Journal of Theoretical and Applied Information Technology
31st August 2023. Vol.101. No 16
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REVOLUTIONIZING FOETAL CARDIAC ANOMALY DIAGNOSIS: UNLEASHING THE POWER OF DEEP LEARNING ON FOETALECHO IMAGES

DIVYA M O[1], M S VIJAYA[2]
[1]Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, Tamilnadu, India
[2] Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, , Tamilnadu, India
[1]divyammo@gmail.com, [2]msvijaya@psgrkcw.ac.in

ABSTRACT

The use of Artificial Intelligence (AI) has amplified in various fields, with remarkable results in medicine in recent times. Despite the potential of AI in the medical field, there are still many unexplored areas due to data unavailability. One such area is cardiac foetal anomaly diagnosis, which is poorly diagnosed globally with a rate of only 50%. The complexity of the task requires a high level of expertise to understand minute hints and conduct thorough exams for accurate image captures. In this research, the FoetalEcho_V01 dataset was used for foetal cardiac anomaly diagnosis, consisting of pre-classified ultrasound images representing 15 different anomalies and a class representing normal heart images. The deep learning models which are efficient in producing potential classifiers for ultra sound scan images are identified. The models are CNN, AlexNet, VGG16 and ResNet50. The best performing deep learning models were used to produce classifiers, and their performance was evaluated. The results showed that the deep learning models performed well on the FoetalEcho_V01 dataset images for diagnosing structural cardiac anomalies in the foetus, with consistent performance as demonstrated by the calculated standard deviation. The results obtained from the research for the FetalEcho_V05 dataset are as follows. The CNN model achieved a precision of 0.94, recall of 0.89, accuracy of 0.90, and F1 score of 0.91. Comparatively, the AlexNet model demonstrated a precision of 0.92, recall of 0.87, accuracy of 0.89, and F1 score of 0.89. The VGG16 model exhibited precision of 0.91, recall of 0.85, accuracy of 0.87, and F1 score of 0.88. Lastly, the ResNet50 model displayed a precision of 0.93, recall of 0.90, accuracy of 0.93, and F1 score of 0.93. Among these models, the CNN model emerged as the best classifier for the FetalEcho_V05 dataset, with its superior performance in terms of precision, recall, accuracy, and F1 score.

Keywords: *Cardiac Heart Defect, Classification, Deep Learning, Diagnosing Foetal Cardiac Anomalies, Prenatal Diagnosis*

1. INTRODUCTION

AI is being widely adopted in the medical field as its applications have demonstrated remarkable results in various medical tasks. Ultrasound imaging technology (USIT) is a widely used diagnostic imaging tool in medicine because of low cost, portability, and ability to produce real-time images. USIT is undergoing rapid evolution, but still facing challenges such as high variability in results and poor image quality control. However, recent advancements in computational power and miniaturization of USIT devices have opened up new opportunities for using advanced image processing to address these challenges.

The overall process of foetal cardiac anomaly diagnosis typically involves the following phases. Routine prenatal screening, such as ultrasound, is performed to detect any potential abnormalities in the foetal heart. If anything, suspicious is detected, the next step is a more detailed evaluation. A specialized ultrasound, known as foetal echocardiography, is performed to confirm the diagnosis of a foetal cardiac anomaly. This test provides a detailed image of the foetal heart and allows the healthcare provider to identify any structural abnormalities. If there is any anomaly diagnosed, a foetal medicine specialist, pediatric cardiologist, or both may be consulted to manage the foetal cardiac anomaly based on the severity. The experts will review the results of all the tests and use their expertise to make the most appropriate decision on the medical intervention with respect to the findings. Once the diagnosis is confirmed, the healthcare team will provide counselling and information to the parents about the anomaly, including potential outcomes and available treatment options. They will develop a plan based on the diagnosis for prenatal monitoring, delivery, and postnatal care. Regular prenatal monitoring is important to ensure the health of the foetus and mother. This may include additional ultrasounds, non-invasive tests, or invasive procedures, such as foetal blood sampling, as deemed necessary by the healthcare team. Delivery and postnatal care will be planned based on the severity of the cardiac anomaly and the overall health of the foetus and mother. In some cases, delivery may need to be planned for a tertiary care centre with a specialized neonatal intensive care unit (NICU) and pediatric cardiology team. After birth, the infant will be closely monitored and evaluated by a pediatric cardiologist to determine the appropriate management and treatment plan. This may include medications, surgery, or other medical procedures. Close follow-up and monitoring will be necessary to ensure the best outcomes for the infant.

The mid-trimester foetal anomaly screening ultrasound is considered a crucial part of prenatal care. During the scan, the ultrasound includes views of the foetal heart to detect any malformations. Despite this, many CHDs still go undetected before birth [6],[7]. The rates of detection vary greatly worldwide, with some countries only detecting 14% of severe cases. Within countries, there is also considerable variation in detection rates [9],[10]. Infants who are not diagnosed before birth are less prospective to endure heart surgery are prospective to face an antagonistic enduring neurological consequence, and may not make it to surgery at all. Accurate prenatal diagnosis enables parents to make informed decisions about continuing the pregnancy and can provide opportunities for medical intervention in certain needy cases [16].

Manual diagnosis of foetal cardiac anomalies has several shortcomings. The accuracy of manual diagnosis depends on the expertise and experience of the person performing the scan. A less experienced sonographer may miss or misinterpret important findings. Along with this, the anomaly diagnosis are limited by the quality of the images obtained during the ultrasound. Factors such as the position of the foetus, foetal movement, and maternal obesity can impact the quality of the images and make it more difficult to diagnose abnormalities. The manual diagnosis of foetal cardiac anomalies requires a thorough and time-consuming evaluation of multiple images, which can be challenging and may result in delay in diagnosis. Manual diagnosis of foetal cardiac anomalies is limited to the information that can be obtained through the details visible in the ultrasound images. Other tests, such as MRI or Doppler flow studies, may be necessary to confirm the diagnosis and gather more information which is not advisable during pregnancy. The manual diagnosis of foetal cardiac anomalies is subject to human error and can lead to false negative or false positive results.

These limitations highlight the need for advanced technologies and techniques to improve the accuracy and reliability of foetal cardiac anomaly diagnosis. For example, machine learning algorithms are being developed and used to help automate the process of foetal cardiac anomaly diagnosis and improve its accuracy [5].

The ultrasound of the foetal heart is considered a reliable diagnostic tool when performed by experienced professionals. However, there is still potential for improvement, especially in increasing the diagnosis rate of CHDs prenatally [19]. The amalgamation of AI technology has the potential to provide more accurate prognoses and quantify various cardiac metrics, but there are challenges that must be overcome before its widespread use is accepted. It is crucial that healthcare professionals in the field of foetal cardiology become familiar with the benefits and limitations of AI.

The application of AI in foetal heart imaging has been approached through traditional machine learning image classification methods. Despite their widespread use, these methods have limitations such as low accuracy and weak adaptability [8]. The current approach split up feature abstraction and classification into discrete steps.

Deep-learning methods have shown to be more effective in achieving higher accuracy in detection, recognition, or segmentation tasks in computer vision when large annotated image datasets are available, as compared to traditional machine learning algorithms. However, in medical image processing and computer-aided diagnosis specifically for ultrasound images, the availability of annotated data is limited which could be handle in a systematic way to get the best out of deep learning (DL). With the arrival of DL techniques in numerous robotic vision applications, improved accuracy in detection, recognition, and segmentation has been observed when large amounts of annotated image data are present [4]. However, the field of USIT image processing and computer-assisted judgment has very limited availability of labelled data. This raises the question of how effective these DL techniques will be in such scenarios. There are questions about the relative effectiveness of deep-learning methods versus conventional machine-learning methods, and how conventional machine-learning methods compare with deep-learning methods when applied to the same dataset. DL models have a strong ability to learn and integrate feature extraction and classification processes, resulting in improved accuracy in image classification tasks. This integration allows for a more comprehensive approach to image classification, leading to better results. [20].

