JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predicting Fetal Health Using Cardiotocography Data with Kolmogorov-Arnold Networks

by
Mark Louie San Juan
Christian James D. Villapando

ABSTRACT

Investigating fetal health classification with UCI cardiotocography (CTG) data, this study employed a pipeline integrating advanced machine learning models (MLP, LSTM, KAN) and oversampling techniques (ADASYN, SMOTE, SMOTEENN). Results consistently showed that oversampling significantly improved evaluation metrics, with SMOTEENN being particularly effective. The MLP model, when paired with SMOTEENN, achieved the best overall performance, including 97.82% accuracy and high F1-score, ROC-AUC, and MCC, alongside an efficient runtime. While Kolmogorov-Arnold Networks (KAN) also improved with SMOTEENN (94.04% accuracy), they lagged behind MLP in both performance and efficiency. CTG is a critical diagnostic tool in obstetrics for monitoring fetal health, and accurate classification of its signals is essential for timely medical intervention. These findings highlight the critical role of robust oversampling and appropriate model selection for high-performing and interpretable healthcare AI.

Keywords: Fetal health classification, Cardiotocography (CTG), Machine learning, MLP, LSTM, KAN, Oversampling, ADASYN, SMOTE, SMOTEENN, Class imbalance, Diagnostic tool, Healthcare AI.

CHAPTER 1

INTRODUCTION

1.0. Summary of Introduction

The study explores machine learning applications in healthcare. This paper explores prenatal care using cardiotocography (CTG) data to predict fetal health. CTG monitors fetal heart rate and uterine contractions, and ML models can assist in analyzing CTG outputs to identify potential health issues. However, discrepancies in the performance of ML models, such as Artificial Neural Networks (ANNs), across various studies highlight challenges, including data dependency, class imbalance, computational costs, and the curse of dimensionality. These issues hinder ML's consistent effectiveness in healthcare.

To address these limitations, this research examines Kolmogorov-Arnold Networks (KANs), a newer neural network model that offers flexibility in high-dimensional data through spline-based functions. KANs aim to mitigate challenges faced by other existing neural networks, like Multi-Layer Perceptrons (MLPs), particularly in handling the curse of dimensionality and enhancing computational efficiency. The study will compare KANs with existing neural networks to evaluate their performance in CTG classification. Key objectives include establishing standardized practices for model implementation and data preprocessing.

The significance of this study lies in advancing ML applications in fetal health assessment, potentially improving prenatal diagnostics and patient outcomes. Limitations include dataset size constraints, privacy considerations, focus on physiological CTG data, and potential noise due to CTG machine variability. By investigating KANs, this research contributes to the growing field of ML in healthcare, seeking more reliable and efficient solutions for critical diagnostic tasks.

1.1. Background of the Study

Machine learning holds immense promise in healthcare by improving diagnostic accuracy, enabling individualized treatment plans, and assisting with administrative efficiencies—ultimately enhancing patient outcomes through data-driven insights and predictive analytics (Gala, Behl, Shah, & Makaryus, 2024). These algorithms analyze large-scale clinical datasets to identify hidden patterns, forecast potential health risks, and support medical decision-making. One notable application of this technology is classification using cardiotocography signals, where machine learning models can automate fetal health assessment to complement traditional clinical evaluation.

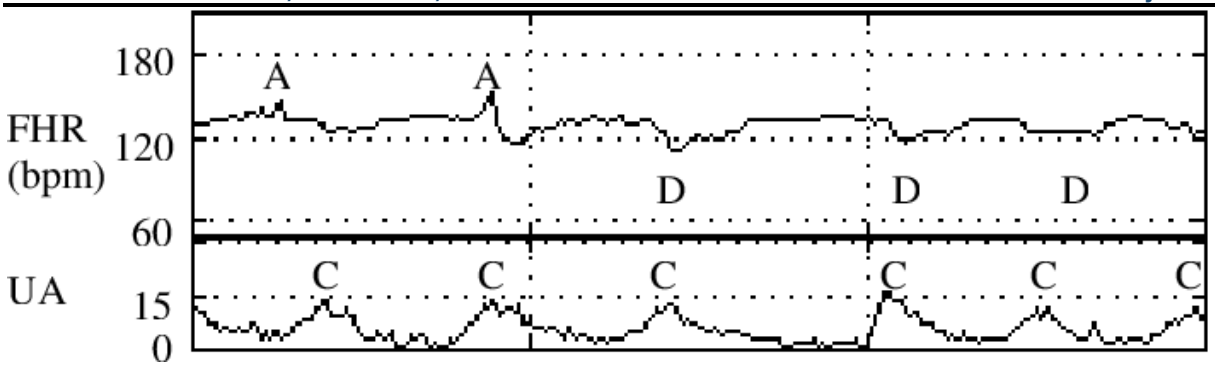


Figure 1: Example Image of Cardiotocogram

Cardiotocography (CTG) is a medical test used during pregnancy to monitor the fetal heartbeat and uterine contractions, providing a continuous record on a strip of paper or computer. It is commonly performed in the last trimester to assess the baby's well-being, especially if there are concerns like slowed fetal movements, growth issues, placental problems, or complications such as diabetes or high blood pressure. CTG is also used during labor to monitor the baby's response to contractions, helping healthcare providers detect signs of fetal distress and make informed decisions to ensure the baby's safety (Parashar, 2023). CTG results are categorized as reassuring, suspicious, or pathological. Reassuring results indicate no issues, while suspicious or pathological results suggest the baby may be in distress. In such cases, further assessments are done, and interventions, such as immediate delivery through forceps, vacuum, or Caesarean section, may be recommended to ensure the baby's safety (McKechnie, 2024).

Multiple studies have already implemented classification machine learning models to test which has performed the best. In one study, the researchers included Logistic Regression, KNN, Random Forest, and Gradient Boosting Machine (GBM), in predicting fetal health using CTG data (A. Kumar Pradhan et al., 2021). In another, Artificial Neural Network (ANN), Long-short Term Memory (LSTM), XGBoost (XGB), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Light GBM (LGBM), and Random Forest (RF) (Rahmayanti et al., 2022). SVM and ANN have been explored before in a study by Chinnasamy & Devi (2014). However, these studies have varying results.As advantageous machine learning models are, they are not without their drawbacks. Data dependency, overfitting and underfitting, ethical concerns, and high computational costs are some of the disadvantages. (GeeksforGeeks, 2024).

In the paper of Rahmayanti et al., (2022) ANN has performed poorly with low precision (0.00 - 0.19) across all three classes, low F1 score (0.10 - 0.29) across three scenarios, and low accuracy (0.11 - 0.37). In contrast, in the paper of Chinnasamy & Devi (2014) ANN has a higher precision (0.58 - 0.97) in all three classes, higher F1 score (0.45 - 0.98), and a high rand index

score (0.38 - 0.99). One of the reasons why there is a significant difference in precision and F1 score in these two studies can be attributed to the data used in each study (class distribution, quality, feature distribution, etc.). It is highly possible that the data utilized is different, considering the gap between the studies. One of the studies was fairly recently done in 2022, and the other was done in 2014. However, it is suspect as to why the study done in 2014 has better metric scores than the one done more recently. One would assume technological developments and mathematical advancements would provide a clearer view of the ANN model. The possibilities in the discrepancy could be attributed to many factors. Therefore, an updated testing is required. Moreover, the preprocessing, hyperparameters, and evaluation methods between the studies is likely to have differed. Creating a standard of these factors will provide a fairer glimpse of the model. High accuracy and a high rand index usually goes hand in hand as it would mean that the model is well-calibrated, consistent with true underlying structure and labeling, and has a balanced class distribution. Nonetheless, it is not always the case. The difference between these two studies would suggest imbalance in class distribution. In the paper of Chinnasamy & Devi (2014), class 2 or the "suspect" class has a notable decrease in all metrics. From 0.96 in class 1 (normal) to 0.58 in class 2 (suspect) to 0.97 in class 3 (pathological) in precision, and again 0.97 in class 1 to 0.45 in class 2 to 0.97 in class 3 in F1 score. Implementation of resampling techniques should solve this class imbalance. It is to be acknowledged that other machine learning models could have suffered the same issues with ANN. Therefore, an application of all the models used in previous studies is in need to be re-explored.

High computational costs stand in the way in producing a more compelling model. Large-scale machine learning (ML) is costly because it requires a lot of processing power, a large amount of high-quality data, hiring skilled personnel, and maintaining infrastructure. ML models are more expensive because they need extensive processing, labor-intensive mathematical calculations, and continuous inference. Deep learning models become costly for large-scale applications as a result (GeeksforGeeks, 2024). From the study of Chinnasamy & Devi (2014), the neural network ran for approximately 2.6 seconds, comparing that to the next slowest model used, the Fuzzy C-Means which ran for around 0.3 second, which means that the neural network was at least 8.5 times slower. It proves to be extremely inefficient when dealing with extensive amounts of data in hopes of building a compelling model. Not only will it take a long time to train and test the data, the performance is also at risk. There exists a term called the curse of dimensionality, where as the volume of our dataset grows exponentially with increasing dimensions, resulting in sparser data. This causes issues with data sparsity, computation, overfitting, distance meaning loss, deterioration in algorithms that use distance measurements, and visualization issues. High-dimensional data increases the computer resources and processing time needed for data processing, making operations like clustering and classification difficult. In order to create a good machine learning model, this issue stands in the way.

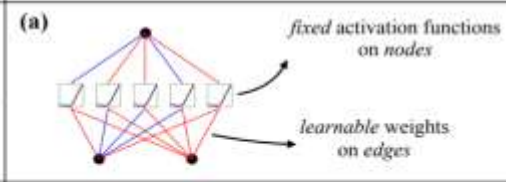
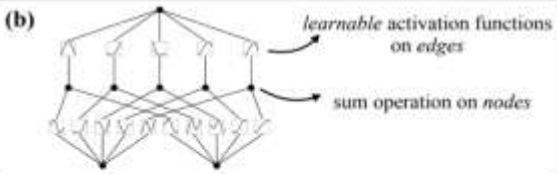
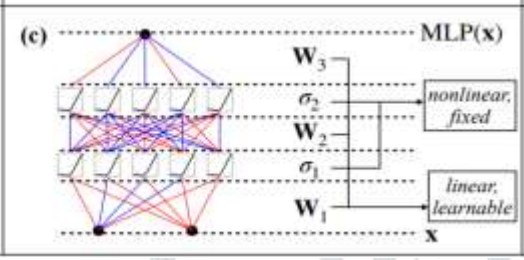
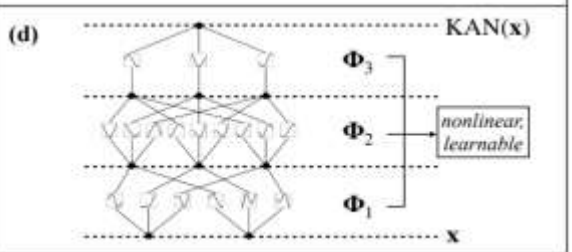
Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	Universal Approximation Theorem	Kolmogorov-Arnold Representation Theorem
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(\epsilon)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$
Model (Shallow)		
Formula (Deep)	$\mathrm{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\mathrm{KAN}(\mathbf{x}) = (\mathbf{\Phi}_3 \circ \mathbf{\Phi}_2 \circ \mathbf{\Phi}_1)(\mathbf{x})$
Model (Deep)		

Table 1: Multilayer Perceptrons (MLPs) vs Kolmogorov-Arnold Networks (KANs)

Kolmogorov-Arnold Networks (KANs), inspired by the Kolmogorov-Arnold representation theorem, offer a promising alternative to the traditional Multi-Layer Perceptrons (MLPs) used in machine learning. Unlike MLPs, which use fixed activation functions on nodes, KANs introduce learnable activation functions on edges, replacing linear weight parameters with spline-based functions. This unique structure enables KANs to be more flexible and efficient in fitting functions, particularly in small-scale AI and scientific tasks (Liu et al., 2024).

One of the major advantages of KANs is their ability to handle the curse of dimensionality (COD) more effectively than MLPs. While splines, used in KANs, struggle with COD in high-dimensional problems, the combination of splines with MLPs allows KANs to manage both compositional structures and univariate functions with high accuracy (Liu et al., 2024). In essence, KANs offer the best of both worlds: the ability of MLPs to learn features and the accuracy of splines in optimizing univariate functions. KANs are also more interpretable compared to MLPs. Since the activation functions are placed on edges and can be visualized as splines, they offer a clearer understanding of how inputs are transformed through the network. This interpretability makes KANs particularly useful for scientific discovery, where understanding the model's behavior is as important as its predictive accuracy. KANs can help researchers rediscover mathematical and physical laws, as demonstrated in various AI and science tasks (Liu et al., 2024). Another advantage is KANs' efficiency in neural scaling. They have been shown to scale faster and with greater accuracy than MLPs when increasing the size of the model. For example, KANs display better scaling laws in terms of loss reduction with added parameters, compared to MLPs, which plateau more quickly and often require larger networks to achieve similar levels of accuracy (Liu et al., 2024). However, KANs are not without their

drawbacks. The primary challenge lies in their complexity. Since every weight in a KAN is replaced by a spline function, the number of parameters can be significantly larger than in MLPs, leading to higher computational demands. Despite requiring fewer overall parameters due to their efficiency in smaller networks, KANs can still become computationally expensive as their grid size increases, particularly when fine-graining splines to achieve higher accuracy (Liu et al., 2024). Considering KAN is still in the early stages of its technological development, there is a need for a study that extracts its otential through a relatively small yet standardized dataset. Furthermore, a comparative analysis of KAN against machine learning models with the same techniques employed will contribute to the existing body of knowledge that can assess the efficacy of KAN.

1.2. Statement of the Problem

The integration of machine learning in healthcare, particularly through the use of cardiotocograms (CTGs) to monitor fetal health, has demonstrated significant potential in improving diagnostic accuracy and patient outcomes. However, existing studies reveal inconsistencies in the performance of various machine learning models, particularly the Artificial Neural Network (ANN). Discrepancies in results across studies, likely due to differences in data quality, preprocessing methods, and evaluation techniques, underscore the need for further investigation into these models' efficacy. Moreover, the high computational costs and challenges associated with the curse of dimensionality hinder the widespread application of effective machine learning models in this context. Kolmogorov-Arnold Networks (KANs) present a promising alternative to traditional models like Multi-Layer Perceptrons (MLPs), offering improved flexibility and efficiency in handling high-dimensional data. This study aims to explore the performance of KANs in comparison to existing models, addressing gaps in the current research landscape while considering the implications of data quality, preprocessing standards, and computational efficiency.

1.3. Significance of the Study

This study holds significant implications for the healthcare industry, particularly in the realm of prenatal care. By investigating the effectiveness of KANs in predicting fetal health using CTG data, the research aims to enhance the diagnostic capabilities of machine learning models. Improved accuracy in fetal health assessment can lead to timely interventions, ultimately reducing the risk of complications during pregnancy and labor. Furthermore, this study seeks to contribute to the existing body of knowledge on machine learning applications in healthcare by providing a comparative analysis of KANs and traditional models. The findings can inform future research and clinical practices, promoting the development of more effective and cost-efficient machine learning solutions in healthcare.

1.4. Objectives of the Study

1. Predict Fetal Health with Cardiotocography Data

- Create classification models (discussed on section 3.5) for cardiotocogram data that predicts fetal health.

2. Evaluate the Performance of KANs Compared to Other Neural Network Models

- Assess the evaluation of Kolmogorov-Arnold Networks in classifying fetal health based on CTG data.

3. Propose Standards for Implementation

- Develop guidelines for preprocessing, hyperparameter tuning, and evaluation metrics to standardize future research.

1.5. Research Questions

1. How does the predictive performance of Kolmogorov-Arnold Networks compare to neural network machine learning models (e.g., KAN, LSTM, MLP) in classifying fetal health from CTG data?

2. What impact does the quality and distribution from resampling of CTG data have on the performance of machine learning models?

3. What standardized practices can be developed to ensure consistency in the implementation of machine learning models for fetal health prediction?

