

Trustworthiness of Four-dimensional Ultrasound and Artificial Intelligence in Improving KANET Test for Detection of Fetuses at Neurorisk

Lemana Spahić¹, Asim Kurjak², Milan Stanojević³, Almir Badnjević⁴, Lejla Gurbeta Pokvić⁵

Received on: 24 December 2023; Accepted on: 30 January 2024; Published on: 28 March 2024

ABSTRACT

Background: Fetal neurological impairment disorders, encompassing conditions like cerebral palsy, epilepsy, and autism spectrum disorder, can result from various factors affecting fetal nervous system development. Timely diagnosis of these disorders is challenging but crucial for early intervention. Recent advancements in deep learning and ultrasound technology present an opportunity to develop a tool for early detection.

Objective: This study aims to leverage convolutional neural networks (CNNs) to analyze fetal neurobehavioral movements in ultrasound images, with the goal of aiding in the early detection of neurological impairment disorders.

Materials and methods: The study utilized a dataset of 3D ultrasound images extracted from 4D recordings of fetuses undergoing the Kurjak Antenatal Neurodevelopmental Test (KANET) during the third trimester. The methodology relies on the application of deep learning, more specifically convolutional neural networks (CNN) for the purpose of recognizing characteristic fetal movements.

Results: The custom CNN architecture achieved an overall accuracy of 93.83%. The system was visualized by means of designing a graphical user interface that includes the developed model that works in the background every time a frame of a recorded 4D ultrasound video is deemed to be parsed through the system. Notably, distinguishing between facial and hand-to-face movements proved challenging. This pilot study lays the foundation for AI-based fetal neurological risk assessment, providing a promising tool for the early detection of fetal neurological impairment disorders.

Conclusion: While acknowledging limitations such as class imbalance and the absence of differentiation between specific facial expressions, the study demonstrates the potential of AI in enhancing prenatal care. Future work will involve expanding the dataset, conducting real-time clinical validations, and further refining the model. The research holds implications for improving outcomes for affected children and making advanced diagnostic capabilities accessible in diverse healthcare settings.

Keywords: Convolutional neural networks, Fetal neurological risk, 4D ultrasound, KANET, TRUEAID.

Donald School Journal of Ultrasound in Obstetrics and Gynecology (2024): 10.5005/jp-journals-10009-2011

INTRODUCTION

Since the dawn of human civilization, we have been cloaked in ignorance about the intricate mechanism of action of our own bodies. We knew nothing of the complex systems and biosignals that operated within us, as our understanding is limited to the tools we have at our disposal. Exploring the brain's development takes us into the world of fetal growth. Every aspect of intelligence and consciousness begins in the womb. Here, the brain's story unfolds, influenced by both genetics and environment. The evaluation of fetal behavior gives the opportunity to recognize the difference between normal and abnormal neurological development and even an early diagnosis of different structural or functional central nervous system abnormalities. Fetal neurological impairment disorders are a group of conditions that affect the development of the nervous system in the fetus. Similar to neurological impairment disorders, fetal disorders of this kind can occur

¹Department for Medical Devices and Artificial Intelligence, Research Institute Verlab for Biomedical Engineering, Medical Devices and Artificial Intelligence, Sarajevo, Bosnia and Herzegovina; Department of Applied Computing in Medicine, Research and Development Center for Bioengineering BioIRC, Kragujevac, Serbia

²Department of Obstetrics and Gynecology, School of Medicine, University of Zagreb, Zagreb, Croatia; Sarajevo Medical School, Sarajevo School of Science and Technology, Sarajevo, Bosnia and Herzegovina

³Department of Obstetrics and Gynecology, Faculty of Medicine, Sveti Duh General Hospital, University of Zagreb, Zagreb, Croatia

⁴Department of Social Pharmacy and Pharmaceutical Legislation, Faculty of Pharmacy, University of Sarajevo, Bosnia and Herzegovina

⁵Department for Medical Devices and Artificial Intelligence, Research Institute Verlab for Biomedical Engineering, Medical Devices and Artificial Intelligence, Sarajevo, Bosnia and Herzegovina

due to a variety of factors, including genetic abnormalities and environmental factors. However, an additional risk factor for fetal neurological impairment disorders is complications during pregnancy and delivery. There are many types of fetal neurological impairment disorders, each with its own unique set of symptoms and causes. Some common examples include cerebral palsy (CP), intellectual disabilities, epilepsy, and autism spectrum disorder.

It has been proven by many studies that most neurological disorders like CP develop prenatally, while postnatal and intrapartum factors are not that important. Analysis of fetal behavior by four-dimensional (4D) ultrasound has been standardized in clinical practice, and the Kurjak antenatal neurodevelopmental test (KANET) is one example of such a test, followed by a postnatal neurological assessment. KANET stands as a peak advancement in the field of noninvasive prenatal neurodevelopmental testing. This test, a result of the work of a consortium led by Dr Asim Kurjak, marks a significant leap in perinatal neurology. KANET is not just a test; it's proof of the fusion of cutting-edge technology and deep clinical understanding, offering an unprecedented insight into fetal neurobehavior. At the core of KANET's groundbreaking approach is the use of 4D ultrasound technology. KANET represents a fundamental shift in how we perceive and analyze fetal development in real-time. The 4D ultrasound stands out for its ability to provide dynamic, multidimensional visualizations of the fetus *in utero*, pushing the boundaries of our understanding of fetal neurological development.^{1,2}

What sets KANET apart is its capacity for real-time, in-detail analysis. The 4D ultrasound creates a dynamic observational platform, turning every fetal movement and behavioral moment into a subject of detailed study. This is not just observation; it is a deep dive into the spectrum of neurobehavioral indices that serve as proxies for the complex neurodevelopmental processes unfolding within the fetal environment. KANET operates on the hypothesis that specific fetal movements and behaviors are direct reflections of the integrity and functionality of the developing neural structures and pathways.³ Each of the eight types of fetal movements assessed in the original KANET framework represents some of the intricate and profound developmental processes occurring within the womb (Fig. 1).

