East Journal of Computer Science
ISSN: 3079-9406

Volume 2, Issue 1
Publisher: East Publication & Technology
DOI: https://doi.org/10.63496/ejcs.Vol2.Iss1.257

Automated Fetal Health Classification Using Auto-Sklearn Approach for Enhanced Clinical Sensitivity

Yasir Hussein Shakir*[1], Reem Ali Mutlag[1], Eshaq Aziz Awadh AL Mandhari[2]

[1]College of Graduate Studies (COGS), University Tenaga Nasional (UNITEIN), Kajang, Malaysia.
[2]Graduate School of Technology, Asia Pacific University of Technology and Innovation (APU), Kuala Lumpur, Malaysia.
*Corresponding Author Email: yasserhessein19855@gmail.com.

Received: 26/01/2026, **Revised**: 17/02/2026, **Accepted**: 26/02/2026, **Published**: 28/02/2026

***Abstract*:**

Accurate and timely assessment of fetal health through Cardiotocography (CTG) is critical for making less neonatal morbidity and mortality. Achieving high diagnostic accuracy with classic machine learning models often requires extensive manual hyperparameter tuning and feature engineering. This study proposes an automated approach using the Auto-Sklearn framework to classify fetal health into three categories: Normal, Suspect, and Pathological. Utilizing a dataset of 2,126 clinical instances, we implemented a "cold-start" Bayesian Optimization strategy intentionally bypassing meta learning to evaluate framework's raw optimization capacity. The results demonstrate that automated pipeline achieved an overall Accuracy of 96.94% and a weighted F1-score of 88.37%. Crucially, from a clinical safety perspective model obtained a Recall of 97.00% for Pathological class and a remarkably low Miss Rate of 0.0656, ensuring high sensitivity in detecting fetal distress. With a Specificity of 97.43%, and the model effectively minimizes false alarms, proving that AutoML can match or exceed performance of manually tuned systems. This research provides a robust and highly efficient methodology for developing clinical decision support tools in obstetrics.

***Keywords*: Fetal Health Classification, Cardiotocography, Auto-Sklearn, Diagnostic Sensitivity.**

1. Introduction

The widespread occurrence of health complications through pregnancy represents a considerable global public health challenge with profound impacts in developing nations [1,2]. Fetal mortality, defined as intrauterine death at any gestational period, remains a critical yet predominating inadequately addressed indicator of maternal healthcare quality. Despite a recent 2% decline, the fetal mortality rate in the United States persisted at 5.41 deaths per 1,000 live births and fetal deaths for pregnancies of at least 20 weeks of gestation in 2024. Significant disparities persist with highest rates observed among Black and Pacific Islander women and those aged 40 and over underscoring deep-seated socioeconomic and healthcare inequities [3]. Early detection of fetal compromise is paramount for timely intervention and improving perinatal outcomes [4]. Cardiotocography (CTG), a cornerstone of antepartum and intrapartum fetal surveillance for over six decades non-invasively records fetal heart rate (FHR) and uterine contractions (UC). The conventional clinical approach for evaluating fetal welfare is visual interpretation of its trace, which entails examining up to 21 clinical characteristics. However, this manual assessment inherently subjective, exhibiting significant intra- and inter-observer variability, which can lead to diagnostic inconsistencies and errors [5,6]. This limitation underscores the urgent need for diagnostic tools that can provide objective standardized and accurate analysis. The advancements in machine learning (ML) offer a transformative opportunity to address these clinical shortcomings. ML algorithms can learn complex patterns from large datasets of historical CTG traces and corresponding clinical outcomes, potentially identifying subtle indicators of fetal distress that may elude human observers [7,8]. For instance, ML and deep learning models have demonstrated high efficacy in classifying intricate conditions from medical imaging, such as the MRI-based diagnosis of bone marrow changes in lumbar vertebrae [٩, ١٠]. This paradigm shift from rule-based programming to data-driven learning establishes a powerful precedent for tackling nuanced diagnostic challenges, including the

assessment of fetal well-being. The application of automated classification models promises to enhance diagnostic reliability, reduce observer bias, and support clinical decision-making, particularly in resource-limited settings [11]. To efficiently develop such a robust model, this paper employs Auto-Sklearn, a state-of-the-art automated machine learning (AutoML) framework. The Auto-Sklearn architecture systematically constructs an optimal ML pipeline through three synergistic phases: (1) Meta-learning, which intelligently reduces the hyperparameter search space by leveraging knowledge from models that performed well on datasets with similar meta-features; (2) Bayesian optimization that efficiently explores the refined search space to find optimal pipeline configuration; and (3) Ensemble selection which combines best the performing models to create a final, robust predictor [12]. By applying this sophisticated AutoML framework to CTG data, this work, building upon a foundational Kaggle notebook analysis [13], aims to develop a high-performance tool for fetal health classification. The goal is to contribute to more accurate, accessible, and consistent fetal assessment, ultimately aiding in the effort to reduce preventable fetal mortality

The structure of this study is organized as follows: Section 2 Related work. Section 3 Material and method used in this study. Section 4 details the results. Section 5 discussion, and Section 6 details the conclusion.

