JHIT
JOURNAL OF HEALTH INSTITUTES OF TÜRKİYE
TÜRKİYE SAĞLIK ENSTİTÜLERİ BAŞKANLIĞI DERGİSİ

Review

TÜSEB 2023;6(3):125-153 • **DOI:** 10.5578/tusebdergisi.20239602

Artificial Intelligence in Fetal Health Diagnosis: A Systematic Literature Review

Adem Kuzu, Yunus Santur

Department of Software Engineering, Fırat University Faculty of Technology, Elazığ, Türkiye

Cite this article as: Kuzu A, Santur Y. Artificial intelligence in fetal health diagnosis: A systematic literature review. TÜSEB 2023;6(3):125-153.

ABSTRACT

Obstetricians commonly use Non-Stress Test to evaluate fetal well-being during the antepartum and intrapartum periods by measuring fetal heart rate and uterine contractions of the mother. Non-Stress Test is also used to diagnose fetal distress at an early stage. Early diagnosis and treatment can increase the fetus's survival rate and improve their quality of life. The fetal heart rate and uterine contraction signals obtained from the Non-Stress Test are recorded on a paper called a trace. Obstetricians interpret the trace to make decisions about the fetus's condition. However, traditional analysis of Non-Stress Test takes time, and there are differences in interpretation among experts. Newly qualified doctors and midwives are more prone to making mistakes and incorrect decisions. To overcome the differences in the interpretation of Non-Stress Test analysis and to automate the process to minimize diagnostic errors, machine learning and deep learning models have been increasingly used in recent years. In this study, the literature of the past five years is researched, and numerical expressions, tables, and graphs related to this are presented.

Keywords: Non-stress test, machine learning, deep learning, fetal heart rate, uterine contractions

ÖZET

Fetal Sağlık Teşhisinde Yapay Zekâ: Sistematik İnceleme

Doğum uzmanları, doğum öncesi dönemde ve doğum esnasında fetal iyilik halini değerlendirmek için fetal kalp atım hızı ile annenin uterus kasılmalarını ölçen Non-Stress Test uygulamasını yaygın olarak kullanırlar. Non-Stress Test, fetal sıkıntıyı erken teşhis etmek için de kullanılmaktadır. Erken teşhis ile yapılacak tedavi, fetüsün yaşama bağlılığını ve yaşam kalitesini olumlu yönde arttırmaktadır. Non-Stress Test ile elde edilen fetal kalp atım hızı ve uterus kasılma sinyalleri trase adı verilen kâğıda aktarılır. Doğum uzmanları traseye bakarak fetüsün durumu hakkında yorumda bulunurlar. Buna karşın Non-Stress Test'inin geleneksel yöntemlerle analiz edilmesi zaman almaktadır. Ayrıca analizin yorumlanmasında uzmanlar arasında farklılıklar olmaktadır. Özellikle görevlerine yeni başlamış olan doktorlar ve ebeler, hata yapmaya ve yanlış kararlar vermeye daha yatkınlardır. Bu amaçla Non-Stress Test analizlerinin yorumlanmasındaki farklılıkların üstesinden gelmek ve bu görevi otomatikleştirerek tanılama hatalarının oranlarını en aza indirgemek amacıyla son yıllarda makine öğrenme ve derin öğrenme modelleri sıklıkla kullanılmaya başlanmıştır. Bu çalışmamızda son beş yıla ait literatür araştırılarak bunlara ait sayısal ifadeler, tablolar ve grafikler gösterilmiştir.

Anahtar kelimeler: Non-stress test, makine öğrenme, derin öğrenme, fetal kalp hızı, uterus kasılmaları

Responsible Author

Adem Kuzu

Department of Software Engineering, Fırat University Faculty of Technology, Elazığ-Türkiye

e-mail: ademkuzu@gmail.com

Received: 23.04.2023

Accepted: 22.11.2023

Available Online Date: 23.12.2023

INTRODUCTION

The period from the 22nd week before birth to covering the first week after birth, including the delivery process, is called the perinatal period (1). This period is a very risky process for the pregnant woman, fetus and the newborn, and infant mortality during this period is frequently observed and increased in high-risk pregnancies (2). Fetal deaths that occur during the perinatal period are referred to as perinatal mortality (PM) (3). Deaths that occur after the 20th week of pregnancy and fetal deaths between 400-500 grams are also considered within the scope of PM (4). Additionally, abandoned babies and cases of infanticide also fall within the scope of PM (5).

During the perinatal period, a number of tests are used to monitor and assess the health of both the mother and the fetus. These tests allow early detection of possible problems in the fetus and the necessary interventions to ensure the healthy development of the fetus.

In the perinatal period before delivery, when there are no uterine contractions and no stress on the fetus, Non-Stress Test (NST) is widely used to assess fetal health (6). The NST provides information on fetal health for a short period of time and is frequently repeated (7). As shown in Figure 1, fetal heart rate (FHR) in section A and uterine contraction (UC) signals in section B are monitored with NST. As shown in Figure 2, they are recorded on a piece of paper called a tracing.

The upper part of the chart contains FHR time series values ranging from 30 to 240. The lower part contains UC time series values, which represent the number of contractions in a 10-minute period and range from 0 to 99 (8). In addition to UC, there are three types of FHR decelerations (slowing) that can occur, including acceleration (speeding up) and early, late and variable decelerations (9). When changes in UC pressure and FHR are interpreted together during the NST process, fetal health can be described as either normal (N) or pathological (P), or as N, suspicious (S) or P (10).

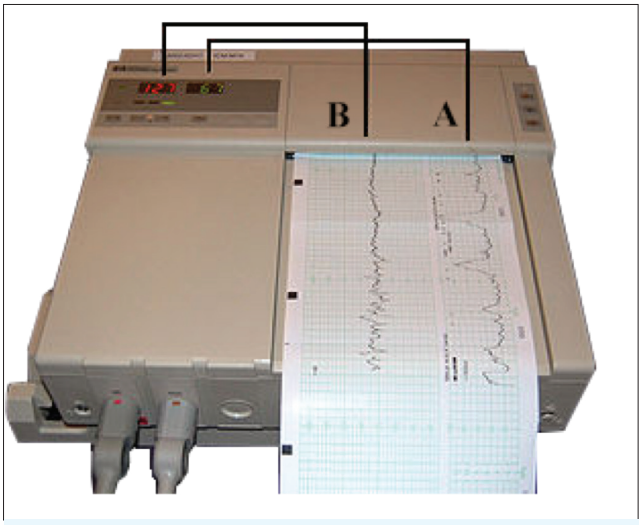

Figure 1. NST device and tracing paper.

As shown in Figure 3a, the reference value of FHR for a normal fetus is considered to be between 110 and 160, with bradycardia occurring if it is below 110 and tachycardia if it is above 160, as shown in Figure 3b (11). An FHR between 160 and 180 beats per minute is considered mild tachycardia while a heart rate above 180 beats per minute is considered severe tachycardia (12,13). Fetal tachycardia at a rate of 160 beats per minute is observed with a frequency of 0.4% (14). An FHR of less than 100 bpm is considered fetal distress (15). If fetal tachycardia persists, it leads to fetal distress and possible fetal demise.

PM affects the level of development of countries (16). The rate of PM is lower in developed countries, around 8-10 per thousand or less, while it is higher in developing countries, around 30-40 per thousand or even more. In 2020 this rate was 0.22% in Germany, 0.27% in the UK, 0.35% in Romania and 0.5% in Türkiye. In contrast, the rate was 3.52% in Afghanistan, 4.04% in Pakistan and an average of 2.7% in Sub-Saharan Africa, as shown in Figure 4 (17).

Obstetricians try to make decisions about fetal health by looking at the tracing paper obtained with NST. Due to the dynamics associated with FHR, it is almost impossible to make a reliable visual interpretation, and there can be inter-observer variability. In addition, medical devices generally support the printing of images on paper, which makes in-depth analysis and data processing difficult (18). Although medical professionals have attempted to create an automated interpretation of NST, they are still unable to make decisions, especially about suspicious fetal conditions, and make accurate predictions (19). In the last decade, computer-aided diagnostic systems based on machine learning (ML) techniques have been developed to support medical decisions (20). Thus, in clinical applications, there is an increasing number of studies that will help specialists by accurately interpreting NST predictions and reducing errors caused by subjective judgements (21). These studies will also help countries achieve PM rates at the level of developed countries.

This study discusses the studies conducted on ML and deep learning (DL) models to classify fetal health information from FHR and UC data as N, S and P. The studies conducted on ML and DL models analyse the errors that can occur in predicting fetal health and the extent to which early

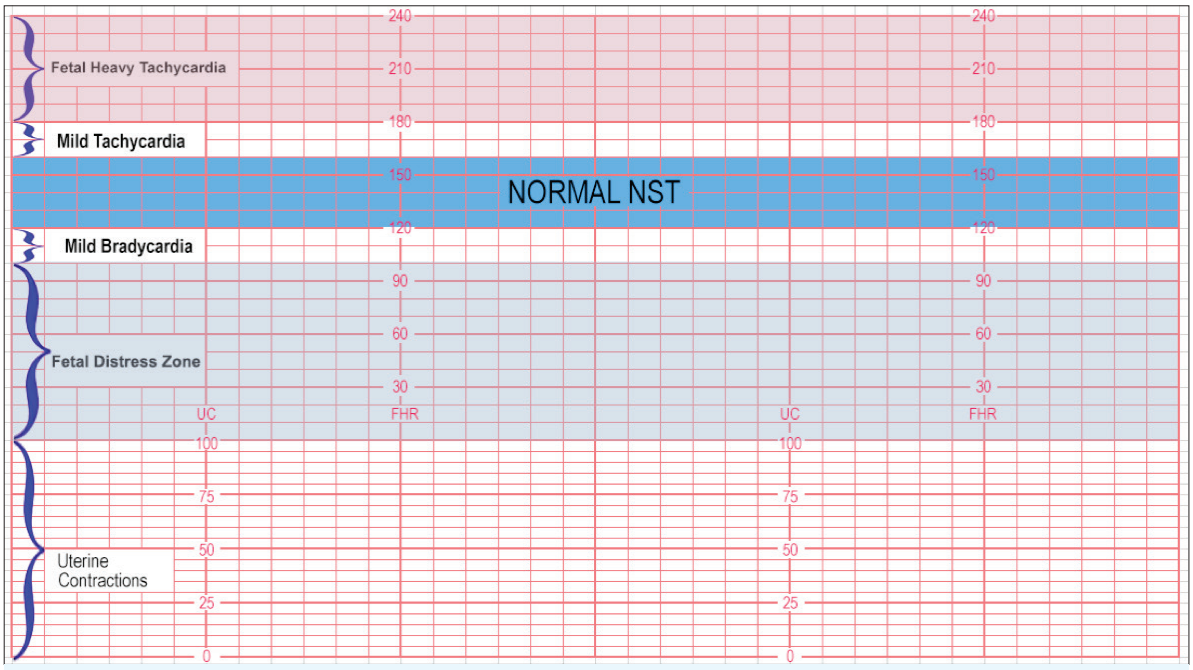

Figure 2. Trace paper on which FKH and UK signals are transferred.

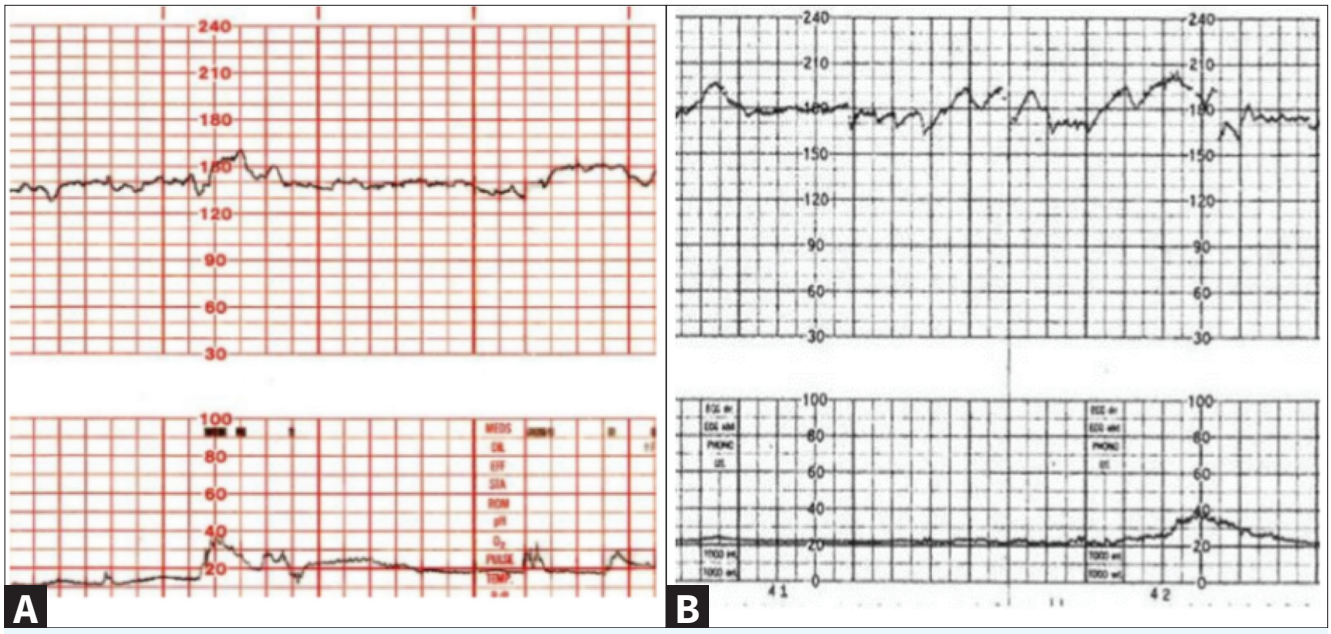

Figure 3. Normal NST and fetal tachycardia. **A.** Normal NST, **B.** Fetal tachycardia NST.

diagnosis is useful. The models used in the analysis, the methods used, and their accuracy rates are described individually with numerical expressions. A summary of the literature on previous studies is presented in the following section, followed by an analysis of the experimental studies and the results obtained in the following sections.

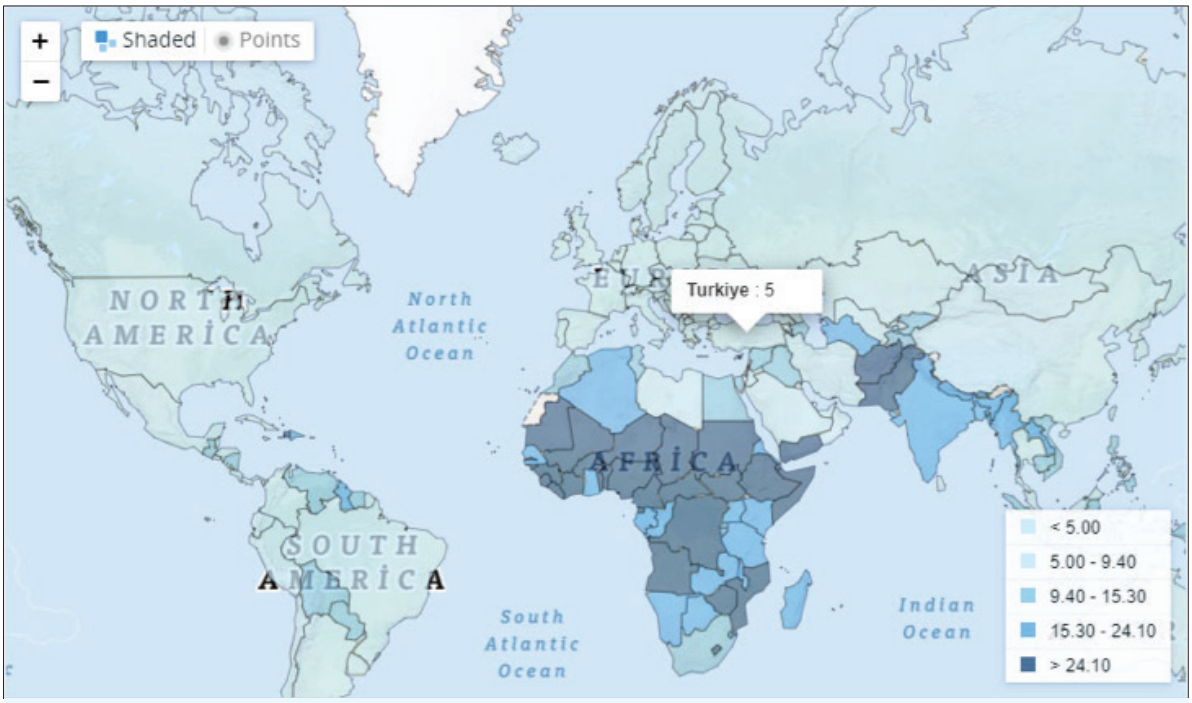

Figure 4. Mortality rate, newborn (per 1000 live births) (WHO).