These movements, though seemingly simple, are the descriptors of neural pathways being formed, muscles being tested, and a primordial consciousness experiencing its existence for the first time.⁵⁻¹⁰

This test is performed by highly skilled professionals, fetal and pediatric neurologists, based on the results of prenatal ultrasound screening. These screening tests have high accuracy, but their implementation in healthcare depends on the existence and availability of skilled medical professionals. This presents a challenge in assuring equitable and high-quality healthcare services for everyone. In countries where there are trained specialists to recognize

Corresponding Author: Lemana Spahić, Department for Medical Devices and Artificial Intelligence, Research Institute Verlab for Biomedical Engineering, Medical Devices and Artificial Intelligence, Sarajevo, Bosnia and Herzegovina; Department of Applied Computing in Medicine, Research and Development Center for Bioengineering BioIRC, Kragujevac, Serbia, e-mail: lemanaspahic@gmail.com

How to cite this article: Spahić L, Kurjak A, Stanojević M, *et al.* Trustworthiness of Four-dimensional Ultrasound and Artificial Intelligence in Improving KANET Test for Detection of Fetuses at Neurorisk. Donald School J Ultrasound Obstet Gynecol 2024;18(1):6–16.

Source of support: Nil

Conflict of interest: None

these diseases, this number is much lower. However, the race against time in terms of educating specialists around the world is certainly a lost battle without the application of today's technological advances. One of the components of the KANET test is the assessment of general movement "gestalt perception," which has been the subject of artificial intelligence (AI) application postnatally, but up to now, it has not been done prenatally.

The future of assessing fetal neurological risk lies in the convergence of technological innovations, genomic insights, and ethical advancements.¹¹ In this evolving landscape, the focus extends beyond diagnostic precision to encompass the holistic well-being of the child and family. It's a future where technology, ethics, and humanity converge, ensuring that each child, regardless of neurological risk status, is born into a world of optimized care, support, and opportunity. AI's potential in the assessment of fetal neurological risk is not just about better accuracy. It's about reshaping our ethical views and principles. The question is not just if AI can help but how it changes the human experience of these disorders.

Artificial Intelligence in Medicine

All AI methodologies have a certain niche for which they are most suitable. However, the most complex architectures are observed in neural networks. That, and their close resemblance to the human brain, had made neural networks the basis of deep learning (DL) involved in current scientific endeavors toward creating machines that are highly accurate and reliable.¹²

In general, AI represents all computer programs that are capable of mimicking processes that usually require human cognitive processes. It is a very broad field that continues to expand with the advent of computational technologies and capabilities. The terminology of AI is often used interchangeably with terms such as machine learning (ML) and artificial neural network (ANN), as well as DL, which is not entirely correct. The relationships between these terms, as well as their corresponding definitions, are shown in Figure 2.

The complexity of the field does not end with this classification. In addition to containing ANN and DL as

subfields, ML encompasses all programming based on statistical techniques that enable computers to make predictions based on recognizing patterns without explicit instructions on how to perform the prediction.¹³

Deep learning (DL) using ANNs has gained popularity in recent years due to its ability to handle complex and high-dimensional data for various tasks, including classification. ANNs are particularly effective for classification tasks because

they can automatically learn and extract relevant features from complex data. Medical data, even for conditions with an established diagnostic procedure, are very complex due to the high redundancy of parameters and their patient-specific variability. ANNs are capable of capturing complex and nonlinear relationships between input features and output labels, allowing them to model complex decision boundaries between different classes.


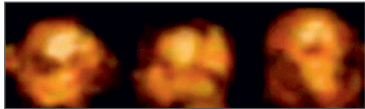

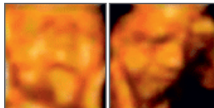
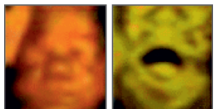
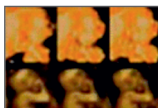

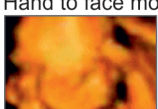

Sign	Score			Sign score
	0	1	2	
Isolated head anteflexion 	Abrupt	Small range (0–3 times of movements)	Variable in full range, many alternation (>3 times of movements)	
Cranial sutures and head circumference 	Overlapping of cranial sutures	Normal cranial sutures with measurement of HC below or above the normal limit (–2SD) according to GA	Normal cranial sutures with normal measurement of HC according to GA	
Isolated eye blinking 	Not present	Not fluent (1–5 times of blinking)	Fluency (>5 times of blinking)	
Facial alteration (grimace or tongue expulsion) 	Not present	Not fluent (1–5 times of alteration)	Fluency (>5 times of alteration)	
Mouth opening (yawning or mouthing) 	Not present	Not fluent (1–3 times of alteration)	Fluency (>3 times of alteration)	
Isolated hand movement 	Cramped	Poor repertoire	Variable and complex	
Isolated leg movement 	Cramped	Poor repertoire	Variable and complex	
Hand to face movements 	Abrupt	Small range (0–5 times of movement)	Variable in full range, many alternation (>6 times of movements)	
Fingers movements 	Unilateral or bilateral clenched fist, (neurological thumb)	Cramped invariable finger movements	Smooth and complex, variable finger movements	
Gestalt perception of GMs	Definitely abnormal	Borderline	Normal Total score	