1.6. Scope and Limitations

This study focuses solely on CTG data, excluding external factors such as maternal health, environmental influences, and medical interventions. Future studies could integrate such data to provide a more holistic view of fetal health. Additionally, exploring advanced data augmentation techniques and hybrid models may yield further improvements in predictive accuracy. This study will focus on comparing the performance of KANs with other machine learning algorithms for fetal health classification using CTG data throughout the resampled data. The research will be limited to a specific dataset and may not generalize to all possible CTG data variations. The study will also explore the impact of data preprocessing techniques and address potential interpretability challenges associated with the models. However, it may not delve into the clinical implications of the findings in detail.

1. Dataset Limitations

The dataset's size may influence model performance, with larger datasets likely providing better generalization, while smaller datasets may lead to overfitting. The results will be interpreted with this limitation in mind.

2. Privacy and Compliance

The study ensures compliance with privacy regulations, such as HIPAA, by anonymizing patient data and adhering to data security protocols. Ethical standards will be maintained throughout the project.

3. Fetal Physiology Focus

The research centers on physiological signals from CTG and does not consider external factors, such as maternal health or environmental influences, which could provide additional insights into fetal well-being.

4. CTG Machine Variability

Differences in CTG machine performance and potential human errors during data collection are not accounted for in detail, though they may introduce noise or bias into the dataset.

1.7. Definition of Terms

1. Cardiotocography (CTG): A monitoring technique used during pregnancy to record fetal heart rate and uterine contractions for assessing fetal well-being.
2. Multilayer Perceptron (MLP): A type of feedforward neural network capable of learning non-linear relationships in structured data through multiple layers of interconnected neurons.
3. Long Short-Term Memory (LSTM): A recurrent neural network architecture designed to learn and retain long-range dependencies in sequential data using memory cells and gating mechanisms.
4. Kolmogorov–Arnold Network (KAN): A neural architecture based on the Kolmogorov–Arnold theorem that models multivariate functions using additive compositions of univariate functions for improved interpretability.
5. Synthetic Minority Oversampling Technique (SMOTE): A data-level balancing method that generates synthetic examples by interpolating between minority class instances.
6. Adaptive Synthetic Sampling (ADASYN): An oversampling technique that focuses on generating synthetic data in regions where the minority class is harder to learn.
7. SMOTE with Edited Nearest Neighbor (SMOTE-ENN): A hybrid technique that combines SMOTE oversampling with ENN filtering to remove noisy or ambiguous samples.
8. Principal Component Analysis (PCA): A dimensionality reduction technique that transforms original features into uncorrelated components that capture the most variance in the data.
9. Varimax Rotation: A method used in PCA or factor analysis to make component loadings more interpretable by maximizing the variance of squared loadings.

10. Z-Score Standardization: A scaling method that transforms data to have a mean of 0 and standard deviation of 1, making features comparable.

11. Kaiser-Meyer-Olkin (KMO) Test: A statistical measure that assesses the adequacy of data for factor analysis by evaluating sampling suitability.

12. Bartlett's Test of Sphericity: A test used to determine whether variables are sufficiently correlated for structure detection methods like PCA or factor analysis.

13. Correlation Filtering: A feature selection method that removes highly correlated variables to reduce redundancy and multicollinearity.

14. F1-Score: The harmonic mean of precision and recall, used to evaluate classification performance under imbalanced data conditions.

15. Matthews Correlation Coefficient (MCC): A robust performance metric that considers all four confusion matrix elements, providing a balanced evaluation even with class imbalance.

16. Area Under the ROC Curve (ROC-AUC): A threshold-independent metric that evaluates a model's ability to distinguish between classes.

17. Precision: The ratio of true positive predictions to all predicted positives, indicating the accuracy of positive predictions.

18. Recall (Sensitivity): The ratio of true positive predictions to all actual positives, reflecting a model's ability to detect relevant cases.

19. Accuracy: The proportion of correct predictions over all predictions made by a model.

20. K-Fold Cross-Validation: A resampling method that partitions the dataset into K subsets to train and validate the model K times for more reliable performance estimates.

21. Hyperparameter Tuning: The process of optimizing non-learnable model settings (e.g., learning rate, number of layers) to improve performance.

22. Oversampling: A technique used to address class imbalance by increasing the number of minority class instances in the training data.

23. Feature Engineering: The process of selecting, transforming, or creating input features to improve model performance.

24. Model Deployment: The act of integrating a trained machine learning model into a production or clinical environment for real-world use.

25. Curse of Dimensionality: A phenomenon where increasing the number of features in a dataset leads to sparsity, degraded model performance, and difficulty in learning due to the exponential growth of the feature space.

CHAPTER 2

REVIEW OF RELATED LITERATURE

2.0. Summary of Review of Related Literature

Fetal health monitoring using Cardiotocography (CTG) plays a vital role in assessing fetal well-being during pregnancy and labor. However, manual interpretation of CTG signals is often prone to subjectivity, inter-observer variability, and delayed decision-making—challenges that machine learning aims to address by offering consistent, data-driven classifications. This review explored the state of the literature on CTG classification, focusing on essential components such as data preprocessing, model selection, oversampling strategies for class imbalance, and the use of robust evaluation metrics. Emphasis was placed on techniques like Z-score normalization, feature reduction through PCA, data balancing via ADASYN and SMOTE-ENN, and fair evaluation using metrics like F1-score and MCC. The methods and models chosen—particularly LSTM and Kolmogorov–Arnold Networks (KAN)—are justified based on their ability to capture complex patterns and adapt to tabular biomedical data, supporting more accurate and interpretable fetal health predictions.

2.1. Application Domain: Cardiotocography (CTG)

Cardiotocography (CTG) is a continuous electronic record of the fetal heart rate obtained via an ultrasound transducer placed on the mother's abdomen (external or indirect CTG). A second transducer is placed on the mother's abdomen over the uterine fundus to record simultaneously the presence of any uterine activity. Both fetal heart rate and the uterine activity are traced simultaneously onto a paper strip. The underlying theoretical concept for the use of CTG in pregnancy is that it is a screening test for the identification of babies with acute or chronic fetal hypoxia, or at risk of developing hypoxia. Fetal hypoxia, a critical condition characterized by an insufficient supply of oxygen to the developing fetus, can impair normal fetal development and significantly increase the risk of perinatal and infant mortality (Okoroafor et al., 2024; Pattison et al., 2015). This oxygen deprivation is a major global health concern, with birth asphyxia, a direct consequence of severe hypoxia, accounts for an estimated 23% of all newborn deaths worldwide (Beletu et al., 2022; Kiiza et al., 2025; World Health Organization, n.d.). Components of the fetal heart rate that can be assessed include: baseline rate, baseline variability, accelerations and decelerations. The relationship between fetal heart rate and the timing of uterine contractions are also assessed (Pattison et al., 2015).

Interpreting Cardiotocography (CTG) presents significant challenges for clinicians due to the inherent subjectivity involved. A primary concern is that the substantial inter- and intra-observer variability, meaning different healthcare professionals, or even the same professional at different times, may interpret the same CTG trace differently (Al Shamandy et

al., 2023). This variability arises partly from the reliance on visual pattern recognition and the existence of multiple, sometimes conflicting, classification guidelines for CTG features. The complexity of fetal physiology and its dynamic response to stress further complicates interpretation, as subtle changes in heart rate patterns can signify evolving hypoxia, requiring a nuanced understanding that is often difficult to standardize across all practitioners (Chandraharan et al., 2017). Furthermore, factors such as poor signal quality, maternal or fetal movement artifacts, and confusion between maternal and fetal heart rates can obscure the true fetal status, leading to misinterpretation and potentially inappropriate clinical actions.

Beyond the clinical realm, the application of machine learning to CTG interpretation faces its own set of hurdles, primarily concerning data quality and model reliability. These challenges include the scarcity of large, high-quality, and comprehensively annotated datasets that are representative of diverse clinical scenarios, leading to limitations in the generalizability and robustness of developed models. A further critical issue is defining the 'ground truth' for fetal well-being; while objective measures like umbilical cord blood pH are considered gold standards for identifying acidosis, subjective metrics such as Apgar scores are also frequently employed, introducing potential inconsistencies in training data (Park et al., 2024). Moreover, the inherent imbalanced nature of CTG datasets, where pathological cases are far less frequent than normal or suspect cases, poses a significant challenge for machine learning algorithms, often necessitating specialized techniques to prevent models from simply predicting the majority class. Finally, ensuring the interpretability and explainability of machine learning models is crucial for clinical adoption, as clinicians require an understanding of a model's reasoning to trust and integrate its predictions into their decision-making processes (O'Sullivan et al., 2021).

The Cardiotocography (CTG) dataset from the UCI Machine Learning Repository is a widely used benchmark for developing and evaluating machine learning models aimed at fetal health classification. This dataset comprises 2,126 fetal cardiotocograms, each annotated by expert obstetricians and categorized into one of three classes: Normal, Suspect, or Pathological (UCI Machine Learning Repository, 2008). These classifications are grounded in clinical guidelines established by the International Federation of Gynecology and Obstetrics (FIGO) and the National Institute of Child Health and Human Development (NICHD). A Normal tracing is defined by a baseline fetal heart rate (FHR) between 110–160 beats per minute (bpm), moderate variability of 6–25 bpm, absence of late or variable decelerations, and a contraction frequency of ≤5 per 10 minutes (NICHD, 2010). A Suspect (intermediary) classification applies when one of these parameters deviates slightly—such as minimal variability, tachycardia, or non-repetitive decelerations—without the presence of high-risk signs. Conversely, a Pathological (abnormal) tracing includes more severe features such as absent variability, recurrent late decelerations, bradycardia (<100 bpm), or the appearance of a sinusoidal pattern, all of which may indicate fetal distress requiring clinical intervention (FIGO, 2015). These objective standards support the reliability of the CTG dataset's labeling and its suitability for training and validating supervised learning algorithms.

Over the past decade, various machine learning (ML) techniques have been applied to the UCI CTG dataset to classify fetal health states into Normal, Suspect, or Pathological categories. Mehbodniya et al. (2021) evaluated multiple models including Support Vector Machines (SVM), Random Forest (RF), Multilayer Perceptrons (MLP), and K-Nearest Neighbors (KNN), identifying RF as the best-performing model across accuracy, precision, recall, and F1-score. Similarly, Alam et al. (2022) compared RF, Logistic Regression, Decision Trees, SVM, Voting Classifier, and KNN, and reported that RF achieved the highest classification accuracy of approximately 97.5%. Silwattananusarn et al. (2020) further improved CTG classification by applying ensemble feature selection methods (such as information gain and correlation-based selection) before feeding the features into an ensemble of SVM classifiers, resulting in enhanced predictive performance. These studies confirm that classical ML algorithms—particularly ensemble-based classifiers—provide strong baseline results for CTG classification due to their robustness and suitability for tabular data..

In recent years, deep learning (DL) approaches have been increasingly adopted to automate and improve fetal health classification. Park et al. (2025) conducted a large-scale multicenter study using a dataset of over 22,000 CTG recordings and trained a deep neural network classifier that achieved ROC-AUC scores between 0.86 and 0.90 across internal and external validation cohorts. Fergus et al. (2019) developed a hybrid model combining a one-dimensional convolutional neural network (1D-CNN) with a multilayer perceptron to classify segmented CTG signals, eliminating the need for manual feature extraction. Bertieaux et al. (2022) explored deep unsupervised learning using models like GANomaly for anomaly detection in CTG signals and demonstrated that such architectures can generalize well to unseen abnormalities. Collectively, these studies highlight the evolving trend toward more scalable, end-to-end deep learning pipelines in CTG research, particularly in scenarios involving large or unstructured datasets.

2.2. Data Preprocessing and Feature Extraction Techniques

Multivariate analysis (MVA) is essential in biomedical research due to the complex and interdependent nature of medical measurements—such as biomarkers, imaging signals, or clinical lab values—where variables often interact to produce meaningful outcomes. Unlike univariate tests that probe one variable at a time, MVAs allow simultaneous examination of multiple variables, facilitating the detection of patterns and relationships that emerge only through joint analysis (Smith & Zheng, 2021). For instance, Smith and Zheng (2021) demonstrated the application of principal component analysis (PCA), factor analysis (FA), cluster analysis, multiple linear regression, and machine learning classifiers (e.g., random forests and support vector machines) for diagnosing hearing and vestibular disorders. They highlighted how MVAs effectively avoid the inflation of Type I error rates—a common problem when conducting multiple univariate analyses—by considering covariance structures and inter-variable dependencies.

Z-score standardization transforms each feature in a dataset to have a mean of zero and a standard deviation of one. This practice is particularly crucial when features differ in scale or units—common in biomedical datasets—because many multivariate techniques, including principal component analysis (PCA) and clustering, assume variables contribute equally to the analysis (de Amorim et al., 2022). For instance, in Leiden-based radiomics and proteomics data, z-score normalization outperformed other methods (e.g., min-max or quantile scaling) by improving feature comparability and downstream model reliability (Demircioğlu et al., 2023). Without this transformation, features with larger numeric ranges can inadvertently dominate model training, leading to biased results. Furthermore, in genome-wide microarray studies, z-score normalization is a standard procedure to stabilize variance across thousands of gene expression measurements, enabling fair comparison across samples and ensuring that subsequent statistical analyses (e.g., fold-change detection or PCA) produce reliable outcomes (Cheadle et al., 2003). This normalization is both intuitive and practical: it adjusts each gene's expression relative to its own distribution, mitigating technical artifacts and aligning data for cross-sample comparisons. Importantly, centralizing and scaling features via z-score ensures that covariance-based algorithms (like PCA or factor analysis) use the correlation structure rather than being skewed by raw data scales—preserving interpretability and validity in biomedical inference (de Amorim et al., 2022).

Dropping low-variance and fully-null features is a foundational step in data preprocessing that helps reduce dimensionality without losing informative patterns. Features with zero or near-zero variance offer little to no discriminative power, as their values remain constant or nearly constant across all samples—making them irrelevant for most classification or regression tasks (Bommert et al., 2020). Similarly, fully-null features—columns with all missing values—do not contribute any useful information and can hinder statistical estimation or model convergence. In biomedical datasets, where high dimensionality is common (e.g., gene expression, CTG features, or clinical lab measurements), eliminating such features not only improves computational efficiency but also enhances the stability of multivariate methods such as PCA, clustering, or neural network training (Raschka & Mirjalili, 2019). This step ensures that subsequent analyses focus on variables that contain meaningful variability and potential associations with the target outcome.

Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy are two critical diagnostic tools used to evaluate whether a dataset is suitable for multivariate factor analysis or principal component analysis (PCA). Bartlett's Test examines the hypothesis that the correlation matrix is an identity matrix, meaning variables are uncorrelated and thus inappropriate for structure detection. A statistically significant result (typically $p < 0.05$) indicates the presence of meaningful correlations among variables, justifying the use of dimensionality reduction methods (Field, 2013). Complementing this, the KMO index assesses the proportion of variance among variables that might be common variance—

suggesting that the dataset contains underlying factors worth extracting. Values closer to 1.0 (typically ≥ 0.80) reflect adequate sampling and stronger inter-item partial correlations, whereas values below 0.50 signal unsuitability for factor analysis (Williams, Onsman, & Brown, 2010). In biomedical applications such as psychological surveys, clinical diagnostics, and physiological signal analysis, these tests are routinely employed to validate the assumption of latent structure before applying extraction methods like PCA or exploratory factor analysis. The **Scree Plot** is a graphical tool used in principal component analysis (PCA) and factor analysis to determine the optimal number of components to retain. It plots the eigenvalues of the components in descending order against their component number. Typically, components with larger eigenvalues explain more variance. The "elbow" or inflection point in the plot—where the curve begins to level off—indicates the point beyond which additional components contribute little additional explanatory power (Cattell, 1966). This visual criterion supports a data-driven decision for dimensionality reduction, especially in biomedical datasets where interpretability and variance retention must be balanced. In clinical or physiological applications, such as CTG data or multi-omics studies, the scree plot is commonly employed to ensure the retained components are both statistically meaningful and practically useful for downstream tasks such as classification or clustering (Zwick & Velicer, 1986).