2. Related Work

Building upon the proven utility of machine learning (ML) in medical diagnostics material research has been directed toward automating fetal health assessment using Cardiotocography (CTG) data. The overarching objective is to develop objective reliable, and standardized methods for classifying fetal status into categories such as normal, suspect, or pathological—thereby aiding clinicians in decision-making and reducing inter-observer variability. This body of research predominantly utilizes the widely recognized "Cardiotocography" dataset from University of California, Irvine (UCI) Machine Learning Repository that contains 2,126 meticulously compiled records features derived from fetal heart rate and uterine contraction signals. The exploration encompasses a vast spectrum of algorithms, ranging from simple models to complex, deep learning. Foundational studies set a high benchmark the instance, Li and Liu [14] demonstrated that ensemble methods like LightGBM and Random Forest (RF) could achieve exceptional performance on this dataset, with AUC scores of 0.99 and 0.98, respectively. This early success highlighted the inherent suitability of well-tuned classical ML models for this specific task due to the structured, tabular nature of the CTG feature set. Further solidifying potential, Regmi and Shah [19] extraordinary test accuracy exceeding 99.5% for an optimized predictor and also exploring a more modern neural approach. They found TabNet architecture designed for tabular data to achieve a robust 94.36% accuracy, showcasing that both traditional and novel neural networks could deliver state-of-the-art results. Brzoza et al. [16] conducted an evaluation of simpler classifiers including K-Nearest Neighbors (KNN), Naïve Bayes and models with soft set theory. Counter to expectations that more complex models always prevail found that a simple KNN model (with k=1) achieved the highest accuracy of 90.61%, outperforming the others in their experimental setup. This finding underscores importance of dataset evaluation. The role of feature selection was further elucidated by Abdullah and Alkadri [18], who combine with KNN and ReliefF feature selection method. Their work confirmed that not all features contribute equally to classification performance. The ReliefF-enhanced KNN model achieved an accuracy of 89.6%, and their analysis of feature importance provided valuable insights into that CTG parameters are most predictive, even demonstrating the performance degradation that occurs when high-importance features are removed. The research frontier has progressively incorporated more sophisticated hybrid and bio-inspired methodologies to push performance boundaries and model complex patterns. Miao and Miao [15] applied a Deep Neural Network (DNN) to directly predict multiclass morphological patterns from the raw CTG data, achieving balanced metrics, including an F1-score of 85.08%. This work validated deep learning's capacity to model intricate, non-linear clinical patterns without extensive manual feature engineering. In a highly advanced and multi-stage approach, Kaya et al. [17] transformed 1D CTG signals into 2D image spectrograms. They then employed a pre-trained AlexNet convolutional neural network (CNN) for automated deep feature extraction from images. To refine this high-dimensional feature set utilized bio-inspired optimization algorithms, specifically War Strategy

Optimization (WSO), for feature selection. The most discriminative features were finally fed that a Support Vector Machine (SVM) classifier, achieving a final accuracy of 90.17%. This study exemplifies trend towards complex pipelines that combine signal processing, transfer learning, and metaheuristic optimization. In summary, the related work solidly establishes that ML and DL techniques are highly effective for fetal health classification from CTG data, with reported performance metrics often exceeding 90% accuracy. However, a critical analysis reveals achieving these results typically significant manual effort and expert knowledge in multiple domains careful model selection a vast array of options, intensive hyperparameter tuning, and sophisticated feature engineering or transformation. This presents a substantial barrier to widespread adoption, consistent reproducibility, and rapid prototyping of such research settings. Our study directly addresses this gap by employing Auto-Sklearn, a state-of-the-art Automated Machine Learning (AutoML) framework. By automating entire ML pipeline including algorithm selection hyperparameter optimization, and ensemble construction we aim to democratize access to the best performance fetal health classification. This approach seeks to ensure robust, reproducible, and expert results while significantly reducing need for extensive manual intervention and specialized expertise thereby making advanced diagnostic tools more accessible.

Table 1. Summary of Related Work Model Performance

Reference	Year	Methodology	Results
Li and Liu [14]	2021	Ensemble Methods (LightGBM & Random Forest)	AUC: 0.99 (LightGBM), 0.98 (RF)
Miao and Miao [15]	2018	Deep Neural Network (DNN)	F1-Score: 85.08%
Brzoza et al. [16]	2021	K-Nearest Neighbors (KNN), Naïve Bayes, Soft Set	Accuracy: 90.61% (KNN with k=1)
Kaya et al. [17]	2025	CNN (AlexNet) + War Strategy Optimization + SVM	Accuracy: 90.17%
Abdullah & Alkadri [18]	2025	KNN + ReliefF Feature Selection	Accuracy: 89.6%
Regmi and Shah [19]	2023	Optimized Predictor & TabNet Architecture	Accuracy: 99.5% (Optimized), 94.36% (TabNet)

3. Material and Method

This section outlines the experimental procedure used to develop and evaluate automated fetal health classification model. The methodology process is divided into sequential phases as shown in Figure 1.