Fig. 1: Kurjak antenatal neurodevelopmental test (KANET) scoring system⁴

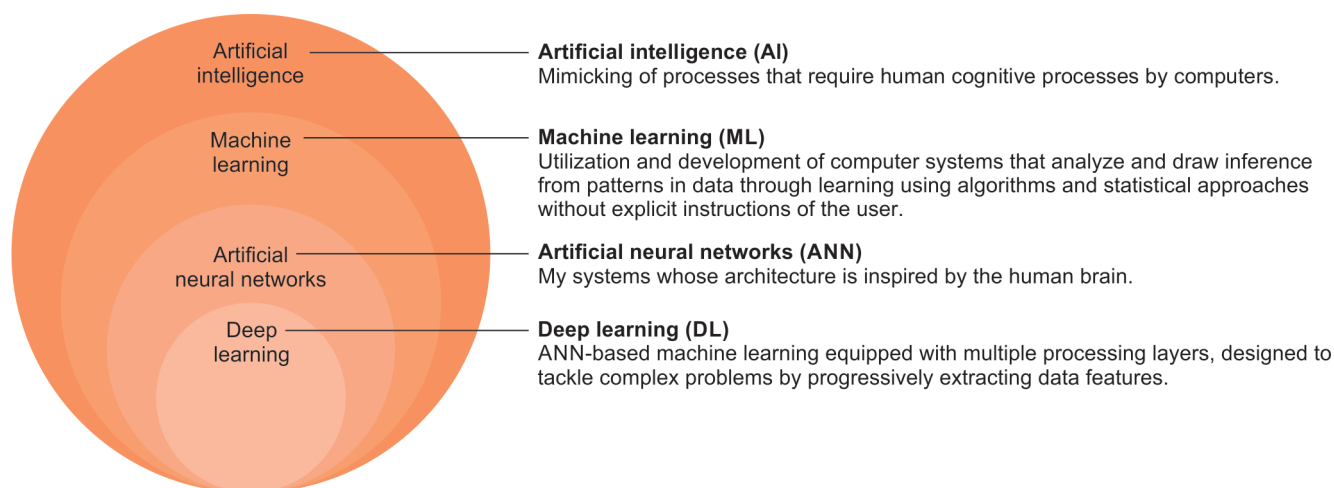


Fig. 2: Relationship between AI and its subfields

Artificial neural networks (ANNs) are computational models inspired by the human brain's structure and function. They consist of layers of nodes or "neurons" connected by "synapses." Each connection has a weight, representing its strength or influence. ANNs are popular in ML and AI for tasks like classification, regression, and pattern recognition. An ANN typically consists of an input layer, one or more hidden layers, and an output layer. Each layer has multiple neurons. The input layer receives the initial data, while the output layer produces the final output. The hidden layers perform complex computations to transform the input into the output.

During training, the network adjusts the weights and biases of the nodes to minimize the error between the predicted and actual output. The number of layers, nodes, and activation functions can vary depending on the task and the type of data being processed. The architecture of ANNs allows for the integration of multiple layers of nonlinear processing, which can capture increasingly abstract and complex representations of the data. The hidden layers of the network are especially important for learning these representations, which can be difficult to define manually. Another advantage of ANNs is their ability to generalize to new, unseen data. The ability of ANNs to learn these representations makes them particularly effective for feature extraction and selection, which is a crucial step in many classification tasks. During training, the network learns to generalize patterns in the training data to new, unseen data, allowing it to make accurate predictions on new samples. This generalization is particularly important for classification tasks, as the goal is often to make accurate predictions on new, previously unseen data. ANNs are particularly effective at feature extraction and selection, which is crucial for many classification tasks. In contrast, decision trees, random forest, and Naive Bayes algorithms rely on predefined features or feature selection techniques, which can limit their performance when dealing with complex data.¹⁴

The mathematics behind ANNs involves linear algebra, calculus, and statistics.¹⁴ Each neuron computes a weighted sum of its inputs, adds a bias, and then applies

an activation function. Mathematically, this process can be represented as:

$$y = f\left(\sum_{i=1}^n w_i x_i + b\right) \quad (1)$$

Where:

- y is the neuron's output,
- f is the activation function,
- w are the weights,
- x are the inputs, and
- b is the bias.

Activation functions introduce nonlinearity into the network, enabling it to learn complex patterns. Common activation functions include the sigmoid, tanh, and Rectified Linear Unit.

Artificial intelligence (AI) continues to revolutionize the field of medical diagnosis, with advancements in ML, particularly DL, leading the charge. These technologies have proven instrumental in enhancing the accuracy, speed, and efficiency of diagnosing a variety of medical conditions. The integration of AI in healthcare has been a subject of ongoing research and development over the past few years.¹⁵ AI systems, particularly ML and DL algorithms, have demonstrated unprecedented capabilities in diagnosing diseases, sometimes outperforming human clinicians.¹⁶

In addition to clinical decision support (CDS) systems (DSS), the application of AI and DSS extends toward the management and maintenance of medical equipment. As medical equipment stands at the forefront of medical decision-making, it is of utmost importance to ensure its performance and accuracy. The European Commission has stipulated the importance of this by introducing postmarket surveillance as mandatory in the new medical device regulation introduced in 2017 and put in force in 2022.¹⁷⁻¹⁹

Postmarket surveillance of medical devices²⁰ has been proven useful in case studies conducted in Bosnia and Herzegovina, where a large number of medical devices

have been deemed inaccurate on the basis of performance inspection.^{21–25} As a result of performing postmarket surveillance, a vast amount of data was collected, and the team from Verlab has decided to utilize it and design algorithms capable of predicting medical device failure on the basis of their performance throughout the years.^{26–30} Transcending the diagnostic challenges and ensuring safe and reliable measurements made by medical devices, the following paragraphs will briefly describe the applications of AI as DSS for aiding in diagnosis, treatment, and prognosis of the leading causes of mortality and comorbidity worldwide.

