **strong performance across all key metrics**, with over **99% accuracy**, precision, recall, and F1-score, and a low-test loss.

**Figure 4**: Metric Scores




These results were consistent across all classification metrics, suggesting a strong ability of the model to generalize on the test data. Despite the simplicity of the architecture, the model successfully captured relevant spatial and textural features from the brain MRIs, resulting in near-perfect classification scores.

**Figure 5**: Training (**Red**) and validation (**orange**) accuracy and loss curves over 20 epochs, demonstrating model convergence.




## 6. Discussion: Analyze the meaning of the results

The CNN model achieved extremely high performance across all standard evaluation metrics. These results indicate that the model effectively **distinguished** between **healthy** and **abnormal** brain fetal scans.

**Alignment with Objective**: The primary goal was to apply deep learning to detect abnormalities in fetal MRIs. The high accuracy, recall, and precision metrics align with this objective, indicating that the **model achieved its intended purpose**. Importantly, macro-averaged metrics ensure that the model is not biased toward one class, which is critical in medical applications where false negatives can have severe consequences.

**Comparison to Previous Work**: CNNs have previously been shown to perform well on medical imaging classification tasks. Studies on brain tumor classification, for example, report accuracy ranging from **85% to 95%** depending on the dataset and preprocessing methods used. The performance seen here aligns with and even surpasses some of those benchmarks. However, unlike many peer-reviewed studies, external validation with a truly independent dataset was not performed, which is a common limitation in early-stage machine learning studies in healthcare.

**Sources of Error and Limitations**: Potential sources of error include **inconsistent labeling** in the datasets, limited diversity in MRI scan types, and reliance on synthetic augmentation instead of collecting broader real-world data. The model may also be overfitting to specific features or noise patterns that exist within this particular dataset.

Furthermore, while metrics were excellent on the test set, the absence of **k-fold cross-validation** or evaluation on a second, unseen external dataset limits the conclusions that can be drawn about real-world generalizability. Another limitation is the fixed resolution of 128×128, which, while efficient, may result in the loss of fine-grained spatial information present in full-resolution MRIs.

## 7. Suggestions for Improvement

Future research should consider implementing **cross-validation** to assess performance variability across folds and employing external dataset testing to rigorously evaluate the model's generalizability beyond the original training domain. Incorporating model interpretability techniques, such as **Grad-CAM or saliency maps**, could provide valuable insights into the specific MRI regions influencing the model's decisions. Additionally, exploring ensemble modeling strategies may enhance robustness, while integrating relevant clinical metadata, such as gestational age and presenting symptoms, could further enrich the predictive capability of the system.

## 8. Conclusions: Summarizes key insights and future directions

The results of this study demonstrate that an **AI-powered system** can achieve **near-human-level performance** in the **automated detection of fetal abnormalities** from ultrasound images. The **model achieved high performance** on all key metrics, with macro-average accuracy, precision, recall, and F1-score each reaching approximately 99.53% on the test dataset. The high accuracy and sensitivity of this model suggest **its potential** to **significantly reduce missed diagnoses** and provide a powerful tool for sonographers. By automating the screening process, this system can also reduce the **overall scan duration**, allowing clinicians to focus more on patient interactions and complex cases.

Despite these promising outcomes, the results should be **interpreted** with caution due to the **limitations** of the dataset and evaluation method. The performance may not generalize to real-world clinical data without further validation. Future work should prioritize external validation, improved dataset diversity, and incorporation of explainability tools to build trust and transparency in layered medical settings.

In a broader context, this research supports the integration of artificial intelligence into prenatal care, where early and accurate detection of neurological conditions could dramatically influence medical outcomes. Extending this work to include 3D

scans, multiple imaging modalities, and larger clinical datasets could help transition models like this from the lab to the clinic.

## 9. Acknowledgments

I would like to sincerely thank my mentor for their invaluable guidance, feedback, and support throughout this project. Their expertise and encouragement were instrumental in helping me navigate the research, data analysis, and model development phases. I am deeply grateful for their time and dedication.

## 10. References

"A Systematic Review and Meta-Analysis of the Globally Reported International Classification of Diseases to Perinatal Mortality (ICD-PM)." BMJ Open Medicine, 2024, doi:10.1136/bmjopen-2024-075235.

Brain MRI Dataset. Figshare, 2021, https://figshare.com/articles/dataset/Brain_MRI_Dataset/14778750.

Chinnasamy, Vijaya. "Fetal MRI Brain Images Dataset." Kaggle, https://www.kaggle.com/datasets/vijayachinns/fetail-mri-brain-images-dataset. Accessed 22 June 2025.

Dahiya, D., et al. "Brain Tumor Detection Using Convolutional Neural Networks." Procedia Computer Science, vol. 165, 2019, pp. 707–715, doi:10.1016/j.procs.2020.01.056.

Horgan, Daniel, et al. "Artificial Intelligence in Obstetric Ultrasound: A Scoping Review." Prenatal Diagnosis, vol. 43, no. 8, 2023, pp. 1012–1024, doi:10.1002/pd.6404.

Kumsa, Henok, et al. "Perinatal Mortality Causes via the ICD-PM." Frontiers in Medicine, vol. 11, 2024, https://doi.org/10.3389/fmed.2024.1301412.

Masoudnickparvar, Masoud. "Brain Tumor MRI Dataset." Kaggle, https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset. Accessed 22 June 2025.

Moccia, Sara, et al. "Application and Progress of Artificial Intelligence in Fetal Ultrasound." Frontiers in Medicine, vol. 10, 2023, https://doi.org/10.3389/fmed.2023.1121186.

Passos, Justine, et al. "Artificial Intelligence-Based Analysis of Fetal Growth Restriction in a Prospective Obstetric Study." BMC Pregnancy and Childbirth, vol. 24, 2024, https://doi.org/10.1186/s12884-024-06098-6.

Paul, Aditya, and Anil Kumar. "Brain Tumor Detection Using Deep Learning Techniques: A Comparative Study." Brain Sciences, vol. 9, no. 9, 2019, p. 231, https://www.mdpi.com/2076-3425/9/9/231.

Sriraam, Natarajan, et al. “A Comprehensive Review of AI-Based Algorithms for Fetal Facial Anomaly Detection (2013–2024).” Artificial Intelligence Review, 2025, https://doi.org/10.1007/s10462-025-10512-1.

Zenodo. “Fetal Brain MRI Scans Dataset.” Zenodo, https://zenodo.org/records/8055666.

“NINS 2022 Dataset Including 5285 T1-Weighted Brain MRI Images.” ResearchGate, https://www.researchgate.net/figure/NINS-2022-dataset-that-include-5285-T1-weighted-brain-MRI-images_tbl1_381370226. Accessed 22 June 2025

“Normal Fetal MRI (30 Weeks).” Radiopaedia, published 8 Apr. 2016, radiopaedia.org/cases/normal-fetal-mri-30-weeks. Accessed 25 June 2025.