Our previous research using the same FetlEcho_V01 dataset was implemented with conventional machine learning models. The diagnostic perfection for foetal cardiac anomaly for those models were recorded. The paper aims to develop classifiers for foetal cardiac anomalies using the best-performing DL models with USIT images for diagnosis. This research area is under-explored, and limited technical implementation is available. Therefore, the study begins by evaluating machine learning models, as a parent group of DL, for CHD classification. The initial step is to identify the best-performing models for CHD classification, and then further improvement can be made based on the results. The research

on fetal cardiac anomaly diagnosis using artificial intelligence holds great significance in contributing to the well-being of future generations. By harnessing the power of AI, we aim to improve the accuracy and effectiveness of diagnosing fetal cardiac anomalies, ultimately working towards building a healthier generation and makes a contribution towards a social cause. This research endeavor has the potential to make a profound impact on healthcare practices and pave the way for advancements in prenatal care, ensuring a brighter and healthier future for unborn children and their families.

2. LITERATURES

This research has reference to different categories of research papers. The first category of research falls under those published by medical practitioners, which proves the need for an automated system for the foetal cardiac anomaly diagnosis which also mentions the disadvantages of the present system. The second category of research is those which have taken traditional machine learning algorithms for diagnosis from USIT. The third category of research referred in this research is those who have adopted DL models for diagnosis using USIT.

Brattain et. AL [1] predicted that, based on recent advancements, traditional machine learning algorithms for USIT will continue to advance and will be a significant trend in ultrasound-based diagnoses in the near future. A traditional machine learning algorithm-based intelligent diagnostic assistant system will incorporate USIT as one of its many inputs. Through multimodal and multiscale observations over time, the system will create clinical-viable quantitative models. This collective machine intelligence will be able to observe data, provide guidance, assess new information, and assist with decision making, potentially leading to substantial improvements in clinical workflow and patient outcomes.

Huang et. Al [2] have identified several supervised DL models that have been specifically developed for analyzing foetal ultrasound images and videos. For instance, Temporal HeartNet has the capability to automatically predict important parameters such as visibility, viewing plane, location, and orientation of the foetal heart in ultrasound videos, with a focus on plane-based detection.

On the other hand, Baumgartner et. Al [3] have reported the development of SonoNet, which can detect various foetal structures in ultrasound videos through bounding boxes, including the brain, spine, abdomen, and four standardized transverse scanning planes of the foetal heart, such as the four-chamber view (4CV), three-vessel view (3VV), right ventricular outflow tract (ROVT), and left ventricular outflow tract (LOVT).

Arnaout et. Al [14] used plane-based detection to screen for congenital heart disease (CHD) by detecting the foetal heart. They utilized U-net for segmentation of several anatomical structures such as the thorax, heart, spine, and four cardiac chambers. The team then utilized these segmentations to calculate typical foetal cardiothoracic measurements.

The impending are some relevant statistics which can establish the importance of improving the diagnosis rate of foetal cardiac anomalies. Heart defects that are present at birth, known as CHDs, are one of the most prevalent form of inborn malformation, accounting for one-third of the reported cases [12]. In Europe, around 36,000 live births are affected each year, with a prevalence rate of 7 out of 1,000 live births [12][13].

Congenital heart disease (CHD) is a leading cause of death in infants during their first year of life [14]. The prenatal screening for CHD varies significantly depending on the country's healthcare system. In the absence of organized screening, the screening rate is only 17.9%, but it increases to 55.6% when two or three ultrasounds are performed systematically (citation 15). In France, three ultrasounds are recommended during pregnancy, but on average, 5.5 ultrasounds are performed, with varying rates across different regions, ranging from 47.3% to 71% [16], [17], [18]. Although the benefits of prenatal CHD screening in terms of morbidity and mortality are not clear from some studies, a prenatal diagnosis among neonates with CHD has been linked to lower preoperative risk factors for heart surgery [14]. Additionally, prenatal diagnosis can improve a child's prognosis regarding neurocognitive development and morbidity [18]. Prenatal screening also allows for genetic investigations for some CHDs with a genetic component. Finally, training has been shown to increase the effectiveness of prenatal heart disease screening.

2.1 Model Selection

The popular DL architectures that have been used in ultrasound image classification include CNN, ResNet50, AlexNet, and VGG16. These architectures have been pre-trained on large image datasets and have demonstrated to be active in numerous USIT image classification tasks.

Convolutional Neural Networks (CNNs) have shown to be very successful in classifying images, including ultrasound images. CNNs have the capability to automatically learn hierarchical representations of images, which are crucial for capturing the complex relationships between pixels and the underlying structure in the image. ResNet50 is a residual network with 50 layers, which makes it capable of handling deep networks without encountering the vanishing gradient problem. ResNet50 has been used in several research studies for ultrasound image classification, including a study by Y. Liu et al. titled "Automatic Detection of Abnormal Pregnancy using Deep Convolutional Neural Network" (IEEE Transactions on Medical Imaging, 2018).

AlexNet is one of the earliest and most well-known CNN architectures, and it has been used for various image classification tasks, including ultrasound image classification. A study by D. D. Pham et al. titled "Automatic Ultrasound Image Segmentation Using Convolutional Neural Network" (IEEE Journal of Biomedical and Health Informatics, 2017) used AlexNet for automatic segmentation of ultrasound images. VGG16 is a deep CNN architecture with 16 layers, which has been used for various image classification tasks, including ultrasound image classification. A study by A. Cetin et al. titled "Automated Diagnosis of Fetal Anomalies using Convolutional Neural Network" (IEEE Transactions on Medical Imaging, 2019) used VGG16 for automatic diagnosis of fetal anomalies in ultrasound images.

These studies speak aloud about the need of automated diagnosis of foetal cardiac anomalies and suggest that DL has the potential to play an important role in the detection and decide upon the medical interventions into the foetal cardiac anomalies. These studies demonstrate the potential of DL techniques to improve the accuracy and consistency of foetal cardiac anomaly detection, which could have significant benefits for patient care. This research will explore the proven DL models for the foetal cardiac anomaly ultrasound dataset and identify the best performing model and best hyperparameter vectors, amongst the best.

These studies demonstrate the potential of deep learning techniques to improve the accuracy and consistency of USIT analysis, which can be referred for foetal cardiac anomaly detection with USIT. Ultimately this could have significant benefits for patient care and to improve the pregnancy outcome.

2.2 Data Collection, Dataset Creation and Pre-processing

The FetalEcho_V05 dataset has been created as a part of a research project to identify different structural heart defects in foetuses, which is a handcrafted dataset. The images used in the dataset have been collected from various sources such as foetal specialty clinics, research repositories of clinical experts, and foetal medicine forums. The research repository used was Radiopedia, where clinical experts publish their research, and other sources such as different chapters of Fetal Medicine foundations (India, UK), Society of Fetal Medicine, and the Karnataka chapter. The process flow is depicted in Fig.1. The images collected have been manually classified and validated by a clinical expert to ensure accuracy.

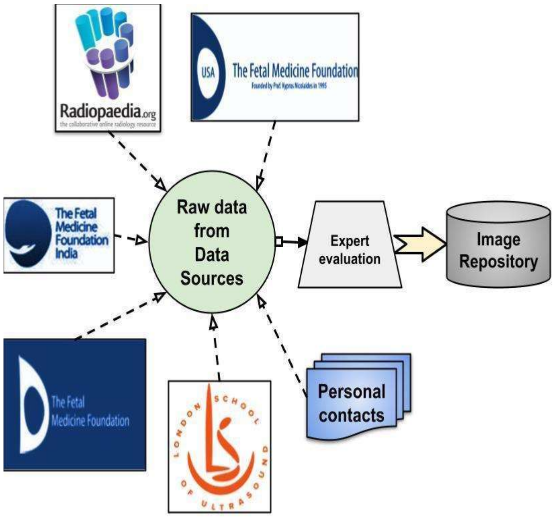

Radiopaedia.org
The Fetal Medicine Foundation
The Fetal Medicine Foundation India
Raw data from Data Sources
Expert evaluation
Image Repository
The Fetal Medicine Foundation
London School of Ultrasound
Personal contacts

Fig.1. Data Collection Process

The FetalEcho_V05 dataset consists of 1600 images, classified into 16 categories of structural heart defects. These defects include Ventricular Septal Defect, Atrial Septal Defect, AV Septal Defect, Tetralogy of Fallot, Truncus Arteriosus, Transposition of Great Arteries, Single Ventricle, Double Outlet RV, Hypoplastic Right Heart Syndrome, Hypoplastic Left Heart Syndrome, Aortic Coarctation or Hypoplasia, Aortic Atresia or Stenosis, Pulmonary Stenosis, Ebstein Anomaly, Ectopia Cordis, and Pentalogy of Cantrell. Fig.2. shows some sample images from FetalEcho_V05. All 16 categories have approximately equal representation in the FetalEcho_V05 dataset, making it a comprehensive resource for the study of structural heart defects in foetuses. The standardization of the images and the validation by a clinical expert ensure that the data in the dataset is accurate and can be used as a reliable source for future research and analysis.