Artificial intelligence (AI) is very useful in predicting heart failure using electronic health records (EHR) and real-time cardiac monitoring data. ML algorithms can analyze vast datasets, including clinical, laboratory, and imaging data, to identify early signs of heart failure, enabling proactive management.^{31,32} Another application of AI in the field of cardiology is ML technologies, which are employed for the prediction, classification, and outcome prediction of stroke. They analyze clinical data, imaging, and genetic information to classify stroke types, predict occurrences, and project recovery outcomes, significantly enhancing patient care.³³

Artificial intelligence (AI)—based DSS in healthcare is an integral tool that assists clinicians and healthcare professionals in making informed and accurate decisions. These systems leverage a combination of technologies, data, and algorithms to provide insights and recommendations, enhancing the quality and efficiency of healthcare delivery. Healthcare DSS integrates a vast array of data sources, including EHRs, laboratory results, and medical imaging data. For instance, Kawamoto et al.³⁴ demonstrated that the integration of clinical data into DSS significantly improves clinical practice and patient outcomes. These systems utilize advanced algorithms and AI to analyze complex datasets, offering personalized recommendations for patient care. CDS systems, a subset of DSS, are particularly notable for their role in diagnosis and treatment. They analyze patient-specific data to provide evidence-based recommendations. A study by Osheroff et al.³⁵ highlighted the role of CDS in reducing medical errors, improving healthcare quality, and reducing costs. However, the implementation of DSS in healthcare is not without challenges.

Ethical and privacy concerns are paramount, underscoring the intricate balance between technological advancement and ethical considerations. The ethical implications of using DSS were analyzed by Ammenwerth et al.,³⁶ shedding light on a spectrum of concerns that are as diverse as they are complex. One of the primary concerns, as mentioned by both the European Union AI Act and the United Kingdom regulation, is data privacy. With DSS integrating vast amounts of sensitive patient data, the risk of unauthorized access and data breaches is a significant concern. Patients' confidential information, including medical histories, diagnoses, and treatment

plans, must be safeguarded with the utmost integrity. The systems must comply with legal frameworks like the Health Insurance Portability and Accountability Act in the United States or the General Data Protection Regulation in Europe, which impose stringent measures to protect patient data. In addition to privacy, security is another important aspect. The infrastructure supporting DSS must be fortified against potential cyber-attacks and unauthorized access. The integrity of the data and the systems is crucial not just for the privacy of the individuals but also for the accuracy and reliability of the decision support provided. A breach could not only compromise privacy but also the quality of healthcare delivery. The potential for bias in algorithmic recommendations is also a pressing ethical issue. Algorithms are designed and trained by humans and can inadvertently inherit biases present in the training data or the designers. This can lead to skewed or unfair recommendations, impacting certain patient groups disproportionately. It underscores the need for transparency, fairness, and accountability in the design and implementation of algorithms in DSS. The issue of informed consent also looms large. Patients must be adequately informed about how their data will be used and must have the autonomy to consent or decline. Transparency in the usage of data and the decisions made by DSS is integral to building trust and ensuring ethical standards.³⁶

This paper looks at AI in the assessment and management of fetal risk as part of a bigger story that combines medical science, engineering, and ethics. It calls for a reevaluation of our current views and an interdisciplinary discussion. The main question is how AI can improve diagnostics and redefine our understanding of existence in the face of these disorders.

MATERIALS AND METHODS

Dataset

The dataset for the development of the AI-empowered DSS, entitled Trustworthy AI System for Fetal Neurological Risk Assessment and Diagnostic Support (TRUEAID), consisted of three-dimensional (3D) ultrasound images. For the trustworthy AI system for TRUEAID development, a total of 10,452 samples were acquired from 2021 to 2023. The images were extracted from 4D ultrasound recordings of fetuses made during the KANET test—a prenatal test for evaluation of the neurological development of fetuses during the third trimester of pregnancy. KANET is performed on both healthy and suspected pathological pregnancies. However, the pool of data on normally developing fetuses in healthy pregnancies is larger; the data retrieved for the purpose of development of the TRUEAID system was acquired from these cases. The dataset for the development of the diagnostic support system for the purpose of this project was acquired courtesy of Dr Panos Antsaklis from Alexandra Maternity Hospital in Athens. This KANET testing

center in Athens represents the busiest testing center in the KANET network; hence, it has the most extensive scope of data at its disposal.

The methodology employed in this study is systematic and adheres to rigorous scientific standards, encompassing four pivotal stages: data preprocessing, data augmentation, convolutional neural network (CNN) model development, and interface development, contained a series of steps that had to be performed in sequence in parallel with evaluation of the results of those steps (Flowchart 1).

As the first step in the development of an AI-based algorithm for classification is the classification of data into corresponding categories, a graphical user interface (GUI) was developed to facilitate this step. The GUI is shown in Figure 3. The raw dataset is loaded, and each image is classified into its corresponding category on the basis of the information provided by gynecologists who have performed the KANET test.

Upon classification of data, it was necessary to perform anonymization. Each raw image contained a lot of additional information in addition to the useful image frame. Hence, it was necessary to crop every image to fit a certain frame and to remove as much background as possible in order not to hinder

the classification power. Due to the fact that all images that are used as input to the diagnostic support system must be of the same size, the images were cropped to a 512×512 frame using a Python script that performed cropping on the edge of every image on the basis of color intensity of useful image frame pixels. However, not all useful image frames were of the same size; hence, padding with black pixels was necessary for some images in order to achieve the same image size.