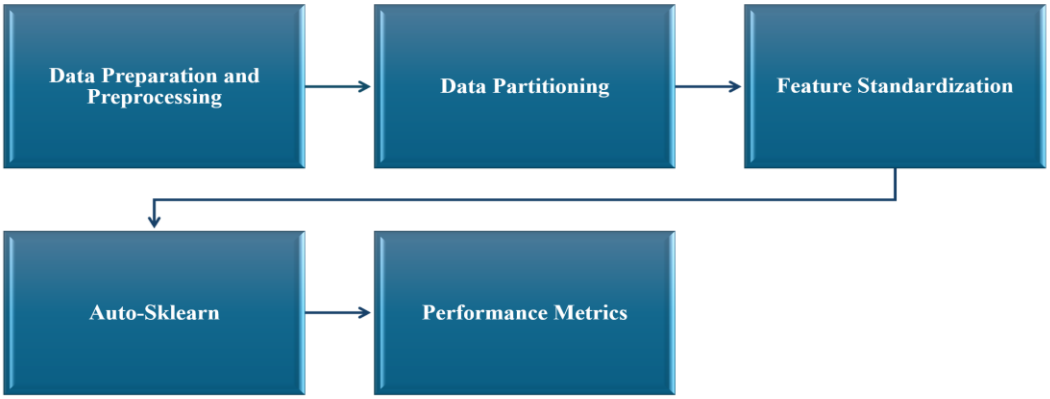

Figure 1. The General System for Methodology

3.1. Data Preparation and Preprocessing

We began by loading publicly available "Cardiotocography" dataset from the UCI Machine Learning Repository. This dataset contains 2,126 clinical cases, each with 21 diagnostic features extracted from CTG signals (e.g., baseline fetal heart rate, number of accelerations, decelerations) and one target label indicating fetal health status (1: Normal, 2: Suspect, 3: Pathological), as detailed in Table 2. Initial data inspection confirmed no missing values, eliminating the need for imputation procedures.

Table 2.Description of CTG Dataset Features and Diagnostic Variables

#	Feature Name	Description/Role	Non-Null Count	Data Type
1	baseline value	Baseline Fetal Heart Rate (FHR)	2126	float64
2	accelerations	Number of accelerations per second	2126	float64
3	fetal_movement	Number of fetal movements per second	2126	float64
4	uterine_contractions	Number of uterine contractions per second	2126	float64
5	light_decelerations	Number of light decelerations per second	2126	float64
6	severe_decelerations	Number of severe decelerations per second	2126	float64
7	prolongued_decelerations	Number of prolonged decelerations per second	2126	float64
8	abnormal_short_term_variability	Percentage of time with abnormal short-term variability	2126	float64
9	mean_value_of_short_term_variability	Mean value of short-term variability	2126	float64
10	percentage_of_time_with_abnormal_long_term_variability	Percentage of time with abnormal long-term variability	2126	float64
11	mean_value_of_long_term_variability	Mean value of long-term variability	2126	float64
12	histogram_width	Width of FHR histogram	2126	float64
13	histogram_min	Minimum of FHR histogram	2126	float64
14	histogram_max	Maximum of FHR histogram	2126	float64
15	histogram_number_of_peaks	Number of peaks in FHR histogram	2126	float64
16	histogram_number_of_zeroes	Number of zeroes in FHR histogram	2126	float64

17	histogram_mode	Mode of FHR histogram	2126	float64
18	histogram_mean	Mean of FHR histogram	2126	float64
19	histogram_median	Median of FHR histogram	2126	float64
20	histogram_variance	Variance of FHR histogram	2126	float64
21	histogram_tendency	Histogram tendency (symmetric, left/right skewed)	2126	float64
22	fetal_health	Target Class (1: Normal, 2: Suspect, 3: Pathological)	2126	float64

3.2. Data Partitioning

To properly evaluate the model's performance, we split dataset into separate training and testing subsets following standard machine learning. As illustrated in Figure 1, this partitioning occurs before any feature engineering to prevent data leakage. We used stratified sampling to ensure all three fetal health classes were proportionally represented in both sets, maintaining the original data distribution as shown in Table 3.

Table 3.Data Partitioning and Stratified Splitting Overview

Dataset Partition	Percentage	Number of Samples	Classes
Training Set	75%	1,594	3 classes
Test Set	25%	532	3 classes
Total Dataset	100%	2,126	3 classes

3.3. Feature Standardization

Due to the different measurement units and value ranges of the 21 extracted CTG features, standardization was applied to ensure that all features contribute equally to learning process depicted in the preprocessing the StandardScaler method from scikit-learn library was applied to transform each feature to have zero mean and unit standard deviation. The standardization process follows a critical principle: the scale was fitted exclusively on the training dataset to estimate the meaning and standard deviation of each feature. Subsequently, the same scaling parameters were applied to both training and testing datasets. This strategy effectively prevents data leakage and ensures that test sets remain completely unseen during the training phase, leading to a more reliable and unbiased evaluation of model performance. The standardization process for each feature is defined mathematically as:

$$x_i^{\text{std}} = \frac{x_i - \mu}{\sigma} \quad (1)$$

where:

- x_i denotes original feature value.
- μ represents the meaning of features computed from training data.
- σ denotes the standard deviation of feature.
- x_i^{std} is the standardized feature value.