The accuracy of a model is of utmost importance, particularly in the medical field where lives are at stake. Achieving the best accuracy requires a dataset that provides sufficient detailing of USIT outcomes for training purposes. To achieve this, appropriate pre-processing techniques must be applied to the UIST images. Selecting the most suitable pre-processing technique for the dataset used for training is a major challenge. The accuracy of the model is directly linked to the choice made on the pre-processing techniques. The importance of pre-processing of images cannot be overstated since any error can be fatal. The performance of the model must be monitored continually, and pre-processing techniques may need to be updated as new data sets become available to ensure the model provides accurate diagnoses [21]. Since Preprocessing of image dataset is a crucial step in preparing data for machine learning models, entire research was done to identify the best suitable preprocessing technique/s for FetalEcho_V05 dataset. The research took three primary pre-processing techniques and found the best primary pre-processing technique along with fusions and robust models for this purpose. The purpose of preprocessing is to transform raw image data into a format that can be easily understood by machine learning algorithms. Rescaling involves resizing the images to a fixed size to make them uniform in size. This can be done by resizing the image to a specific pixel value or scaling it by a certain factor. Normalization is used to scale the pixel values of the images to a standard range. This helps in reducing the effect of brightness and contrast variations in the images. Filtering technique involves applying filters to the image to enhance or remove specific features. This can be useful in removing noise, blur or enhancing the edges of objects in the image. The FoetalEcho_V01 dataset has undergone pre-processing as well. The objective of the experiments conducted in the previous research was to determine the most effective pre-processing techniques for a newly created dataset FetalEcho, using conventional AlexNet. The initial dataset without any pre-processing, produced a 49% accuracy rate. However, pre-processing techniques greatly improved the accuracy with the same set of hyperparameters. The accuracy increased to 87% after noise removal, 88% after blur removal, and 69% after sharpening. When both blur and noise removal were used, the accuracy increased to 89%. Combining blur removal and sharpening resulted in an accuracy of 79%, while noise removal and sharpening produced an accuracy rate of 53%. However, when all three pre-processing techniques were applied, the accuracy was only 57%. These results indicate that sharpening is not suitable for the FetalEcho dataset, and its behavior changes when combined with noise and blur removal.Overall, the combination of noise and blur removal was found to be the most effective pre-processing technique for the FetalEcho dataset, in conjunction with the given hyperparameters. As per the results obtained in the previous research, the blur removal and noise removal were showing best performance for classification. Hence the same pre-processing method had been adapted in this study. The images have been standardized by fine-tuning the resolution of the images. The resolution has been set to 256 * 256 * 3 for all experiments, to ensure that all images in the dataset are of the same size and distribution. Fig.3 shows some sample output after preprocessing the USIT images.

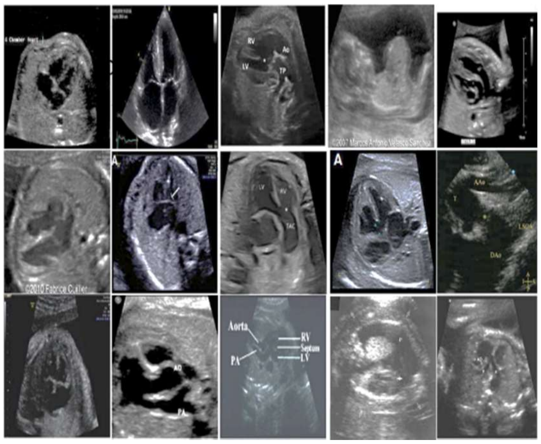


Fig.2. Sample Images from the dataset

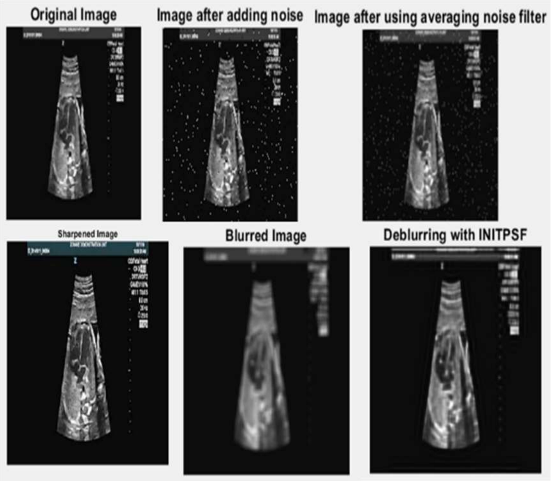


Fig.3. Sample Images After Pre-Processing

3. METHODOLOGY FOR FETAL CARDIAC ANOMALY DIAGNOSIS USING DL MODELS

These CNN architectures have proven to be effective in ultrasound image classification due to their ability to learn and capture complex relationships between pixels in the images. The pre-trained weights and the hierarchical representations learned by these networks make them suitable for tasks such as ultrasound image classification, where the images are often high-dimensional and complex.

Hence, In this work, the above-mentioned models are used for diagnosis and the classification performance of DL models for the FoetalEcho_V01 dataset is figured out. As best performing algorithms with ultra sound scan images, CNN, VGG16, AlexNet, ResNet50 models are considered for experiments. The selection of the above-mentioned models are done based on the results shown by the algorithms in related research done for doing similar diagnosis. The overview of research is featured in Fig.4.

3.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are a specific type of neural network architecture that is designed to analyze and process data that has a grid-like structure, such as images. These networks comprise several layers that perform specific functions, including convolutional layers, activation layers, pooling layers, and fully connected layers.

In a convolutional layer, a set of filters (also called kernels or weights) are applied to the input image, producing a set of feature maps. The convolution operation is defined as:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau).g(t-\tau)\ d\tau \qquad (1)$$

where f is the input image and g is the filter. The output of the convolution operation is called a feature map.

Equation 1, represents the convolution operation between two functions, f and g. The convolution operation takes two functions and produces a third function, which represents the amount of overlap between the two input functions as thxey are shifted relative to each other.

The symbol * represents the convolution operation. The input function f is multiplied element-wise by the filter function g, then integrated over all values

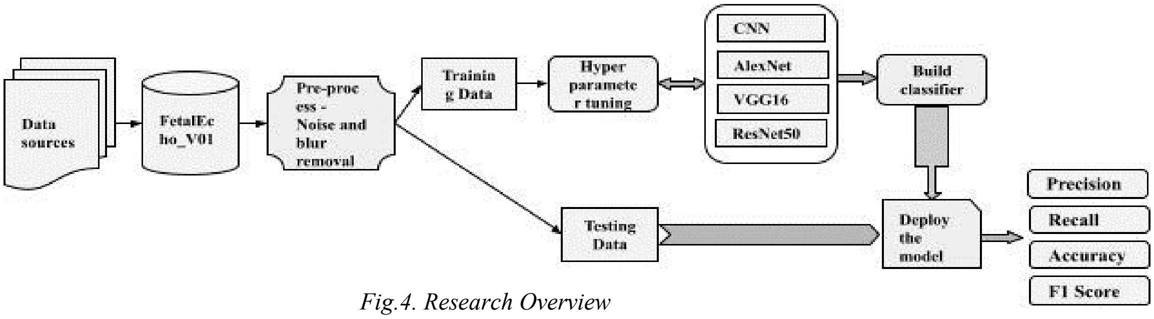


Fig.4. Research Overview

of tau. The output of the convolution operation, (f * g)(t), is a single scalar value that represents the overlap between the two functions at a particular value of t.

In the context of convolutional neural networks, f is typically the input image and g is the filter (also called kernel or weight). The convolution operation extracts features from the input image by sliding the filter over the image and computing the element-wise product between the filter and the overlapping part of the image. This process is repeated for multiple filters to produce multiple

feature maps, which capture different aspects of the input image.