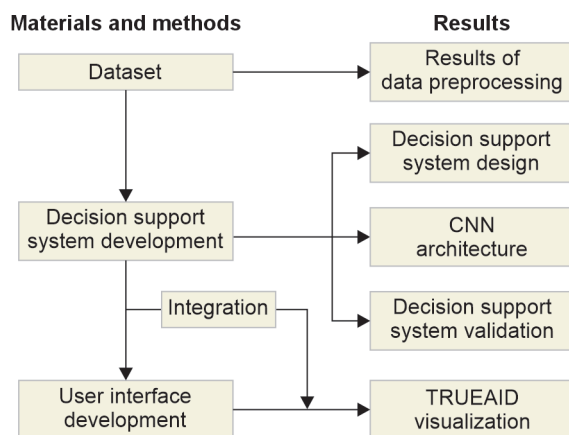
Before the images could be used for the development of the expert system, it was necessary to perform image quality enhancement procedures in order to make extraction of the relevant features as easy as possible for the AI system. For this purpose, a set of filters was applied in order to denoise the images, removing unnecessary low-contrast differences in the images and increasing the pixel intensity variation in each image. Even though the differences between raw and processed images cannot be seen with the naked eye, the performance of the system is significantly improved following image quality enhancement procedures.

Once preprocessed, the dataset was analyzed to check for significant class imbalance in the dataset. As class imbalance was noticed, dataset augmentation was performed for the minority classes, all classes except the "facial movement class." Augmentation was performed using a Python script that rotates every image in different angles up to 60° angle to account for as many variations as possible and contribute to the generalization power of the model.

Artificial Intelligence System Development

Convolutional neural networks (CNN) have proven their usefulness for computer vision problems and image classification models. CNNs automatically learn and create hierarchical patterns in the data, which is particularly useful for complex tasks like image recognition. They apply filters/kernels to input images to create feature maps that identify patterns like edges, textures, and more complex patterns in deeper layers. CNNs use shared weights across their architecture, reducing the number of parameters to be learned and making the model more efficient. Weight

Flowchart 1: Methodology workflow



Meaning of numerical classes:

- 0—Pathological
- 1—Face
- 2—Hand-to-face
- 3—Characteristic thumb
- 4—Legs

Fig. 3: Graphical user interface (GUI) for image classification

assignment and weight sharing are of particular usefulness in image recognition tasks as they lead to translation invariance, meaning the network can recognize patterns irrespective of their position in the input space. As discussed previously, medical images are susceptible to noise; hence, CNN's resistance to noise and distortions in the input image makes them robust in varied conditions.

The main characteristic of a CNN is its depth, meaning the number of convolutional layers and the number of neurons used in each convolutional layer to perform classification. Over 100 iterations of the CNN were performed in order to test a multitude of different combinations of:

- The number of convolutional layers.
- Number of filters.
- Number of neurons in each layer.
- Activation functions of convolutional layers.
- Activation functions of fully connected layers.

Artificial Intelligence System Evaluation

The main instruments used to determine the CNN performance are training and validation loss and accuracy values at the end of the training and training graphs. However, in order to evaluate the performance of a classifier, it is instrumental to show and evaluate the confusion matrix of internal system validation during training and to derive true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each class. These parameters are defined as:

- TP: The number of correct predictions for the particular class.
- TN: The number of correct predictions for all other classes.
- FP: The number of incorrect predictions where other classes were predicted as the particular class.
- FN: The number of incorrect predictions where the particular class was predicted as some other class.

Table 1 shows an example of a confusion matrix where we evaluate the TP, TN, FP, and FN for class 2 of a multiclass classifier. In this case, TP is represented at the intersection of the predicted label and the true label. FP are the values in the "class 2" column but outside the "class 2" row. Thus, FP = sum of all values in the "class 2" column - TP. FN are the values in the "class 2" row but outside the "class 2" column. Thus, FN is the sum of all values in the "class" row - TP. TN are the remaining values but can be calculated as total - FP - FN - TP.

True positive (TP), TN, FP, and FN are then used to calculate the system's performance metrics. Conventionally used performance metrics are sensitivity (recall), specificity,

precision or positive predictive value (PPV), negative predictive value (NPV), and accuracy. These parameters are defined by mathematical formulas (Eqs. 2–8).

Sensitivity (recall) or TP rate (TPR) is the proportion of positive instances that are correctly classified. It indicates the proportion of actual positives that are correctly identified.

$$\text{sensitivity (recall)} = \frac{TP}{TP + FN} \quad (2)$$

Specificity or TN rate (TNR) is the proportion of negative instances that are correctly classified. Specifically, it measures the proportion of actual negatives that are correctly identified.

$$\text{specificity} = \frac{TN}{TN + FP} \quad (3)$$

Precision or PPV indicates the proportion of positive identifications that were actually correct. In other words, of all the instances classified as positive, how many were actually positive?

$$\text{precision} = \frac{TP}{TP + FP} \quad (4)$$

Negative predictive value (NPV) is the proportion of negative instances among the instances that are predicted as negative. In other words, NPV indicates the probability that a predicted negative result is indeed negative.

$$\text{NPV} = \frac{TN}{TN + FN} \quad (5)$$

Accuracy measures the proportion of all classifications that were correct.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

One of the primary challenges in many real-world classification problems is class imbalance, where one class significantly outnumbers the other(s). In such cases, accuracy can be misleading. For instance, in a dataset where 95% of instances are negative, and 5% are positive, a naive classifier that predicts everything as negative would achieve a 95% accuracy rate, even though it's not making any useful prediction.

While accuracy provides a general measure of classification performance, it doesn't always provide a complete picture, especially in cases with imbalanced classes or when different types of classification errors carry different costs or implications. Even though it gives an overall measure of how many instances are correctly classified, it doesn't differentiate between the types of errors being made. Two models with the same accuracy might have very different numbers of FPs and FNs. The F1 score and Matthews correlation coefficient (MCC) can provide a more comprehensive understanding of a model's performance in such situations.