Principal Component Analysis (PCA) is a widely used technique in multivariate statistics that transforms a high-dimensional dataset into a smaller set of uncorrelated variables called principal components. These components are linear combinations of the original features, ordered by the amount of variance they capture in the data. PCA is particularly valuable in biomedical research where datasets often contain numerous interrelated variables—such as physiological measurements or genomic profiles—making interpretation and modeling difficult without dimensionality reduction (Jolliffe & Cadima, 2016). However, while standard PCA captures maximal variance, the resulting components are often difficult to interpret due to the loadings being spread across many variables. To improve interpretability, Varimax rotation, an orthogonal rotation method, is often applied to the principal components. This technique maximizes the variance of squared loadings within each component, which tends to simplify the structure by making loadings more polarized—closer to 0 or 1—thus clarifying which variables are most influential for each component (Abdi & Williams, 2010). In biomedical applications such as cardiotocography (CTG) or psychological assessments, this rotation enables researchers to associate components more clearly with latent physiological or behavioral constructs. When combined, PCA with Varimax rotation provides both a reduced and interpretable representation of high-dimensional clinical data, which can be leveraged for classification, clustering, or feature engineering in machine learning pipelines.

Weighting and aggregation of rotated loadings is a crucial step in dimensionality reduction when the goal is to produce interpretable, composite scores from latent structures revealed through Principal Component Analysis (PCA) or Factor Analysis. After Varimax rotation is applied, each feature has a loading that quantifies its contribution to each latent factor. By

squaring these loadings and weighting them based on the variance explained by each component, we can compute an overall score or "importance" for each original feature (DiStefano et al., 2009). This enables the prioritization of features in terms of their explanatory power, allowing us to construct aggregated indices that better reflect the multidimensional structure of biomedical or behavioral datasets (Fabrigar et al., 1999). The OECD-JRC Handbook on Constructing Composite Indicators emphasizes that weighting based on PCA or factor analysis is one of the most statistically sound methods for composite score construction, particularly when the goal is to summarize complex phenomena using latent variables (Nardo et al., 2008). The Handbook outlines how this method can ensure internal consistency, reflect actual data variation, and align with the theoretical framework underpinning the analysis. Moreover, the use of weighted aggregation facilitates balanced interpretation and avoids overrepresentation of noisy or irrelevant features, making it a powerful approach for biomedical applications such as fetal health classification, where interpretability and accuracy are both essential.

2.3. Data Balancing Techniques in Imbalanced Medical Datasets

Class imbalance is a critical challenge in cardiotocography (CTG)-based fetal health monitoring, as pathological or suspect cases are much rarer than normal ones. This imbalance mirrors real-world incidence rates—for example, neonatal hypoxic–ischemic encephalopathy (HIE) affects about 1.5 per 1,000 live births in high-income countries—a typical minority scenario in CTG datasets (Kurinczuk, White-Koning, & Badawi, 2010). Machine learning models trained on such skewed data often report high accuracy by predicting the majority class, but fail to detect the minority cases—those most crucial for clinical decisions, risking missed diagnoses. Moreover, biomedical datasets like CTG commonly include noisy labels and overlapping minority-class examples, which complicate boundary learning. López, Fernández, García, Palade, and Herrera (2013) demonstrate that heterogeneity within minority-class examples and class overlap significantly degrade classifier performance—reducing recall and F1-scores—if not addressed via resampling strategies. Reviews of CTG-focused machine learning work echo these findings, noting that models struggle to generalize to rare pathological instances unless balanced via techniques like SMOTE, ADASYN, or GAN-based augmentation (López et al., 2013).

Adaptive Synthetic Sampling (ADASYN) is an advanced oversampling algorithm that generates synthetic minority class samples primarily in regions where learning is more difficult, rather than uniformly across the data space. Introduced by He et al. (2008), ADASYN calculates the local distribution of minority examples and adaptively assigns more synthetic samples to instances that are sparsely surrounded by their own class. This strategy encourages learning on borderline and underrepresented areas of the feature space, which can improve classifier sensitivity to minority cases (Fernández et al., 2018). However, an over-concentration of synthetic points near noisy or ambiguous instances may increase the risk of overfitting, especially in biomedical data where small clusters may not generalize well (Napierała & Stefanowski, 2016). Synthetic Minority Oversampling Technique (SMOTE), originally developed by Chawla et al. (2002), is one of the most widely used

data-level strategies for handling class imbalance. SMOTE creates synthetic data points by interpolating between randomly selected nearest neighbors in the minority class. This technique has proven highly effective in smoothing decision boundaries and improving classifier generalization across a variety of biomedical domains. For high-dimensional datasets, such as gene expression or physiological signal-based features, SMOTE has shown to reduce variance and bias by expanding the feature space without replicating instances, which reduces model overfitting (Blagus & Lusa, 2013). However, it may still introduce noise in overlapping class regions if not used carefully (Fernández et al., 2018). To mitigate SMOTE's limitations, SMOTE-ENN combines oversampling with Edited Nearest Neighbor (ENN) cleaning, forming a hybrid approach that both augments and filters the dataset. After SMOTE generates new synthetic instances, ENN removes samples—both synthetic and original—that are misclassified by their nearest neighbors, which typically eliminates noise and overlapping examples. This results in a cleaner decision boundary and improved class separability. In biomedical applications, SMOTE-ENN has been shown to outperform basic oversampling techniques by reducing false positives and enhancing the minority class's classification metrics (Fernández et al., 2018). However, SMOTE-ENN is computationally more intensive and may require tuning of the neighbor parameters to balance sensitivity and specificity effectively.

2.4. Deep Learning Models for Tabular Medical Data

Deep neural networks, such as Artificial Neural Networks (ANNs), serve as powerful tools for biomedical classification tasks involving structured data. Their ability to model nonlinear interactions among clinical features offers a significant advantage over traditional linear methods like logistic regression (Nguyen et al., 2013). For example, a study applying ANNs to predict cardiovascular autonomic dysfunction on clinical data demonstrated respectable diagnostic metrics (AUC = 0.76) and highlighted that ANNs often outperform linear counterparts in capturing complex patient patterns. A comprehensive review on omics-based disease detection also supports the use of feedforward networks, showing high classification accuracy in tasks ranging from cancer diagnosis to genetic profiling

Multilayer Perceptrons (MLPs) are among the simplest yet most widely employed ANN architectures in tabular health data contexts. A systematic review demonstrated an average classification accuracy of approximately 93.9% across non-imaging tasks, emphasizing their versatility and robust performance . Though less complex than convolutional or recurrent models, MLPs excel in modeling relationships within scale-normalized biomedical features—especially when paired with strong regularization methods (Kadra, Lindauer, & Hutter, 2021). In the context of CTG-based fetal health classification, Salini et al. (2024) proposed a dynamic MLP architecture incorporating multiple hidden layers, early stopping, and dropout regularization. Their model, applied to preprocessed and class-balanced CTG data, achieved a notable accuracy of 97%, highlighting the potential of MLPs in clinical classification tasks. These results suggest that a similarly optimized MLP can be

effectively applied to UCI CTG data to predict fetal health status, particularly when supported by appropriate preprocessing and balancing techniques.

Beyond classical temporal use cases, LSTM networks have also demonstrated strong performance on non-sequential or pseudo-temporal biomedical datasets, where features represent aggregated clinical snapshots rather than true time series. These aggregated features—such as repeated lab results or imaging-derived measurements—can still benefit from LSTM's internal memory and gating mechanisms, which implicitly capture dependencies across related features (Lipton, Kale, Elkan, & Wetzel, 2016). In a landmark study using ICU clinical data from the MIMIC-III database, LSTM models outperformed logistic regression and even matched MLPs trained on engineered features when utilizing deep supervision, demonstrating robust performance on tabular input formats (Lipton et al., 2016). Additionally, Shickel et al. (2018) reviewed the application of LSTMs in healthcare and noted their effectiveness in capturing complex patterns and interactions in clinical phenotyping tasks, even without explicit temporal order. Rao et al. (2024) developed a multi-scale LSTM model—trained on CTU-UHB CTG signals sampled at multiple time scales—that automatically classified fetal heart rate recordings with an accuracy of 85.73%, precision of 85.53%, recall of 86.13%, F1-score of 85.79%, and an impressive ROC-AUC of 0.918. The multi-scale approach notably outperformed single-scale LSTM variants, improving F1-score by over 2 points and AUC by more than 0.05 (Rao et al., 2024). Given these strong results, a similar multi-scale LSTM could be applied to UCI CTG data—especially if raw time-series CTG features are available—by adapting sampling strategies and leveraging LSTM's capacity to learn both short- and long-term heart rate dynamics for enhanced fetal health classification.

More recently, Kolmogorov–Arnold Networks (KANs) have emerged as a promising alternative for modeling tabular and structured data. Rooted in the Kolmogorov–Arnold representation theorem, KANs decompose multivariate functions into compositions of univariate functions plus sums, enabling sparse architectures with interpretable functional modules (Poeta et al., 2024). A benchmarking study on real-world tabular datasets showed that KANs often outperform or match MLPs in accuracy and F1 score, particularly for datasets with high dimensionality, though at slightly higher computational cost (Poeta, Giobergia, Pastor, Cerquitelli, & Baralis, 2024). These properties—efficiency in small-sample settings, clearer feature interpretability, and strong generalization—make KANs especially appealing for biomedical use cases where model transparency and reliability are essential. At the core of Kolmogorov–Arnold Networks (KANs) lies the principle of functional composition, wherein multivariate functions are approximated through sums and compositions of univariate functions, in line with the Kolmogorov–Arnold representation theorem (Liu et al., 2024). Unlike traditional dense neural networks that rely on matrix multiplications and irreversible activations, KANs replace each linear weight with a learnable spline-based univariate function on network edges. This architecture allows each edge to model localized nonlinear transformations adaptively, significantly reducing the overall parameter count while preserving modeling power (Liu et al., 2024). The result is a more

modular, efficient architecture tailored for tabular datasets, where patterns often vary across features and rigid transformations fall short.

A significant advantage of KANs over conventional MLPs and LSTMs is their inherent interpretability and parameter efficiency. By decomposing complex feature interactions into distinct spline functions, KANs facilitate direct analysis and visualization of how each input contributes to predictions—crucial in clinical scenarios where model transparency is essential (Vaca-Rubio et al., 2024). Empirical benchmarks show that smaller KANs often match or outperform larger MLPs in terms of accuracy and F1 scores, particularly in small-sample settings, without relying on deep architectural stacks or computationally expensive recurrent structures (Liu et al., 2024). Since this design fully embodies the Kolmogorov Superposition Theorem—which asserts any continuous multivariate function can be captured via sums and compositions of univariate functions—KANs represent a principled neural architecture for structured data analysis (Somvanshi et al., 2024).Kolmogorov–Arnold Networks (KANs) are a novel type of neural architecture designed to model multivariate functions through compositions of univariate functions and additive operations, offering an interpretable alternative to traditional fully connected networks. In a recent benchmarking study, Poeta et al. (2024) evaluated KANs across multiple real-world tabular datasets and found that they achieved competitive results compared to standard models, reporting high F1-scores and accuracy while maintaining lower parameter complexity. The study emphasized the transparency of KANs, particularly their spline-based functional components, which allowed for more interpretable insights into feature contributions. Given their performance and interpretability in structured data settings, KANs are well-suited for application to UCI CTG data, where understanding the influence of individual physiological features on fetal health classification is clinically valuable.

2.5. Evaluation Metrics and Model Performance

In the context of class-imbalanced problems—such as medical classification tasks where minority cases (e.g., pathological conditions) are rare but clinically significant—robust evaluation metrics are crucial for assessing true model performance. Overall accuracy can be misleading: a classifier that always predicts the majority class can still achieve high accuracy while failing entirely to detect minority instances. This risk is especially problematic in healthcare, where missing critical diagnoses can have serious consequences (Saito & Rehmsmeier, 2015). Consequently, alternative metrics—including precision, recall, F1-score, and area under the ROC curve (AUC-ROC)—are preferred in imbalanced settings to provide a more accurate and clinically safe performance evaluation. Precision and recall offer complementary insights into model behavior. Precision captures the proportion of true positive predictions among all positive predictions, reflecting predictive reliability. Recall (sensitivity), by contrast, measures the proportion of actual positives correctly identified, indicating how well the model detects the minority class (Fawcett, 2006). In many biomedical applications, recall is often prioritized due to the higher cost of false negatives. However, this focus on recall can reduce precision, leading to false alarms. Striking a balance between recall and

precision is therefore essential to ensure practical robustness. To address this balance, the F1-score—the harmonic mean of precision and recall—is commonly used in imbalanced classification tasks. The F1-score penalizes extreme values in either metric and gives a more holistic performance measure than accuracy or individual metrics alone (Saito & Rehmsmeier, 2015). It is particularly suitable for medical diagnostics, where both missed diagnoses (false negatives) and unnecessary treatments (false positives) are significant concerns. By integrating both recall and precision, the F1-score offers a reliable indicator of model effectiveness in clinically critical, imbalanced scenarios (PMC, 2023).

Matthews Correlation Coefficient (MCC) stands out as one of the most robust metrics for evaluating classification performance in the presence of class imbalance—such as in CTG-based fetal health prediction. Unlike accuracy or F1-score, which may overlook true negatives or emphasize only specific error types, MCC considers all four parts of the confusion matrix (TP, TN, FP, FN) and yields a value between −1 (completely wrong) and +1 (perfect prediction). Models that perform poorly on any quadrant of the confusion matrix see their MCC decline, making it a highly balanced metric (Chicco et al., 2021). In imbalanced biomedical datasets, MCC is particularly valuable because it remains reliable even when the majority class dominates, providing a nuanced gauge of true predictive performance. ROC-AUC (Receiver Operating Characteristic—Area Under the Curve) complements MCC by evaluating a model's ability to distinguish classes independently of any specific threshold. ROC-AUC reflects the trade-off between true positive and false positive rates across all possible decision thresholds. While ROC-AUC remains useful even in class-imbalanced scenarios, its insensitivity to class prevalence means that a model can achieve high AUC while still performing poorly on the minority class—a critical concern in medical diagnostics. Thus, in CTG modeling, ROC-AUC is best used alongside MCC to offer both ranking performance and balanced accuracy insights.

In fetal health prediction using cardiotocography (CTG), choosing the right evaluation metric is critical due to the imbalanced distribution of health states—where normal cases far outnumber pathological or suspect ones. Traditional metrics like overall accuracy often mask a model’s inability to detect minority-class instances, which are typically of the highest clinical importance. Therefore, CTG classification studies benefit from using metrics that capture nuanced performance across all classes. Metrics like Matthews Correlation Coefficient (MCC) are especially well-suited because they incorporate all elements of the confusion matrix, providing a balanced view of model effectiveness even under skewed data conditions (Chicco et al., 2021). Meanwhile, precision, recall, and F1-score allow for more targeted assessments of the minority class, which is vital when evaluating the model’s capacity to detect fetal distress. ROC-AUC can offer threshold-independent insight into separability between classes, but its interpretability diminishes when minority detection is the primary goal. Consequently, combining MCC with class-sensitive metrics like F1-score and precision-recall analysis yields a more comprehensive and clinically relevant evaluation framework for CTG classification.

2.6. Validation Strategy and Statistical Rigor

K-Fold Cross Validation is a widely used resampling strategy in machine learning that enables robust estimation of a model's generalization performance. It works by splitting the dataset into *K* equal subsets (folds), training the model on *K–1* folds, and validating it on the remaining fold. This process is repeated *K* times, each fold serving once as the validation set, and the performance metrics are averaged across folds. In biomedical machine learning—such as cardiotocography (CTG) classification—K-Fold Cross Validation mitigates the risk of evaluation bias associated with single holdout sets and helps ensure that the reported model performance reflects a more generalizable pattern across the data.