The activation layer applies an activation function, such as ReLU (rectified linear unit), to the feature map to introduce non-linearity into the network. The activation function is defined as:

$f(x) = \max(0,x)$ (2)

Pooling layers downsample the feature map, reducing its spatial dimensions, while preserving the most important information. The most commonly used pooling operation is max pooling, defined as:

$Y = \max(x_1, x_2, \ldots..x_n)$ (3)

where x is the input feature map and y is the output of the pooling operation.

Finally, fully connected layers combine the features from all locations in the previous layer to make a prediction. The output of a fully connected layer can be calculated as:

$y=Wx+b$ (4)

where W is the weight matrix, x is the input to the layer, and b is the bias term. The predicted label is then determined by applying a softmax activation function to the output of the fully connected layer. Fig.5 pictorially represented the CNN architecture.

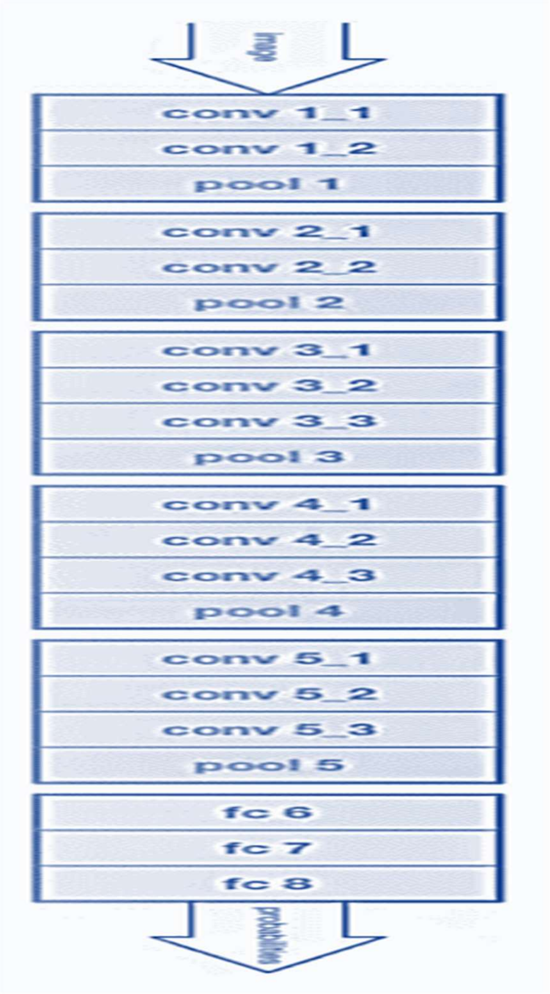


Fig.5. CNN architecture

3.2 VGG16

VGG16 is a Convolutional Neural Network (CNN) architecture that was introduced in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" in 2014. It consists of multiple convolutional layers, activation layers, pooling layers, and fully connected layers.

The convolutional layers apply filters to the input image to extract features. The activation layer applies a non-linear activation function, such as ReLU (rectified linear unit), to the output of the convolutional layer to introduce non-linearity into the network. The activation function is defined as:

$f(x) = \max(0,x)$ (5)

The pooling layer downsamples the feature map to reduce its spatial dimensions, while preserving the most important information. The most commonly used pooling operation is max pooling, defined as:

$Y = \max(x_1, x_2, \ldots..x_n)$ (6)

where x is the input feature map and y is the output of the pooling operation.

The fully connected layer combines the features from all locations in the previous layer to make a prediction. The output of a fully connected layer can be calculated as:

y=Wx+b (7)

where W is the weight matrix, x is the input to the layer, and b is the bias term. The predicted label is then determined by applying a softmax activation function to the output of the fully connected layer.

In the case of VGG16, the architecture consists of 13 convolutional layers and 3 fully connected layers. It uses only 3x3 convolutional filters, which allows for a greater depth of the network compared to networks that use larger filters. This deeper network allows for the extraction of more complex features from the input image. The VGG16 architecture has been widely used for image classification tasks and has achieved state-of-the-art results on various benchmark datasets. Fig.6 pictorially represents the VGG16 architecture.

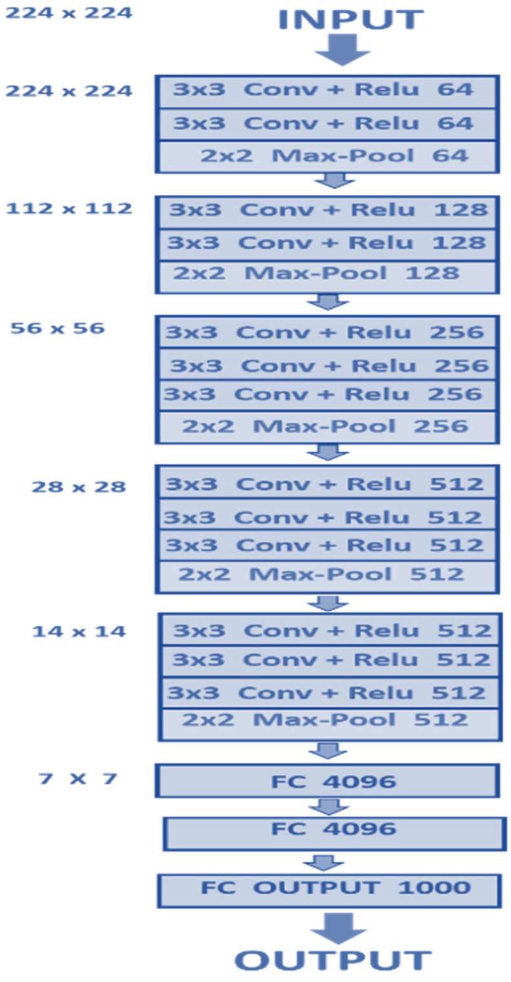


Fig.6. VGG16 architecture

3.3 AlexNet

AlexNet is a Convolutional Neural Network (CNN) architecture that was introduced in the paper "ImageNet Classification with Deep Convolutional Neural Networks" in 2012. It is one of the first DL models to achieve high accuracy on the ImageNet dataset, which is a large-scale image classification task.

The architecture of AlexNet consists of several convolutional layers, activation layers, pooling layers, and fully connected layers.

The convolutional layer applies filters to the input image to extract features. The filter is defined as a set of weights, which are learned during the training process. The output of the convolution operation can be calculated as:

y = f(W*x+b) (8)

where W is the weight matrix, x is the input to the layer, b is the bias term, and f is the activation function. In AlexNet, the activation function used is ReLU (rectified linear unit), which is defined as:

f(x)=max(0,x) (9)

The pooling layer reduces the spatial dimensions of the feature map while preserving important information. The most commonly used pooling operation is max pooling, defined as:

$Y = \max(x_1, x_2, \ldots..x_n)$ (10)

where x is the input feature map and y is the output of the pooling operation.

The fully connected layer combines the features from all locations in the previous layer to make a prediction. The output of a fully connected layer can be calculated as:

y=Wx+b (11)

where W is the weight matrix, x is the input to the layer, and b is the bias term. The predicted label is then determined by applying a softmax activation function to the output of the fully connected layer.

In AlexNet, the architecture consists of 5 convolutional layers, 3 pooling layers, and 3 fully connected layers. The network uses a combination of large and small filters and also includes normalization and dropout layers to improve its robustness and prevent overfitting. The AlexNet architecture was a major breakthrough in the field of DL and set the foundation for the development of more complex models. Fig.7 pictorially represents the AlexNet architecture.