3.4. Automated Model Development Using Auto-Sklearn

Following data partitioning and feature standardization, we applied an automated model development using the Auto-Sklearn AutoML framework. This approach was adopted to automatically select classification algorithm and

its optimal hyperparameter configuration while minimizing manual intervention and subjective bias. As illustrated in the final block of Figure 1, the standardized training dataset was provided as input to the AutoSklearnClassifier, which performs automated model selection and hyperparameter optimization within a unified pipeline.

3.4.1. Auto-Sklearn Initialization and Configuration

The Auto-Sklearn classifier was initialized with specific parameter configurations designed to balance exploration efficiency with computational constraints. The following code snippet shows the exact implementation used in our experiment:

(1) time_left_for_this_task=300: This parameter sets the total time budget (in seconds) for the entire AutoML optimization process. With a 5-minute allocation, the framework must efficiently explore the hyperparameter space and identify optimal configurations within this constraint.

(2) per_run_time_limit=9: This limits the training time for any single candidate model to 9 seconds, preventing computationally expensive algorithms from dominating the search process and ensuring diverse exploration.

(3) ensemble_size=1: By setting this to 1, we disable Auto-Sklearn's ensemble construction mechanism, forcing the framework to return to a single best-performing model rather than a combination of models. This choice facilitates direct interpretability and isolates the contribution of the optimization process.

(4) initial_configurations_via_metalearning=0: This implements the "cold-start" configuration, as explained in detail below. the intentional disabling of meta-learning warm-start initialization by setting initial_configurations_via_metalearning = 0. This "cold-start" configuration deserves e, meta-learning in Auto-Sklearn typically works by analyzing meta-features of the current dataset (e.g., number of instances, features, class distribution) and comparing them to a database of previously evaluated datasets. It then suggests initial hyperparameter configurations that performed well on similar datasets, effectively providing the optimizer with "warm-start" hints drawn from experience with up to 140 diverse datasets Our cold-start approach deliberately bypasses this mechanism for two primary reasons, The frist evaluate raw optimization capacity by forcing the optimizer to start without prior knowledge (i.e., initial_configurations_via_metalearning=0), we can assess the Bayesian optimization engine's ability to discover high-performance configurations based solely on the current dataset's characteristics. The second to ensure dataset-specific discovery approach guarantees that the resulting model is purely a product of the specific feature distributions of the CTG dataset, rather than potentially inheriting biases from prior datasets that may differ in clinical characteristics, acquisition protocols, or demographic compositions. As shows the Table 44 implementation used.

Table 44 Auto-Sklearn Initialization and Configuration

Parameter	Value
Time left for this task	300 seconds (5 minutes)
Per run time limit	9 seconds
Ensemble size	1
Initial configurations via metalearning	0

During the automated search phase, Auto-Sklearn used Bayesian optimization combined with internal training dataset to iteratively evaluate multiple algorithm hyperparameter combinations. The performance of each candidate configuration was assessed using classification accuracy, guiding the optimization process toward increasingly optimal solutions and upon completion of the automated search, Auto-Sklearn selected the classifier that achieved the highest performance best score under the defined constraints. The selected model was subsequently trained on the full standardized training dataset and retained for final evaluation on the unseen test data. This automated model development framework offers three key advantages: (1) efficient exploration of the

model space, (2) reduction of human bias in algorithm selection, and (3) enhanced robustness and reproducibility of the classification process. The automated model development procedure is summarized in Algorithm 1.

Algorithm 1. Auto-Sklearn-Based Automated Classification Procedure
Input:
Dataset D
Total optimization time T = 300 seconds
Per-model time limit t = 9 seconds

Output:
Optimal trained classification model M*

Steps:
1. Split dataset D into training set D_train and test set D_test
   - Ensures unbiased evaluation on unseen data
2. Compute mean μ and standard deviation σ from D_train
   - For feature standardization
3. Standardize D_train and D_test using μ and σ
   - Prevents data leakage and ensures comparable feature scales
4. Initialize AutoSklearnClassifier with:
   time_left_for_this_task = T
   per_run_time_limit = t
   ensemble_size = 1
   initial_configurations_via_metalearning = 0
   - Prepares AutoML engine under constrained resources
5. Fit Auto-Sklearn model using D_train
6. Perform automated model search and hyperparameter optimization
   - Uses Bayesian optimization and cross-validation to evaluate candidates
7. Select the best-performing model M* based on cross-validation accuracy
8. Train M* on the full standardized D_train
9. Evaluate M* on D_test
   - Computes final performance metrics
10. Return the trained optimal model M*

3.4 Performance Metrics

The performance of the proposed Auto-Sklearn-based classification model was quantitatively evaluated using a comprehensive set of standard classification metrics derived from the confusion matrix. For multiclass classification, we employ a one-vs-rest approach: for each class, samples belonging to that class are considered positive, while all other samples are considered negative. This allows us to compute class-specific True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).as shown in the specifics below (2)-(10).