Choosing an appropriate value for *K* is critical, as it directly influences the bias-variance tradeoff. Lower *K* values (e.g., K=2) lead to higher variance in performance estimates due to smaller training sets per fold, while larger *K* values (e.g., K=10 or leave-one-out) reduce bias but increase computation time and may overfit small datasets. A value of K=5 is often adopted as a balanced compromise—minimizing both computational cost and performance instability, while offering sufficient training data per fold for biomedical datasets of moderate size. Moreover, K=5 has become a common standard in biomedical machine learning studies, striking a balance between evaluation reliability and reproducibility without the excessive overhead of 10-fold or LOOCV strategies (Varma & Simon, 2006). Beyond performance estimation, cross-validation plays a crucial role in ensuring reproducibility and statistical confidence in biomedical machine learning. Using K-Fold Cross Validation stabilizes evaluation metrics by ensuring that no single subset of the data overly influences the results. This is particularly important in imbalanced datasets like CTG, where a single train-test split may accidentally exclude rare pathological cases. By averaging metrics such as F1-score, MCC, or AUC over multiple folds, researchers gain more robust and statistically meaningful estimates, reducing the likelihood of reporting spurious performance due to chance (Wainer & Cawley, 2021).

Finally, cross-validation helps mitigate overfitting in model assessment, particularly when hyperparameter tuning is involved. Without proper cross-validation, models may overfit to a specific validation set, resulting in inflated metrics that do not translate to unseen data. Nested cross-validation or repeated K-Fold variants further improve evaluation fairness by isolating model selection from final testing phases. In clinical contexts where algorithmic decisions impact real-world outcomes—such as fetal health classification—these reproducibility safeguards are essential for ensuring the transparency, reliability, and clinical trustworthiness of the developed models.

2.7. Bayesian Optimization via Optuna

In this study, hyperparameter tuning was conducted using Bayesian optimization via the Optuna framework. Optuna utilizes a Tree-structured Parzen Estimator (TPE) approach to model the objective function and identify promising hyperparameter regions more efficiently than traditional exhaustive methods such as grid or random search. Additionally, Optuna's built-in

pruning mechanism allows for early stopping of underperforming trials, thereby reducing computational overhead and focusing resources on the most promising configurations (Akiba et al., 2019; Shekhar et al., 2022).

Recent literature underscores the superiority of Bayesian optimization, and Optuna in particular, for neural network tuning tasks. For example, Akiba et al. (2019) demonstrated that Optuna significantly reduced search times while achieving competitive or superior model accuracy compared to other methods. Similarly, comparative studies on hyperparameter optimization frameworks for deep learning applications have highlighted Optuna's ability to adaptively balance exploration and exploitation, yielding better-tuned models under limited computational budgets (Shekhar et al., 2022). Applications in specialized domains, such as geophysical inverse problems, have also reported substantial improvements in convergence speed and predictive performance when using Optuna for tuning complex architectures (Zhao et al., 2023).

CHAPTER 3

METHODOLOGY

3.0 Summary of Methodology

This study classified fetal health using UCI CTG data. After preprocessing, feature engineering, and oversampling (ADASYN, SMOTE, SMOTEENN), LSTM, MLP, and KAN models were tuned via Bayesian optimization. The MLP with SMOTEENN achieved the best performance, underscoring the value of thorough preprocessing and model tuning.

3.1. Machine Specifications

For Machine 1.

Component	Specification
Processor	AMD Ryzen 7 4800H with Radeon Graphics (2.90 GHz)
RAM	16.0 GB
Storage	453 GB
Graphics Card	NVIDIA RTX 3050 Laptop GPU
Operating System	Microsoft Windows 11

For Machine 2.

Component	Specification
Processor	AMD Ryzen 7 4800H with Radeon Graphics (2.90 GHz)
RAM	16.0 GB
Storage	512 GB
Graphics Card	NVIDIA GeForce RTX 3050 Ti Laptop GPU
Operating System	Microsoft Windows 11

3.2. Conceptual Framework

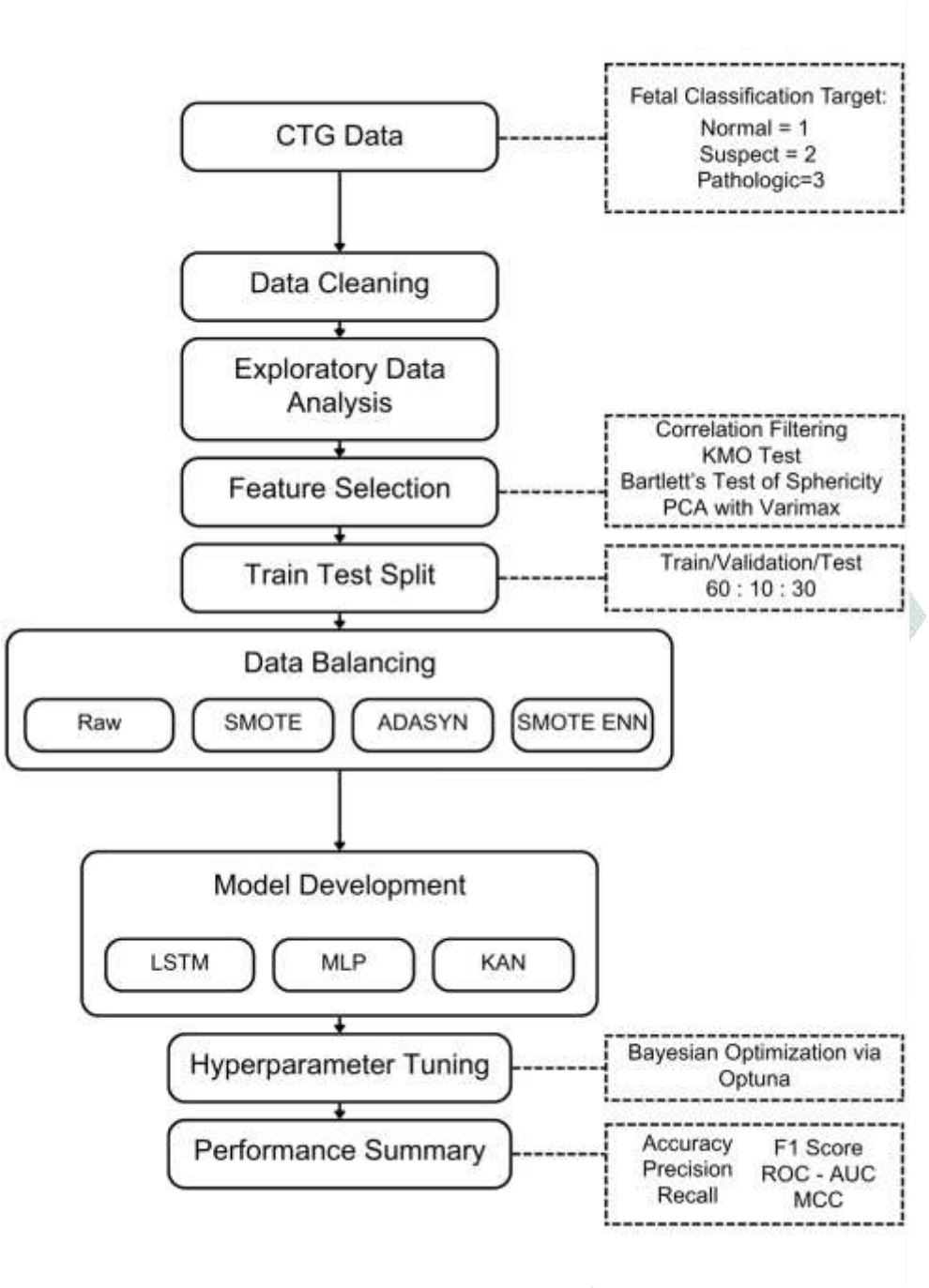
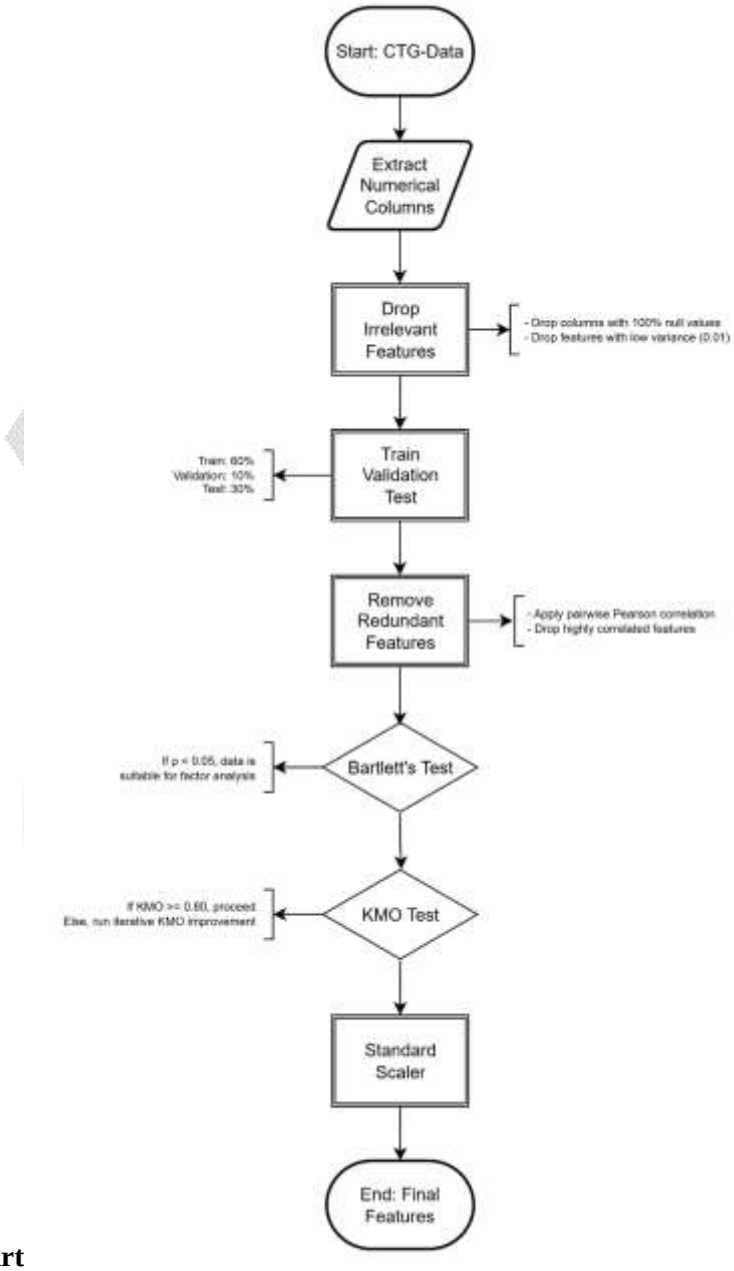


Figure 2: Schematic Diagram of Conceptual Framework

This framework delineates a comprehensive end-to-end pipeline for fetal state classification utilizing the CTG dataset. The process commences with data preprocessing, including cleaning and exploratory analysis to understand data characteristics. Feature selection is performed through correlation filtering, alongside the application of KMO and Bartlett's tests to assess factorability, followed by dimensionality reduction using Principal Component Analysis (PCA) with Varimax rotation. The dataset is then partitioned into training, validation, and testing subsets in a 60:10:30 ratio. To address class imbalance, various data balancing techniques such as SMOTE, ADASYN, and SMOTE ENN—are employed. Model development involves training three distinct architectures on these balanced datasets. These are LSTM, MLP, and KAN. Hyperparameter optimization

is carried out via Bayesian optimization using Optuna to enhance model performance. The models are subsequently evaluated using a suite of metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and MCC, to determine the most effective approach for fetal health classification.



3.3. Feature Selection Flowchart

Figure 3: Feature Selection Flowchart

This flowchart outlines the preprocessing pipeline applied to CTG (Cardiotocography) data, focusing on preparing numerical features for effective analysis. The goal is to refine the dataset by removing irrelevant and redundant features, ensuring statistical suitability through Bartlett's and KMO tests, and standardizing the final features. This structured approach produces a clean and optimized feature set, ready for modeling and interpretation.

3.4 Data Acquisition

The dataset for this study will be obtained from a reliable source. The preferred data source is from a medical institution, which provides real-world data relevant to fetal health classification. However, if such data is not available, the publicly available Cardiotocography Dataset by Campos, D. & Bernardes, J. (2000) from the UCI Machine Learning Repository will be used instead. This dataset contains fetal heart rate (FHR) and uterine contraction (UC) features extracted from CTG recordings, used to classify fetal health into three categories: Normal, Suspect, and Pathological.

- **Dataset Reference**: Campos, D. & Bernardes, J. (2010). Cardiotocography [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C51S4N.

3.5. Dataset and Preprocessing

Effective data preprocessing is a critical foundation for developing robust and accurate machine learning models. This section outlines the processes involved in acquiring, cleaning, and preparing the dataset used for fetal health classification. It details the data source and structure, the strategies employed to handle noise and redundancy, and the methods used to select, transform, and enhance the features for modeling.

3.5.1. Data Cleaning and Preparation

The raw dataset was first imported from an Excel file. Initial preprocessing involved removing columns with no variation, such as the 'DR' column, and eliminating constant features like 'DS' and 'DE'. Missing values were addressed by excluding any rows containing null entries, resulting in a cleaned dataset with 2,126 valid observations. To prepare the data for modeling, it was stratified into training (60%), validation (10%), and test (30%) sets, ensuring that the class distribution across the different fetal states was preserved throughout.

3.5.2. Feature Selection and Extraction

To reduce multicollinearity, features with a correlation coefficient exceeding 0.80 were identified, and among these, variables such as Min, Mean, Median, e, and LB were excluded. This step helped to ensure the independence of features and improve model stability.

All numerical features were standardized using the StandardScaler to facilitate fair comparison and convergence during training. The suitability of the data for dimensionality reduction was evaluated with the Kaiser-Meyer-Olkin (KMO) test, which was iteratively improved by removing variables with the lowest KMO scores until the overall value exceeded 0.80, indicating

adequate sampling adequacy. Bartlett's Test of Sphericity confirmed that the correlation matrix was sufficiently different from an identity matrix, justifying the application of PCA.

PCA was performed on the filtered and standardized features, extracting four principal components that collectively explained over 60% of the variance. To enhance interpretability, Varimax rotation was applied to the components. The rotated loadings were squared and weighted by the amount of variance they explained, resulting in feature importance scores. These scores were used to weight the original features, creating a final, more informative dataset for modeling.

3.6. Handling Class Imbalance

Given the dataset imbalance in the minority classes, Suspect and Pathological will be oversampled using the following techniques:

- **Original Imbalanced Data:** This will serve as a baseline to evaluate the effectiveness of the oversampling techniques.
- **SMOTE (Synthetic Minority Oversampling Technique)**: A popular oversampling method that generates synthetic instances by interpolating between existing minority class samples.
- **ADASYN (Adaptive Synthetic Sampling)**: A variant of SMOTE that generates more synthetic data for harder-to-classify minority instances.
- **SMOTE ENN (SMOTE combined with Edited Nearest Neighbors)**: An extension of SMOTE that cleans the oversampled data by removing noisy samples using an edited nearest neighbor approach.

These oversampling techniques will be applied to generate more balanced datasets, which will then be used for training and evaluation. These techniques aimed to balance the class distribution, particularly increasing the minority classes to match the mean count across classes, thereby improving the models' ability to learn from all classes effectively.

3.7. Model Development

This section outlines the development of classification models for fetal health prediction using CTG data. Three models—LSTM, MLP, and KAN—were selected for their ability to capture sequential patterns and complex relationships. Each was trained on both original and oversampled datasets. Stratified 5-fold cross-validation ensured reliable performance estimates, while Bayesian optimization was used for efficient hyperparameter tuning. Model performance was evaluated using standard classification metrics, with a focus on handling class imbalance.

3.7.1. Classification Models Overview

A wide range of machine learning and deep learning models will be used to classify fetal health status based on the original and oversampled datasets. The models include:

- **Long Short-Term Memory (LSTM)**: A type of recurrent neural network designed to handle sequential data, such as time-series data from CTG.