Literature Review

Kuzu et al. proposed a batch learning-based prediction method for the classification of fetal health (normal, suspicious, pathological) using fetal heart rate acceleration obtained from a cardiotocography (CTG) dataset and NST tests in their study (22). In this research, they developed binary and multiclass classification models using LR, RF, and extreme gradient boosting (XGBoost) algorithms, and applied these models to the UCI dataset. They specified that the examined dataset consists of a total of 22 columns and 2126 data points. In their experimental work, they created a hybrid network model and conducted 10-fold cross-validation tests using RF for bagging, XGBoost for boosting, and LR for stacking. They employed accuracy, F_1 score, precision, and recall metrics for model evaluation. As a result of their studies, they have achieved 100% accuracy with XGBoost.

Hephzibah et al. aimed to predict fetal status, a critical factor in preventing fetal deaths (23). They used a common dataset containing 23 features and 2126 samples. They applied a novel ensemble classifier consisting of XGBoost and random forest (RF) classifiers using the widely used CTG technique to monitor fetal heart rate and uterine contractions. The results obtained showed that the ensemble classifier exhibited a high accuracy level of 96%, along with precision and F_1 score performance, outperforming individual XGBoost and RF classifiers. They found that this classifier effectively improved the identification of true positives while minimising false positives and negatives. They also highlighted the effectiveness of their approach compared to other classifiers when combined with adaboost and gradient boosting (GB) techniques.

Cao et al. investigated the use of CTG, a technique used to monitor fetal heart rate and uterine contraction signals during pregnancy, to evaluate antenatal intrauterine monitoring for assessing fetal intrauterine safety status and reducing perinatal morbidity and mortality (24). They highlighted the importance of CTG in detecting severe asphyxia, a leading cause of neonatal death and disability, which can result in neural damage and sequelae. The study used a University of California Irvine (UCI) dataset of 2126 samples to evaluate the performance of different ML methods in CTG classification and to support clinicians' clinical judgement. Their experiments showed that XGBoost achieved the highest accuracy (91%) while other models also showed good accuracy (ranging from 83% to 90%). In addition, XGBoost had the best precision, recall and F_1 score results. Based on these results, the study concluded that ML methods can be applied to CTG data to assess fetal health status, providing clinicians with an objective assessment tool.

Kaliappan et al. aimed to improve and determine the best performing algorithm among ML models, including decision tree (DT), RF, support vector machines (SVM), K-nearest neighbor (KNN), Gaussian Naïve Bayes, adaboost, GB, voting classifier (VC) and neural networks (NN), by applying various cross-validation (CV) techniques such as K-fold, Hold-Out, Leave-One-Out, Leave-P-Out, Monte Carlo, Stratified K-fold and Repeated K-fold (25). They used 22 features related to fetal heart rate obtained from the clinical CTG in 2126 patients. In addition, exploratory data analysis (EDA) was performed to gain detailed insight into the features. Using CV techniques with GB and voting classifier models, they achieved 99% accuracy.

Sheakh et al. aimed to present a risk factor analysis using ML approaches to reduce maternal and fetal mortality (26). They used datasets containing a total of 2126 and 1488 records of pregnant women in the third trimester with a total of 21 features. In this study, they evaluated various ML algorithms such as SVM, logistic regression (LR), Naive Bayes, DT, RF, KNN and XGBoost. To evaluate the performance of different classification algorithms, they used accuracy, precision and recall metrics. As a result, it was observed that the RF algorithm achieved the highest accuracy rate of 99.98% compared to all other algorithms.

Akmal et al. proposed an ML model based on feature extraction [autoencoder (AE)], feature selection [recursive feature elimination (RFE)] and Bayesian optimisation to diagnose and classify different fetal conditions (N, S, P) (27). They evaluated the model on a publicly available CTG dataset and highlighted its potential application as a decision support tool for pregnancy management. The proposed model achieved high performance metrics, particularly when used in conjunction with RF, achieving an accuracy of 96.62% for fetal status classification and 94.96% for CTG morphological pattern classification. They highlighted that the proposed model accurately predicted 98% of suspect cases and 98.6% of P cases in the dataset.

Mehbodniya et al. aimed to predict fetal health from CTG data using ML algorithms (28). They evaluated the impact of various factors on CTG data and used algorithms such as SVM, RF, multilayer perceptrons (MLP) and KNN. They reported that the dataset consisted of 21 fetal health monitoring features. The performance of the models was assessed using accuracy, precision, recall, F_1 score and support metrics. In the RF algorithm, they achieved higher accuracy, precision, recall and F_1 score compared to others. They also found that the second-best performing algorithm was SVM with an accuracy of 93% and the same F_1 score.

Hardalaç et al. based on the literature research, aimed to classify a CTG dataset consisting of 2126 records and 21 features divided into three classes, namely 1655 N, 295 S and 176 P, using different algorithmic methods based on literature research, with the highest accuracy in three stages (29). As they misinterpreted 26.4% of the S conditions in the whole dataset, they reduced the dataset to 17 features, reducing the misinterpretation rate to 15.8%. When they fine-tuned the dataset using Bayesian optimisation, they reduced the misprediction rate of S conditions by 60.22% in multiple training and testing conditions. They correctly predicted 98.8% of N conditions and 100% of P conditions with the proposed model. They stated that they could correctly predict N conditions without hyperparameter optimization, but used feature elimination and hyperparameter optimisation to predict S and P conditions with high accuracy. They obtained the highest accuracy of 97.20% from the RF classifier in their experiments.

Hussain et al. aimed to classify CTG data with high accuracy by combining the AlexNet architecture with SVM and compared their hybrid model with RF, GoogleNet, DenseNet and NiftyNet algorithms (30). In their study, they achieved higher reliability compared to other models in terms of 99.72% accuracy, 96.67% precision and 99.6% specificity metrics obtained with their proposed AlexNet SVM. In addition, they stated that their proposed model showed promising results in terms of time computation and classification accuracy in CTG data.

Chaturvedi et al. analysed the performance of different algorithms in classifying fetal well-being (31). They divided the dataset of 2126 samples into 0.75 and 0.25 proportions for training and testing, with 1655 classified as N, 295 as S and 176 as P. In their experiments, they achieved accuracy rates of 86% with RF, 97% with XGBoost, 96% with KNN and 99% with SVM. They found that SVM and XGB outperformed the other algorithms.

Piri et al. aimed to model a clinical decision support system to assist obstetricians in distinguishing N cases from S and P cases using information extracted from FHR tracings (32). Due to the quantitative nature of the variables, they used the multi-objective differential evolution algorithm for mining numerical association rules (MODENAR), the multi-objective genetic algorithm (MOGA) and the multi-objective evolutionary algorithm, namely the non-dominated sorting genetic algorithm-II with crowding distance (QAR-CIP-NSGA-II), which they applied for the first time to the relational analysis of FHR tracings. They explained that appropriate intervals of numerical features were automatically adjusted,

so there was no need for discretisation as in the other techniques. After analysing all the FHR data, they found that MODENAR was superior to MOGA in terms of average confidence, average intelligibility, average coverage, average recall and average precision factor. They mentioned that, according to the results of their experiments, MODENAR outperformed multi-target rule sets in terms of basic performance indicators.

Rahmayanti et al. aimed to compare the performance of artificial neural network (ANN), long short-term memory (LSTM), adaptive neuro fuzzy inference system (ANFIS), DT, K-nearest neighbour (KNN), light GB machine (LGMB) and RF algorithms in accurately predicting high-risk fetuses in three stages (33). They also investigated the effect of pre-processing the data prior to modelling. They reported achieving accuracy rates ranging from 89% to 99% with XGB, SVM, KNN, LGBM and RF algorithms. They stated that they consistently achieved the best performance with LGMB in all three scenarios based on their experiments.

Tokmak et al. investigated the effects of dimensionality reduction techniques, including principal component analysis (PCA), autoencoder (AE) and stacked autoencoder (SAE), on ML methods using antenatal CTG data used to determine fetal condition (34). They trained RF, Naïve Bayes (NB), SVM and deep neural networks (DNN) classifiers by addressing the imbalance of the dataset with different techniques. They found that they achieved similar results with RF and DNN in classification without using symmetric minority over-sampling technique (SMOTE) on the original dataset, but PCA was more successful in dimensionality reduction. They achieved the best accuracy rate of 95.2% with the PCA model of the DNN classifier. After the SMOTE algorithm method, they achieved F_1 score, precision and recall rates of 95.2%, 95.3% and 95.2% respectively. In the results of their experiments, they stated that the performance of the classifiers was higher when using SMOTE and that the DNN model performed better than the other algorithm they used in their experiments.

Spairani et al. conducted an experimental study on a hybrid approach to classify N and P fetuses by handling heterogeneous data (20). They binary categorised each entry of the dataset as 0 for N cases and 1 for P cases. They removed the intervals and corruptions as signal loss in the dataset. They applied preprocessing steps in MATLAB 2021a. They compared the performance obtained separately with MLP and convolutional neural networks (CNN) with the performance obtained with the CNN-MLP hybrid model. According to the results of their experiments, they found that their proposed CNN-MLP model was the best performing architecture with a performance of 80.1%.

Das et al. aimed to develop a robust classification model that can identify the fetal status during the first and second stages of labour (35). They used SVM, RF, MLP and two-stage bagging classifiers in their experiments. They used the dataset from the Czech Technical University and Brno University Hospital, which consisted of 552 records collected using the OB TraceVue system between 27 April 2012 and 6 August 2012. They found that the classification performance they obtained using the SVM and RF methods in the first and second stages of labour was higher than that obtained using the MLP and bagging methods. For suspicious cases, they obtained accuracy rates of 97.4% and 98% for the SVM and RF algorithms, respectively, along with sensitivity rates of 96.4% and specificity rates of 98%. Based on their experimental results, they concluded that the proposed classification model is efficient and can be integrated into a decision support system.

Sharma et al. studied the enhanced binary bat algorithm (EBBA) on a CTG dataset consisting of 2126 fetal recordings (36). They compared the accuracy of EBBA with the quantum grey wolf optimisation (qGWO) and genetic algorithm (GA) methods. They stated that they achieved the highest accuracy rates by extracting the minimum number of features possible. They mentioned that the EBBA, qGWO and GA algorithms selected 11, 15 and 12 features respectively. As a result of their study, they stated that their proposed EBBA classifier achieved an accuracy rate of 96.21% and outperformed the GA and qGWO algorithms in terms of optimised feature selection.

Aslam et al. used the RF, SVM, KNN and GB algorithms to predict whether the fetus is healthy or not (37). They used two datasets with different time intervals in their experiments. They performed feature selection using RFE and grid search techniques. In their experimental results, they obtained the highest accuracy results in all measurements from the RF algorithm. They also noted that the proposed model would help doctors predict whether the fetus is healthy and provide the necessary medical treatment at the appropriate time to protect the fetus' health.

Jebadurai et al. investigated the impact of filter-based feature selection techniques on classification methods, as well as the analysis of correlation-based filtering techniques based on Pearson, Spearman and Kendall methods (38). They also conducted studies using statistical filtering techniques such as mutual information, chi-square, analysis of variance (ANOVA) and receiver operating characteristic

(ROC) - area under the curve (AUC). They observed a 3% improvement in the performance of Gaussian Naive Bayes (GNB) and KEYK in statistical feature selection techniques, and a 4% improvement in the performance of decision trees (DT) and SVM in correlation-based techniques. They also found that the statistical techniques of VARAN and ROC-AUC increased accuracy by up to 92%, and similarly Spearman correlation provided improved performance measures compared to other correlation techniques. In conclusion, they stated that the accuracy of GNB classification increases when statistical feature selection techniques are applied but remains unchanged in correlation-based filtering techniques.

Aslam et al. reviewed the results and analyses of several ML models for classifying fetal health status (39). To focus on the diagnosis of prenatal risks, they used the CTG dataset obtained from the UCI ML database. The dataset was provided by the Biomedical Engineering Institute of Portugal and the Faculty of Medicine of the University of Porto in September 2010. They stated that there were no missing attributes in the dataset, which was obtained periodically from 1980 to 1995 and from 1995 to 1998, and that the class distribution consisted of a total of 2126 records, with 1655 N, 295 S and 176 P. They divided the dataset into 0.77 training and 0.33 test sizes and used stratified sampling. They used RF, LR, DT, SVM classifiers, voting classifier (VC) and KNN methods for classification. As a result of their experiments, they achieved an accuracy rate of 97.51% with the RF model.

Singh et al. aimed to determine the best classification model for predicting fetal health and early diagnosis by comparing different classifier models (40). They used the CTG dataset consisting of 2126 data samples classified into three classes: N, S and P. In their experiments, they evaluated each model for accuracy, sensitivity, precision and F_1 score metrics for a similar dataset and found that the XGB and LGBM classifier models showed a high accuracy rate of 95.14%.

Dadario et al. conducted a statistical study on a dataset to reduce the risk of maternal and fetal death (41). They used a CTG dataset consisting of 2126 records represented by 21 features and labelled with 1655 N, 295 S or 196 P case numbers. They fitted the dataset using Gaussian process regression and created a post-processed LGBM model using CV. They evaluated the performance of the model using the area under the sensitivity specificity curve metric. According to their experiments, they found that the best model was the CV group LGBM model, which provided an accuracy rate of 95.82%.

Manikandan et al. proposed methods for predicting infant mortality in the early stages of pregnancy (42). They used ML models such as DT, NB, RF and KNN to classify the CTG dataset consisting of 2126 records into N, S and P cases. They investigated the techniques and effectiveness of basic classifiers using community learning methods such as bagging and boosting. Based on their experiments, they reported that the RF classifier classified the dataset into N, S and P with 96.617% accuracy.

Dutta et al. aimed to investigate the accuracy of ML algorithm techniques in identifying high-risk fetuses from CTG data (43). They used the CTG dataset of 2126 pregnant women, consisting of 78% N, 14% S and 8% P classes, available in the UCI ML repository. They applied SMOTE to improve the imbalance of the dataset. They trained DT, SVM, KNN, RF and linear SVM methods on the CTG dataset. They attempted to obtain the sensitivity, precision and F_1 score for each class and the overall accuracy of each model in predicting N, S and P fetal conditions. They used the Matthews correlation coefficient (MCC) and Cohen's kappa (k) statistical parameters for model validation. Based on their experiments, they achieved the best accuracy rate of 98.01% using the SMOTE-based RF model.

Fasihi et al. have proposed a one-dimensional convolutional neural network (1-D CNN) that reduces computational complexity to improve the accuracy of fetal state detection (44). To evaluate the performance of their proposed architecture, they used four different datasets obtained from the UCI ML repository, including a CTG dataset with three fetal states: N, S and P. In their initial experiments, they used three different architectures: 1-D CNN, 1-D CNNI, and their proposed 1-D CNNII. In their second experiment, they compared the accuracy of CNNII with DT, ANN, LR, SVM, KNN and deep belief network (DBN) classifiers. They reported that 1-D CNNII was more efficient than other classifiers and achieved higher accuracy in fetal condition assessment compared to previous studies.

Pradhan et al. aimed to investigate how well ML models perform in predicting fetal health using CTG dataset (45). They used different classifiers such as LR, KNN, RF and GB and evaluated their performance in terms of accuracy, precision, recall and F_1 score. In their experimental results, they found that the RF algorithm achieved the highest accuracy of 99% among the classifiers.