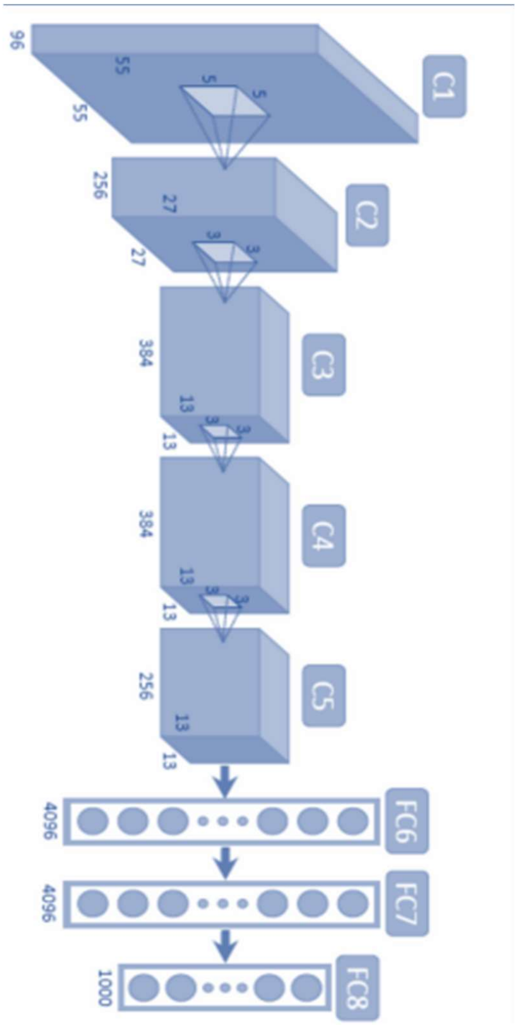


Fig.7. Alexnet Architecture

3.4 ResNet50

ResNet50 is a Convolutional Neural Network (CNN) architecture that was introduced in the paper "Deep Residual Learning for Image Recognition" in 2015. It is a variant of the ResNet architecture and is notable for its extremely deep network (up to 152 layers) while still being able to achieve high accuracy on image classification tasks.

The key concept behind the ResNet architecture is the residual connection. A residual connection allows the network to learn the residual between the input and desired output, rather than the desired output itself. This allows the network to learn complex functions even when it has a large number of layers. The residual connection can be defined as:

$$y = x + F(x,W) \qquad (12)$$

where x is the input to the layer, F(x,W) is the mapping function of the layer, W is the weight matrix, and y is the output of the residual connection.

The residual connection is implemented by adding the input to the output of a sequence of convolutional, activation, and batch normalization layers. The activation function used in ResNet50 is ReLU (rectified linear unit), which is defined as:

$$f(x) = \max(0,x) \qquad (13)$$

The batch normalization layer normalizes the activations of the previous layer, which helps to stabilize the training process and prevent overfitting. The output of the batch normalization layer can be calculated as:

$$x^\wedge = (X - \mu)/\sigma \qquad (14)$$

where x is the activations of the previous layer, μ is the mean of the activations, and σ is the standard deviation of the activations.

In ResNet50, the architecture consists of multiple residual blocks, each containing several convolutional, activation, and batch normalization layers. The network also includes pooling layers to reduce the spatial dimensions of the feature maps, and fully connected layers to make the final prediction. The ResNet50 architecture has been widely used for image classification tasks and has achieved state-of-the-art results on various benchmark datasets. Fig.8 pictorially represented the ResNet50 architecture.

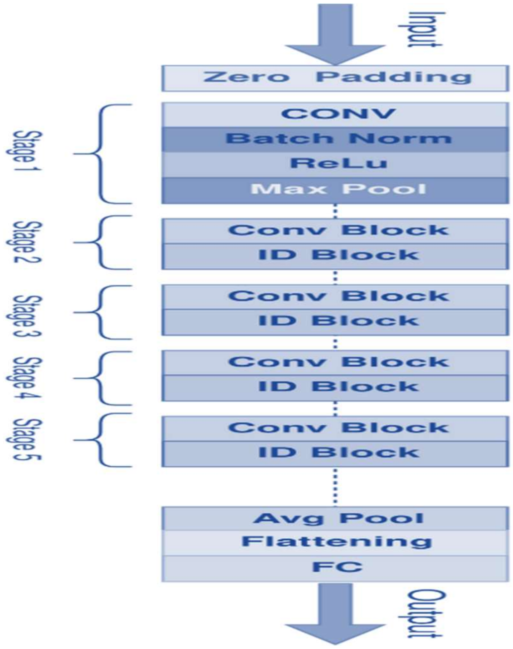


Fig.8. Resnet50 Architecture

3.5 Model Building

The set of pre-processed images of FetalEcho_V05 is divided into training and testing data. The training dataset consists of 80% of the instances and is used to train the model. The supervised pattern classification algorithms such as CNN, AlexNet, VGG16 and ResNet50 are used to learn the patterns from the train images with the help of hyper-parameters. Hyperparameters are parameters that are not learned from the data during model training, but instead must be set before training begins. The choice of hyperparameters can have a significant impact on the performance of a model. Train-test split ratio determines the proportion of data that will be used for training and testing the model, this can vary depending on the size and complexity of the dataset. Learning rate controls how quickly the model adjusts its parameters in response to errors during training. Optimization algorithm determines the method used to update the model's parameters during training. Activation function determines the non-linear function that is applied to the outputs of each layer in the model. Loss function determines the objective function that the model tries to minimize during training. Popular loss functions include Mean Squared Error (MSE), Cross-Entropy, and Binary Cross-Entropy. Number of hidden layers determines the number of layers in the neural network that are not the input or output layers. Dropout rate determines the probability of dropping out a neuron during training, which helps prevent overfitting. Number of iterations per epoch determines how many iterations are run per epoch, which can affect the speed of training. Kernel filter size determines the size of the convolutional filters that are applied to the input data in a convolutional neural network. Pooling size determines the size of the pooling window that is used to down sample the feature maps in a convolutional neural network. Batch size determines the number of training samples that are used in each batch during training. Epoch determines the number of times the entire training dataset is passed through the model during training.

Finally the image classifiers based on CNN, AlexNet, VGG16 and ResNet50 DL models have been developed with appropriate hyperparameter settings by training the preprocessed image set. The recognition capability of the classifiers to diagnose an unlabelled ultra sound scan image as a normal heart image or an image with anomaly and figure out the anomaly is the outcome of the model building. The performance of the independent models have been evaluated for foetal cardiac anomaly detection with the help test set.

3.6 Performance evaluation metrics

The performance measures used in this research for evaluating the performance of each model are, Precision, recall[11], accuracy and F1 score.

Precision is a metric used to evaluate the accuracy of a binary classification model, which is a model that predicts one of two possible classes (positive or negative) for each input. A high precision score indicates that the model is making very few false positive predictions, which means that the model is accurate when it predicts a positive case.

Recall is a metric used to evaluate the effectiveness of a binary classification model, which is a model that predicts one of two possible classes (positive or negative) for each input. Recall measures how well the model can identify positive cases out of all the actual positive cases in the data. A high recall score indicates that the model is able to correctly identify most of the positive cases in the data, which means that the model is effective at identifying positive cases.

Accuracy is a metric used to evaluate the overall performance of a classification model, which is a model that predicts one of several possible classes for each input. Accuracy measures how often the model correctly predicts the correct class out of all the predictions it makes. A high accuracy score indicates that the model is making a high percentage of correct predictions, which means that the model is performing well.

F1 score is a measure of a classification model's accuracy, which takes into account both precision and recall. It is the harmonic mean of precision and recall, with a value between 0 and 1, where 1 is the best possible F1 score. F1 score is a measure of how well a model can predict the positive class while minimizing false positives and false negatives.

4. Experiment and Results

The previous research steered using the same FetalEcho_V05 dataset. The work delivered the results as in Table.1, using the selected traditional machine learning (TML) models. The selected TML models were used for creating classifiers for foetal cardiac anomaly classification. The performance of the classifiers with respect to precision, recall, F1 Score, accuracy and their standard deviation(SD) and geometric mean(GM) observations of the previous research experiments are outlined in Table.1.