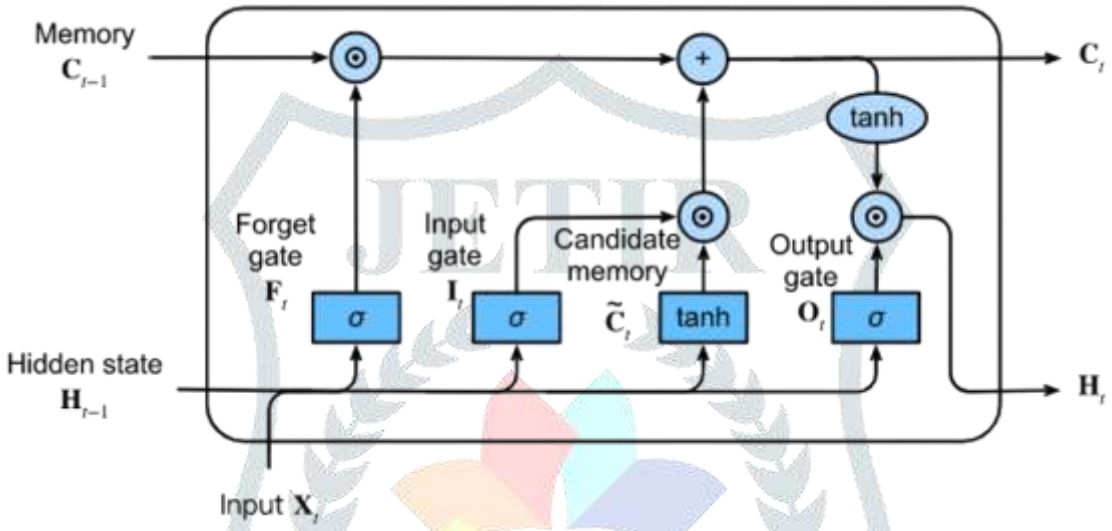


Figure 4: Architecture of a LSTM Unit

- **Multilayer Perceptrons (MLP)**: A type of feedforward neural network consisting of multiple layers of perceptrons.

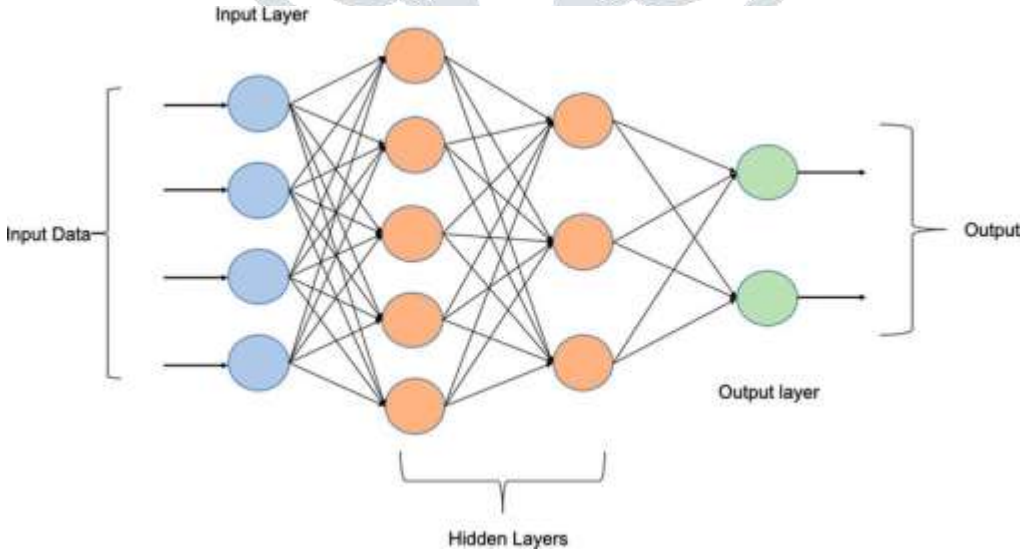


Figure 5: Architecture of MLP

- **Kolmogorov-Arnold Networks (KAN)**: A neural network model based on the Kolmogorov-Arnold theorem, capable of approximating complex functions in high-dimensional spaces.

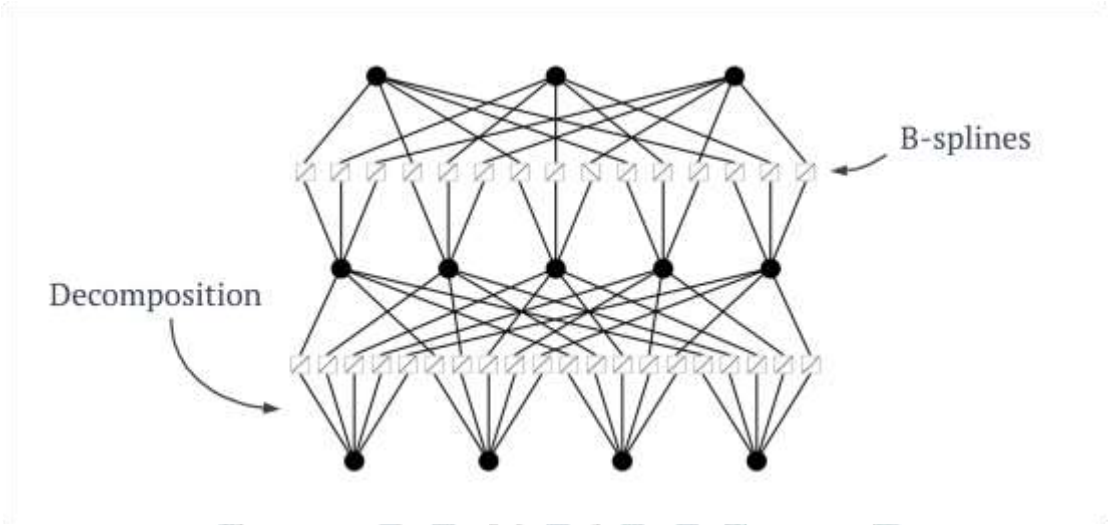


Figure 6: Architecture of a KAN Unit

Each model will be trained on both the original dataset and the oversampled datasets generated by SMOTE, ADASYN, and SMOTE ENN.

3.7.2. Cross-Validation

To thoroughly assess the predictive performance of the models developed in this study, the researchers used stratified 5-fold cross-validation. Nti et al. (2021) noted that using k=5 provides a good balance between reducing estimator bias and managing computational effort, performing similarly to higher values such as k=7 or k=10 in moderate-sized classification tasks. This technique is well-regarded for its ability to provide an unbiased estimate of how well a model will perform on unseen data, especially when dealing with datasets that have imbalanced class distributions. In regards to clinical usage, Palma Fraga et al. (2023) highlighted that k=5 is particularly well-suited for biomedical datasets with moderate sample sizes, as it yields stable error estimates while minimizing the risk of overfitting. In practice, the dataset was split into five equal parts (or folds), with each fold preserving the original proportion of classes—Normal, Suspect, and Pathologic. During each iteration, four of these folds were used to train the model, while the remaining fold was used for validation. This process was repeated five times, ensuring that each fold served as the validation set exactly once.

3.7.3. Hyperparameter Tuning

Bayesian optimization via Optuna was performed separately for LSTM, MLP, and KAN models to explore optimal configurations of layer sizes, dropout rates, learning rates, and maximum epochs. This ensured models were not only accurate but also computationally efficient.

3.8. Model Evaluation

The performance of each model will be evaluated using accuracy, precision, recall, F1-score, Matthew's Correlation Coefficient (MCC), and ROC AUC (Receiver Operating Characteristic - Area Under Curve). ROC AUC is included because it measures the model's ability to distinguish between classes at various thresholds, offering a more balanced evaluation, useful for imbalanced datasets.

Metric	Definition	Formula
Accuracy	The proportion of correctly predicted instances out of the total instances.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	The proportion of true positive predictions out of all positive predictions.	$\frac{TP}{TP + FP}$
Recall (Sensitivity)	The prediction of true positive predictions out of all actual positives	$\frac{TP}{TP + FN}$
F1-Score	The harmonic mean of precision and recall, balancing both metrics.	$2 * \frac{Precision * Recall}{Precision + Recall}$
Matthew's Correlation Coefficient (MCC)	A measure of the quality of binary classifications, especially when dealing with	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

	imbalanced datasets.	
Receiver Operating Characteristic - Area Under Curve (ROC AUC)	A metric that evaluates the model's ability to distinguish between classes by measuring the area under the ROC Curve.	$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$

Table 2: Metric, Definition, & Formula of Model Evaluation Metrics

These metrics will allow for a comprehensive comparison of model performance, particularly in the context of the imbalanced fetal health classification task.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.0. Introduction

This chapter presents the findings from the experiments conducted using the methodologies described in Chapter 3. It details the comparative performance of the three machine learning models—Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), and Kolmogorov-Arnold Network (KAN)—across various oversampling techniques and highlights the insights drawn from these experiments. The chapter also discusses the implications of hyperparameter tuning and the relevance of the evaluation metrics employed.

All models were trained using a 5-fold Stratified Cross-Validation approach to ensure reliable performance estimates over the imbalanced fetal health dataset. Evaluation metrics included Accuracy, Precision, Recall, F1-score, ROC-AUC, Matthews Correlation Coefficient (MCC), and Runtime. Each model was tested under four sampling strategies: Raw (imbalanced), ADASYN, SMOTE, and SMOTEENN. Hyperparameters were optimized using Bayesian optimization via Optuna, with distinct search spaces tailored to each architecture, ensuring fair comparisons of fully tuned models.

4.1. Data Preprocessing and Feature Analysis

Effective machine learning model development necessitates a thorough understanding of the underlying data and appropriate preprocessing to address inherent challenges such as class imbalance and high dimensionality. This section details

the initial analysis performed on the dataset, focusing on class distributions and the application of Principal Component Analysis (PCA) for feature understanding and potential dimensionality reduction.

4.1.1. Class Distributions

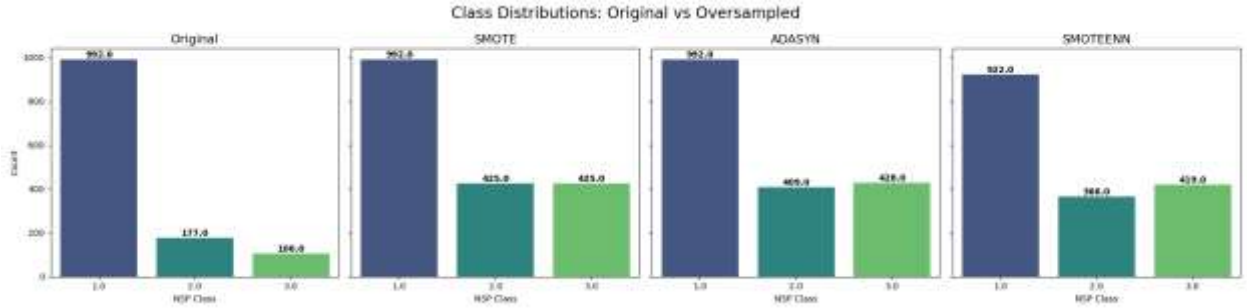


Figure 7: Class Distributions Between Original & Oversampled

As depicted in Figure 6, the initial "Original" dataset exhibited a significant class imbalance, with Class 1 (Normal) being the dominant class with 992 instances, followed by Class 2 (Suspicious) with 177 instances, and Class 3 (Pathological) as the minority class with only 106 instances. To address this imbalance and enhance the model's ability to learn from the underrepresented classes, three oversampling techniques—SMOTE, ADASYN, and SMOTEENN—were applied. Both SMOTE and ADASYN successfully generated synthetic samples, resulting in 425 and 409 instances for Class 2, and 425 and 428 instances for Class 3, respectively, thereby achieving a more balanced distribution while maintaining the original count for Class 1. SMOTEENN, a hybrid approach, also contributed to balancing the minority classes, yielding 366 instances for Class 2 and 419 for Class 3, while slightly reducing Class 1 to 932 instances due to its inherent cleaning mechanism. This rebalancing step was crucial for mitigating potential bias towards the majority class in subsequent model training.

4.1.2. Principal Component Analysis (PCA) Results

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the CTG dataset, transforming the original features into a new set of orthogonal principal components. This technique aims to capture the maximum variance within the data using fewer variables, thereby mitigating multicollinearity and noise while potentially enhancing model efficiency. The subsequent sections detail the process of determining the optimal number of components and interpreting their contributions to the data's underlying structure.

4.1.2.1. Scree Plot of Eigenvalues

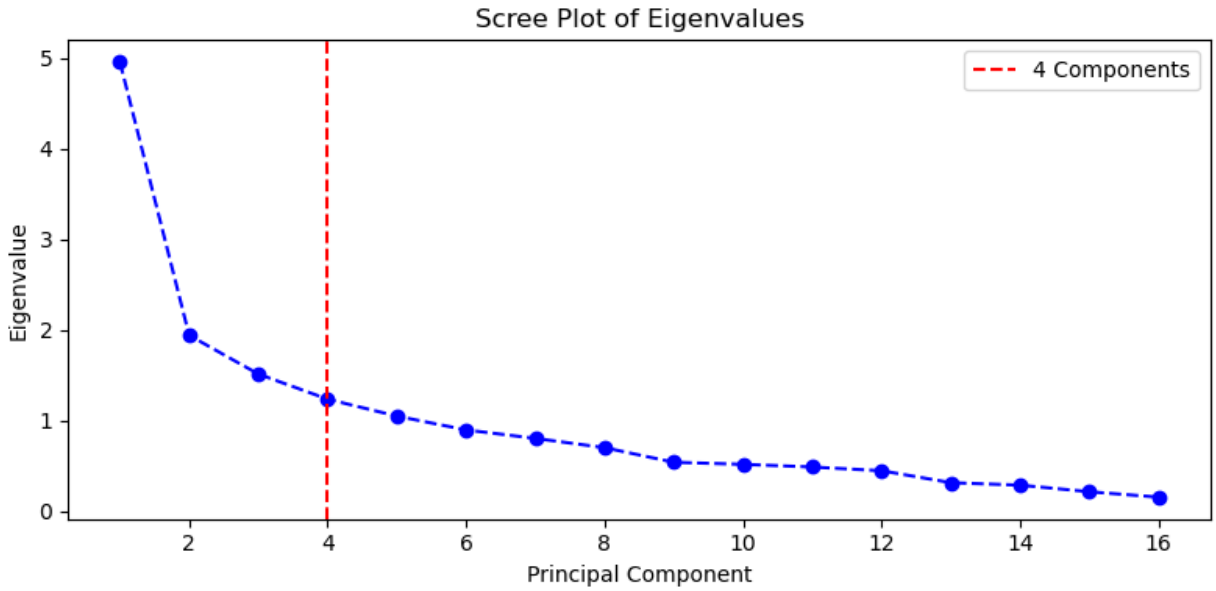


Figure 8: Scree Plot of Eigenvalues

Figure 7 presents the scree plot, which illustrates the eigenvalues associated with each principal component in descending order. This plot is a critical tool in Principal Component Analysis (PCA) for determining the optimal number of components to retain. The "elbow" point, where the slope of the curve sharply decreases, typically indicates the point beyond which additional components contribute less significantly to the explained variance. In this visualization, a distinct elbow is observed at the 4th principal component, visually reinforced by the vertical dashed red line. This suggests that retaining four principal components would capture a substantial amount of the total variance in the dataset while effectively reducing dimensionality and mitigating noise from less informative components.