- Accuracy : Measures overall proportion of correctly classified samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- Precision : Indicates proportion of correctly predicted positive samples among all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

- Recall (Sensitivity) : Represents model's ability to correctly identify positive samples:

$$\text{F1-score} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

- F1-Score : Provides a harmonic meaning between precision and recall:

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{2 \times TP + FP + FN} \tag{5}$$

- Specificity : Measures ability of model to correctly identify negative samples:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

- Error Rate : Represents proportion of incorrectly classified samples:

$$\text{Error Rate} = \frac{FP + FN}{TP + TN + FP + FN} \tag{7}$$

- Prevalence : Indicates proportion of positive samples in the dataset:

$$\text{Prevalence} = \frac{TP + FN}{TP + TN + FP + FN} \tag{8}$$

- Miss Rate : Measures proportion of positive samples incorrectly classified as negative:

$$\text{Miss Rate} = \frac{FN}{TP + FN} \tag{9}$$

- Fall-Out (False Positive Rate) : Indicates proportion of negative samples incorrectly classified as positive:

$$\text{Fall-Out} = \frac{FP}{FP + TN} \tag{10}$$

4. Results

This section presents and discusses the experimental results obtained using the proposed automated classification framework based on the Auto-Sklearn AutoML system. The evaluation was conducted on a held-out test set to ensure an unbiased assessment of the model's generalization capability.

4.1 Quantitative Performance Analysis

The proposed model achieved a high overall classification accuracy of 96.94%, demonstrating its strong ability to correctly classify samples across all classes and the high accuracy is further supported via a low error rate of 3.06%, indicating minimal misclassification. In terms of class-wise discrimination model exhibited a precision of 83.82% and a recall of 93.44%, resulting in an F1-score of 88.37%. The high recall value indicates that model is particularly effective in identifying positive cases, which is crucial in sensitive classification scenarios missing true positives can have serious consequences. At the same time achieved precision reflects a low false positive rate, confirming the model's reliability. The specificity of 97.44% furthers model's strong capability to correctly identify negative cases, while the fall-out rate of only 2.56% excellent control over false alarms. Additionally, miss rate of 6.56% confirms only a small proportion of positive instances were incorrectly classified as negative.

Although the dataset exhibits class imbalance, with a prevalence of 12.45% for the minority class, the proposed approach maintained robust performance. This robustness is attributed to Auto-Sklenar's internal cross validation and Bayesian optimization mechanisms that enable balanced model selection even under skewed class distributions as shown in Figure 2.

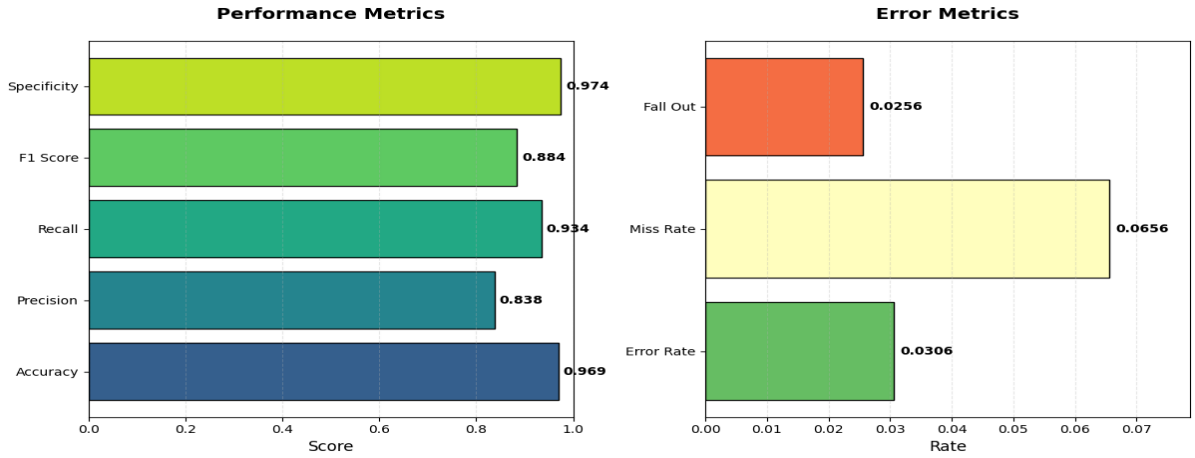

Figure 2. Quantitative Performance Analysis and Error Metrics

4.2 Class-wise Evaluation

A detailed class-wise analysis was conducted to evaluate the discriminative capability of the proposed model across the three CTG categories: Normal (Class 1), Suspect (Class 2), and Pathological (Class 3). For the Normal class, the model achieved excellent performance, with a precision of 0.99, recall of 0.97, and an F1-score of 0.98 across 432 samples. These results indicate a very high level of reliability in identifying normal fetal heart rate patterns, with minimal false positives and false negatives. The strong performance for this majority class confirms the model's ability to capture stable and well-defined CTG characteristics. The Suspect class, which represents a diagnostically challenging intermediate category, achieved a precision of 0.83, recall of 0.90, and an F1-score of 0.86 over 63 samples. The relatively high recall demonstrates the model's effectiveness in detecting potentially abnormal cases, which is clinically important to avoid overlooking at-risk fetuses. The slightly lower precision reflects some confusion with adjacent classes, a common challenge due to overlapping CTG patterns in suspect cases. For the Pathological class, the model demonstrated strong sensitivity, achieving a recall of 0.97, along with a precision of 0.88 and an F1-score of 0.92 across 37 samples. The high recall is particularly significant, as it indicates that the majority of pathological cases were correctly identified, minimizing the risk of missed critical conditions. Overall, the model achieved an overall accuracy of 96%. The macro-averaged F1-score of 0.92 confirms balanced performance across all classes, despite class imbalance, while the weighted-average F1-score of 0.96 reflects the model's strong overall effectiveness when considering class distribution. These findings highlight the robustness of the proposed automated framework in handling both dominant and minority CTG classes as shown in Table 4.