Table.1 Performance Metrics For Machine Learning Models

Models	**Precision**	**Recall**	**Accuracy**	**F1 Score**
SVM	0.76	0.71	0.75	0.74
KNN	0.68	0.63	0.64	0.64
NB	0.69	0.61	0.63	0.64
RF	0.75	0.68	0.65	0.67
GM	0.72	0.66	0.67-	0.67
SD	0.04	0.05	0.06	0.05

In this research, the DL toolbox in MATLAB was utilized to construct and link neural network layers for classifying 1600 USIT images into 16 categories. The classification performance for the selected DL models were assessed using Accuracy, Precision, Recall, and F1-score matrices over FetalEcho_V05 dataset. Once the dataset is ready for further processing, the selected pretrained models are used for building classifiers for classifying foetal echo USIT images and the classifier building is done using FetalEcho_V05 image dataset. After setting up the dataset, the most important phase of classifier development is to identify the perfect combination of hyper-parameters. For DL models, the hyper-parameters play a very important role in deciding the outcome for the experiments. The hyper-parameter fixing is done after several trials and finally the best setting for developing the best performing classifier will be recognized. Each DL models has undergone iterative fine tuning with different combinations of hyper parameters. The train-test split ratio for CNN is 80%/20% and for AlexNet, Vgg16 and ResNet50 is 70%/30%. The Learning rate for CNN is 0.001 and for AlexNet, Vgg16 and ResNet50 is 0.0001. The optimization algorithm selected for CNN is Stochastic Gradient Descent (SGD), for AlexNet it is Local Response Normalization, for Vgg16 it is Momentum SGD and for ResNet50 is Adaptive Moment Estimation. The -activation function selected for training for CNN, AlexNet, Vgg16 and ResNet50 is ReLU. The loss function selected for training for CNN is Softmax Cross-Entropy and for AlexNet, Vgg16 and ResNet50 is Categorical cross entropy. The number of hidden layers for CNN, AlexNet, Vgg16 and ResNet50 are 10,40,50 and 10 respectively. The drop-out rate used in training for CNN, AlexNet, Vgg16 and ResNet50 are 0.5, 0.4,0.4 and 0.3 respectively. The number of iterations per epoch selected for CNN, AlexNet, Vgg16 and ResNet50 is 20. The Kernel filter size selected for CNN, AlexNet, Vgg16 and ResNet50 are 3*3, 11*11, 3*3, 3*3 respectively. The pooling size selected for CNN, AlexNet, Vgg16 and ResNet50 is Max-pooling. The Batch size selected for CNN, AlexNet, Vgg16 and ResNet50 is 64. The number of epoch suitable for CNN, AlexNet, Vgg16 and ResNet50 are 100. After various iterations, the best hyperparameter tuning phase concluded with the best set of hyper-parameters for each model. The tuned hyper-parameters for training the DL all the models are delineated in Table.2.

The results for the DL models are as follows, CNN has got the precision of 0.94, recall of 0.89, Accuracy of 0.90, F1 score of 0.91. The AlexNet model has results as follows, precision of 0.92, recall of 0.87, Accuracy of 0.89, F1 score of 0.89. The VGG16 model recorded with the following values: precision of 0.91, recall of 0.85, Accuracy of 0.87, F1 score of 0.88. The ResNet50 model got the following results: precision of 0.93, recall of 0.90, Accuracy of 0.93, F1 score of 0.93. The best classifier for FetalEcho_V05 dataset is CNN. The results are summarised in Table.3. For finding the consistency of performance for the models, two descriptive statistic measures are identified, they are SD and GM. For precision, Recall, Accuracy and F1 score, the GM and standard deviation are also calculated. The mean value will help compare the performance metrics with the middle value of all models in the class.

Table.2. Hyper Parameter For Each Model

Models	**Train-test split ratio**	**Learning rate**	**Optimization algorithm**	**Activation function**	**Loss function**	**Number of hidden layers**	**The drop-out rate**	**Number of iterations per epoch**	**Kernel or filter size in convolutional layers**	**Pooling size**	**Batch size**	**Epoch**
CNN	0.8-0.2	0.001	Stochastic Gradient Descent (SGD)	ReLU	Softmax Cross-Entropy	10	0.5	20	3*3	Max-pooling	64	100

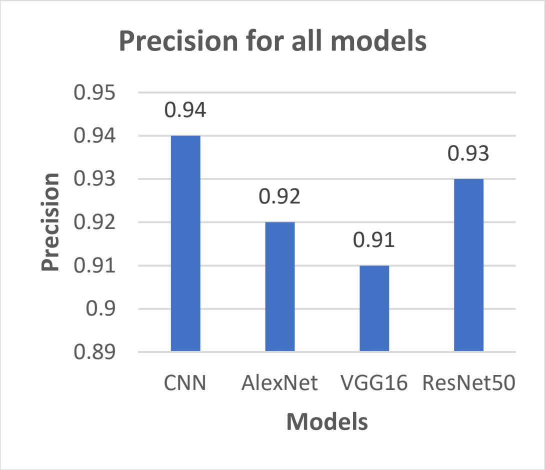
AlexNet	0.7-0.3	0.0001	Local Response Normalization	ReLU	Categorical cross entropy	40	0.4	20	11*11	Max-pooling	64	100
VGG16	0.7-0.3	0.0001	Momentum SGD	ReLU	Categorical cross entropy	50	0.4	20	3*3	Max-pooling	64	100
ResNet50	0.7-0.3	0.0001	Adaptive Moment Estimation	ReLU	Categorical cross entropy	10	0.3	20	3*3	Max-pooling	64	100

The SD will help to find the consistency of performance metrics for both classes. These conclusions will give more authentic inferences. The experiment outcomes shows that the DL algorithms are performing extremely well when compared to the manual process.

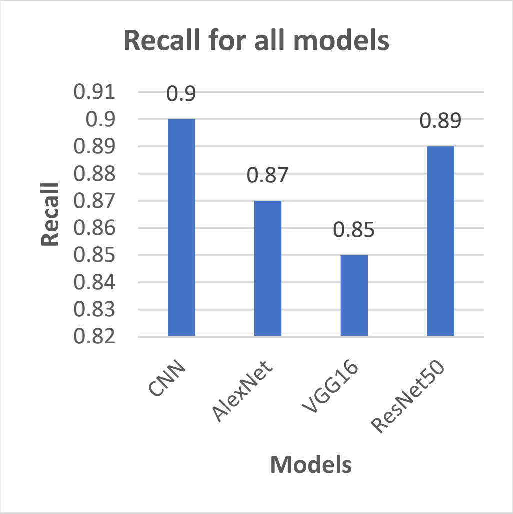
Table.3 Performance Metrics Of DL Models

Models	Precision	Recall	Accuracy	F1 Score
CNN	0.94	0.9	0.93	0.93
AlexNet	0.92	0.87	0.89	0.89
VGG16	0.91	0.85	0.87	0.88
ResNet50	0.93	0.89	0.9	0.91
GM	0.92	0.87	0.90	0.89
SD	0.02	0.02	0.03	0.02

From Table.3, the SD for precision of the models is 0.02. From this, it is clear that the DL models are very consistent. The SD for Recall of the models is 0.02. From this, it is clear that the models are performing consistently with respect to recall. The SD for Accuracy for the models is 0.03. From this, it is clear that the DL models are consistent performance with respect to accuracy. The SD for F1-Score is 0.02. From this, it is clear that the models are consistent with respect to F1-score. Fig.10 shows the SD slope. With reference to the previous research where the traditional models [22] [23] are used for diagnosing, the DL models are performing far better and are able to demonstrate consistency in performance also. Another inference from the experiments is that the best performing model among the four DL models, is CNN. This is very clearly evident in Fig.9(a),(b),(c),(d).