Piri et al. aimed to improve the classification effectiveness of SVM, RF, DT and KNN methods on imbalanced CTG dataset and to find critical features that affect fetal health by using preprocessing techniques such as feature selection and

balancing (46). They stated that the total number of records in the CTG dataset they used was 2126, of which 1655 were N, 295 were S and 176 were P cases. They identified the dataset as unbalanced due to the number of cases and used the SMOTE oversampling technique. From the experimental results, they found that the performance of the balanced dataset improved compared to the unbalanced dataset.

Feng et al. proposed a stacked model based on XGB and RF feature selection to assist obstetricians in CTG interpretation, increase diagnostic accuracy, and conserve medical resources (47). They stated that the CTG dataset contained 21 structural features extracted from 2126 records, and the number of samples with N, S and P fetal conditions were 1655, 295 and 176, respectively. They applied stacking fusion to unbalanced data to produce a robust model. In their experimental results, they measured 96.08% accuracy, 93.36% F_1 score and 0.9883 AUC under the curve value.

Chen et al. proposed a method to improve the classification accuracy of FHR data using the deep forest (DF) algorithm and to assist obstetricians in clinical decision making (48). They used a publicly available FHR dataset from the UCI ML database. The dataset consisted of 2126 cases with 1655 N, 295 S and 176 P samples and 21 features. They performed deep iterations with basic classifiers such as RF, weighted RF (WRF), totally RF (TRF) and gradient boosted decision trees (GBDT). When they compared the results, they measured the accuracy, average F_1 score and AUC values, which were 92.64%, 92.01% and 0.990 respectively. They found that the best results among all the models compared were 91.64%, 88.92% and 0.9493 respectively. They concluded that their proposed model has good classification results, which are essential for clinical decision making, healthy fetal development and safe delivery for pregnant women.

Kasım attempted to classify a CTG dataset as benign and malignant with N, S and P using the extreme learning machine (ELM) algorithm (49). He used the publicly available CTG dataset from the UCI ML database for his proposed method. He evaluated the performance of the method using accuracy, F_1 score, Cohen kappa, precision, and sensitivity metrics. The experiments resulted in a binary classification accuracy of 99.29% and a multi-classification accuracy of 98.12%. He stated that as a result of his studies, high classification accuracy can be achieved with both binary and multi-classification analysis of the CTG dataset.

Dwivedi et al. proposed the LGBM algorithm to classify fetal health (50). They stated that the dataset they used was categorised as N, S and P and consisted of 2126 samples with 21 features. They balanced the dataset using the SMOTE technique. As a result of their experiments, they obtained accuracy of 0.9561, sensitivity of 0.9056, Cohen kappa of 0.8792, precision of 0.9552, AUC of 0.9864, F_1 score of 0.9550 and MCC of 0.8805 with a model processing time of 2 minutes.

Li et al. applied twelve ML models individually to the CTG dataset (51). They used the soft voting integration method to integrate the best four models to create a blender model and compared it with the stacking integration method. They stated that their dataset consisted of a total of 2126 CTG examples classified by experts, including N, S and P. They used accuracy, precision, sensitivity, F_1 score and Cohen kappa metrics to measure performance. As a result of their experiments, they achieved an accuracy rate of 95.9%, an AUC of 0.988, a sensitivity of 0.916, a precision of 0.959, an F_1 score of 0.958 and a Cohen kappa value of 0.886 with the blender model.

Haweel et al. worked on a probabilistic neural network (PNN) based method for fetal condition classification (52). They compared the performance of their proposed PNN classifier with legendre neural network (LNN) and volterra neural network (VNN) classifiers. They stated that the CTG dataset they used consisted of 21 features and 2126 records. They presented three initial states with 21 and 10 features to the PNN classifier. As a result of their experiments, they reported that the PNN classifier outperformed the LNN and VNN classifiers in terms of mean square error, overall classification accuracy, computation time and computational complexity. They achieved an overall accuracy of 99.74% for their proposed PNN classifier.

Avuçlu worked on a hybrid model consisting of KNN, DT, NB and SVM algorithms to predict fetal status from CTG recordings obtained from FHR and UC signals (53). They automatically processed 2126 fetal CTG recordings and recorded the necessary diagnostic features. According to their results, they improved the accuracy of diagnosis by 34% using the hybrid model they created. They achieved success rates of 98.3925% and 94.4175% in the training and test datasets respectively. In their experiments, they achieved 100% classification accuracy, sensitivity and specificity for NB and DT ML algorithms.

Bhowmik et al. aimed to analyse pre-delivery CTG dataset and develop an efficient tree-based ensemble learning classifier model to predict fetal health status (54). They adopted and developed the stacking approach and tried to apply different machine learning algorithm (MLA) techniques to the CTG dataset and determine their performance. They stated that the dataset they used is a publicly available standard dataset from the UCI ML database, labelled by

three professional obstetricians. They explained that the dataset consists of 21 features and has 2 target variables for classifying fetal health status, one for patterns (1-10) and the other for N, S and P. They stated that the dataset contains 2126 observations, including 1655, 295 and 176 samples belonging to the N, S and P classes, respectively, and that it is an unbalanced dataset since the N class contains 77.85% of all examples. They applied 10-fold CV. As a result of their experiments, they achieved an accuracy of 96.05% with their proposed model.

Rayhana et al. aimed to improve the diagnostic accuracy of obstetricians in interpreting fetal heart rate signals by automating the process (55). They used five different ML models and the UCI-CTG dataset. They used accuracy, sensitivity, precision and F_1 score as performance measures. They generated ten models for comparison. They found that XGB with all features and RF with selected features performed better than other methods. They achieved 96.7% accuracy and 0.963 F_1 score with the XGB all features model and 95.6% accuracy and 0.963 F_1 score with the RF model. They stated that the F_1 scores in the P class were very high and realistic at 0.963 for both models.

Jayashree et al. aimed to analyse the automatic prediction of fetal health from the UCI CTG dataset and improve the fetal risk prediction rate using an optimised technique such as GA-DVM (56). They automatically processed the 2126 fetal CTG dataset and calculated the relevant diagnostic features. They found that the CTG dataset was classified both morphologically and into N, S and P, so they could use it for either 10-class or 3-class experiments. In the experiments, they found that the features selected by the optimised GA-SVM provided higher accuracy than those selected by GA, and that SVM outperformed NB, RF and MLP techniques.

Kannan et al. aimed to explore uncertain information in the CTG dataset and evaluate the performance of classifiers based on rules, trees, and functions to classify CTG data into N, S, P categories (57). They used a dataset consisting of 2126 samples with 23 features and applied particle swarm optimisation (PSO) in preprocessing. They evaluated the performance of the classifiers using accuracy, sensitivity, F_1 score, precision, and ROC evaluation parameters in the WEKA program. They obtained the highest accuracy of 99.57% with the RF classifier and found that a tree-based approach performed better than other approaches.

Marvin et al. aimed to analyse and classify CTG signals consisting of 2126 real data with low uncertainty and high accuracy (58). They applied GB, LGBM, DT and categorical boosting models by extracting their features on the dataset. As a result of their experiments, they obtained 99%, 100% and 97% accuracy, precision and recall rates respectively with the LGBM classification model.

Amin et al. aimed to develop a good and efficient classifier to assist physicians in diagnosing FHR using MLA (59). They explained that their proposed fuzzy diagnostic method does not only improve the performance of simple neural networks but also outperforms other algorithms. Using the WEKA application, they distributed the UCI-CTG dataSET with 2126 samples and 21 input features in their experiments. They also visualised the results using a box plot. As a result of their experiments, they observed that they achieved higher performance than other models with 95.1% accuracy, 94.95% precision, 95.2% recall and 95.1% F_1 score in classifying CTG dataSET using the recursive neural network (RNN) model.

Debjani et al. aimed to analyse the presence of fetal heart disease by optimising the ELM with a new activation function (60). In their study, they used the UCI-CTG dataset with 23 original features and 2126 instances. They stated that the best features from the CTG dataset were selected using a genetic algorithm (GA). They measured and compared ELM using metrics such as accuracy, sensitivity, specificity, precision, F_1 score, AUC, and computation time. In the experiments, they achieved an accuracy rate of over 95% with ELM having sigmoid and square root activation functions.

Arif et al. attempted to predict fetal health using the classification and regression tree (CART) method, a variant of the DT algorithm (61). They used the UCI-CTG dataset containing 2126 observations with 22 attributes. After analysing the dataset, they separated it into three variables, N, S and P, and obtained 19 nodes in the classification tree, which they measured according to their weights. In their experiments, they obtained an accuracy rate of 98.7% with CART.

Nandipati et al. aimed to perform feature selection and classification on a derived dataset using R-based CARET and Python-based Scikit learning packages (62). They used the publicly available UCI-CTG dataset consisting of 2126 samples with 1655 N, 295 S and 176 P cases with 23 features. They mentioned that they reduced the dataset to 771 examples by randomly removing 300 cases from the N class due to the imbalance in the dataset, which could reduce the performance of the model. They set the data set ratio to 0.70:0.30 and performed 10-fold CV. After testing with KNN, SVM, RF, NB, MLP, bagging and boosting classification algorithms, they obtained the highest accuracy of 97.87% using the RF and NB methods.

Piri et al. focused on the evolutionary MOGA method to extract important factors causing fetal death through cardiotocographic analysis of fetal assessment (63). They used a CTG dataset consisting of 21 features and 2126 records in their study. They found that the data was unbalanced, with more N cases than S and P cases, and used MOGA for feature selection and 10-fold CV to avoid overfitting. As a result of their experiments, they obtained accuracy rates of 81%, 87%, 93%, 92%, 90%, 85% and 94% from the LR, SVM, RF, DT, KNN, GNB and XGB models respectively when using the reduced dataset.

Das et al. aimed to determine the most appropriate feature set and the most effective ML technique to accurately predict fetal status (64). They used the UCI-CTG dataset, which contains 2126 data points, each represented by 37 features. They mentioned that the dimensionality of the feature set was reduced by the physicians using various automatic methods. The resulting datasets were classified using different ML algorithms. In their experiments, when the feature set was reduced using the maximum relevance - minimum redundancy (MRMR) method, they achieved the highest accuracy of 99.91% and a Cohen kappa measure of 0.997 with the RF algorithm.

Kadhim et al. analysed a CTG dataset consisting of 2126 records with 21 features and three different groups using a NB classifier integrated with the firefly algorithm (FA) (65). They proposed the FA method to find the optimal subset of features and minimise the classification time while maximising the accuracy performance. As a result of their experiments, they achieved an accuracy rate of 86.5474% using the NB classifier with the FA algorithm.

Ramla attempted to analyse the natural structure of the data using methods such as linear discriminant analysis (LDA), RFE, forward and backward elimination, and ridge regression (66). She investigated the effect of feature selection on the KA-J48 classifier. She used the open-source UCI-CTG dataset consisting of 2126 records in her study. As a result of the experiment, she achieved a higher accuracy rate of 86.46% with the ridge regression method compared to other methods.

Avuçlu et al. aimed to develop an application for faster and more accurate interpretation of FHR results (67). They used the CTG dataset consisting of 2126 records and the NB algorithm for classification in their experiments. In their experimental results, they achieved a classification accuracy of 97.18% and a pass rate of 95.68% using the NB ML algorithm.

John et al. used different classifiers to predict fetal condition (68). They attempted to classify a dataset containing 1655 N and 176 P classes using the WEKA application. They measured the performance of the models using sensitivity, specificity and accuracy metrics. As a result of their experiments, they reported that the stacking model predicted P-fetal condition with a success rate of 98.9%.

Fei et al. proposed the FCM-ANFIS method with 1655 N, 295 S, 176 P, 2126 cases and 21 features to classify the CTG dataset, which is the hybrid model of Fuzzy C-Means (FCM) and ANFIS methods (69). They divided the dataset into a random training set and a test set according to the ratio of 0.7:0.3. They used data visualization and Spearman correlations. As a result of their experiments, they obtained an accuracy of 96.39% with the FCM-ANFIS method.

Ricciardi et al. aimed to provide an ML approach that would help doctors in the decision-making process when assessing fetal well-being (70). They used a dataset of 370 CTG recordings. They used the SMOTE technique to balance the dataset and applied CV. As a result of their experiments, they achieved an accuracy rate of 91.1% with the RF algorithm.

Islam et al. aimed to classify a CTG dataset consisting of 2126 cases including 1655 N, 295 S and 176 P cases using RF, NB and DT algorithms and compare their results (71). They achieved a performance of 82.27% with the NB method and over 90% with the DT method. In the results of their experiments, they obtained a higher accuracy rate of 95.11% based on the RF method compared to other methods, based on accuracy and root mean square error (RMSE).

Silwattananusarn et al. proposed a classifier ensemble model based on ensemble learning and feature selection to improve classification accuracy (72). They evaluated their proposed approach on the CTG dataset consisting of 23 features and 2126 samples. They used four feature selection techniques: relief, correlation based, consistency based and information gain (IG). In the tests they performed with SVM, they achieved an accuracy rate of 99.85% (66).

Kaluri et al. investigated the effect of PCA and LDA on RF, DT, SVM and NB algorithms (73). They applied dimensionality reduction techniques to the UCI-CTG dataset consisting of 2126 samples and 23 attributes. They stated that the performance of the PCA-based classifiers was better than that of the LDA-based classifiers. They also found that the DT and RF classifiers performed better than the other two algorithms, both with PCA and LDA, without using dimensionality reduction.

Bautista et al. attempted to develop an application that could accurately identify datasets for experts (74). They stated that they obtained a dataset consisting of 97 samples

and 23 features from records from a special clinical centre between 17 January 2015 and 21 February 2017. They applied DT, RF, KNN and SVM MLA to the dataset. They used accuracy, precision, recall and F_1 score metrics to measure performance. They reported that the most effective algorithm was the RF decision tree with the highest training score of 0.987 and test score of 0.900 in their experiments.

Thomas et al. aimed to analyse the performance of AE, commonly used to detect outliers in healthcare datasets, as well as 1-D SVM classification techniques and the hybrid model created using these classification techniques (75). They used two datasets, one of which was CTG. They stated that the CTG dataset consists of 10 classes according to its morphological structure and three classes according to the fetal status. They used the F_1 score metric to evaluate the performance of the models they used and proposed. They reported that in their experiments they obtained a higher F_1 score with the hybrid AE-SVM compared to AE and SVM classifiers.

Hoodbhoy et al. evaluated the accuracy of MLA techniques in identifying high-risk fetuses using a CTG dataset (76). They used a dataset consisting of 2126 records and 21 features from the UCI ML database to compare the performance of SVM, KNN, XGB, adaptive boosting, RF, LR, GNB and DT algorithms. On the training dataset, they achieved precision and recall rates of 96% with the XGB technique and 99% or higher with models built by DT and RF. Their experiments showed that the model developed using the XGB technique had the highest overall accuracy of 93% compared to other ML models.

Appaji et al. used ML methods including DT, RF and adaptive boosting to classify a dataset consisting of 2126 samples with 23 features from the UCI database (77). They also visualised the information obtained, stating that this would help doctors to treat patients. The first 22 features were considered as input variables, and the 23rd feature was considered as an output variable, which could take the values N, S or P. They normalised the dataset and divided it into training and test sets in a ratio of 0.75:0.25. In their experiments they obtained F_1 values of 96.40% with DT, 93.46% with RF and 81.34% with adaptive boosting.

Afridi et al. aimed to improve the performance of DT, KNN, LR, SVM, RF and NB classification algorithms used to predict N, S and P cases from CTG dataset (78). They stated that their CTG dataset consisted of 2126 samples classified into three fetal states and had 23 features. They used the preprocessing technique to reduce the feature set by eliminating features with lower correlation values. They conducted two separate experiments to analyse the effect of their feature selection technique by comparing the full feature dataset with the reduced feature dataset. They found that using correlation-based feature selection (CFS) negatively affected the overall performance of all classifiers except KNN and RF. They stated that using the preprocessing technique improved the performance of all classifiers except KNN and RF. They obtained higher results with NB achieving 85.88% accuracy, 94.60% precision, 85.90% recall and 89.50% F_1 score compared to other classifiers.

Piri et al. proposed a classification based on association (CBA) model for accurate prediction of fetal health status (79). They used a UCI-CTG dataset with 21 features and a total of 2126 samples, including 1665 N, 295 S and 176 P cases. They divided the dataset into two parts for training and testing, and preprocessed it with discretisation, feature selection, correlation-based classification, and CV techniques. They performed classification studies using CBA-M1 and CBA-M2 algorithms and compared their results with LR, SVM, KNN, XGB, DT, RF and GNB classifiers. They found that RF and XGB performed better than the proposed model and other classifiers, with 94% accuracy according to the results of their experiments.