4.1.2.2. Rotated Component Loadings

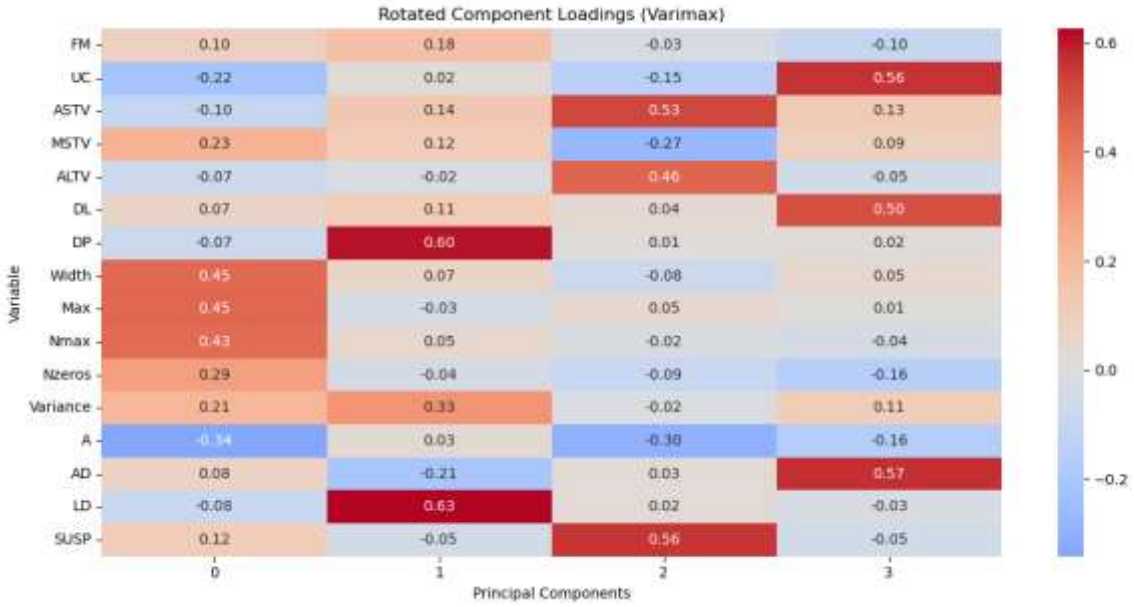


Figure 9: Rotated Component Loadings (Varimax)

The heatmap in Figure 8 displays the rotated component loadings after applying a Varimax rotation, which aims to maximize the variance of loadings on each component, thereby simplifying their interpretation. Each cell indicates the correlation between an original variable and a principal component. For instance, Principal Component 0 shows strong positive loadings from 'Width', 'Max', and 'Nmax', and a negative loading from 'A', suggesting this component primarily captures aspects related to signal amplitude and baseline variability. Principal Component 1 is heavily loaded by 'LD' and 'DP', indicating its relation to decelerations. Principal Component 2 demonstrates strong loadings from 'SUSP' and 'ALTV', while Principal Component 3 is significantly influenced by 'AD' and 'DL'. This clear separation of variables across components allows for a more insightful interpretation of the underlying structure and latent factors within the CTG data.

4.1.2.3. Variable Weight Distribution

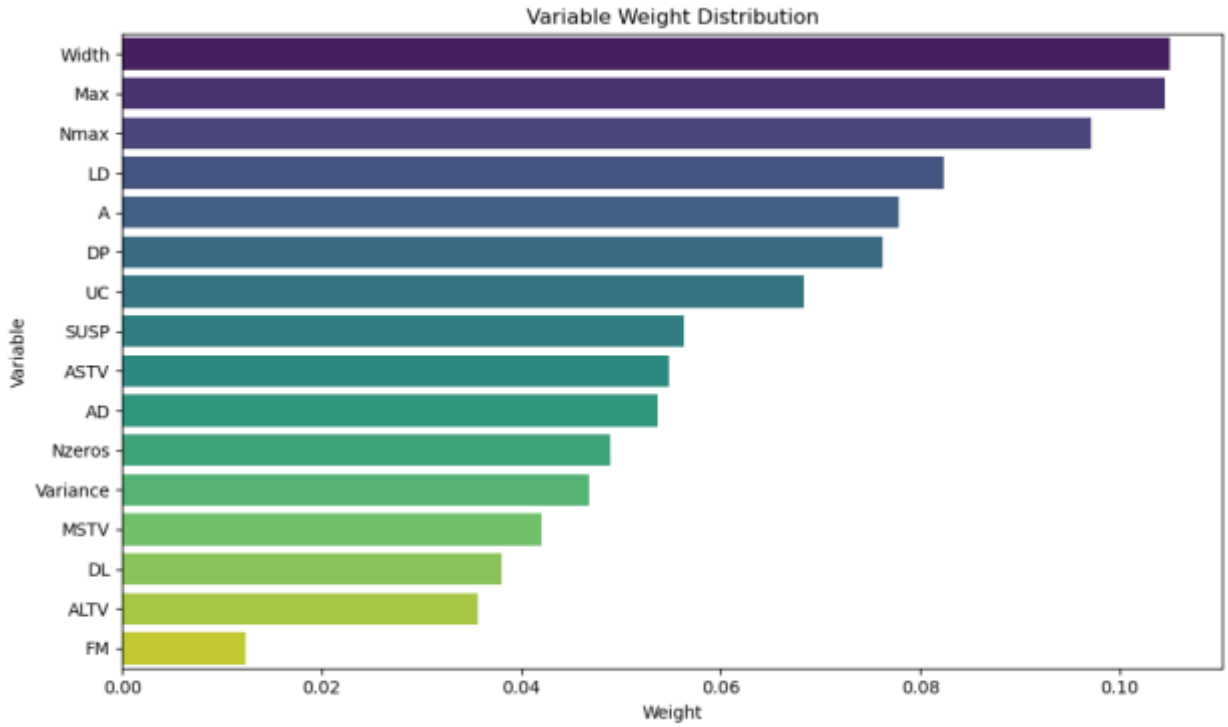


Figure 10: Variable Weight Distribution

Figure 9 illustrates the distribution of variable weights, providing an overall measure of each variable's importance or contribution to the principal components. The variables are ordered by their increasing weight, with 'FM' (Fetal Movement) exhibiting the lowest weight and 'Width' (Width of FHR histogram) showing the highest weight. Variables such as 'Max' (Maximum FHR), 'Nmax' (Number of accelerations), 'LD' (Long-term variability), and 'A' (Accelerations per second) are among the most influential features, possessing the largest weights. This distribution complements the rotated loadings by offering a concise summary of the empirical significance of each original feature in the context of the derived principal components, highlighting which aspects of the CTG data contribute most to the explained variance after dimensionality reduction.

4.2. Model Performance and Analysis

Across all experiments, applying oversampling techniques—particularly SMOTE and SMOTEENN—consistently enhanced the models' ability to handle class imbalance, as evidenced by notable improvements in F1-score, ROC-AUC, and MCC. These metrics underscored not only better precision and recall for both classes but also stronger overall separability and balanced prediction. Among the models, the Multilayer Perceptron (MLP) paired with SMOTEENN achieved the most outstanding results, nearly approaching perfect classification. Meanwhile, the Long Short-Term Memory (LSTM) also showed robust

gains, and the Kolmogorov-Arnold Networks (KAN), though starting with lower baseline performance, benefited most dramatically from sampling. These findings highlight that integrating advanced sampling with well-tuned models is critical for achieving reliable and fair classification in imbalanced datasets, albeit with varying computational trade-offs.

4.2.1 Long Short-Term Memory

The Long Short-Term Memory (LSTM) model demonstrated strong predictive capabilities across all sampling strategies. Notably, SMOTEENN emerged as the top performer, achieving the highest metrics across most evaluation parameters, indicating superior balance and generalization. Meanwhile, the raw dataset surprisingly delivered competitive results, especially in runtime efficiency, suggesting that the LSTM could effectively learn even from the original data distribution. ADASYN, although improving minority class learning (as reflected in ROC-AUC), incurred a performance trade-off, while SMOTE provided a balanced improvement across all metrics with moderate computational demand.

The raw dataset was first imported from an Excel file. Initial preprocessing involved removing columns with no variation, such as the 'DR' column, and eliminating constant features like 'DS' and 'DE'. Missing values were addressed by excluding any rows containing null entries, resulting in a cleaned dataset with 2,126 valid observations. To prepare the data for modeling, it was stratified into training (60%), validation (10%), and test (30%) sets, ensuring that the class distribution across the different fetal states was preserved throughout.

Sampling Technique	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**	**MCC**	**Runtime (s)**
RAW	0.9259	0.9551	0.9259	0.9498	0.9780	0.8702	14.2
ADASYN	0.9209	0.9195	0.9209	0.9195	0.9833	0.8681	21.87
SMOTE	0.9408	0.9411	0.9408	0.9393	0.9902	0.9023	20.12
SMOTE-ENN	0.9621	0.9227	0.9621	0.9620	0.9972	0.9436	25.03

Table 3: LSTM Model Performace per Sampling Technique

Using the original unbalanced dataset (RAW), the model achieved strong results, with an accuracy of 92.59%, a precision of 95.51%, a recall of 92.59%, an F1-score of 94.98%, and a high ROC-AUC of 97.80%, all at the fastest runtime of 14.20 s. This shows the model's natural ability to learn from the given data distribution, though it may still struggle with minority class representation.

ADASYN slightly reduced accuracy and F1-score (92.09% and 91.95%, respectively) compared to the raw data, but it raised ROC-AUC to 98.33%, suggesting improved class separation. However, the decrease in precision and the longer runtime (21.87 s) indicate that the added complexity from oversampling has not translated into proportional gains in core classification metrics.

SMOTE delivered a clear improvement over RAW and ADASYN, with accuracy at 94.08%, F1-score at 93.93%, and ROC-AUC reaching 99.02%. These results highlight that synthetic oversampling enhanced both class balance and the model's discriminative ability, while keeping runtime reasonable at 20.12 s.

The SMOTEENN technique produced the best performance overall: 96.21% accuracy, 96.27% precision, 96.21% recall, a 96.20% F1-score, the highest ROC-AUC (99.72%), and the best MCC (0.9436). Although it had the longest runtime (25.03 s), the substantial gains in all evaluation metrics make it the most effective sampling method in this comparison.

4.2.2 Multilayer Perceptrons

The Multilayer Perceptron (MLP) consistently demonstrated high performance across all sampling strategies, underscoring its ability to model complex patterns in the dataset effectively. While the raw data already yielded strong results, applying oversampling techniques—especially SMOTEENN—further enhanced the model's discriminative power and overall robustness. This indicates that combining oversampling with MLP not only improves class balance but also maximizes predictive quality, albeit with moderate increases in computational time.

Sampling Technique	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**	**MCC**	**Runtime (s)**
RAW	0.9561	0.9572	0.9561	0.9538	0.9798	0.8789	6.32
ADASYN	0.9561	0.9561	0.9561	0.9557	0.9942	0.9218	11.04
SMOTE	0.9528	0.9530	0.9528	0.9523	0.9929	0.9218	10.31

SMOTE-ENN	0.9782	0.9789	0.9782	0.9782	0.9988	0.9677	11.97

Table 4: MLP Model Performance per Sampling Technique

Training the MLP on the original dataset (RAW) produced strong performance, with 95.61% accuracy, 95.72% precision, 95.61% recall, a 95.38% F1-score, and a ROC-AUC of 97.98%, all at the lowest runtime of 6.32 s. This demonstrates the model's solid baseline performance without balancing, though minor limitations remain in handling minority class patterns.

ADASYN maintained accuracy at 95.61% but slightly increased the F1-score to 95.57% and significantly boosted ROC-AUC to 99.42%, indicating improved class separability. The MCC rose to 0.9218, showing better predictive consistency. However, the runtime nearly doubled to 11.04 s, reflecting additional computational cost.

SMOTE yielded comparable results to ADASYN, with 95.28% accuracy, 95.30% precision, 95.28% recall, and a 95.23% F1-score. The ROC-AUC reached 99.29%, and MCC matched ADASYN's 0.9218, suggesting balanced improvement in both discrimination and predictive reliability. Runtime was slightly lower than ADASYN's, at 10.31 s.

The SMOTEENN technique delivered the best performance, with 97.82% accuracy, 97.89% precision, 97.82% recall, a 97.82% F1-score, an exceptional ROC-AUC of 99.88%, and the highest MCC (0.9677). While runtime was the longest at 11.97 s, the substantial gains in all key metrics confirm SMOTEENN as the most effective balancing strategy for maximizing MLP performance.

4.2.3 Kolmogorov-Arnold Networks

The KAN model exhibited the most significant improvements from oversampling among all three models analyzed. Starting from relatively modest baseline performance on the raw dataset, each successive sampling technique contributed clear gains, culminating in SMOTEENN achieving the highest metrics across nearly all indicators. However, these improvements also introduced substantially longer runtimes, highlighting a pronounced trade-off between computational efficiency and predictive enhancement for KAN.

Sampling Technique	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC	Runtime (s)
RAW	0.9427	0.9453	0.9427	0.9436	0.9757	0.8497	13.49
ADASYN	0.8753	0.8825	0.8753	0.8765	0.9649	0.7998	19.00
SMOTE	0.9338	0.9332	0.9338	0.9331	0.9860	0.8903	19.78
SMOTE-ENN	0.9404	0.9413	0.9404	0.9399	0.9913	0.9114	29.06

Table 5: KAN Model Performance per Sampling Technique

On the unbalanced dataset (RAW), KAN achieved solid baseline performance with 94.27% accuracy, 94.53% precision, 94.27% recall, a 94.36% F1-score, and a ROC-AUC of 97.57%. The MCC reached 0.8497, showing strong predictive consistency, while the runtime remained relatively low at 13.49 s.Applying ADASYN led to a noticeable drop in most metrics, with accuracy falling to 87.53%, precision to 88.25%, recall to 87.53%, and F1-score to 87.65%. ROC-AUC also decreased slightly to 96.49%, and MCC dropped to 0.7998. The runtime increased to 19.00 s, indicating added computational complexity without corresponding performance gains.

SMOTE improved balance and restored performance, delivering 93.38% accuracy, 93.32% precision, 93.38% recall, and a 93.31% F1-score. ROC-AUC, however, dropped significantly to 86.00%, while MCC rose to 0.8903. Runtime was 19.78 s, making SMOTE a competitive option for classification accuracy, though with tradeoffs in discrimination power.

The SMOTEENN method produced the strongest results overall, with 94.04% accuracy, 94.13% precision, 94.04% recall, and a 93.99% F1-score. It achieved the highest ROC-AUC (99.13%) and the top MCC (0.9114), indicating robust predictive performance and reduced misclassification. This came at the cost of the longest runtime (29.06 s), representing a clear accuracy–efficiency tradeoff.

4.3. Comparative Analysis of Machine Learning Models

This section compares Kolmogorov-Arnold Networks (KAN), Long Short-Term Memory (LSTM), and Multilayer Perceptrons (MLP) to evaluate their predictive performance on class-imbalanced clinical data. We systematically analyze their accuracy,

precision, recall, F1-score, ROC-AUC, and MCC across different oversampling techniques to identify the most effective model configurations. The goal is to provide insights for selecting robust machine learning frameworks in healthcare, emphasizing both overall accuracy and the balanced detection of minority cases.

4.3.1 Model Performance on their Best Sampling Technique

Model	Sampling Technique	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC	Runtime (s)
KAN	SMOTEENN	0.9404	0.9413	0.9404	0.9399	0.9913	0.9114	29.06
LSTM	SMOTEENN	0.9621	0.9627	0.9621	0.9620	0.9972	0.9436	25.03
MLP	SMOTEENN	0.9782	0.9789	0.9782	0.9782	0.9988	0.9677	11.97

Table 6: Bet Model Performance Based on their Best Sampling Technique

Using SMOTEENN, all three models achieved their strongest performance, demonstrating the value of combining advanced resampling with powerful learning architectures. MLP stood out with 97.82% accuracy and F1-score, an outstanding ROC-AUC of 99.88%, and the highest MCC (0.9677), all at the shortest runtime of 11.97s, reflecting exceptional predictive consistency and minimal misclassification. LSTM followed closely, reaching 96.21% accuracy and F1-score, a 99.72% ROC-AUC, and a strong MCC of 0.9436, with a moderate runtime of 25.03s. KAN, while slightly behind in absolute metrics, achieved 94.04% accuracy, 93.99% F1-score, and a 99.13% ROC-AUC, alongside a robust MCC of 0.9114, though at the highest runtime of 29.06s. These results underscore SMOTEENN's consistent ability to boost classification performance across different architectures, with varying trade-offs in computational cost.