Table 1: Example confusion matrix

True label	Predicted label			
Class 1	TN	FP	TN	TN
Class 2	FN	TP	FN	FN
Class 3	TN	FP	TN	TN
Class 4	TN	FP	TN	TN

The F1 score is the harmonic mean of precision and sensitivity (recall) (Eq. 2). It gives a balanced measure between precision and recall when the class distribution is uneven. By using the F1 score, we ensure that both FPs (precision aspect) and FNs (recall aspect) are taken into account. It's especially useful when FNs and FPs have different impacts on the classification power. It is very useful for imbalanced datasets. As mentioned, the F1 score can provide a more realistic measure of a classifier's performance on imbalanced datasets by balancing the importance of precision and recall.

$$F1\ score = 2 \times \frac{(precision \times sensitivity)}{precision + sensitivity} \quad (7)$$

Matthews correlation coefficient (MCC) is a metric that gives a balanced measure even when the class sizes are highly imbalanced. MCC is essentially a correlation coefficient between the observed and predicted binary classifications. It returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 indicates no better than a random prediction, and -1 means a total disagreement between the prediction and the actual class. Unlike accuracy, which only considers TPs and TNs, MCC takes into account TPs, TNs, FPs, and FNs. This gives a more holistic view of the classifier's performance. MCC is especially useful in datasets where the classes are of very different sizes.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (8)$$

Graphical User Interface Development

A suitable user interface is essential for the effective utilization of AI DSSs in healthcare. The implementation of appropriate user interfaces for AI DSSs in healthcare aligns with the current guidelines and regulations. The regulation emphasizes the necessity for user-friendly interfaces, ensuring the safety and performance of medical devices, including those incorporating AI technology, within the healthcare system. Therefore, once developed and proven useful, the TRUEAID DSS required a GUI in order to make it comprehensible for medical experts.

RESULTS

The results of training the DSS can be represented in a multitude of ways. The first results of training that are obtained as early as during the training process itself are the performance metrics such as loss and accuracy. The loss function is plotted throughout CNN training, and the graphs are monitored for convergence. As mentioned in the methodology section, CNN had over 100 iterations until an architecture with satisfactory performance was achieved. However, the most effective way of testing the performance of a CNN is by monitoring the performance metrics such as specificity, sensitivity, precision, NPV, accuracy, F1 score, and MCC.

Table 2 represents a summarized version of the training process and achieved performance parameters for different CNN iterations.

It can be deduced from Table 2 that iteration 126 was the point at which the training process that underpinned significant changes made to the CNN architecture was halted. After that, minor fine-tuning was performed in order to inspect the robustness of the developed system.

A glance at the performance metric reveals that the model demonstrates remarkable adeptness in discerning negative instances for each class. This uniform excellence in specificity ensures that the classifier rarely mislabel instances from other classes. Recall somewhat varies between the classes. While classes such as "face" and "legs" are emblematic of state-of-the-art detection of TPs, other classes exhibit nuanced performances. The "hand-to-face" class displays a decent but not exemplary recall of 0.894, suggesting that there's a subset of this class the model is not capturing effectively. "Thumb," although good, still leaves some room for improvement with a sensitivity of 0.937. In terms of precision, "face" and "legs" are the pinnacles of perfection, with scores at 1.00, reinforcing that their positive predictions are consistently accurate. "Thumb" showcases commendable precision at 0.929, indicating its predictions are mostly on the mark. However, "hand-to-face" presents a precision of 0.902, which, though good, signals that occasional FPs do creep into its predictions. The accuracy metric provides a comprehensive view of the model's overall correctness. "Face" and "legs" have impeccable accuracy,

Table 2: Summary performance evaluation during CNN training

CNN iteration	Specificity (TNR)	Sensitivity (recall/TPR)	Precision (PPV)	NPV	Accuracy	F1 score	MCC
No. 1	0.92	0.30	0.30	0.94	0.69	0.30	-0.02
No. 20	0.91	0.19	0.18	0.82	0.78	0.19	0.06
No. 60	0.99	0.87	0.99	0.918	0.97	0.91	0.86
No. 85	1.00	0.51	1.00	1.00	0.68	0.68	0.58
No. 126	0.99	0.95	0.96	0.98	0.94	0.96	0.94

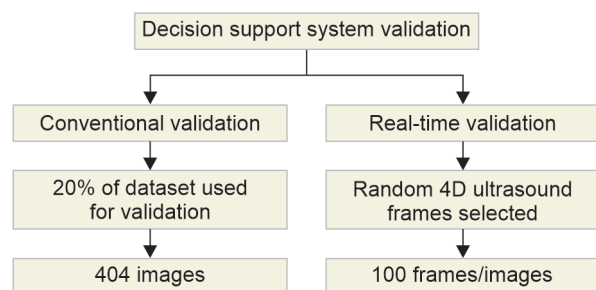
but other classes, such as "thumb" and "hand-to-face," also tread the higher echelons with scores of 0.972 and 0.953, respectively. These scores reinforce the model's overall reliability, but they also underscore the importance of examining other metrics for a nuanced understanding. The F1 score captures the harmonic mean of precision and recall. Stellar performances by "face" and "legs" are echoed here, emphasizing their balanced prowess. "Thumb," too, with an F1 score of 0.948, showcases a healthy balance between precision and recall. However, the "hand-to-face" class, with its F1 score of 0.898, reflects the earlier observed imbalance. The MCC for class "legs" has a perfect score of 1.00. "Face" and "thumb" also resonate excellently with MCC values of 0.996 and 0.915, respectively, highlighting their near-perfect classification. However, "hand-to-face," with its MCC of 0.874, underpins the challenges it faces in delivering a balanced classification.

The DSS was validated in two distinct manners (Flowchart 2). The first validation was performed using conventional methodology, where 20% of the initial dataset was used for subsequent validation of the system. The second mode of validation was real-time validation, where 4D ultrasound recordings, whose frames were not previously presented to the network, were parsed at random time points to evaluate the system performance.

The results of the conventional validation are represented in Table 3.

When analyzing the performance evaluation table, it can be seen that the performance is comparable with the system's internal validation. The performance metrics do not deviate significantly and are within $\pm 2\%$ in comparison with the internal validation of the last CNN architecture.