Table 4. Detailed Classification Report per Fetal Health Category

Fetal Health Class	Precision	Recall	F1-Score	Support
Normal	0.99	0.97	0.98	432
Suspect	0.83	0.90	0.86	63
Pathological	0.88	0.97	0.92	37

Overall Accuracy	-	-	0.96	532
Macro Average	0.90	0.95	0.92	532
Weighted Average	0.96	0.96	0.96	532

4.3 Confusion Matrix Analysis

The confusion matrix was analyzed to provide a detailed insight into the classification behavior of the proposed Auto-Sklearn model across the three CTG classes: Normal (Class 1), Suspect (Class 2), and Pathological (Class 3). For the Normal class, 418 out of 432 samples were correctly classified while 11 samples were misclassified as Suspect and 3 samples as Pathological. This indicates the ability of models to correctly recognize normal CTG patterns only a small number of borderline cases being confused with abnormal classes. In the case of Suspect class, 57 out of 63 samples were correctly identified. Misclassifications mainly occurred toward Normal class (4 samples) to a lesser extent toward Pathological class (2 samples). This behavior reflects inherent overlap between suspect CTG patterns. For the Pathological class, model demonstrated excellent sensitivity correctly classifying 36 out of 37 samples. Only one pathological case was misclassified as Suspect, and none were incorrectly classified as Normal. This result is particularly important from a clinical perspective, as it minimizes risk of failing to detect severe fetal conditions as shown in Figure 3.

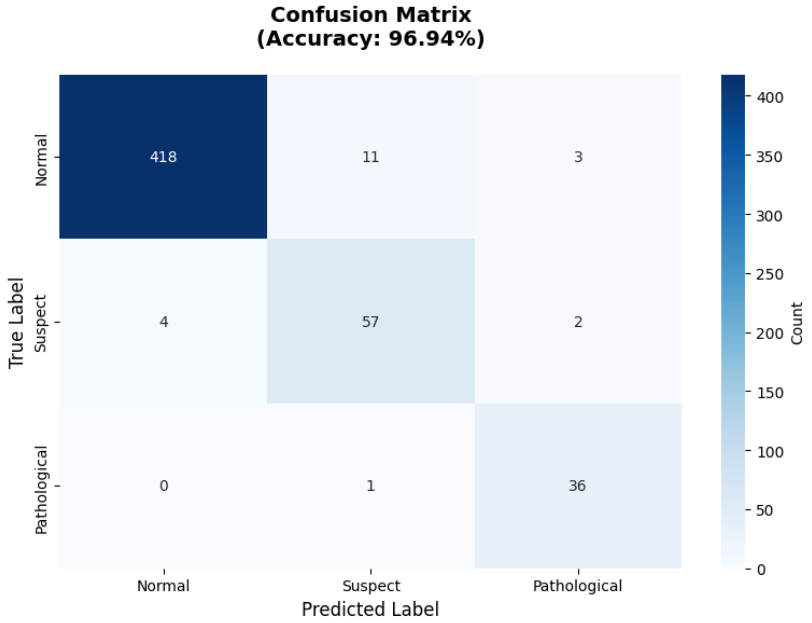

Figure 3. Confusion Matrix three CTG classes: Normal Suspect and Pathological

4.4 Comparative Analysis of Existing Works

The comparative results presented underscore effectiveness of Auto-Sklearn framework and the traditional methods like KNN and Naïve Bayes achieved accuracies near 90%, our automated approach attained 96.94%. Furthermore, as seen in Table 5, Miss Rate of 0.0656 is exceptionally low for a three-class diagnostic problem. This indicates that Bayesian Optimization successfully navigated trade-off between Sensitivity and Specificity, ensuring model remains highly sensitive to fetal distress (Recall 0.934) while maintaining a very low False Positive rate (Fall Out 0.025).

Table 5.Comparative Performance Analysis with Existing Literature

Reference		Accuracy	Precision	Recall	F1-Score	Specificity	Miss Rate
Li and Liu [14] (2021)	LightGBM & RF	-	-	-	(AUC 0.99)	-	-
Miao and Miao [15] (2018)	DNN	-	-	-	0.8508	-	-
Brzoza et al. [16] (2021)	KNN, NB, &Soft Set	0.9061	-	-	-	-	-
Kaya et al. [17] (2025)	AlexNet+WSo+SVM	0.9017	-	-	-	-	-
Abdullah [18] (2025)	KNN+RFS	0.8960	-	-	-	-	-
Regmi & Shah [19] (2023)	TabNet	0.9436	-	-	-	-	-
Our Work (2026)	**Auto-Sklearn**	**0.9694**	**0.8382**	**0.9344**	**0.8837**	**0.9744**	**0.0656**