Xue, in his study, used neural network (NN) and RF methods to classify fetal status by analysing the UCI CTG dataset (80). It was mentioned that the CTG consisted of 2126 fetal FHR measurements and 23 features. Since there was an imbalance in the CTG dataset, a weighted grading method was applied to the dataset. As a result of their experiments, they achieved accuracy rates of 88.84% and 91.85% in the training and test datasets respectively with the RF algorithm.

Amin et al. attempted to measure the accuracy and time consumption by classifying the UCI-CTG dataset consisting of 21 features and 2126 samples using rough neural network simulation (81). In the WEKA application, they applied the supervised learning model of the rough neural network method to the CTG dataset in three stages: preprocessing, training, and testing. In the results of their experiments, they explained that the rough neural network method classified the CTG dataset with higher accuracy than other algorithms according to the accuracy rate at the appropriate time.

Iraji aimed to design an intelligent model using DL to predict the condition of the fetus with high accuracy and low error (82). He stated that the CTG dataset he used was automatically processed by the SisPorto 2.0 program and consisted of 2126 fetal records with 21 diagnostic features, of which 1655 were N, 295 were S and 176 were P. He applied different topologies of an adaptive neuro-fuzzy inference system (MLP-ANFIS) using deep ANFIS models to his dataset.

He mentioned that the ANFIS model could predict the output using the inputs. As a result of the experiment, he achieved the highest accuracy of 96.77% with 21 features.

Okwuchi et al. aimed to classify fetal health in terms of both fetal status and morphological models using a community learning model consisting of seven different methods (83). They stated that their dataset consisted of 21 features and 2126 samples. The model consisting of GB, RF and SVM methods was superior to the other classification methods they used. In their experiments, they obtained accuracy measurements of 98.6% with GB, 97.4% with RF and 95.8% with SVM.

Potharaju et al. aimed to improve the classification accuracy of learning algorithms by applying preprocessing techniques to CTG data (84). They stated that the UCI CTG dataset used for this study consists of 2126 samples from three classes, which are 295, 1655 and 176, respectively. They addressed the imbalance in the dataset using SMOTE and created a new balanced dataset with a total of 4773 samples consisting of 1622 samples for the first class, 1655 for the second class and 1496 for the third class. They used the KNN algorithm in the SMOTE method and set the CV coefficient of the dataset to five. They explained that the classification performance on the balanced dataset was better than on the unbalanced dataset. In the results of their experiments, they achieved the highest accuracy of 99.05% with the IBk algorithm for fetal status classification using Jrip, Ridor, J48, NBStar, IBk and Kstar.

Sevani et al. proposed a statistical approach using the F_1 score-based feature selection method to overcome unbalanced data and multi-class output (85). They used the SVM classifier to implement the F_1 score method. The CTG dataset they used for the experiment consisted of 21 features and 2126 samples from three classes, N, S and P. They also tested the compatibility of the F_1 score method with other datasets. Their experiments showed that the classifier achieved an accuracy of 94.35% with 21 features and 99.91% with 8 features.

Vani conducted an experiment on a decision support system using DL-based neural networks to determine the health status of fetuses from the CTG dataset (86). They used the CTG dataset with 21 features and 2126 examples and used SVM and DNN models with a ratio of 0.7:0.3 for training and testing. With the SVM model, they obtained precision and F_1 score performance measures of 93% and 81%, respectively. With the three-layer DNN model, they stated that they achieved significantly improved performance in detecting P conditions compared to SVM, with a Gmean metric of 91% and a sensitivity metric of 89%.

Kaur et al. proposed an MLA-based perinatal hypoxia diagnosis system for larger datasets (87). In this system, they applied SVM, RF and LR models to the CTG dataset obtained from a study at the University of Porto and available in the UCI ML repository. They automatically processed the dataset consisting of 2126 fetal records. They compared SVM and RF methods with Sparks, which makes the MLA easily scalable. As a result of their experiments, they achieved 97% accuracy with Spark RF.

Alkhasawneh used the hybrid cascade forward neural network and elman neural network (HECFNN) algorithm to classify six datasets, along with the CTG dataset (88). The experimental results were analysed, and the results of elman neural network (ENN) and cascade forward neural network (CFNN) were compared. In the experimental results, Alkhasawneh reported that the accuracy of HECFNN with the CTG dataset was 99.25%, which was higher than the accuracy obtained by CFNN and ENN. Furthermore, the proposed HECFNN model was reported to produce higher accuracies when compared to other different methods found in the literature.

Bhuiyan et al. attempted to design a healthcare technology that could predict medical outcomes for any patient based on their past and current medical data (89). To achieve this, they used CTG and diabetic datasets. They found that the CTG dataset consisted of 3000 patient observations and 21 features, and after pre-processing, the patient observation consisted of 2126 samples. They divided the datasets into training and test datasets in a ratio of 0.7:0.3. They applied MLA such as ANN, RF, SVM, C5.0 and NB to both datasets. In their experiments, they achieved an accuracy rate of 93.34% with RF after applying a genetic algorithm.

Ramla et al. proposed the CART method, a variation of the DT algorithm, to predict fetal health status in high-risk pregnancies (90). They mentioned that the CTG dataset they used to apply the CART method consisted of 2126 fetal records, which were automatically processed. In their experiments, they obtained accuracy rates of 88.87% using entropy and 90.1% using the Gini index.

Deressa et al. aimed to create a model to predict fetal health by applying GA, SVM, ANN, KNN, RF and C4.5 decision tree classifiers to the UCI-CTG dataset (91). They achieved the highest accuracy with RF, which reached 99.18% accuracy in their experiments.

Uzun et al. aimed to efficiently classify the UCI-CTG dataset consisting of 23 real features and 2126 different fetal signal recordings using the ELM method and compared this

method with previous studies in the literature (92). In addition to the ELM ML method, they also used sigmoid, sinus, hardlim, tribas, radbas and tansig activation functions in their studies. In the results of their experiments, the highest accuracy rate was obtained from the PCA-14 ELM algorithm with 84.3%, depending on the number of 2000 hidden layer neurons and the Hardlim activation function. They also explained that while the accuracy values can reach 99% in 2-class and 3-class, this rate does not exceed 88% in 10-class.

Akbulut et al. in their study, tested and compared nine binary classification algorithms to predict fetal health status (93). They used 80% of a clinical dataset consisting of 96 pregnant women to train the proposed model, and 20% to test the selected model. They achieved the best performance measurements with 89.5% accuracy, 75% F_1 score and 95% AUC using the RF model.

Li et al. proposed their own model to improve the classification accuracy of FHR recordings (94). They stated that they obtained a total of 4473 data sets consisting of 3012 N, 1024 S, and 437 P as a result of collaboration with a specialty hospital. They conducted a comparative experiment and used a feature extraction method based on basic statistics to extract features of FHR. In the results of their experiments, they obtained classification accuracies of 79.66% with SVM, 85.98% with MLP and 93.24% with CNN.

Miao et al. proposed an alternative and improved artificial intelligence approach using DL-based classification models for fetal assessment (95). They stated that their proposed model consisted of a DL-based training classification and prediction model. In their experiments, they used the CTG dataset consisting of 2126 samples and 21 features. As a result of their experiments, they obtained an average of 88.02% accuracy, 84.30% sensitivity, 85.01% precision and 0.8508 F_1 score with the DNN model they developed.

Fetal Health Diagnosis with Machine/Deep Learning

Our article discusses the work done in the field of CTG between 2018 and 2023. In these studies, different datasets consisting of 2126 datasets and 23 UCI datasets and 262, 399, 370, 4473, 97 and 14000 datasets were used in XGBoost, RF, DF, LGBM, CART, NB, GNB, LR, DT, SVM, AdaBOOST, GB. Classification results are discussed using methods such as, VC, CNN, RNN. The graphical representation of the number of studies per country and their performance is explained according to the methods used. The number and performance of studies performed with ML and DL methods are given in detail. The most commonly used methods and their performance are also included. The metrics used to measure performance are shown. Studies using CV and the impact of CV on performance measurement are explained. A summary of the different literature on ML and DL classification based on fetal heart rate (FHR) using different datasets is presented in Table 1 and Table 2. The number of features and the choice of ML methods can have a significant impact on the success of NST classification. In addition, studies using CV have shown higher and more consistent accuracy rates compared to simple approaches, as it reduces bias and errors due to data scatter and fragmentation. CV has become increasingly important in recent years. In addition, the SMOTE, which facilitates the production of synthetic data, was used in 10 studies. Meanwhile, PCA was used in three studies, which allowed for a reduction in size while preserving significant variance in the data set.

The studies of different algorithms and models on the dataset consisting of 2126 records are shown in Table 1.

Studies conducted on different datasets, including 97, 262, 370, 399, and 14000 medical records, are presented in Table 2.

The number of literature studies using ML and DL models obtained from our research, by year, is shown in Figure 5. It is observed that the number of studies has increased in recent years.

Figure 6 shows the number and average of ML and DL models per year, as well as the average accuracies for each year. The average accuracies of ML methods have increased each year, depending on the models and techniques used.

Figure 7 shows the average of DL and ML by year and the average of all. It was observed that the performance of ML increased compared to DL. It has been observed that methods such as performance metrics, feature selection, feature extraction and cross validation applied to the models used are effective in obtaining clearer performances. As the number of such studies increases, it is certain that accuracy rates will increase. As ML was used in all 6 studies in 2023, the DL average was set to 0. Therefore, the average of the DL models is shown as 77.25%.

As can be seen in Figure 8, one of the reasons for the increased use of ML in recent years is that it provides results in less time than DL. This is because it breaks the data down into smaller steps, rather than end-to-end, and thus produces faster and more reliable results. In addition, the fact that the training process is faster has increased the use of ML by users.

Table 1. Studies on the CTG dataset

Ref.	Method	Number of Features
Feng et al. (2021)	RF	11
Ramla (2020)	J48-Lasso	17
Akmal et al. (2023)	RF	21
Sheakh et al. (2023)	RF	21
Chen et al. (2021)	DF	21
Alkhasawneh (2019)	HECFNN	21
Tokmak et al. (2022), Miao et al. (2018)	DNN	21
Kadhim et al. (2020)	FA-NB	21
Fei et al. (2020)	FCM-ANFIS	21
Okwuchi et al. (2019)	GrA	21
Mehbodniya et al. (2022)	RF	21
Rahmayanti et al. (2022), Dwivedi et al. (2021)	LGBM	21
Iraji (2019)	MLP- ANFIS	21
Piri et al. (2022)	MODENAR	21
Haweel et al. (2021)	PNN	21
Sharma et al. (2022)	RF	21
Hardalaç et al. (2022)	RF	21
Kaur et al. (2019)	RF	21
Piri et al. (2021)	RF	21
Bhuiyan et al. (2019)	RF	21
Amin et al. (2021)	RNN	21
Amin et al. (2019)	PNN	21
John et al. (2020), Bhowmik et al. (2021)	STACKNG	21
Sevani et al. (2019), Vani (2019)	SVM	21
Kasım (2021)	SVM, MLP, ELM	21
Piri et al. (2020), Hoodbhoy et al. (2019)	XGB	21
Piri et al. (2019)	XGB-RF	21
Kuzu et al. (2023)	XGB	22
Kaliappan et al. (2023)	GB, VC	22
Ramla et al. (2018)	CART	22
Jebadurai et al. (2022)	DT	22
Avuçlu (2021)	DT, NB	22
Dadario et al. (2021)	LGBM	22
Avuçlu et al. (2020)	NB	22
Deressa et al. (2018), Pradhan et al. (2021), Islam et al. (2020)	RF	22
Hephzibah et al. (2023)	Ensemble	23
Qingjun et al. (2023)	XGBoost	23
Hussain et al. (2022)	ALEXNET – SVM	23
Li et al. (2021)	BLENDER	23
Arif et al. (2020)	CART	23
Fasihi et al. (2021)	CNN	23

Table 1. Studies on the CTG dataset (continue)		
Ref.	**Method**	**Number of Features**
Appaji et al. (2019)	DT	23
Panda et al. (2021), Uzun et al. (2018)	ELM	23
Jayashree et al. (2021)	GA-SVM	23
Thomas et al. (2020)	HYBRID	23
Potharaju et al. (2019)	IBK	23
Marvin et al. (2021)	LGBM	23
Afridi et al. (2019)	NB	23
Kannan et al. (2021), Dutta et al. (2021), Nandipati et al. (2020), Aslam et al. (2022), Manikandan et al. (2021), Xue (2019)	RF	23
Silwattananusarn et al. (2020), Chaturvedi et al. (2022), Kaluri et al. (2020)	SVM	23
Rayhana et al. (2021)	XGB	23
Das et al. (2020)	RF	37
Singh et al. (2022)	LGBM	-

Table 2. Studies conducted on other datasets			
Ref.	**Method**	**Dataset**	**Number of Features**
N. Aslam et al. (2022)	RF	262	10
Das et al. (2022)	RF	399	11
Li et al. (2018)	ESA	4473	22
Ricciardi et al. (2020)	RF	370	17
Bautista et al. (2020)	RF	97	23
Akbulut et al. (2018)	RF	97	22
Spairani et al. (2022)	CNN-MLP	14000	15

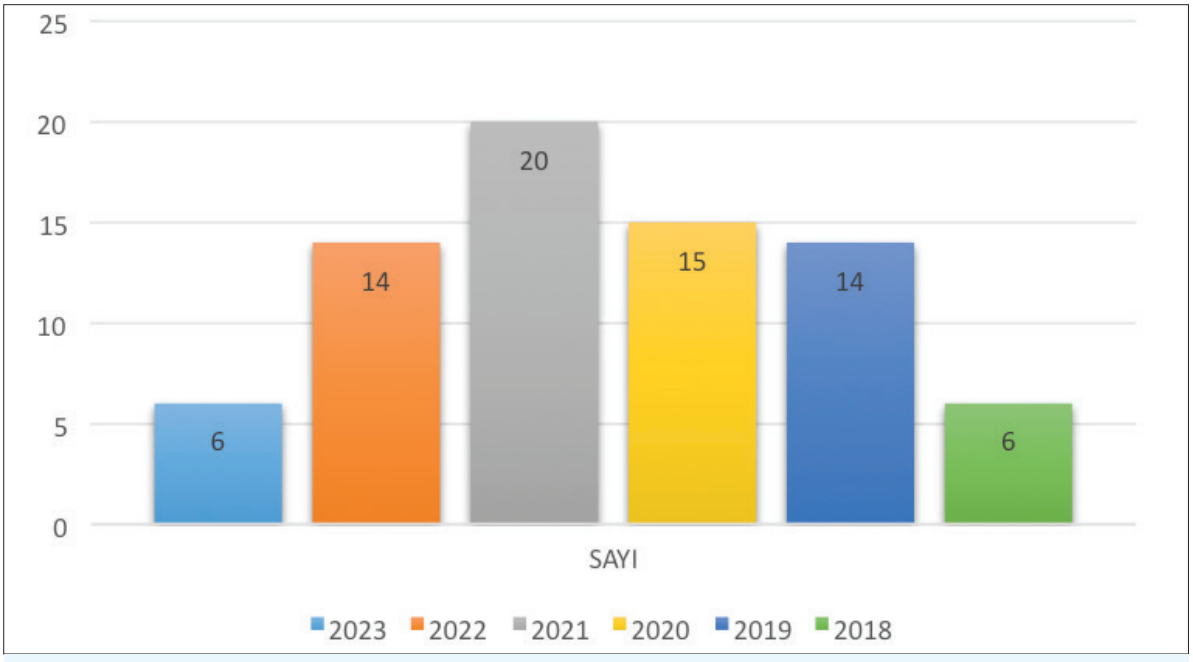

Figure 5. Distribution of literature studies by year.