Flowchart 2: Block diagram of system validation



The second validation option is dedicated to the evaluation of system performance in real-time. Hence, it required the development of an interactive decision support interface and visualization of the entire system.

The TRUEAID system was visualized by designing a GUI that includes the developed model that works in the background every time a frame of a recorded 4D ultrasound video is deemed to be parsed through the system. In case any of the characteristic movements belonging to classes "face," "hand-to-face," "legs," or "thumb" are recognized, the class of movement is displayed on the screen. In case none of these movements are present in the parsed frame, the frame is deemed as "pathological."

As can be seen from Figure 4, the interactive TRUEAID system allows for loading of a video, playing it, and pausing it at a certain frame to parse that frame through the developed CNN. The physician can scroll through the entire video, take any frame, and parse it through the CNN (Fig. 4).

Upon reaching the desired frame, the physician presses the "pause and predict" button, and the frame is parsed through the CNN. This process takes up to 10 seconds, as the model was optimized for performance in both high and low computational power settings. Once the frame is parsed, the result is displayed in the form of a pop-up window (Figs 5 to 7).



Fig. 4: Scrolling through the loaded 4D ultrasound video

Table 3: Validation performance evaluation

Class	Specificity (TNR)	Sensitivity (recall/TPR)	Precision (PPV)	NPV	Accuracy	F1 score	MCC
Face	0.98	1.00	0.95	1.00	0.94	0.97	0.97
Hand-to-face	0.97	0.89	0.89	0.97	0.95	0.89	0.87
Legs	1.00	0.94	1.00	0.98	0.98	0.97	0.96
Thumb	0.98	0.93	0.95	0.97	0.97	0.94	0.90
Overall	0.98	0.94	0.95	0.98	0.96	0.95	0.89



Fig. 5: Result of TRUEAID when hand-to-face movements are recognized



Fig. 7: Result of DSS when no characteristic movements are recognized



Fig. 6: Result of TRUEAID when facial movements are recognized

CONCLUSION

The overall aim of this work was to establish new knowledge in the prediction of risk for neurological disorders and to explore the feasibility and trustworthiness of a new technology for this purpose. Achievement of this goal shall contribute to the advancement in the management of noncommunicable diseases and a very slow process of the digital transition. The specific objective of the project was to develop TRUEAID—a trustworthy AI system for fetal neurological risk assessment and diagnostic support. This project is driven by ambition and commitment to bring AI technology to actual use in obstetrics and gynecology for the prevention and detection of neurological risk in order to improve the well-being of affected populations (pregnant women, mothers and infants, and children with neurological conditions). Once developed and proven, TRUEAID can be used anywhere in the world, from low-resource to high-resource settings, enabling better care of affected populations and supporting the fight against noncommunicable neurological diseases.

Trustworthiness of a developed AI system encompasses the following characteristics: transparency, fairness, reliability, and safety. Transparency in AI is a multifaceted concept, encompassing the need for clarity in how AI systems operate and make decisions. This goes beyond the technical realm, involving the communication of AI processes to users and stakeholders in a comprehensible manner. As AI is commonly considered a black box, it is necessary to eliminate the doubt in the processes governing decision-making by the development of interpretable models where the decision-making pathways are not just a black box but accessible and understandable. TRUEAID is developed and described in a transparent way so that every step of the process leading to the final product is thoroughly explained, enabling a better understanding of the entire decision-making process.

Trustworthy AI System for Fetal Neurological Risk Assessment and Diagnostic Support (TRUEAID) solution can revolutionize the detection of neurological risk, enabling early possible and preliminary diagnosis during the antenatal period and allowing investigation of possible treatments in this phase as well.

REFERENCES

1. Kurjak A, Stanojevic M, Andonotopo W, et al. Fetal neurobehavioral development: from the womb to the world. *J Perinat Med* 2007;35(5):377–390. DOI: 10.1515/JPM.2007.108
2. Kurjak A, Stanojević M, Barišić LS, et al. Kurjak antenatal neurodevelopmental test (KANET): a useful tool for fetal neurodevelopmental assessment. *Clinical Management of Infertility: Problems and Solutions*. 2021. pp. 271–301.
3. Kurjak A, Stanojevic M, Azumendi G, et al. The potential of four-dimensional (4D) ultrasonography in the assessment of fetal awareness. *J Perinat Med* 2005;33(1):46–53. DOI: 10.1515/JPM.2005.008
4. Tinjić S. Experiences and Results of the KANET Test Application in Clinical Practice in Tuzla, Bosnia and Herzegovina. *Donald School J Ultrasound Obstet Gynecol* 2019;13(3):94–98. DOI: 10.5005/jp-journals-10009-1595