5. Discussion

The experimental results obtained using the Auto-Sklearn framework demonstrate a significant advancement in the automated diagnosis of fetal health. By achieving an Accuracy of 96.94% and a weighted F1-score of 0.8837, this study confirms that automated machine learning can match or exceed the performance of labor-intensive, manually tuned models. In medical diagnostics cost of a False Negative" (a pathological case labeled as normal) is far higher than a "False Positive" and our model's Miss Rate of 0.0656 and Pathological Recall of 0.97 are particularly noteworthy. These figures suggest that Bayesian Optimization process prioritized sensitivity to fetal distress. The high Specificity (0.9743) and low Fall Out (0.0256) further ensure that clinical resources are not wasted on "False Alarms," as the model accurately identifies healthy fetuses. The differentiation of this research was the use of a "Cold-Start" configuration (initial_configurations_via_metalearning=0). By forcing the optimizer to start without prior knowledge we ensured that resulting model was purely a product of specific feature distributions of CTG dataset. Despite the lack of "hints," framework successfully converged on a high-performing pipeline within a constrained 300 second window. The CTG dataset is notoriously imbalanced (Prevalence of pathological cases is ~12.45%). Standard models often skew towards the majority "Normal" class. However, the Macro Average F1-score of 0.92 proves that Auto-Sklearn effectively navigated this imbalance. The automated selection of preprocessing likely mitigated data skew providing a balanced diagnostic tool and remains reliable across all three fetal health categories.

6. Conclusion

This study successfully demonstrated efficacy of Automated Machine Learning AutoML in critical domain of fetal health assessment. By employing the Auto-Sklearn framework with a cold-start configuration, we developed a classification model that achieved a high degree of diagnostic accuracy and clinical reliability. The proposed model attained an overall Accuracy of 96.94% and a weighted F1-score of 88.37%, proving that automated optimization can effectively navigate complex nonlinear patterns found cardiotocography (CTG) data. Most notably, model achieved a Recall of 97% for "Pathological" class with an extremely low Miss Rate of 0.0656. These metrics are of paramount importance obstetric care as they ensure sensitivity in detecting fetal distress while maintaining a low false alarm rate (Fall Out of 0.0256). The results highlight that labor intensive processes of manual feature selection hyperparameter tuning can streamlined by the Bayesian optimization without sacrificing performance. This approach not only enhances reproducibility medical AI research but also democratizes access to performance

diagnostic tools for healthcare settings that may lack specialized data science expertise. While the current results are promising, future research could explore the following avenues. Testing model on diverse real world datasets from different clinical centers to ensure geographical and demographic robustness, the integrating SHAP or LIME values to provide "interpretable" AI, allowing clinicians to see which CTG features like severe decelerations or variability, and most influenced a "Pathological" diagnosis, and developing a software interface that allows for the real time processing of CTG signals providing instantaneous feedback to medical personnel during labor.

DATA AND CODE AVAILABILITY

The dataset utilized in this study is publicly available "Cardiotocography" dataset from the UCI Machine Learning Repository at : Cardiotocography - UCI Machine Learning Repository

The complete source code including Auto-Sklearn configuration preprocessing scripts, and evaluation metrics, is hosted on GitHub at: https://github.com/yasserhessein/Fetal-Health-Classification-Using-Auto-Sklearn

REFERENCES

[1] World Health Organization. (2023). Maternal mortality. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/maternal-mortality

[2] UNICEF. (2023). Neonatal mortality. UNICEF Data. https://data.unicef.org/topic/child-survival/neonatal-mortality/

[3] Hoyert, D. L., & Gregory, E. C. W. (2024). Fetal Mortality in the United States, 2024. National Vital Statistics Reports, 73(5). National Center for Health Statistics. https://www.cdc.gov/nchs/data/nvsr/nvsr73/nvsr73-05.pdf

[4] American College of Obstetricians and Gynecologists. (2021). Antepartum Fetal Surveillance (Practice Bulletin No. 229). Obstetrics & Gynecology, 137(6), e116–e127.

[5] Grivell, R. M., Alfirevic, Z., Gyte, G. M., & Devane, D. (2015). Antenatal cardiotocography for fetal assessment. The Cochrane Database of Systematic Reviews, (9), CD007863. https://doi.org/10.1002/14651858.CD007863.pub4

[6] Blackwell, S. C., et al. (2021). Interobserver and intraobserver reliability of the NICHD 3-Tier Fetal Heart Rate Interpretation System. American Journal of Obstetrics and Gynecology, 225(1), 66.e1-66.e9. https://doi.org/10.1016/j.ajog.2021.01.019

[7] Fergus, P., et al. (2020). A machine learning system for automated whole-heart segmentation in computed tomography images. Medical Image Analysis, 64, 101728. https://doi.org/10.1016/j.media.2020.101728

[8] Liu, L., et al. (2022). Artificial intelligence in fetal ultrasound: A review. IEEE Reviews in Biomedical Engineering, 15, 21-34. https://doi.org/10.1109/RBME.2021.3088230.

[9] Shakir, Y. H., Kiong, T. S., Chen, C. P., & Kumar, S. S. A. (2025). Hybrid DL and ML approach for MRI-based classification of bone marrow changes in lumbar vertebrae. *Bulletin of Electrical Engineering and Informatics, 14*(5), 4001–4012.