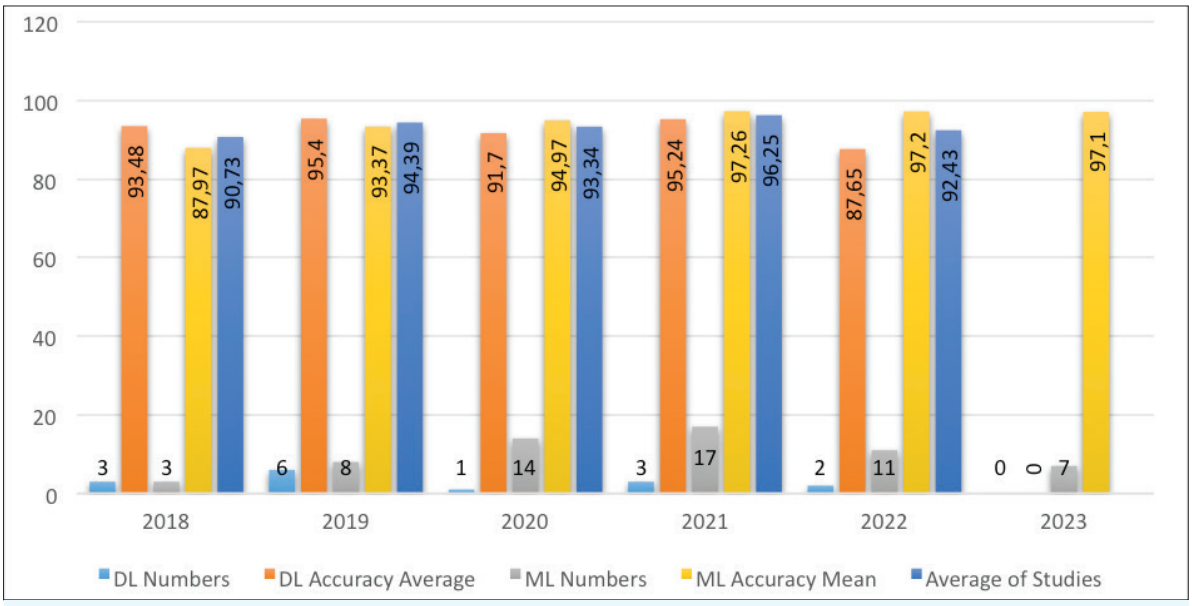

Figure 6. Average accuracy rates by year for ML and DL models.

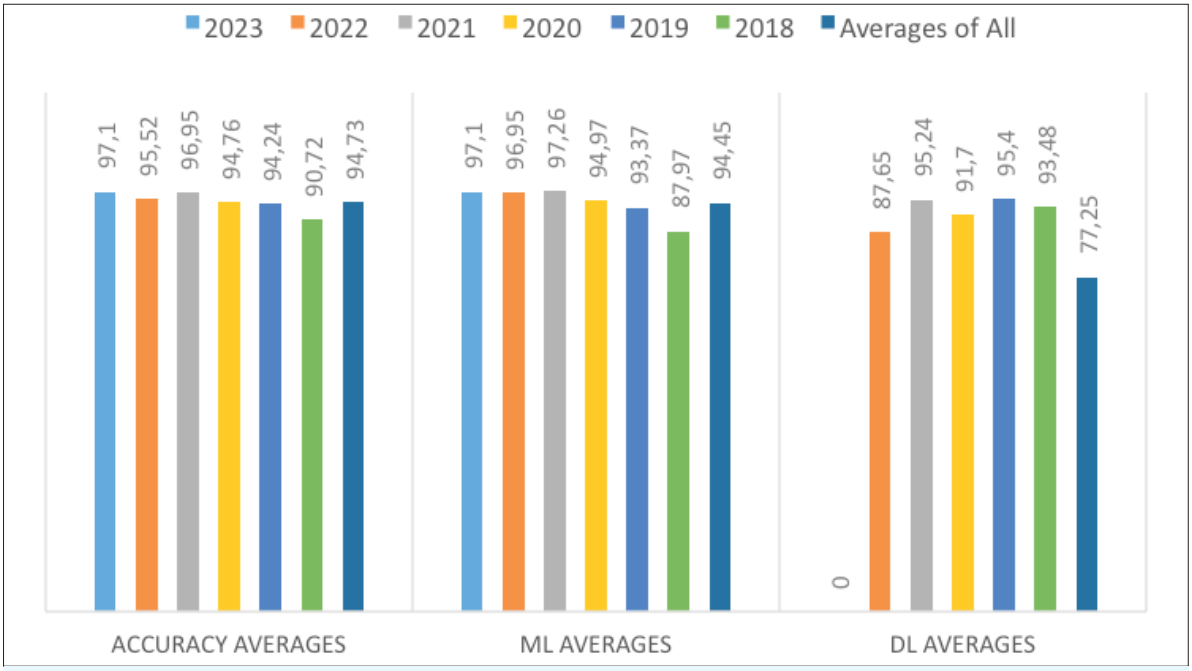

Figure 7. ML and DL counts and accuracy averages.

It was observed that in the classification studies of CTG data using ML and DL methods, India obtained an average accuracy of 95.66% with 27 studies, China obtained an average accuracy of 94.61% with eight studies, Türkiye obtained an average accuracy of 94.765% with eight studies and Bangladesh obtained an average accuracy of 97.09% with six studies. Figure 9 shows the detailed average accuracies and number of references for each country.

We found that the most commonly used classification method was the RF algorithm. However, the highest accuracy rate was achieved by Avuçlu using DT and NB algorithms with 100% (53). On the other hand, Deressa et al. achieved the highest accuracy rate of 99.18% using RF in the same feature set. Figure 10 shows the names and numbers of the models used in the studies.

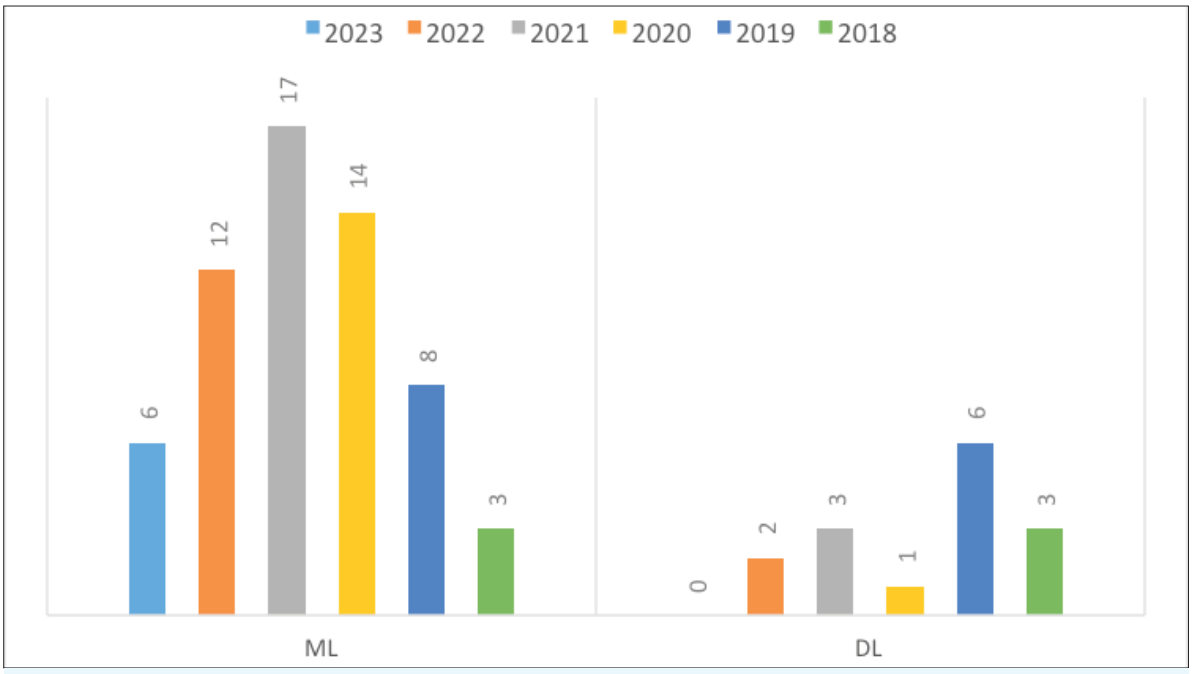

Figure 8. ML and DL numbers by year.

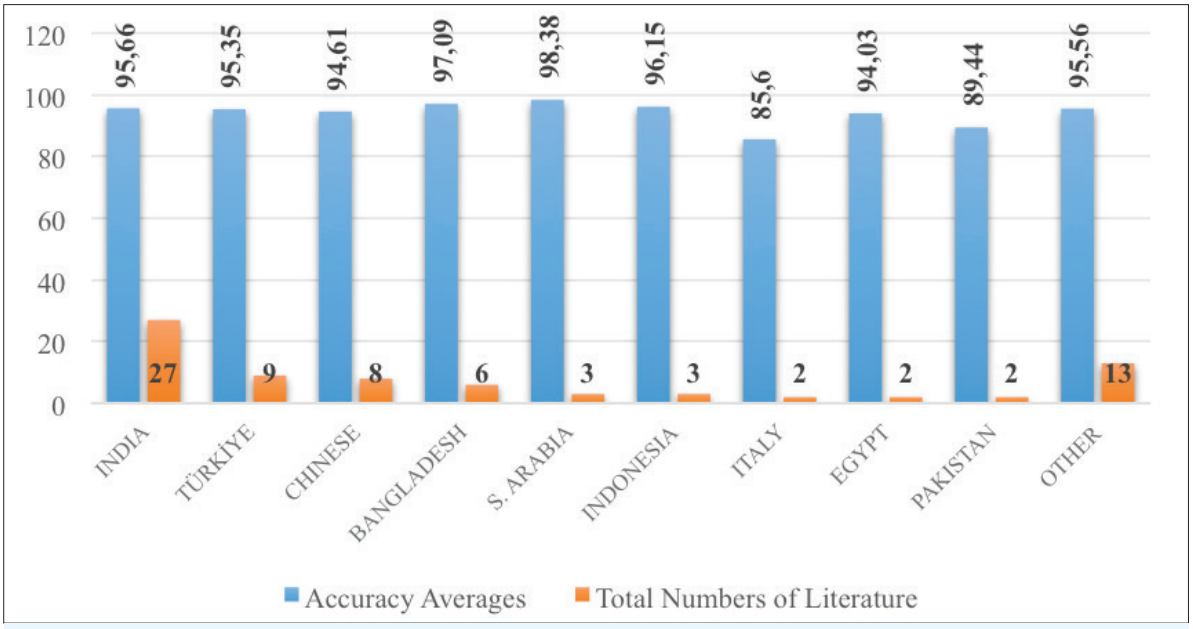

Figure 9. Literature numbers and average accuracy scores obtained by countries.

In this paper, of the 75 studies we searched, we identified 16 studies with an accuracy of 99% and above, which are shown in Table 3. The RF model was used in five studies, the SVM model in three studies, the LGBM model in two studies, the ML model in the other six studies and the DL model in one study. When looking at the features, 21 features were used in five studies, 22 features in four studies, 23 features in six studies and 37 features in one study. It was observed that 22 features were used in two studies and 23 features were used in two studies using cross validation (CV) along with accuracy, precision, recall and F_1 score metrics.

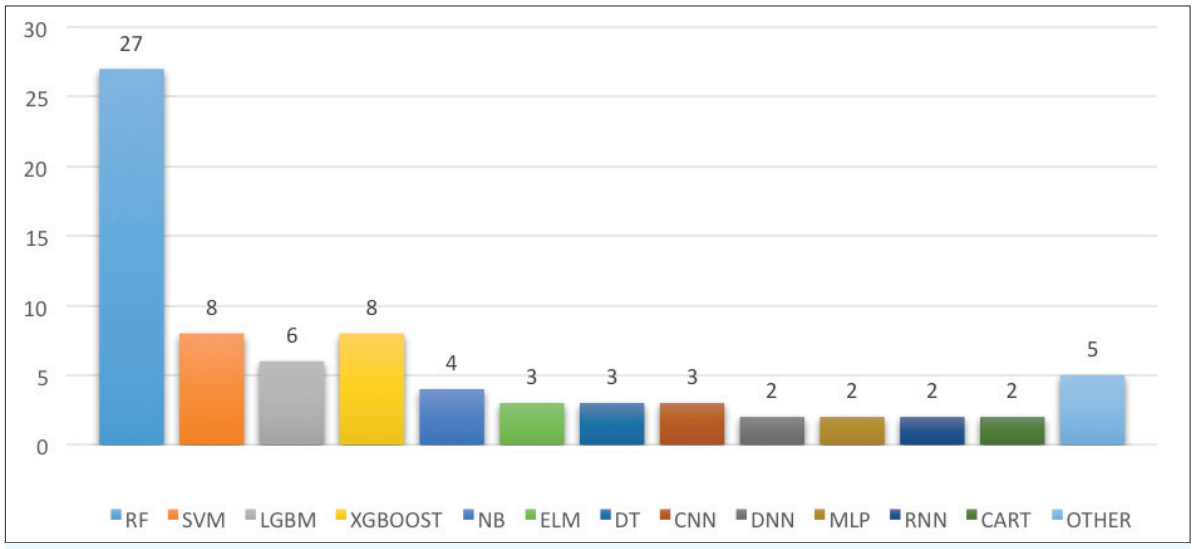

Figure 10. Number of classification methods used.

Table 3. Studies on the number of 2126 dataset

Ref.	Algorithm	Number of Features	Accuracy (%)	Precision	Recall	F_1 Score	CV
Kuzu et al. (2023)	XGB	22	100.00	1	1	1	Used
Avuçlu (2021)	DT, NB	22	100.00	-	-	-	Unused
Sheakh et al. (2023)	RF	21	99.98	0.9959	0.9945		Unused
Das et al. (2020)	RF	37	99.91	0.999	-	0.999	Used
Haweel et al. (2021)	PNN	21	99.74	-	-	-	Unused
Hussain et al. (2022)	ALEXNET – SVM	23	99.72	-	-	-	Unused
Kannan et al. (2021)	RF	23	99.57	0.996	0.996	0.996	Used
Iraji (2019)	MLP- ANFIS	21	99.50	-	-	-	Unused
Silwattananusarn et al. (2020)	SVM	23	99.39	-	-	-	Unused
Alkhasawneh (2019)	HECFNN	21	99.25	-	-	-	Used
Deresa et al. (2018)	RF	22	99.18	-	-	-	Unused
Potharaju et al. (2019)	IBK	23	99.05	-	-	-	Used
Kaliappan et al. (2023)	GB and VC	22	99.00	0.99	0.99	0.99	Used
Chaturvedı et al. (2022)	SVM	23	99.00	0.97	1	0.98	Used
Rahmayanti et al. (2022)	LGBM	21	99.00	-	-	0.98	Used
Pradhan et al. (2021)	RF	22	99.00	0.8	0.73	0.76	Used
Marvin et al. (2021)	LGBM	23	99.00	0.97	1	1	Unused

Kuzu et al. achieved 100% accuracy as a result of their experiments using the polynomial expansion (PE) function for feature extraction along with CV, entropy and normalisation (22). Das et al. on the other hand, achieved 99.91% accuracy by using standard dimensionality reduction algorithms such as PCA, correlation-based feature subset selection, Chi-squared feature selection, MinMax, and CV methods (64). On the other hand, according to Kannan et al. using CV with OneR, ZeroR, RIDOR, and JRIP classifiers resulted in 99.74% accuracy for PNN and 99.25% accuracy for HECFNN using Alkhasawneh balanced the analysed dataset with SMOTE using Jrip, Ridor, J48, NBStar, IBk and Kstar classifiers, achieving 99.05% accuracy with IBK using CV (57,84,88). Kaliappan et al. achieved 99% accuracy with GB and VC using various CV techniques such as Hold-Out, Leave-One-Out, Leave-P-Out, Monte Carlo, Stratified K-fold

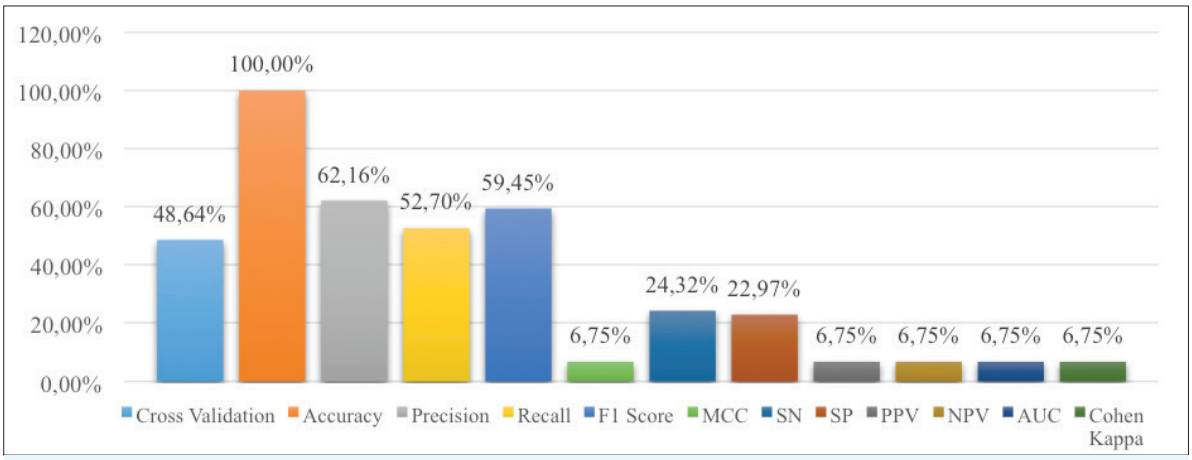

Figure 11. Ratios of performance metrics to number of literature searched.

and Repeated K-fold. In addition, Chaturvedi et al. achieved 99% accuracy by combining SMOTE, PCA and cross-validation in their study (25).