5. Kurjak A, Azumendi G, Vecek N, et al. Fetal hand movements and facial expression in normal pregnancy studied by four-dimensional sonography. *J Perinat Med* 2003;31(6):496–508.
6. Kurjak A, Azumendi G, Andonotopo W, et al. Three- and four-dimensional ultrasonography for the structural and functional evaluation of the fetal face. *Am J Obstet Gynecol* 2007;196(1):16–28. DOI: 10.1016/j.ajog.2006.06.090
7. Kurjak A, Badreldeen A, Azumendi G, et al. Can we improve the assessment of fetal behavior in early pregnancy? *Int J Gynecol Obstet* 2005;88(3):307–313.
8. Kurjak A, Pooh RK, Merce LT, et al. Structural and functional early human development assessed by three-dimensional and four-dimensional sonography. *Fetal Diagnos Ther* 2005;20(6):496–508. DOI: 10.1159/000088164
9. Kurjak A, Chervenak FA. *Donald School Textbook of Ultrasound in Obstetrics and Gynecology*, 4th edition. Jaypee Brothers Medical Publishers; 2017.
10. Kurjak A, Stanojevic M, Andonotopo W, et al. How useful is 4D sonography in perinatal medicine? *J Perinat Med* 2006;34(5):437–450. DOI: 10.1515/JPM.2006.087
11. Spahić L, Mašetić Z, Badnjević A, et al. Artificial intelligence-based ultrasound imaging classification for infant neurological impairment disorders: a review. *Mediterranean Conference on Medical and Biological Engineering and Computing*. Switzerland; Cham: Springer Nature. 2023. pp. 620–627.
12. Liu F, Zhou Z, Samsonov A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology* 2018;289(1):160–169. DOI: 10.1148/radiol.2018172986
13. Yazdani A, Costa S, Kroon B. Artificial intelligence: friend or foe? *Aust N Z J Obstet Gynaecol* 2023;63(2):127–130. DOI: 10.1111/ajo.13661
14. Nu S, Bhokal R. Study of artificial neural network. *Int J Math Trends Technol* 2017;47:253–259. DOI: 10.14445/22315373/IJMTT-V47P535
15. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;2(4):230–243. DOI: 10.1136/svn-2017-000101
16. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med* 2019;25(1):24–29. DOI: 10.1038/s41591-018-0316-z
17. Badnjević A, Pokvić LG, Spahić L. Inspection of medical devices. *Clinical Engineering Handbook*. Academic Press. 2020. pp. 491–497.
18. Badnjević A, Pokvić LG, Spahić L. Pediatric and neonate incubators. *Clinical Engineering Handbook*; Academic Press. 2020. pp. 514–519.
19. Badnjević A, Pokvić LG, Deumić A, et al. Post-market surveillance of medical devices: A review. *Technol Health Care* 2022;30(6):1315–1329. DOI: 10.3233/THC-220284
20. Badnjević A, Gurbeta L, Bošković D, et al. Measurement in medicine—Past, present, future. *Folia Medica* 2015;50(1):43–46.
21. Gurbeta L, Badnjević A, Žunić E, et al. (2015, October). Software package for tracking status of inspection dates and reports of medical devices in healthcare institutions of Bosnia and Herzegovina. In 2015 XXV International Conference on Information, Communication and Automation Technologies (ICAT) (pp. 1–5). IEEE.
22. Gurbeta L, Badnjević A, Dzemic Z, et al. (2016, October). Testing of therapeutic ultrasound in healthcare institutions in Bosnia and Herzegovina. In 2nd EAI International Conference on Future Access Enablers of Ubiquitous and Intelligent Infrastructures (pp. 24–25).
23. Gurbeta L, Dzemic Z, Bego T, Sejdic E & Badnjević A. Testing of anesthesia machines and defibrillators in healthcare institutions. *Journal of Medical Systems*, 2017;41:1–10. DOI: 10.1007/s10916-017-0783-7
24. Gurbeta L, Izetbegović S, Badnjević-Čengić A. Inspection and testing of pediatric and neonate incubators. *Inspection of Medical Devices*. 2018. pp. 221–249.
25. Gurbeta L, Badnjević A. Inspection process of medical devices in healthcare institutions: software solution. *Health Technol* 2017;7(1):109–117. DOI: 10.1007/s12553-016-0154-2
26. Badnjević A, Avdihodžić H, Gurbeta Pokvić L. Artificial intelligence in medical devices: past, present and future. *Psychiatr Danub* 2021;33(suppl 3):S336–S341.
27. Hadžić L, Fazlić A, Hasanić O, et al. Expert system for performance prediction of anesthesia machines. *CMBEBIH 2019: Proceedings of the International Conference on Medical and Biological Engineering*. 16–18 May 2019, Banja Luka, Bosnia and Herzegovina. Springer International Publishing. 2020. pp. 671–679.
28. Hrvat F, Spahić L, Aleta A. (2023) Heart disease prediction using logistic regression machine learning model. Joint conference of the Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON) and the International Conference on Medical and Biological Engineering in Bosnia and Herzegovina (CMBEBIH) (September 2023)
29. Kovačević Ž, Gurbeta Pokvić L, Spahić L, et al. Prediction of medical device performance using machine learning techniques: infant incubator case study. *Health Technol* 2020;10(1):151–155. DOI: 10.1007/s12553-019-00386-5
30. Spahić L, Kurta E, Čordić S, et al. Machine learning techniques for performance prediction of medical devices: infant incubators. In *CMBEBIH 2019: Proceedings of the International Conference on Medical and Biological Engineering*, 16–18 May 2019, Banja Luka, Bosnia and Herzegovina. Springer International Publishing. 2020. pp. 483–490.
31. Šećkanović A, Šehovac M, Spahić L, et al. (2020, June). Review of artificial intelligence application in cardiology. In 2020 9th Mediterranean Conference on Embedded Computing (MECO) (pp. 1–5). IEEE.
32. Spahić L, Softić A, Durak-Nalbantić A, et al. (2023, September). Integrating Machine learning in clinical decision support for heart failure diagnosis: case study. *Mediterranean Conference on Medical and Biological Engineering and Computing*. Switzerland. Cham: Springer Nature. 2023. pp. 696–705.
33. Hrvat, F., Spahić, L., Aleta, A. (2023) Heart disease prediction using logistic regression machine learning model. Joint conference of the Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON) and the International Conference on Medical and Biological Engineering in Bosnia and Herzegovina (CMBEBIH) (September 2023)
34. Kawamoto K, Houlihan CA, Balas EA, et al. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005;330(7494):765. DOI: 10.1136/bmj.38398.500764.8F
35. Osheroff JA, Teich JM, Middleton B, et al. A roadmap for national action on clinical decision support. *J Am Med Inform Assoc* 2007;14(2):141–145. DOI: 10.1197/jamia.M2334
36. Ammenwerth E, Rigby M, Talmon J. *Evidence-Based Health Informatics: Promoting Safety and Efficiency Through Scientific Methods and Ethical Policy*. IOS Press. 2018.