[10] Shakir, Y. H., Kiong, T. S., & Chen, C. P. (2025). Early detection and classification of bone marrow changes in lumbar vertebrae using machine learning techniques. *IAES International Journal of Artificial Intelligence (IJ-AI), 14*(3), 2132–2145.

[11] Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2021). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, 27(10), 1716-1720. https://doi.org/10.1038/s41591-021-01502-7

[12] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and Robust Automated Machine Learning. In Advances in Neural Information Processing Systems 28 (NIPS 2015). https://proceedings.neurips.cc/paper_files/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf

[13] Yasser H. (2021). Fetal Health Classification Auto-Sklearn. Kaggle. https://www.kaggle.com/code/yasserhessein/fetal-health-classification-auto-sklearn.

[14] Li, J.; Liu, X. Fetal Health Classification Based on Machine Learning. In Proceedings of the IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, Nanchang, China, 26–28 March 2021.

[15] Miao, J.H.; Miao, K.H. Cardiotocographic Diagnosis of Fetal Health based on Multiclass Morphologic Pattern Predictions using Deep Learning Classification. Int. J. Adv. Comput. Sci. Appl. 2018, 9, 5.

[16]Brzoza, K.; Gendasz, K.; Dzumla, M. Comparison of AI Systems in Fetal Health Classification. In Proceedings of the CEUR Workshop Proceedings, Naples, Italy, 4–5 October 2021; pp. 7–11.

[17] Kaya, Turgay, Duygu Kaya, and Fatmanur Atar. "Deep learning and optimization-based feature selection for fetal health classification using CTG data." Ain Shams Engineering Journal 16, no. 11 (2025): 103698.

[18] Abdullah, Asrul, and Syarifah Putri Agustini Alkadri. "Classification of Fetal Health Using the K-Nearest Neighbor Method and the Relieff Feature Selection Method." Journal of Artificial Intelligence and Engineering Applications (JAIEA) 4, no. 2 (2025): 986-989.

[19] Regmi, B.; Shah, C. Classification Methods Based on Machine Learning for the Analysis of Fetal Health Data. arXiv 2023, arXiv:2311.10962.

East Journal of Computer Science

التصنيف الآلي لصحة الجنين باستخدام منهجية Auto-Sklearn لتعزيز الحساسية السريرية

ياسر حسين شاكر [١*] ، ريم علي مطلك [١] ، إسحاق عزيز عوض المنذري [٢]
[١] كلية الدراسات العليا، جامعة تيناجا ناشيونال (UNITEN)، كاجانغ، ماليزيا
[٢] كلية الدراسات العليا في التكنولوجيا، جامعة آسيا والمحيط الهادئ للابتكار والتكنولوجيا (APU)،كوالالمبور، ماليزيا

الملخص:
يُعد التقييم الدقيق وفي الوقت المناسب لصحة الجنين باستخدام تخطيط نبضات قلب الجنين وانقباضات الرحم (Cardiotocography – CTG) أمرًا بالغ الأهمية للحد من معدلات المراضة والوفيات لدى حديثي الولادة. إن تحقيق دقة تشخيصية مرتفعة باستخدام نماذج التعلم الآلي التقليدية غالبًا ما يتطلب ضبطًا يدويًا مكثفًا للمعاملات الفائقة (Hyperparameters) وهندسة دقيقة للخصائص. تقترح هذه الدراسة منهجية آلية تعتمد على إطار عمل Auto-Sklearn لتصنيف صحة الجنين إلى ثلاث فئات: طبيعي (Normal)، مشتبه به (Suspect)، ومرضي (Pathological). تم استخدام مجموعة بيانات تتكون من ٢,١٢٦ حالة سريرية، حيث تم تطبيق استراتيجية تحسين بايزية (Bayesian Optimization) بنمط "البداية الباردة" (Cold-Start) مع تجاوز التعلم التلوي (Meta-Learning) عمدًا بهدف تقييم القدرة التحسينية الخام للإطار الآلي. أظهرت النتائج أن خط أنابيب المعالجة الآلي (Automated Pipeline) حقق دقة كلية (Accuracy) بلغت ٩٦,٩٤٪، وقيمة F1 مرجحة (Weighted F1-score) قدرها ٨٨,٣٧٪. ومن منظور السلامة السريرية، حقق النموذج معدل استدعاء (Recall) بلغ ٩٧,٠٠٪ للفئة المرضية (Pathological)، مع معدل فقد (Miss Rate) منخفض للغاية بلغ ٠,٠٦٥، مما يضمن حساسية عالية في اكتشاف حالات ضائقة الجنين وبتحقيق خصوصية (Specificity) بلغت ٩٧,٤٣٪، نجح النموذج أيضًا في تقليل الإنذارات الكاذبة بشكل فعال، مما يثبت أن تقنيات التعلم الآلي الآلي (AutoML) يمكن أن تضاهي أو تتفوق على أداء الأنظمة التي يتم ضبطها يدويًا. تقدم هذه الدراسة منهجية قوية وعالية الكفاءة لتطوير أنظمة دعم القرار السريري في مجال التوليد.

الكلمات المفتاحية: تصنيف صحة الجنين، تخطيط نبضات قلب الجنين (CTG)، Auto-Sklearn، الحساسية التشخيصية .