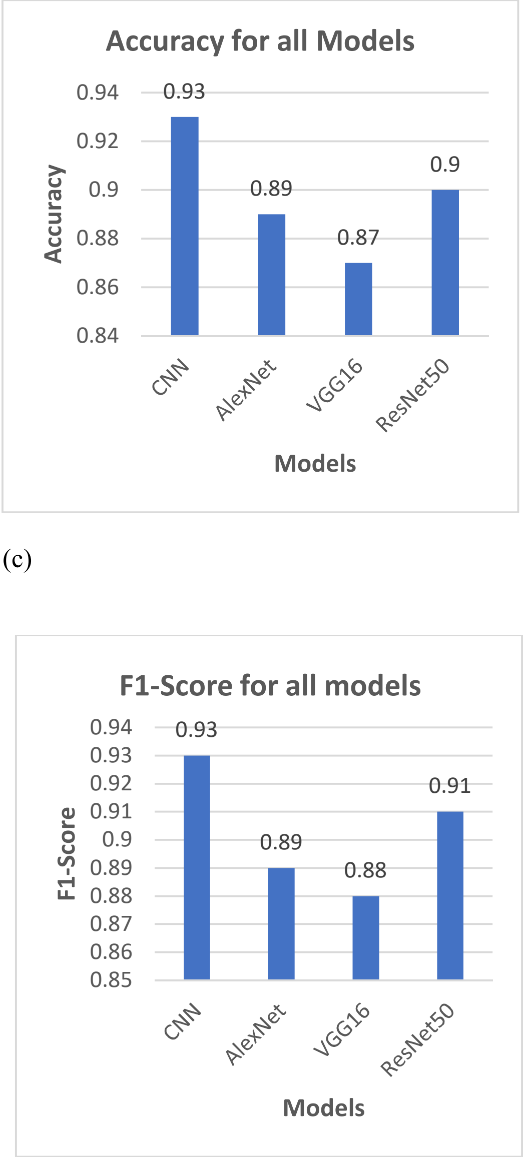


(a)



(b)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(c)

(d)

Fig. 9 The Performance Metrics Indicating The Best Performing Model (A) Precision (B) Recall (C) Accuracy (D) F1 Score

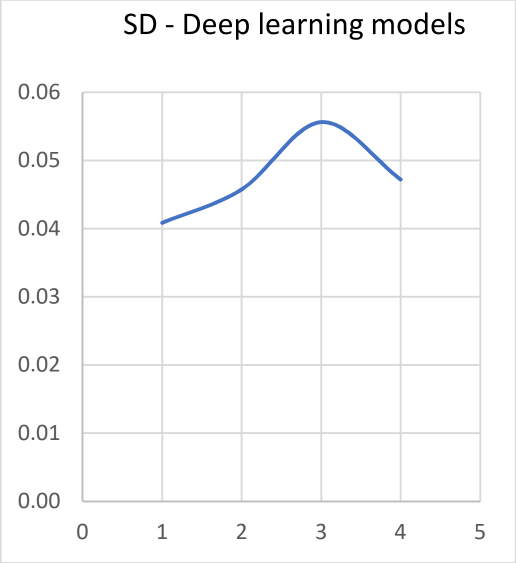


Fig.10 SD For The DL Models

4.1 Discussions

In the previous research the diagnosis of foetal cardiac anomalies using the FetalEcho_V05 dataset was experimented using traditional machine learning models (TML) [24]. It is observed that the performance of the DL models is proved to be the best approach for diagnosing foetal cardiac anomalies using USIT images. The SD measure and the GM for the DL models shows that the DL models demonstrate consistent performance compared to TML models. The following Table.4 shows the comparison of both TML models and DL models.

Table.4 Performance Metrics Of Both TML And DL Models

TML Models (TML)					**DL Models (DL)**				
Models	**Precision**	**Recall**	**Accuracy**	**F1 Score**	**Models**	**Precision**	**Recall**	**Accuracy**	**F1 Score**
SVM	0.76	0.71	0.75	0.74	CNN	0.94	0.9	0.93	0.93
KNN	0.68	0.63	0.64	0.64	AlexNet	0.92	0.87	0.89	0.89
NB	0.69	0.61	0.63	0.64	VGG16	0.91	0.85	0.87	0.88
DT	0.75	0.68	0.65	0.67	ResNet50	0.93	0.89	0.9	0.91

G- Mean	0.72	0.66	0.67	0.67	**G - Mean**	0.92	0.87	0.90	0.90
SD	0.04	0.05	0.06	0.05	**SD**	0.02	0.03	0.03	0.03

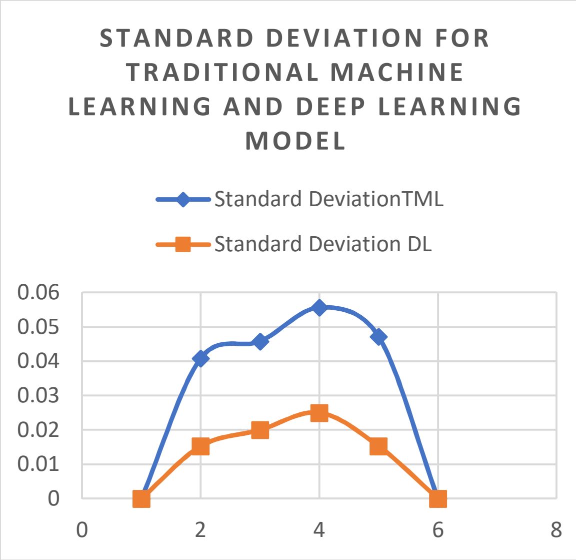


Fig.11 SD For DL Models And TML Models

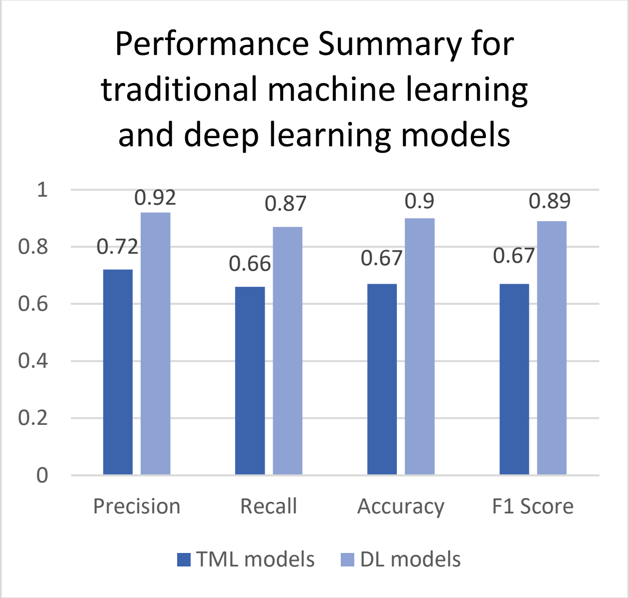


Fig.12 Summary Of Performance Metrics For TML Models And DL Models

Fig.11 shows a sample SD slope for both the TML models and DL models. The SD is a measure of the amount of variation or dispersion in a set of data. It measures the spread of the data from the mean or average value. If the SD is small, it indicates that the data points are close to the mean and tightly clustered together. Fig.9 demonstrates the SD chart for DL models and TML models. The chart clearly shows that the DL models are more consistent than the TML models. The mean precision, recall, accuracy and F1 score are 0.92,0.87,0.90,0.89 for the DL models respectively and 0.72, 0.66, 0.67, 0.67 for the TML algorithms respectively and pillars the same inference.

Fig.12, graphically represents the average of performance metrics demonstrated by both TML and DL classes for better comparison. The graph shows that the average precision of DL models are very high than the TML models. Average precision for TML is 0.72 and average precision for DL is 0.92. This shows that the average precision is very high for DL which shows the percentage of false prediction is very less for DL. The average recall for TML is 0.66 and for DL it is 0.87. This shows that the DL classifiers are having 21% better predictions in terms of recall. This proves the DL models are predicting a very good number of positive predictions when compared to TML models. The average accuracy for 0.90 for the DL class had 0.67 for the TMl class, which shows that 23% accuracy is more for the DL class. The average F1 score for the DL models is 0.89 TML models is 0.67, which clearly indicates that the accuracy for DL class is 22% more than TML. The collated results demonstrate that the DL models are very much more competent than the TML models**.**

4.2 Findings

This research scope was identified from various indications collected from clinical publications, which strongly suggested the automation for foetal cardiac anomaly diagnosis. There is no standard dataset available in public research repositories. Hence the data collection had been done as the first major step. Data was collected from different research repositories, published research by clinical experts and speciality clinics. All the images were standardized in terms of their size and color scheme. Furthermore, they were converted to a single file format and underwent the blur removal and noise removal process. Thus an appropriate preprocessed image dataset has been developed named FetalEcho_V05. The experiments conducted demonstrate the following findings.

The DL models are identified as the best option for foetal cardiac anomaly diagnosis. A comparative study had also been done to weigh up DL models and TML models performance accuracy for diagnosing. The DL models are performing very consistently in diagnosis of Fetal cardiac anomalies. When compared the results with the previous research of TML model implementation for foetal cardiac anomaly diagnosis, DL models are recognized as performing more accurately than TML models and consistently performing as well.