Therefore, the use of techniques such as SMOTE, PCA in combination with CV is important to obtain more accurate and, more importantly, reliable results in terms of performance on unbalanced datasets.

CV is important compared to traditional approaches in evaluating the performance of machine learning methods. This is evident from studies that show how this method assists us in more robustly assessing how well a model can generalize to real-world data. In traditional evaluation methods, a model can achieve high success on training data but exhibit low performance on new and unseen data. Furthermore, this situation can lead to the emergence of overfitting issues.

CV divides the dataset into different parts and employs them as training and validation data. Consequently, it evaluates how the model responds to various data samples and distributions in a more realistic manner. If a model has solely adapted to a specific subset of data or focused on meaningless patterns, CV helps us detect such issues more effectively. In conclusion, CV provides a better measure of a model's overall performance and serves as an important tool for identifying problems like overfitting. Therefore, it offers a more reliable performance evaluation compared to traditional methods.

The accuracy metric is used to measure the success of the model in all studies in performance metrics. However, since it is not sufficient alone, the usage rates of measurements such as precision, recall, CV, F_1 score, which are the average of precision and recall metrics, are also shown in detail in Figure 11 according to the number of literature investigated.

In recent years, the use of CV has gradually increased. As shown in Figure 12, the average accuracy rate of the 34 studies using CV was 95.49%, while this rate was 94.69% in the other 34 studies without CV. It can be seen that CV has a positive effect of about 1% on accuracy rates.

The highest accuracy rate achieved using CV was 99.91%, and the use of PCA was effective in achieving this rate. The average accuracy rates of studies using and not using CV in the last six years are shown in Figure 13.

Feature extraction techniques are used to enhance the performance of machine learning and deep learning models, prevent overfitting, reduce computational costs, and provide better generalization. Table 4 displays the feature extraction techniques employed in the investigated studies.

In machine learning, PCA is used to highlight features by expressing data in a lower-dimensional form, LDA enhances classification performance by emphasizing inter-class differences, ICA separates data into independent components, and relief and Chi-square are employed to determine the significance of features. RFE and random forest feature importance are used for feature selection to improve model performance. Other techniques facilitate feature selection or extraction through different methods.

In deep learning, convolutional neural networks (CNN) are used for visual data, while recurrent neural networks (RNN) are utilized for time series and sequential data. Autoencoders extract essential features from data, while generative adversarial networks (GAN) are employed for data synthesis and model development. Transformer and BERT are used in natural language processing, and ResNet is employed to address issues arising from complex network structures.

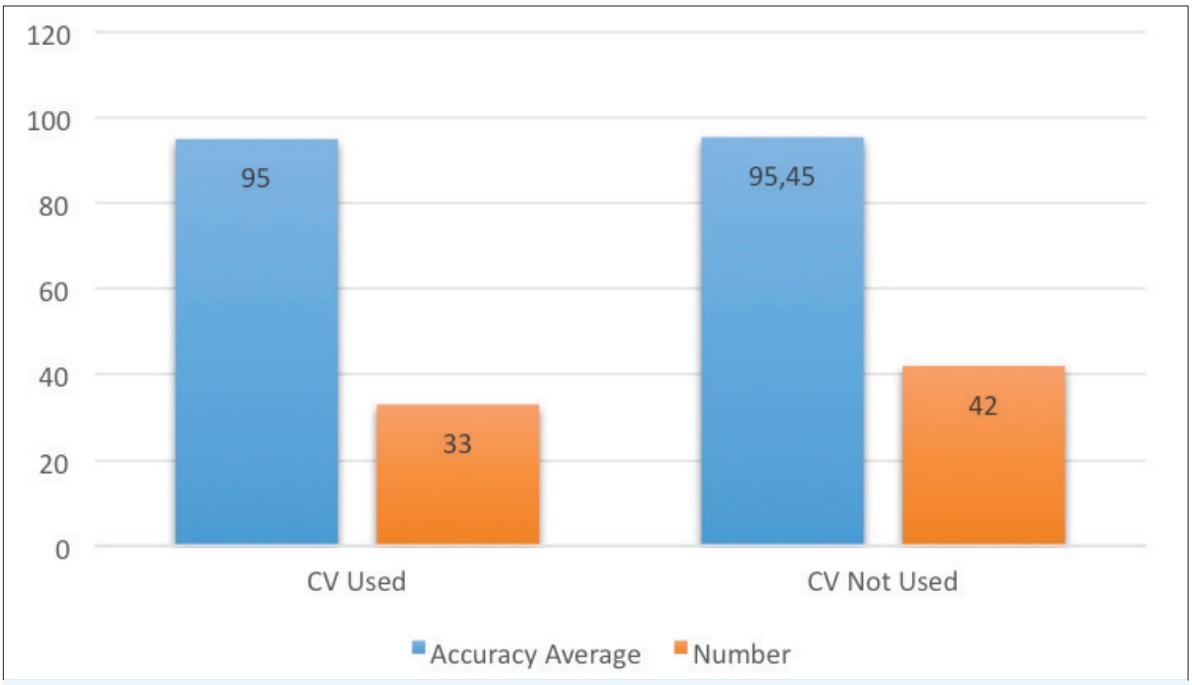

Figure 12. Accuracy averages in terms of CV.

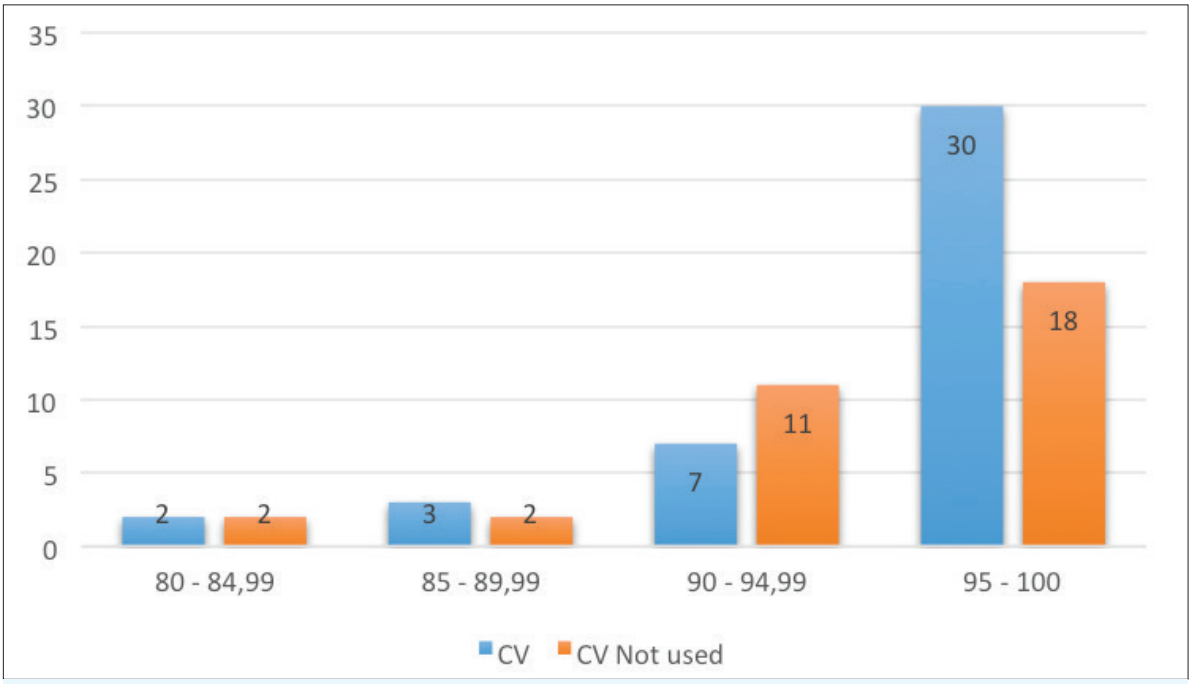

Figure 13. Effect of CV usage on accuracy rate.

Table 4 above summarises the results of 21 studies using different feature extraction techniques and models. The common goal of these studies is to achieve high performance in the data analysis and model building phases. For this purpose, researchers have combined different feature extraction methods and machine learning algorithms.

Table 4. Studies on the CTG dataset

Ref.	Feature Extraction	Method	Accuracy	CV	Model
Kuzu et al. (2023)	Polynomial expansion (PE)	XGB	100	10 CV	EL
Das et al. (2020)	PCA, Chi-square, MRMR	RF	99.91	10 CV	EL
Iraji (2019)	Stacked Autoencoder (SAE)	MLP-ANFIS	99.50	Unused	DL
Silwattananusarn et al. (2020)	ReliefF, CFS, IG	SVM	99.39	Unused	ML
Potharaju et al. (2019)	Chi-square, IG, Relief	IBK	99.05	10 CV	ML
Chaturvedı et al. (2022)	PCA	SVM	99.00	10 CV	ML
Kaliappan et al. (2023)	EDA	GB, VC	99	10 CV	ML
Kaluri et al. (2020)	PCA, LDA	SVM	98.59	Unused	EL
Hardalaç et al. (2022)	RFE	RF	97.20	5 CV	EL
Feng et al. (2021)	RFE	RF	97.20	5 CV	EL
Aslam et al. (2022),	RFE	RF	97.00	CV	EL
Rayhana et al. (2021)	MAMF	XGBoost	96.70	Unused	EL
Akmal et al. (2023)	AE, RFE	RF	96.62	5 CV	ML
Bhowmik et al. (2021)	Chi-square	Stacking Ensemble Learning	96.05	10 CV	EL
Dwivedi et al. (2021)	EDA	LGBM	95.61	5 CV	DL
Tokmak et al. (2022)	PCA, AE	DNN	95.20	10 CV	DL
Piri et al. (2021)	MOALO-CD,	RF	95.00	10 CV	EL
Piri et al. (2019)	MOGA-CD	XGBoost	94.00	Unused	EL
Jebadurai et al. (2022)	Spearman, Pearson, Kendall	DT	92.00	Unused	ML
Ramla et al. (2018)	RFE	J48-Lasso	86.46	10 CV	ML
Afridi et al. (2019)	CFS	NB	85.88	Unused	ML
Uzun et al. (2018)	PCA	ELM	84.30	10 CV	EL

As shown in the table, the highest accuracy rate of 99.91 was achieved in the study by Das et al. (64). This result, achieved by using feature extraction techniques such as PCA, Chi-square and MRMR, was obtained using the RF model. In the work of Iraji, the SAE technique led to the result of the MLP-ANFIS model (82). Silwattananusarn et al. achieved high accuracy through the application of ReliefF, CFS and IG methods combined with the SVM model (72).

When considering the results of cross-validation, it is observed that some studies were performed with CV, while others were marked as “Unused”. CV is a method used to better assess the real-world performance of a model. Therefore, the CV results of the studies provide a more robust understanding of the generalisation capabilities of the models.

The combination of different feature extraction techniques and different machine learning algorithms was particularly effective in significantly improving model performance. Furthermore, the combination of different data mining and analysis techniques to optimise the results contributes to a better understanding and prediction of usable datasets in various fields. Cross-validation results also help to reliably assess the overall performance of models.

The studies that used CV and those that did not use it, along with their accuracy performance, are shown in Tables 5 and Table 6. CV has been preferred due to its more reliable accuracy performance compared to simple methods.

The provided tables demonstrate the wide range of accuracy rates achieved by different machine learning and deep learning models through various methods. While some studies report very high accuracy rates, others exhibit lower rates. This diversity underscores the importance of understanding the specifics of each study, such as methodology, dataset, and the specific problem domain, in order to interpret, assess, and compare the reported results.

The differences in accuracy rates across studies have emerged due to various choices, such as the quality of the dataset and the selection of machine learning or deep learning algorithms. The success of these models is

Table 5. Studies using CV

Ref.	Accuracy	CV
Kuzu et al. (2023)	100.00	10
Das et al. (2020)	99.91	10
Kannan et al. (2021)	99.57	10
Alkhasawneh (2019)	99.25	10
Potharaju et al. (2019)	99.05	10
Chaturvedi et al. (2022)	99.00	10
Rahmayanti et al. (2022)	99.00	5
Pradhan et al. (2021)	99.00	5
Okwuchi et al. (2019)	98.60	5
John et al. (2020)	98.40	10
Dutta et al. (2021)	98.01	Used
Nandipati et al. (2020)	97.87	10
Fasihi et al. (2021)	97.46	10
Hardalaç et al. (2022)	97.20	5
Feng et al. (2021)	97.20	5
N. Aslam et al. (2022)	97.00	Used
Kaur et al. (2019)	97.00	Used
Das et al. (2022)	96.71	5
Panda et al. (2021)	96.45	Used
Fei et al. (2020)	96.39	10
Bhowmik et al. (2021)	96.05	10
Li et al. (2021)	95.90	10
Dadario et al. (2021)	95.82	4
Dwivedi et al. (2021)	95.61	5
Tokmak et al. (2022)	95.20	10
Islam et al. (2020)	95.11	10
Amin et al. (2021)	95.10	5
Piri et al. (2021)	95.00	10
Sevani et al. (2019)	94.35	10
Ricciardi et al. (2020)	91.10	10
Ramla et al. (2018)	90.10	5
Akbulut et al. (2022)	89.50	10
Ramla (2020)	86.46	10
Uzun et al. (2018)	84.30	10
Piri et al. (2019)	84.00	Used

Table 6. Non-CV studies

Ref.	Accuracy
Avuçlu (2021)	100.00
Haweel et al. (2021)	99.74
Hussain et al. (2022)	99.72
Iraji et al. (2019)	99.50
Silwattananusarn et al. (2020)	99.39
Deresa et al. (20218)	99.18
Marvin et al. (2021)	99.00
Sharma et al. (2022)	98.74
Arif et al. (2020)	98.70
Kaluri et al. (2020)	98.59
Kasım (2021)	98.12
Appaji et al. (2019)	97.55
Aslam et al. (2022)	97.51
Avuçlu et al. (2020)	97.18
Rayhana et al. (2021)	96.70
Manikandan et al. (2021)	96.62
Singh et al. (2022)	95.14
Jayashree et al. (2021)	95.10
Piri et al. (2020)	94.00
Bhuiyan et al. (2019)	93.34
Li et al. (2018)	93.24
Hoodbhoy et al. (2019)	93.00
Vani (2019)	93.00
Amin et al. (2019)	92.95
Chen et al. (2021)	92.64
Jebadurai et al. (2022)	92.00
Xue (2019)	91.85
Thomas et al. (2020)	91.70
Bautista et al. (2020)	90.00
Miao et al. (2018)	88.02
Kadhim et al. (2020)	86.55
Afridi et al. (2019)	85.88
Spairani et al. (2022)	80.10
Piri et al. (2022)	-

influenced not only by the algorithms themselves but also by the quality and quantity of data used for training and testing.