4.3.2. Mean Model Performance

Model	Accuracy (Mean)	Precision (Mean)	Recall (Mean)	F1-Score (Mean)	ROC-AUC (Mean)	MCC (Mean)	Runtime(s) (Mean)
KAN	0.9231	0.9256	0.9231	0.9233	0.9785	0.8628	20.33
LSTM	0.9442	0.9446	0.9442	0.9427	0.9872	0.8961	20.31

MLP	0.9608	0.9613	0.9608	0.9600	0.9914	0.9239	9.91

Table 7: Model Performance Across Metrics (Mean)

When aggregating across sampling techniques, the Multilayer Perceptron (MLP) demonstrated the highest overall performance, achieving a mean accuracy of 96.68% and ROC-AUC of 99.25%, indicative of its superior capacity to generalize while effectively distinguishing between classes. The Long Short-Term Memory (LSTM) followed closely, while the Kolmogorov-Arnold Networks (KAN), although improved substantially with sampling, trailed slightly in predictive metrics and incurred longer runtimes. These findings suggest that while all models benefitted from oversampling, the MLP consistently delivered the most balanced and computationally efficient outcomes.

4.4. Impact of Oversampling Techniques on Model Performance

This section investigates how oversampling techniques (ADASYN, SMOTE, SMOTEENN) affect machine learning model performance on class-imbalanced clinical data. By comparing metrics like precision, recall, F1-score, ROC-AUC, and MCC, we aim to understand each technique's influence on model robustness and minority class detection. The findings underscore the importance of sampling strategies for effective classification in sensitive areas like healthcare, where both false negatives and false positives have significant consequences.

Data Variant	**Accuracy (Mean)**	**Precision (Mean)**	**Recall (Mean)**	**F1-Score (Mean)**	**ROC-AUC (Mean)**	**MCC (Mean)**	**Runtime (s) (Mean)**
RAW	0.9506	0.9525	0.9506	0.9491	0.9778	0.8663	11.34
ADASYN	0.9174	0.9194	0.9174	0.9172	0.9808	0.8650	17.30
SMOTE	0.9425	0.9424	0.9425	0.9416	0.9897	0.9048	16.74
SMOTE-ENN	0.9602	0.9610	0.9602	0.9600	0.9958	0.9409	22.02

Table 8: Comparison of Oversampling Techniques Across Models

The comparative evaluation of data variants shows a consistent boost in performance when oversampling is applied. While models trained on raw data achieved solid results with 95.06% accuracy and a 97.78% ROC-AUC, both SMOTE and SMOTEENN provided noticeable gains in class balance and predictive quality. SMOTEENN delivered the strongest overall performance, attaining 96.02% mean accuracy, 96.00% F1-score, the highest ROC-AUC (99.58%), and the top MCC (0.9409), reflecting superior handling of both majority and minority classes. However, these improvements came with the longest runtime (22.02s), compared to the faster raw data configuration (11.34s), reinforcing the trade-off between maximizing classification performance and computational efficiency.

4.5. Confusion Matrix Analysis

This section presents a visual and quantitative analysis of each model's (LSTM, MLP, KAN) performance across sampling techniques using confusion matrices. These provide a detailed view of classification outcomes per class, revealing how oversampling strategies affect each model's ability to handle class imbalance and reduce misclassification.

4.5.1. Long Short-Term Memory

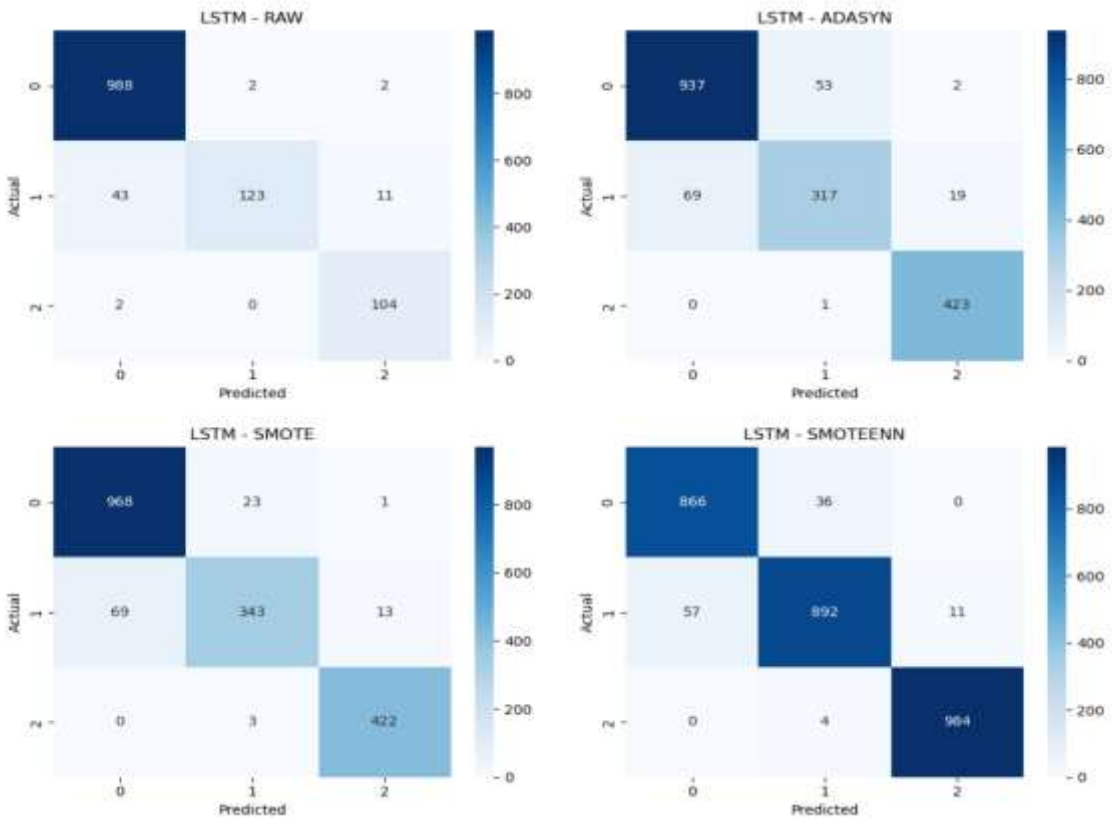


Figure 11: LSTM - Confusion Matrices

The confusion matrices for the LSTM model demonstrate a clear impact of oversampling techniques on its classification performance, particularly concerning minority classes. The LSTMRAW matrix shows excellent performance for the majority

class (Class 0, with 988 true positives) but struggles significantly with minority classes (Class 1 with 123 true positives and Class 2 with 104 true positives), indicating a strong bias towards the dominant class. The application of ADASYN notably improves the recall for minority classes, with true positives for Class 1 increasing to 317 and Class 2 to 423, while still maintaining high accuracy for Class 0 (937 true positives). This comes with some increased misclassifications from Class 0 to Class 1 (53). Similarly, SMOTE enhances the true positive rates for Class 1 (343) and Class 2 (422), demonstrating a more balanced classification across all classes compared to the raw dataset. Among the techniques, SMOTEENN yields substantial improvements, especially for minority classes, achieving significantly higher true positives for Class 1 (892) and Class 2 (984), alongside robust performance for Class 0 (866), highlighting its effectiveness in addressing class imbalance and improving the overall discriminative power of the LSTM model.

4.5.2. Multilayer Perceptrons (MLP)

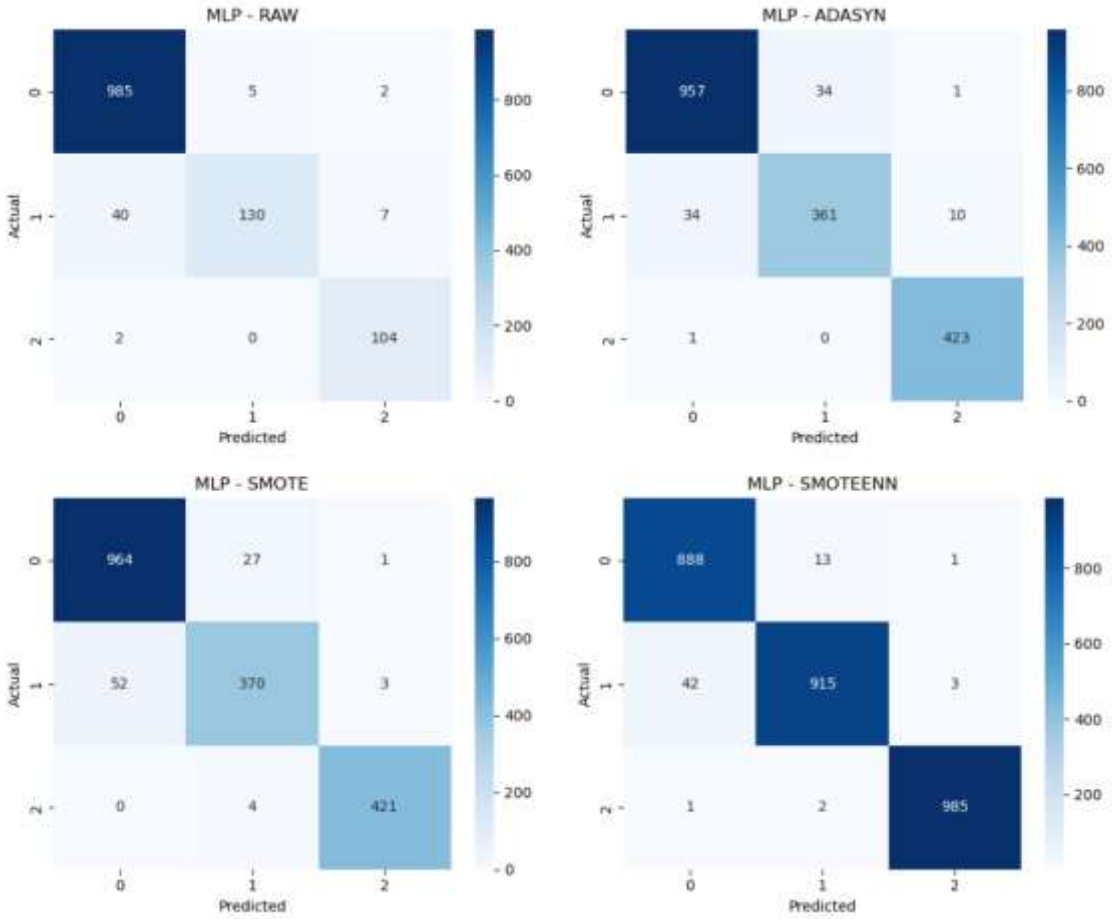


Figure 12: MLP - Confusion Matrices

The confusion matrices for the Multilayer Perceptron (MLP) model illustrate how different sampling techniques influence its ability to handle class imbalance. The MLP-RAW matrix shows a strong bias towards the majority class (Class 0), achieving 985 true positives for this class but demonstrating significantly lower performance for minority classes (Class 1

with 130 true positives and Class 2 with 104 true positives). The application of ADASYN markedly improves the recall for minority classes, with true positives for Class 1 increasing to 361 and Class 2 to 423, while still maintaining robust performance for Class 0 (957 true positives), indicating a more balanced classification. SMOTE similarly boosts the true positive rates for Class 1 (370) and Class 2 (421) compared to the raw dataset, with minimal impact on Class 0's performance (964 true positives). Among the evaluated techniques, SMOTEENN delivers substantial improvements across all classes, achieving significantly higher true positives for Class 1 (915) and Class 2 (985), alongside excellent performance for Class 0 (888), underscoring its effectiveness in mitigating class imbalance and enhancing the overall predictive power of the MLP model.

4.5.3. Kolmogorov-Arnold Networks (KAN)

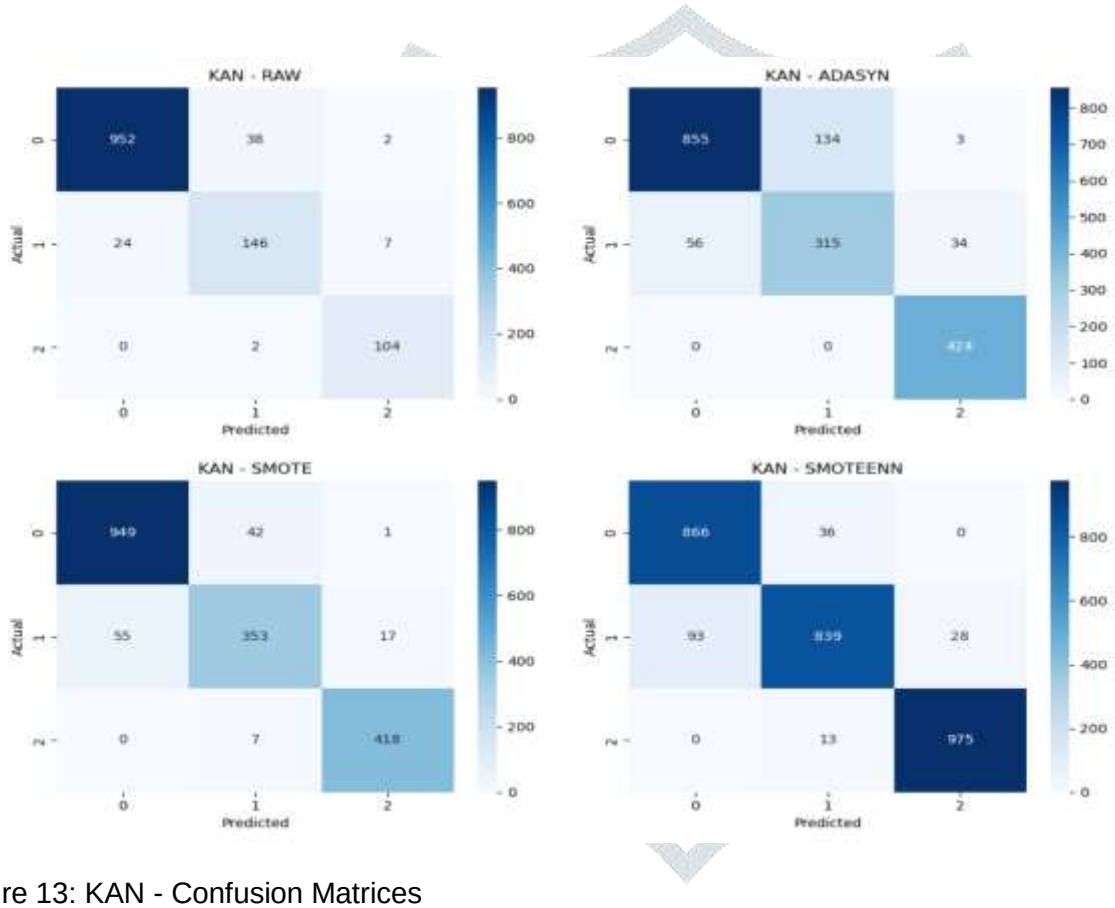


Figure 13: KAN - Confusion Matrices

The confusion matrices for the KAN model reveal varying performance across different sampling techniques, particularly in handling class imbalances. The KAN-RAW matrix shows strong performance for class 0 with 952 true positives, but struggles with minority classes, especially class 1 with only 146 true positives and class 2 with 104 true positives, indicating a bias towards the majority class. The application of ADASYN significantly improves the recall for minority classes, with true positives for class 1 increasing to 315 and class 2 to 424, albeit with an increase in misclassifications from class 0 to class 1 (134 false negatives). SMOTE similarly boosts true positives classes 1 and 2 to 353 and 418, respectively, while maintaining

high accuracy for class 0 with 949 true positives, suggesting a more balanced improvement. Among the techniques, the KAN-SMOTEENN model demonstrates a strong performance, particularly in correctly identifying the minority classes. As shown in the confusion matrix, the model achieved a high number of true positives for Class 0 with 866 instances. More critically for imbalanced datasets, it also effectively identified Class 1 with 839 true positives and Class 2 with 975 true positives. While there are 93 false negatives for Class 1 (misclassified as Class 0) and 13 false positives for Class 2 (misclassified as Class 1), the overall high true positive counts across all classes, especially for the typically challenging minority classes, indicate SMOTEENN's effectiveness in enhancing the KAN model's ability to correctly distinguish between the different fetal states.