The experiment result shows that the mean precision, recall, accuracy and F1 score for DL class representatives, CNN, AlexNet, ResNet50 and VGG16 are very healthy. The SD shows that the classifiers produced are performing very consistently.

In this area there is no technically implemented solution available. Hence, a sequence of continuous research was conducted to collect data, to identify the scope of the work, pre-processing framework development and implementations using TML and DL models. The outcome of the research is very promising, progressive and worth taking to the next level of refinement as the domain is very much relevant to society which is already declared by the clinical experts as per the literature survey.

5. CONCLUSION

This research explores an untapped problem domain where no technical solution is available till date. But there is huge scope for improvement in foetal cardiac anomaly diagnosis using Artificial Intelligent models. This work has identified four representatives from DL models, CNN, AlexNet, VGG16, and ResNet50. These models are trained for creating classifiers for FoetalEcho_V01 dataset which can classify foetal echo USIT images. The performance of the image classifiers have been evaluated using the metrics, precision, recall, accuracy and F1-score. The observations are recorded and are analysed. The analysis clearly shows that DL models are the most appropriate techniques which produce the best experimental outcome. The future enhancements to these models could be, transfer learning techniques to overcome the challenge of dataset size. Also, some handcrafted features can be used individually or can be fused with the existing features to capture the minute patterns from USIT to support the existing features associated with models.

REFERENCES

[1]. Brattain, L. J., Telfer, B. A., Dhyani, M., Grajo, J. R., & Samir, A. E. (2018). Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdominal radiology*, *43*(4), 786-799.

[2]. Huang, W.; Bridge, C.P.; Noble, J.A.; Zisserman, A. Temporal HeartNet: Towards human-level automatic analysis of foetal cardiac screening video. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Philadelphia, PA, USA, 2017; pp. 341–349

[3]. Baumgartner, C.F.; Kamnitsas, K.; Matthew, J.; Fletcher, T.P.; Smith, S.; Koch, L.M.; Kainz, B.; Rueckert, D. SonoNet: Real-Time Detection and Localisation of Foetal Standard Scan Planes in Freehand Ultrasound. IEEE Trans. Med. Imaging 2017, 36, 2204–2215.

[4]. Arnaout, R., Curran, L., Zhao, Y., Levine, J. C., Chinn, E., & Moon-Grady, A. J. (2020). Expert-level prenatal detection of complex congenital heart disease from screening ultrasound using deep learning. medRxiv, 2020-06.

[5]. Dozen, A.; Komatsu, M.; Sakai, A.; Komatsu, R.; Shozu, K.; Machino, H.; Yasutomi, S.; Arakaki, T.; Asada, K.; Kaneko, S.; et al. Image Segmentation of the Ventricular Septum in Foetal Cardiac Ultrasound Videos Based on Deep Learning Using Time-Series Information. Biomolecules 2020, 10, 1526.

[6]. Meng, Q.; Sinclair, M.; Zimmer, V.; Hou, B.; Rajchl, M.; Toussaint, N.; Oktay, O.; Schlemper, J.; Gomez, A.; Housden, J.; et al. Weakly Supervised Estimation of Shadow Confidence Maps in Foetal Ultrasound Imaging. IEEE Trans. Med. Imaging 2019, 38, 2755–2767.

[7]. Brehar, R., Mitrea, D. A., Vancea, F., Marita, T., Nedevschi, S., Lupsor-Platon, M., ... & Badea, R. I. (2020). Comparison of deep-learning and conventional machine-learning methods for the automatic recognition of the hepatocellular carcinoma areas from ultrasound images. *Sensors*, *20*(11), 3085.

[8]. Stanik, C., Haering, M., & Maalej, W. (2019, September). Classifying multilingual user feedback using traditional machine learning and deep learning. In *2019 IEEE 27th international requirements engineering conference workshops (REW)* (pp. 220-226). IEEE.

[9]. Chauhan, N. K., & Singh, K. (2018, September). A review on conventional machine learning vs deep learning. In *2018 International conference on computing, power and communication technologies (GUCON)* (pp. 347-352). IEEE.

[10]. Zhou, X., Wang, S., Xu, W., Ji, G., Phillips, P., Sun, P., & Zhang, Y. (2015, April). Detection of pathological brain in MRI scanning based on wavelet-entropy and naive Bayes classifier. In *International conference on bioinformatics and biomedical engineering* (pp. 201-209). Springer, Cham.

[11]. Snider, E. J., Hernandez-Torres, S. I., & Boice, E. N. (2022). An image classification deep-learning algorithm for shrapnel detection from ultrasound images. *Scientific reports*, *12*(1), 1-12.

[12]. Ogawa, T., Lu, H., Watanabe, A., Omura, I., & Kamiya, T. (2020, October). Identification of normal and abnormal from ultrasound images of power devices using VGG16. In *2020 20th International Conference on Control, Automation and Systems (ICCAS)* (pp. 415-418). IEEE.

[13]. Dolk, H., Loane, M., Garne, E., & a European Surveillance of Congenital Anomalies (EUROCAT) Working Group. (2011). Congenital heart defects in Europe: prevalence and perinatal mortality, 2000 to 2005. *Circulation*, *123*(8), 841-849.

[14]. Van Der Linde D, Konings EEM, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJM, et al. Birth prevalence of congenital heart disease worldwide: A systematic review and meta-analysis. Journal of the American College of Cardiology. 2011; 58(21):2241–7. https://doi.org/10.1016/j.jacc.2011.08.025 PMID: 22078432

[15]. Khoshnood B, Lelong N, Houyel L, Thieulin AC, Jouannic JM, Magnier S, et al. Prevalence, timing of diagnosis and mortality of newborns with congenital heart defects: A population-based study. Heart. 2012 Nov 15; 98(22):1667–73. https://doi.org/10.1136/heartjnl-2012-302543 PMID: 22888161

[16]. Stoll C, Garne E, Clementi M. Evaluation of prenatal diagnosis of associated congenital heart diseases by foetal ultrasonographic examination in Europe. Prenatal Diagnosis. 2001; 21(4):243–52. https://doi. org/10.1002/pd.34 PMID: 11288111

[17]. INSERM/ DRESS. Enquête nationale pe´rinatale Rapport 2016. 2017; Available from: http://www.xn— epop-inserm-ebb.fr/wp-content/uploads/2017/10/ENP2016_rapport_complet.pdf

[18]. Durand I, David N, Blaysat G, Marguet C. Diagnosis of congenital heart disease in a nonselected population in Upper Normandy: retrospective study between 2003 and 2007. Archives de Pediatrie. 2009; 16 (5):409–16. https://doi.org/10.1016/j.arcped.2009.02.013 PMID: 19324538

[19]. Suard, C., Flori, A., Paoli, F., Loundou, A., Fouilloux, V., Sigaudy, S., ... & Bretelle, F. (2020). Accuracy of prenatal screening for congenital heart disease in population: A retrospective study in Southern France. *PloS one*, *15*(10), e0239476.

[20]. Divya, M. O., & Vijaya, M. S. (2023). Artificial Intelligent Models for Automatic Diagnosis of Foetal Cardiac Anomalies: A Meta-Analysis. In *Proceedings of the International Conference on Cognitive and Intelligent Computing* (pp. 179-192). Springer, Singapore.

[21]. Divya, M. O., & Vijaya, M. S. (2023, March). Optimizing Pre-processing for Foetal Cardiac Ultra Sound Image Classification. In Innovations in Bio-Inspired Computing and Applications: Proceedings of the 13th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2022) Held During December 15-17, 2022 (pp. 273-286). Cham: Springer Nature Switzerland.

[22]. Divya, M. O., & Vimina, E. R. (2020). Content based image retrieval with multi-channel LBP and colour features. International Journal of Applied Pattern Recognition, 6(2), 177-193.

[23]. Vimina, E. 1., & Divya, M. O. (2020). Maximal multi-channel local binary pattern with colour information for CBIR. Multimedia Tools and Applications, 79(35-36), 25357-25377.

[24]. Divya, M. O., & Vijaya, M. S. (2023, March). Diagnostic Models for Foetal Cardiac Anomalies Using Pattern Classification and FetalEcho_V01 Dataset (In Press). In Proceedings of the Eighth International Conference on Computing, Communication and Security (ICCCS 2023) : Springer in their Communications in Computer and Information Science series.