The average accuracy of 29 studies using accuracy, recall, precision, and F_1 score metrics together, as shown in Table 7, was found to be 95.21%. Among these 29 studies, the average accuracy of 17 studies using CV ranged from 89.50% to 99.57% with an average of 96.13%, while the average accuracy of 12 studies not using CV ranged from 85.88% to 99% with an average of 93.90%. The highest accuracy in these 29 studies was obtained using the RF method with 10 CV and particle swarm optimization (PSO), with a value of 99.57%.

Table 7. Studies using Accuracy, Precision, Recall, and F1 Score metrics in ML

Ref.	Accuracy (%)	Precision	Recall	F_1 Score	CV
Marvin et al. (2021)	99.00	0.97	1	1	Unused
Kaluri et al. (2020)	98.59	0.99	0.99	0.99	
Kasım (2021)	98.12	0.99	0.99	0.99	
Appaji et al. (2019)	97.55	0.93	0.89	0.964	
Aslam et al. (2022)	97.51	0.99	1	0.99	
Singh et al. (2022)	95.14	0.9107	0.9048	0.9107	
Bhuiyan et al. (2019)	93.34	0.9561	0.9737	0.963	
Jebadurai et al. (2022)	92.00	0.92	0.92	0.92	
Thomas et al. (2020)	91.70	0.943	0.892	0.917	
Bautista et al. (2020)	90.00	0.94	0.75	0.8	
Miao et al. (2018)	88.02	0.85	0.843	0.85	
Afridi et al. (2019)	85.88	0.946	0.859	0.895	
Kuzu et al. (2023)	100	1	1	1	Used
Kannan et al. (2021)	99.57	0.996	0.996	0.996	
Chaturvedi et al. (2022)	99.00	0.97	1	0.98	
Pradhan et al. (2021)	99.00	0.8	0.73	0.76	
Okwuchi et al. (2019)	98.60	0.99	0.99	0.99	
Dutta et al. (2021)	98.01	0.978	0.977	0.975	
Feng et al. (2021)	97.20	0.94	0.89	0.92	
Hardalaç et al. (2022)	97.20	1	1	1	
Kaur et al. (2019)	97.00	0.97	0.99	0.98	
Fei et al. (2020)	96.39	0.9938	0.9698	0.9816	
Li et al. (2021)	95.90	0.959	0.916	0.958	
Dadario [35]	95.82	0.9286	0.8986	0.9128	
Dwivedi et al. (2021)	95.61	0.9552	0.9056	0.955	
Tokmak et al. (2022)	95.20	0.953	0.952	0.952	
Amin et al. (2021)	95.10	0.9495	0.952	0.951	
Piri et al. (2021)	95.00	0.97	0.95	0.95	
Ramla et al. (2018)	90.10	0.89	0.89	0.89	
Akbulut et al. (2018)	89.50	0.75	0.75	0.75	

The given table displays the performance results of various machine learning and deep learning models measured with different metrics. These results indicate that the models were applied to various problem domains to evaluate their performance. Metrics such as “Accuracy,” “Precision,” “Recall,” and “ F_1 Score” are used to assess the model’s performance from different angles, representing important measurements.

Upon examining the results in the table, it can be observed that different studies achieved high accuracy rates with different models. While some studies achieved accuracy rates of 99% or higher, others attained success with lower rates. This discrepancy could stem from factors such as the nature of the dataset used, the selection of features, and the choice of algorithms.

Additionally, we can see that other metrics such as “Precision,” “Recall,” and “ F_1 Score” are also considered. These metrics indicate how well the model separates positive and negative classes, how effectively it captures true positive results, and how few false positive results it produces.

In conclusion, the table illustrates how different machine learning and deep learning models perform in various problem domains. These results serve as a valuable resource for understanding which method might perform better in specific situations.

Table 8. Accuracy rates obtained using AUC

Ref.	Method	Number of Features	Accuracy (%)	AUC	CV
Sharma et al. (2022)	LGBM	21	99.00	0.9822556	5
Singh et al. (2022)	LGBM	23	99.00	0.993	-
Hussain et al. (2022)	RF	23	98.01	0.89	CV
Das et al. (2022)	CNN	23	97.46	0.975	10
Hardalaç et al. (2022)	ELM	23	96.45	0.8214	CV
Tokmak et al. (2022)	Stacking	21	96.05	0.9595	10
Jebadurai et al. (2022)	Blender	-	95.90	0.988	10
N. Aslam et al. (2022)	LGBM	21	95.61	0.9864	5
Piri et al. (2022)	RNN	21	95.10	0.93	5
Aslam et al. (2022)	DF	21	92.64	0.99	-
Chaturvedi et al. (2022)	DT	22	92.00	0.91	-
Spairani et al. (2022)	RF		91.10	0.967	10
Rahmayanti et al. (2022)	LGBM	22	89.50	0.958	10

Table 8 shows the accuracy rates obtained with AUC and the use of CV. ROC curve is an important performance criterion used in classification problems. As ROC-AUC increases, the discrimination performance between classes increases. As AUC approaches 1, the classification performance increases. Among the studies, there is diversity in the utilization of different methods and the characteristics of each study leading to variations in accuracy and AUC values. Especially factors like the size of the dataset, feature selection, algorithm type, and training process can influence these differences. This diversity demonstrates that each study has distinct methodologies, datasets, and problem domains. As a result, the data in the table showcases promising results in the evaluation of fetal health using machine learning and deep learning methods. However, to determine which method performs better under which circumstances, further comprehensive analysis and comparisons are required.

DISCUSSION AND CONCLUSION

We statistically analyzed 75 different literature studies that used at least one algorithm to classify fetal health from NST signals, including studies that analyzed computerized interpretation of NST based on the DL and ML models. Our main finding is that the predictions based on the DL and ML significantly show positive results to assist experts in determining fetal health, demonstrating the importance of DL and ML applications in real clinical practice.

In our research conducted in the last six years, we observed an increase in studies related to NST, and that ML models were used more frequently than DL models in these studies. The results of these studies showed that fetal classification using ML models was successfully achieved with high accuracy rates. This demonstrates the importance of using DL and ML models in the early detection and timely diagnosis of possible fetal illnesses. Additionally, the use of DL and ML models in clinical settings was shown to reduce medical errors. Furthermore, the use of DL and ML models in clinical settings will improve the quality of care for patients by assisting healthcare professionals in the diagnosis and treatment phases.

The diagnosis and monitoring of fetal health is of great importance in modern medical practice. The use of ML and DL models in this area has the potential to provide more accurate and effective results compared to traditional methods. In particular, the implementation of these models in conjunction with wearable technologies has the potential to provide a more convenient and seamless health monitoring experience for pregnant women and fetuses.

However, despite progress in the field, the application of ML and DL models in fetal health is not as widespread as in other medical areas. This is due to several factors. Firstly, there is a need to establish specific standards for the clinical use and reliability of these models. In addition, further research and optimisation of the algorithms and features used in these models is essential to increase their success.

Another key issue is the quantity and quality of data. The effective functioning of ML and DL models relies on comprehensive and representative datasets. These datasets should include different demographic groups and different health conditions. In addition, accurate labelling and data reliability have a significant impact on the accuracy of the

models. Therefore, researchers in this field must prioritise the collection, processing, and organisation of data.

In conclusion, the use of ML and DL models for fetal health diagnosis and monitoring has considerable potential. However, realising this potential requires additional research and effort. By delving deeper into the field, enriching datasets, refining algorithms and achieving reliable results in clinical applications, researchers can promote the effective application of ML and DL models in fetal health.

Despite significant advances in DL techniques in recent years, there are many reasons why they have not been as successful as machine learning methods. One reason is the dependence of DL on high computational power, due to the need for large amounts of data and the construction of complex models. This can hinder the performance of DL when data and computing resources are limited. In addition, the lengthy learning processes and complex management of hyper-parameter settings of DL models can be challenging. Another difficulty with DL is the limited interpretability of model results due to their complex structures, making it difficult to understand which features are significant. On the other hand, machine learning methods offer a number of advantages in comparison, such as the flexibility to work with smaller datasets, lower computational requirements, faster training times, the ability to produce interpretable results with less complex models, and broader applicability across different domains. These factors contribute to the fact that machine learning methods are often preferred.

CV is a method used to evaluate the performance of machine learning models more reliably. This approach involves splitting the dataset into different subsets to conduct multiple model trainings and evaluations. CV enhances performance assessment by measuring how well the model generalizes to real-world data. By mitigating the impact of a single data split, this method aids in better understanding the overall model performance with increased confidence.

The challenges faced by researchers in the field of cardiotography are quite diverse. Among these challenges are ensuring data quality and consistency, analyzing complex heart signals, accounting for different physiological responses among individuals, interpreting findings accurately, considering ethical considerations, managing large amounts of data, integrating various data sources, ensuring clinical applicability, validating, and replicating results, and keeping pace with rapidly advancing technology. These challenges can be considered significant factors that require a careful approach and expertise for researchers to achieve accurate and meaningful results.

Acknowledgement

The outputs of the project supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK, Grant 5220067) were used in this study.

Informed Consent: Patient consent was obtained.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept- YS, AK; Design- YS; Supervision- YS; Data Collection and/or Processing- YS, AK; Analysis and/or Interpretation- YS, AK; Literature Search - AK, YS; Writing- AK, YS; Critical Review- AK, YS.

Conflict of Interest: Authors declare that they have no conflicts of interest or funding to disclose.

Financial Disclosure: Authors declared that this study has received no financial support.

REFERENCES

1. World Health Organization (WHO). Perinatal mortality: A listing of available information. Maternal health and safe motherhood programme. Geneva, 1996.
2. Hacettepe Üniversitesi Nüfus Etütleri Enstitüsü (HÜNEE). Türkiye nüfus ve sağlık araştırması, 2008. Ankara: HÜNEE, Sağlık Bakanlığı AÇSAP Genel Müdürlüğü, Başbakanlık DPT Müsteşarlığı ve TÜBİTAK; 2009.
3. Keith M. Introduction in Keith LM (eds) The developing human clinically oriented embryology. Fifth ed. WB Saunders Company: The Curtis Center Independence Square West 19106. Philadelphia, Pennsylvania; 1993;1-13.
4. Nakamura Y, Hosokawa Y, Yano H, Nakashima N, Nakashima T, Komatsu Y, et al. Primary causes of perinatal death. An autopsy study of 1000 cases in Japanese infants. Human Pathology 1982;13(1):56-61. https://doi.org/10.1016/S0046-8177(82)80139-1
5. Keeling JW. Fetal and perinatal death in Busitill A, Keeling JW (eds). Pediatric forensic medicine and pathology. Chapter 10. 1st ed. London Edward Arnold (Publishers);2009;180-97.
6. Aladağ S, Güven A. Kardiyotokogram verilerinin yapay sinir ağları ile sınıflandırılması. 2014.
7. Ergür AR, Yergök YZ, Başhekim Ç, Ertekin A, Müngen E, Tütüncü l. Yüksek riskli gebelerde morbidite tesbitinde, umbilikal arter doppler çalışmalarının, non-stress test ve biofizik profil skor ile karşılaştırılması. 1996.
8. Bafor EE, Kalu CH, Omoruyi O, Elvis-Offiah UB, Edrada-Ebel R. Thyme (Thymus vulgaris [Lamiaceae]) leaves inhibit contraction of the nonpregnant mouse uterus. J Medicinal Food 2021;24(5):541-50. https://doi.org/10.1089/jmf.2020.0076
9. Ekizler H, Eryılmaz H. İntrapartum fetal monitorizasyon ve hemşirelik girişimleri. Florence Nightingale J Nurs 1994;8(31):52-61.
10. Singh S, Pai S, Sahu B. Study of umbilical coiling index and perinatal outcome. International J Reproduction, Contraception, Obstetrics and Gynecology 2020;9(10):3977-83. https://doi.org/10.18203/2320-1770.ijrcog20204021

11. Aktaş S, Osmanağaoğlu MA. İntrapartum elektronik fetal monitorizasyon uygulaması ve sağlık profesyonellerinin sorumlulukları. Life Sciences 2017;12(1):14-29. https://doi.org/10.12739/NWSA.2017.12.1.4B0009
12. Rooth G, Huch A, Huch R. Guidelines for the use of fetal monitoring. Int J Gynecol Obstet 1987;25:159.
13. Nijhuis IJM, ten Hof J, Mulder EJH, Nijhuis JG, Narayan H, Taylor DJ, et al. Antenatal fetal heart rate monitoring; normograms and minimal duration of recordings. Prenat Neonat Med 1998;3:314-22.
14. Bergmans MGM, Jonker GJ, Kock HCL. Fetal supraventricular tachycardia: Review of the literature. Obstet Gynecol Surv 1985;40:61-8. https://doi.org/10.1097/00006254-198502000-00002
15. Comart N, Yıldırım G, Güngördük K, Aktaş FN, Ark HC. Elektronik fetal kalp hızı monitörizasyonu: Normal monitör, fetal stres, fetal distres ile ilişkili erken neonatal sonuçlar. J Clin Obstetrics Gynecol 2007;17(3):186-95.
16. Korkmaz A, Aydın Ş, Duyan Çamurdan A, Okumuş N, Onat N, Özbaş FS, ve ark. Türkiye'de bebek ölüm nedenlerinin ve ulusal kayıt sisteminin değerlendirilmesi. Çocuk Sağlığı Hastalıkları Derg 2013;56:105-21.
17. World Health Organization (WHO). Mortality rate, neonatal. Available from: https://data.worldbank.org/indicator/SH.DYN.NMRT?view=map (Accessed date: 10.01.2023).
18. Waits GS, Soliman EZ. Digitizing paper electrocardiograms: Status and challenges. J Electrocardiol 2017;50(1):123-30. https://doi.org/10.1016/j.jelectrocard.2016.09.007
19. Chitradevi M, Geetharamani G. Prediction of neonatal state by computer analysis of fetal heart rate tracings: The antepartum arm of the SisPorto® multicentre validation study. Int J Comput Appl 2012;47(14):19-25.
20. Spairani E, Daniele B, Signorini MG, Magenes G. A deep learning mixed-data type approach for the classification of FHR signals. Front Bioeng Biotechnol 2022;10:887549. https://doi.org/10.3389/fbioe.2022.887549
21. Cömert Z, Kocamaz AF, Subha V. Prognostic model based on image-based time-frequency features and genetic algorithm for fetal hypoxia assessment. Comput Biol Med 2018;99:85-97. https://doi.org/10.1016/j.compbiomed.2018.06.003
22. Kuzu A, Santur Y. Early diagnosis and classification of fetal health status from a fetal cardiotocography dataset using ensemble learning. Diagnostics 2023;13(15):2471. https://doi.org/10.3390/diagnostics13152471
23. Hephzibah R, Christinal AH, Jayanthi R, Chandy DA, Bajaj C. A novel ensemble classifier framework for accurate fetal heart rate classification. 2023 4th International Conference on Signal Processing and Communication (ICSPC), Coimbatore, India, 2023;32:1-4. https://doi.org/10.1109/ICSPC57692.2023.10125713
24. Cao Q, Sun H, Wang H, Liu X, Lu Y, Huo L. Comparative study of neonatal brain injury fetuses using machine learning methods for perinatal data. Computer Methods and Programs in Biomedicine 2023;240:107701. https://doi.org/10.1016/j.cmpb.2023.107701
25. Kaliappan J, Bagepalli AR, Almal S, Mishra R, Hu YC, Srinivasan K. Impact of cross-validation on machine learning models for early detection of intrauterine fetal demise. Diagnostics 2023;13(10):1692. https://doi.org/10.3390/diagnostics13101692
26. Sheakh MA, Tahosin MS, Hasan MM, Islam T, Islam O, Rana MM. Child and maternal mortality risk factor analysis using machine learning approaches. In 2023 11th International Symposium on Digital Forensics and Security (ISDFS) 2023;1-6. https://doi.org/10.1109/ISDFS58141.2023.10131826
27. Akmal H, Hardalaç F, Ayturan K. A fetal well-being diagnostic method based on cardiotocographic morphological pattern utilizing autoencoder and recursive feature elimination. Diagnostics 2023;13(11):1931. https://doi.org/10.3390/diagnostics13111931
28. Mehbodniya A, Lazar AJP, Webber J, Sharma DK, Jayagopalan SKK, Sengan S. Fetal health classification from cardiotocographic data using machine learning. Expert Systems 2022;39(6):e12899. https://doi.org/10.1111/exsy.12899
29. Hardalaç F, Akmal H, Ayturan K, Acharya UR, Tan RS. Fetal status classification based on feature elimination and hyperparameter optimization using cardiotocographic data. SSRN 2022;4197627. https://doi.org/10.2139/ssrn.4197627
30. Muhammad Hussain N, Rehman AU, Othman MTB, Zafar J, Zafar H, Hamam H. Accessing artificial intelligence for fetus health status using hybrid deep learning algorithm (AlexNet-SVM) on cardiotocographic data. Sensors 2022;22(14):5103. https://doi.org/10.3390/s22145103
31. Chaturvedi M, Agrawal S, Silakari S. Strategic analysis of different classification algorithms on CTG Data. Ann Agri-Bio Res 2022;27(1):97-101.
32. Piri J, Mohapatra P, Dey R. Investigating association relationship between fetal heart rate parameters from cardiotocography employing multi-objective evolutionary algorithms. Int J Inf Technol 2022;14(4):1923-35. https://doi.org/10.1007/s41870-022-00909-w
33. Rahmayanti N, Pradani H, Pahlawan M, Vinarti R. Comparison of machine learning algorithms to classify fetal health using cardiotocogram data. Procedia Computer Science 2022;197:162-71. https://doi.org/10.1016/j.procs.2021.12.130
34. Tokmak M, Küçüksille EU. Comparative analysis of dimension reduction and classification using cardiotocography data. 2022.
35. Das S, Mukherjee H, Roy K, Saha CK. Fetal health classification from cardiotocograph for both stages of labor-a soft computing based approach. Diagnostics 2022;13:858. https://doi.org/10.3390/diagnostics13050858
36. Sharma P, Sharma K. Optimized classification of fetal state health using GWO and WOA. In 2022 IEEE 9th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT) 2022;573-8. https://doi.org/10.1109/SETIT54465.2022.9875521
37. Aslam MT, Khan MAI, Dola NN, Tazin T, Khan MM, Albraikan AA, et al. Comparative analysis of different efficient machine learning methods for fetal health classification. Applied Bionics and Biomechanics, 2022. https://doi.org/10.1155/2022/6321884

38. Jebadurai IJ, Paulraj GJ, Jebadurai J, Silas S. Experimental analysis of filtering based feature selection techniques for fetal health classification. SJEE 2022;19(2):207-24. https://doi.org/10.2298/SJEE2202207J
39. Aslam N, Khan IU, Aljishi RF, Alnamer ZM, Alzawad ZM, Almomen FA, et al. Explainable computational intelligence model for antepartum fetal monitoring to predict the risk of IUGR. Electronics 2022;11(4):593. https://doi.org/10.3390/electronics11040593
40. Singh V, Agrawal R, Gourisaria MK, Singh PK, Das H. Comparative analysis of machine learning models for early detection of fetal disease using feature extraction. In 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT) 2022;169-75. https://doi.org/10.1109/CSNT54456.2022.9787635
41. Dadario AMV, Espinoza C, Nogueira WA. Classification of fetal state through the application of machine learning techniques on cardiotocography records: Towards real world application. medRxiv 2021.
42. Manikandan M, Vijayakumar P. Improving the performance of classifiers by ensemble techniques for the premature finding of unusual birth outcomes from cardiotocography. IETE J Res 2021;1-11. https://doi.org/10.1080/03772063.2021.1910579
43. Dutta P, Paul S, Majumder M. Intelligent SMOTE based machine learning classification for fetal state on cardiotocography dataset. Research Square 2021. https://doi.org/10.21203/rs.3.rs-1040799/v1
44. Fasihi M, Nadimi-Shahraki MH, Jannesari A. A shallow 1-D convolution neural network for fetal state assessment based on cardiotocogram. SN Computer Science 2021;2(4):1-9. https://doi.org/10.1007/s42979-021-00694-6
45. Pradhan AK, Rout JK, Maharana AB, Balabantaray BK, Ray NK. A machine learning approach for the prediction of fetal health using CTG. In 2021 19th OITS International Conference on Information Technology (OCIT) 2021;239-44. https://doi.org/10.1109/OCIT53463.2021.00056
46. Piri J, Mohapatra P, Dey R. Multi-objective ant lion optimization based feature retrieval methodology for investigation of fetal wellbeing. In 2021 Third international conference on inventive research in computing applications (ICIRCA) 2021;1732-7. https://doi.org/10.1109/ICIRCA51532.2021.9544860
47. Feng J, Liang J, Qiang Z, Li X, Chen Q, Liu G, Wei H. Effective techniques for intelligent cardiotocography interpretation using XGB-RF feature selection and stacking fusion. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2021;2667-73. https://doi.org/10.1109/BIBM52615.2021.9669694
48. Chen Y, Guo A, Chen Q, Quan B, Liu G, Li L, et al. Intelligent classification of antepartum cardiotocography model based on deep forest. Biomedical Signal Processing and Control 2021;67:102555. https://doi.org/10.1016/j.bspc.2021.102555
49. Kasım Ö. Multi-classification of fetal health status using extreme learning machine. Icontech Int J 2021;5(2):62-70. https://doi.org/10.46291/ICONTECHvol5iss2pp62-70
50. Dwivedi P, Khan AA, Mugde S, Sharma G. Diagnosing the major contributing factors in the classification of the fetal health status using cardiotocography measurements: An AutoML and XAI approach. In 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI) 2021;1-6. https://doi.org/10.1109/ECAI52376.2021.9515033
51. Li J, Liu X. Fetal health classification based on machine learning. In 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) 2021;899-902. https://doi.org/10.1109/ICBAIE52039.2021.9389902
52. Haweel MT, Zahran O, Abd El-Samie FE. Polynomial FLANN classifier for fetal cardiotocography monitoring. In 2021 38th National Radio Science Conference (NRSC) 2021;1:262-70. https://doi.org/10.1109/NRSC52299.2021.9509832
53. Avuçlu E. A novel hybrid model for automated analysis of cardiotocograms using machine learning algorithms. Int J Intelligent Systems Appl Engin 2021;9(4):266-72. https://doi.org/10.18201/ijisae.2021473716
54. Bhowmik P, Bhowmik PC, Ali UME, Sohrawordi M. Cardiotocography data analysis to predict fetal health risks with tree-based ensemble learning. I.J. Information Technology and Computer Science 2021;5:30-40. https://doi.org/10.5815/ijitcs.2021.05.03
55. Rayhana MT, Arefina AS, Chowdhury SA. Automatic detection of fetal health status from cardiotocography data using machine learning algorithms. J Bangladesh Academy Sci 2021;45(2):155-67. https://doi.org/10.3329/jbas.v45i2.57206
56. Jayashree J, Vijayashree J, Iyengar NCS. Fetal risk prediction using optimized genetic algorithm-support vector machine based feature selection techniques. 2021.
57. Kannan E, Ravikumar S, Anitha A, Kumar SA, Vijayasarathy M. Analyzing uncertainty in cardiotocogram data for the prediction of fetal risks based on machine learning techniques using rough set. J Ambient Intelligence Humanized Computing 2021;1-13. https://doi.org/10.1007/s12652-020-02803-4
58. Marvin G, Alam MGR. Cardiotocogram biomedical signal classification and interpretation for fetal health evaluation. In 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) 2021;1-6. https://doi.org/10.1109/CSDE53843.2021.9718415
59. Amin B, Salama AA, El-Henawy IM, Mahfouz K, Gafar MG. Intelligent neutrosophic diagnostic system for cardiotocography data. Computational Intelligence and Neuroscience 2021;2021:12. https://doi.org/10.1155/2021/6656770
60. Panda D, Panda D, Dash SR, Parida S. Extreme learning machines with feature selection using GA for effective prediction of fetal heart disease: A novel approach. Informatica 2021;45(3):381-92. https://doi.org/10.31449/inf.v45i3.3223
61. Arif MZ, Ahmed R, Sadia UH, Tultul MSI, Chakma R. Decision tree method using for fetal state classification from cardiotography data. J Advanced Engineering Computation 2020;4(1):64-73. https://doi.org/10.25073/jaec.202041.273
62. Nandipati SCR, XinYing C. Classification and feature selection approaches for cardiotocography by machine learning techniques. J Telecommunication, Electronic Computer Engineering (JTEC) 2020;12(1):7-14.

63. Piri J, Mohapatra P, Dey R. Fetal health status classification using moga-cd based feature selection approach. In 2020 IEEE international Conference on Electronics, Computing and Communication Technologies CONECCT 2020;1-6. https://doi.org/10.1109/CONECCT50063.2020.9198377
64. Das S, Mukherjee H, Obaidullah S, Roy K, Saha CK. Ensemble based technique for the assessment of fetal health using cardiotocograph-a case study with standard feature reduction techniques. Multimedia Tools Appl 2020;79(47):35147-68. https://doi.org/10.1007/s11042-020-08853-2
65. Kadhim NJA, Abed JK. Enhancing the prediction accuracy for cardiotocography (CTG) using firefly algorithm and naive Bayesian classifier. In IOP Conference Series: Materials Science and Engineering 2020;745(1):012101. https://doi.org/10.1088/1757-899X/745/1/012101
66. Ramla M. (2020). Influence of Feature Selection Methods on Cardiotocography Data: A Quantitative Investigation. International J Engineering Advanced Technology (IJEAT) 2019;8. https://doi.org/10.35940/ijeat.D1006.0484S219
67. Avuçlu E, Elen A. Classification of cardiotocography records with naïve bayes. International Scientific Vocational Studies J 2020;3(2):105-10.
68. John M, Shaiba H. Ensemble based foetal state diagnosis. In 2020 6th Conference on Data Science and Machine Learning Applications (CDMA) 2020;129-33. https://doi.org/10.1109/CDMA47397.2020.00028
69. Fei Y, Huang X, Chen Q, Chen J, Li L, Hong J, et al. Automatic classification of antepartum cardiotocography using fuzzy clustering and adaptive neuro-fuzzy inference system. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2020;938-1942. https://doi.org/10.1109/BIBM49941.2020.9313143
70. Ricciardi C, Improta G, Amato F, Cesarelli G, Romano M. Classifying the type of delivery from cardiotocographic signals: A machine learning approach. Computer Methods and Programs in Biomedicine 2020;196:105712. https://doi.org/10.1016/j.cmpb.2020.105712
71. Islam S, Yulita I. Predicting fetal condition from cardiotocography results using the random forest method. In Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar, MSCEIS, Bandung, West Java, Indonesia 2020. https://doi.org/10.4108/eai.12-10-2019.2296540
72. Silwattananusarn T, Kanarkard W, Tuamsuk K. Enhanced classification accuracy for cardiotocogram data with ensemble feature selection and classifier ensemble. 2020.
73. Reddy GT, Kumar Reddy MP, Lakshmanna K, Kaluri R, Singh Rajput D, Srivastava G, et al. Analysis of dimensionality reduction techniques on big data. IEEE Access 2020;8:54776-88. https://doi.org/10.1109/ACCESS.2020.2980942
74. Bautista JM, Quiwa QAI, Reyes RS. Machine learning analysis for remote prenatal care. In 2020 IEEE REGION 10 CONFERENCE (TENCON) 2020;397-402. https://doi.org/10.1109/TENCON50793.2020.9293890
75. Thomas R, Judith JE. Hybrid outlier detection in healthcare datasets using DNN and one class-SVM. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) 2020;1293-8. https://doi.org/10.1109/ICECA49313.2020.9297401
76. Hoodbhoy Z, Noman M, Shafique A, Nasim A, Chowdhury D, Hasan B. Use of machine learning algorithms for prediction of fetal risk using cardiotocographic data. Int J Appl Basic Med Res 2019;9(4):226. https://doi.org/10.4103/ijabmr.IJABMR_370_18
77. Appaji SV, Shankar RS, Murthy KVS, Rao CS. Cardiotocography class status prediction using machine learning techniques. Indian J Pub Health Res Develop 2019;10(8). https://doi.org/10.5958/0976-5506.2019.01961.2
78. Afridi R, Iqbal Z, Khan M, Ahmad A, Naseem R. Fetal heart rate classification and comparative analysis using cardiotocography data and KNOWN classifiers. Int J Grid Distributed Comput (IJGDC) 2019;12:31-42. https://doi.org/10.33832/ijgdc.2019.12.1.03
79. Piri J, Mohapatra P. Exploring fetal health status using an association based classification approach. In 2019 International Conference on Information Technology (ICIT) 2019;66-171. https://doi.org/10.1109/ICIT48102.2019.00036
80. Xue G. The application of machine learning models in fetal state auto-classification based on cardiotocograms. In IOP Conference Series: Earth and Environmental Science 2019;310(5):052007. https://doi.org/10.1088/1755-1315/310/5/052007
81. Amin B, Gamal M, Salama AA, El-Henawy IM, Mahfouz K. Classifying cardiotocography data based on rough neural network. Int J Adv Comp Sci Appl 2019;10(8). https://doi.org/10.14569/IJACSA.2019.0100846
82. Iraji MS. Prediction of fetal state from the cardiotocogram recordings using neural network models. Artificial Intelligence Medicine 2019;96:33-44. https://doi.org/10.1016/j.artmed.2019.03.005
83. Okwuchi I, Carnduff C, Pruthi S. Comparison of machine learning algorithms used for cardiotocography classification considering target labels correlation. 2019.
84. Potharaju SP, Sreedevi M, Ande VK, Tirandasu RK. Data mining approach for accelerating the classification accuracy of cardiotocography. Clinical Epidemiology and Global Health 2019;7(2):160-4. https://doi.org/10.1016/j.cegh.2018.03.004
85. Sevani N, Hermawan I, Jatmiko W. Feature selection based on F-score for enhancing CTG data classification. In 2019 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom) 2019;18-22. https://doi.org/10.1109/CYBERNETICSCOM.2019.8875656
86. Vani R. Weighted deep neural network basedclinical decision support system for the determination of fetal health. Int J Recent Tech Engin (IJRTE) 2019. https://doi.org/10.35940/ijrte.D4378.118419
87. Kaur H, Khullar V, Singh H, Bala M. Perinatal hypoxia diagnostic system by using scalable machine learning algorithms. Int J Innov Technol Explor Eng 2019;8(12):1954-9. https://doi.org/10.35940/ijitee.L2905.1081219

88. Alkhasawneh MS. Hybrid cascade forward neural network with Elman neural network for disease prediction. Arabian J Sci Engin 2019;44(11):9209-20. https://doi.org/10.1007/s13369-019-03829-3

89. Bhuiyan MAR, Ullah MR, Das AK. iHealthcare: Predictive model analysis concerning big data applications for interactive healthcare systems. Applied Sciences 2019;9(16):3365. https://doi.org/10.3390/app9163365

90. Ramla M, Sangeetha S, Nickolas S. Fetal health state monitoring using decision tree classifier from cardiotocography measurements. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) 2018;1799-803. https://doi.org/10.1109/ICCONS.2018.8663047

91. Deressa TD, Kadam K. Prediction of fetal health state during pregnancy: A survey. Int J Comput Sci Trends Tech (IJCST) 2018;6(1):29-36.

92. Uzun A, Kızıltas CE, Yılmaz E. Cardiotocography dataset classification with extreme learning machine. In International conference on advanced technologies, computer engineering and science 2018;224-30.

93. Akbulut A, Ertugrul E, Topcu V. Fetal health status prediction based on maternal clinical history using machine learning techniques. Computer methods and programs in biomedicine 2018;163:87-100. https://doi.org/10.1016/j.cmpb.2018.06.010

94. Li J, Chen ZZ, Huang L, Fang M, Li B, Fu X, Zhao Q. (2018). Automatic classification of fetal heart rate based on convolutional neural network. IEEE Internet of Things J 2018;6(2):1394-401. https://doi.org/10.1109/JIOT.2018.2845128

95. Miao JH, Miao KH. Cardiotocographic diagnosis of fetal health based on multiclass morphologic pattern predictions using deep learning classification. International J Advanced Comp Sci Appli 2018;9(5). https://doi.org/10.14569/IJACSA.2018.090501