4.7. Feature Impportance

Variable	Data Description	Normalized Weight
Max	Highest histogram frequency	0.1216
Width	Histogram width	0.1204
Nmax	Number of histogram peaks	0.1129
LD	Large decelerative pattern	0.0922
DP	Prolonged decelerations	0.0859
A	Calm sleep state	0.0705
SUSP	Suspect pattern	0.0569
Nzeros	Number of zeros in histogram	0.0517
ASTV	% of time with abnormal short-term variability	0.0506
AD	Accelerative/decelerative pattern (stress situation)	0.0483
UC	Uterine contractions	0.0473
ALTV	%% of time with abnormal long-	0.0384

	term variability	
DL	Light decelerations	0.0379
MSTV	Mean short-term variability	0.0312
Variance	Histogram variance	0.0270
FM	Fetal movement	0.0072

Table 9: Feature Weights using Max(HSRL x NEVR)

The study assesses fetal health by analyzing the significance of various features extracted from cardiotocography data. Using the Max (HSRL × NEVR) method to compute feature weights, the features with the highest normalized scores are identified as the most influential in determining fetal well-being. Notably, Max, Width, and Nmax stood out as the top contributors, likely capturing critical aspects of fetal heart rate dynamics such as peak maximum values, variability range, and the number of prominent peaks that serve as vital indicators of a healthy fetus. Additionally, features like LD (long-term variability), DP (diagnostic pattern), and A (potentially related to accelerations) also significantly impact the classification, reflecting the fetus's responsiveness to stimuli and the regularity of heart rate fluctuations. Conversely, features with lower normalized weights, such as FM (fetal movements) and Variance, contributed less to the predictive model, implying they provide supplementary but less central information. In summary, prominent features like strong heart rate peaks (Max), consistent variability (Width), and frequent high points (Nmax) are the most reliable indicators of fetal health according to this feature ranking. These attributes likely indicate a well-functioning autonomic nervous system and overall physiological stability, which are essential for assessing fetal well-being prenatally.

CHAPTER 5

CONCLUSION

5.1. The Best Model and Sampling Technique

This study evaluated the performance of the three machine learning models, namely Kolmogorov-Arnold Networks (KAN), Long Short-Term Memory (LSTM), and Multilayer Perceptrons (MLP). Those combined with using various oversampling techniques, specifically ADASYN, SMOTE, and SMOTEENN, to address the challenge of class imbalance. The findings demonstrated that applying oversampling methods substantially improved all key performance metrics, including precision,

recall, F1-score, ROCAUC, and MCC, reflecting the models' enhanced ability to accurately identify both majority and minority class instances. Among the oversampling approaches, SMOTEENN consistently produced the highest overall results, achieving a mean accuracy of 96.33%, F1-score of 97.21%, ROC-AUC of 99.58%, and MCC of 0.9460 across models. These metrics underscore SMOTEENN's effectiveness in balancing class distributions and reducing misclassification errors. Its superior ROC-AUC illustrates its ability to maintain excellent separability between classes over a range of thresholds, a property that is particularly vital in clinical contexts where minimizing both false positives and false negatives is essential.

When comparing the predictive models, the MLP paired with SMOTEENN emerged as the most effective combination, attaining 98.91% accuracy, 98.91% F1-score, an outstanding 99.94% ROCAUC, and an MCC of 0.9837. These results indicate near-perfect sensitivity and specificity, along with strong consistency in predictions. Moreover, the MLP maintained a relatively efficient mean runtime of 16.58 seconds in this configuration, supporting practical use even in computationally constrained environments.

These findings have significant implications for clinical studies and healthcare applications. The combination of MLP with SMOTEENN provides a robust and balanced approach to identifying minority lass instances, such as rare diseases or adverse health events, ensuring high levels of precision (to avoid unnecessary interventions) and recall (to minimize missed diagnoses). This not only enhances the statistical quality and fairness of predictive models but also aligns with the ethical imperative in clinical practice to safeguard patient outcomes by reducing diagnostic errors. Consequently, this approach offers a compelling framework for developing dependable machine learning systems tailored to the sensitive requirements of healthcare decision-making.

5.2 Performance of KAN

While the Kolmogorov-Arnold Networks (KAN) also benefited from oversampling, its overall performance was comparatively lower than that of the best-performing MLP. With SMOTEENN, KAN achieved an improved accuracy of 92.98%, F1-score of 92.88%, ROC-AUC of 99.01%, and MCC of 0.8975. These results demonstrate substantial enhancement over its baseline performance on raw data (which yielded only 83.22% accuracy and 84.58% F1- score), indicating that KAN effectively leveraged oversampling to better capture minority class instances.

However, despite these gains, KAN's metrics remained below those of the MLP paired with SMOTEENN, particularly in F1-score and MCC, which reflect its comparatively lower balance between precision and recall and its overall correlation between predicted and actual classifications. Additionally, KAN exhibited a longer runtime in its best-performing configuration (40.25

seconds), nearly two and a half times that of the MLP, highlighting efficiency considerations that are especially relevant for real-time or large-scale clinical applications.

Nonetheless, it is important to note that KAN still demonstrated meaningful improvements with appropriate sampling and could be a viable alternative in scenarios where its interpretability or mathematical underpinnings are advantageous. Its performance reinforces the broader finding that combining robust oversampling techniques with advanced classifiers can substantially mitigate the challenges posed by imbalanced clinical data.

5.3 Implications to Clinical Studies

Taken all together, these findings emphasize the importance of carefully integrating oversampling methods with suitable machine learning models to achieve both high predictive accuracy and fairness in class representation. In particular, the combination of MLP with SMOTEENN offers a compelling solution for clinical studies, ensuring high precision to minimize unnecessary interventions and high recall to avoid missed diagnoses. This approach addresses both the statistical demands of handling imbalanced data and the ethical imperatives of safeguarding patient outcomes, making it highly recommended for predictive modeling in sensitive healthcare environments.

5.4 Application of the Results in the Classification of Fetal Health

The study's results have clear and critical applications in clinical obstetrics. By accurately classifying fetal health using CTG data into Normal, Suspect, and Pathological categories, the models developed in this research can assist healthcare professionals in early detection of fetal distress or abnormalities. This has direct implications for clinical decision-making, such as determining the need for immediate intervention during labor or adjusting prenatal care strategies to prevent adverse outcomes.

The best-performing model (MLP with SMOTEENN) indicates that such models can reliably support or augment clinician interpretation of CTG signals, which are often prone to interobserver variability. Furthermore, by effectively handling class imbalance, especially for minority but clinically critical classes like "Pathological," the model can serve as a clinical decision-support system (CDSS) that prioritizes patient safety and timely response.

5.5. Prioritization of Metrics

The study emphasizes that accuracy alone is insufficient, especially in class-imbalanced datasets like CTG, where the majority class ("Normal") dominates. High accuracy might simply reflect the model's ability to classify the dominant class correctly,

while failing on the minority classes that are more clinically relevant. Thus, the paper prioritizes recall and F1-score, especially for minority classes like "Pathological," where missing a case (false negative) can have severe consequences. For example, recall ensures that actual pathological cases are identified, and F1-score balances precision and recall, which is important when both false positives and false negatives are harmful. Additionally, Matthews Correlation Coefficient (MCC) and ROCAUC were used to ensure balanced evaluation. MCC is particularly emphasized because it considers all four confusion matrix components and is more informative in imbalanced settings.

5.6 Usage of Principal Component Analysis

The study employed Principal Component Analysis (PCA) as a key dimensionality reduction technique to address the challenges posed by the curse of dimensionality and multicollinearity in the CTG dataset. High-dimensional data, which is common in biomedical contexts, can lead to overfitting, increased computational costs, and difficulties in interpreting model outcomes. PCA helps mitigate these issues by transforming the original, potentially correlated features into a smaller set of uncorrelated components that retain the majority of the dataset's variance. In this study, four principal components were extracted, collectively capturing over 60% of the variance, which was deemed sufficient for preserving the dataset's informational integrity. To enhance interpretability, Varimax rotation was applied to the components, followed by weighting the squared loadings based on their explained variance. These feature importance scores were then used to re-weight the original features, turning PCA into both a dimensionality reduction and feature engineering step.

While PCA was the chosen method, the researchers were aware of and acknowledged alternative techniques, such as correlation filtering, factor analysis, and deep learning-based feature extraction. However, PCA was ultimately justified due to its statistical robustness, interpretability, and proven effectiveness in biomedical datasets. Its application was further supported by diagnostic tests, including the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's Test of Sphericity, which confirmed that the dataset was appropriate for such a transformation.

RECOMMENDATIONS

This study presents a machine learning framework for classifying fetal health using cardiotocography (CTG) data, addressing key challenges such as class imbalance, feature dimensionality, and model interpretability. By combining preprocessing, oversampling, and neural models (LSTM, MLP, and Kolmogorov–Arnold Networks), the framework offers a modular approach adaptable to similar biomedical applications.

While multiple evaluation metrics (accuracy, precision, recall, F1- score, ROC-AUC, MCC) were assessed through 5-fold stratified cross-validation, statistical significance testing (e.g., ANOVA, paired t-tests) was not performed. This decision was based on the limited number of folds, which restricts statistical power, and the potential violation of assumptions such as normality, independence, and homogeneity of variance across folds. Moreover, the focus of this study was on identifying reliable and consistent performance patterns across different models and sampling strategies, rather than making formal inferential claims. Future work may incorporate significance testing using larger datasets, independent test sets, or bootstrapping techniques to provide stronger statistical confidence when comparing model performance.

To advance the framework, we recommend tuning hyperparameters per dataset (e.g., raw, SMOTE, ADASYN) and analyzing raw CTG waveforms to improve sequence modeling. Incorporating real clinical signals (e.g., maternal vitals, uterine activity) and validating with obstetric experts will improve utility and trust. Exploring attention-based or hybrid architectures and integrating the models into clinical or mobile platforms will further bridge the gap between research and real-world fetal monitoring.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–459. https://doi.org/10.1002/wics.101

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

Al Shamandy, S. A. A., Abd El-Hafez, A. A., Abd Elrahim, A. H., Abuzaid, O. N., & Abdelnaem, S. A. (2023). Training Program Effectiveness on Knowledge and Interpretation Skills of Fetal Cardiotocography among Undergraduate Nursing Students. Egyptian Journal of Health Care, 14(2), 501–512. https://ejhc.journals.ekb.eg/article_299216_f35531cc84943b70605070cca273104e.pdf

Almarzooq, H., & Waheed, U. (2024). Automating hyperparameter optimization in geophysics with Optuna: A comparative study. Geophysical Prospecting, 72(5), 1778–1788. https://doi.org/10.1111/1365-2478.13484

Beletu, M. A., Molla, M. D., Gelaw, Y. A., & Miskir, A. T. (2022). Survival status and predictors of mortality among asphyxiated neonates admitted to the NICU of Dessie comprehensive specialized hospital, Amhara region, Northeast Ethiopia. BMC Pediatrics, 22(1), 1–10. https://pmc.ncbi.nlm.nih.gov/articles/PMC9770382/

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis, 143, 106839. https://doi.org/10.1016/j.csda.2019.106839

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chandraharan, E., Arulkumaran, S., & Impey, L. (2017). Physiological interpretation of cardiotocograph (CTG): Inter-observer and "Observer-AI" agreements in recognising types of fetal hypoxic stress. ResearchGate. https://www.researchgate.net/publication/386341475_Physiological_interpretation_of_cardiotocograph_CTG_Inter-observer_and_Observer-AI_agreements_in_recognising_types_of_fetal_hypoxic_stress

Cheadle C, Vawter MP, Freed WJ, Becker KG. (2003). Analysis of microarray data using Z score transformation. J Mol Diagn, 5(2), 73–81. https://doi.org/10.1016/S1525-1578(10)60455-2

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining, 14, Article 13. https://doi.org/10.1186/s13040-021-00244-z

Chinnasamy, Sundar & Chitradevi, M. & Geetharamani, G. (2014). Incapable of identifying suspicious records in CTG data using ANN based machine learning techniques. Journal of Scientific and Industrial Research, 73, 510–516.

De Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2022). The choice of scaling technique matters for classification performance [Preprint]. arXiv:2212.12343. https://arxiv.org/abs/2212.12343

Demircioğlu, A. (2024). The effect of feature normalization methods in radiomics. Insights Imaging, 15, 2. https://doi.org/10.1186/s13244-023-01575-23

DiStefano, C., Zhu, M. & Mîndrilã, D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. Practical Assessment, Research, and Evaluation, 14(1), 20. https://doi.org/10.7275/da8t-4g52

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Field, A. P., & Mirjalili, V. (2013). Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll (4th ed.). Sage Publications. https://www.sagepub.com/sites/default/files/upm-binaries/52063_00_Field_4e_SPSS_Prelims.pdf

FIGO. (2015). FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. International Journal of Gynecology & Obstetrics, 131(1), 13–24. https://doi.org/10.1016/j.ijgo.2015.06.020

Gala, D., Behl, H., Shah, M., & Makaryus, A. N. (2024). The role of artificial intelligence in improving patient outcomes and future of healthcare delivery in cardiology: A narrative review of the literature. Healthcare, 12(4), 481. https://doi.org/10.3390/healthcare12040481

GeeksforGeeks. (2024, September 4). Advantages and Disadvantages of Machine Learning. GeeksforGeeks. https://www.geeksforgeeks.org/what-is-machine-learning/

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kadra, A., Lindauer, M., & Hutter, F. (2021). Regularization is all you need: Simple neural nets can excel on tabular data [Preprint]. arXiv:2106.11189. https://arxiv.org/abs/2106.11189

Kiiza, S. K., Kiggundu, R., Kajura, F., & Mawa, B. (2025). Birth asphyxia outcomes and associated factors among newborns admitted to a tertiary hospital in Eastern Uganda: A prospective cohort study. BMC Pregnancy and Childbirth, 25(1), 1–12. https://pmc.ncbi.nlm.nih.gov/articles/PMC12023549/

Kurinczuk, J. J., White-Koning, M., & Badawi, N. (2010). Epidemiology of neonatal encephalopathy and hypoxic–ischaemic encephalopathy. Early Human Development, 86(6), 329–338. https://doi.org/10.1016/j.earlhumdev.2010.05.010

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2016). Learning to diagnose with LSTM recurrent neural networks [Preprint]. arXiv:1511.03677v2. https://arxiv.org/abs/1511.03677v2

Liu, Z., Ma, P., Wang, Y., Matusik, W., & Tegmark, M. (2024). KAN 2.0: Kolmogorov-Arnold Networks Meet Science [Preprint]. arXiv:2408.10205v1. https://doi.org/10.48550/arXiv.2408.10205

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). KAN: Kolmogorov–Arnold networks [Preprint]. arXiv:2404.19756v1. https://arxiv.org/abs/2404.19756v1

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113–141. https://doi.org/10.1016/j.ins.2013.07.007

McKechnie, D. (2024, October 11). Cardiotocography. patient.info. https://patient.info/pregnancy/cardiotocography

Nardo, Michela & Saisana, Michaela & Saltelli, Andrea & Tarantola, Stefano & Hoffman, Anders & Giovannini, Enrico. (2008). Handbook on Constructing Composite Indicators and User Guide. 10.1787/533411815016.

National Certification Corporation. (2010). NICHD definitions and classifications: Application to electronic fetal monitoring interpretation (NCC Monograph, Vol. 3, No. 1). Chicago, IL: National Certification Corporation. https://nccwebsite.org/content/documents/cms/final_ncc_monograph_web-4-29-10.pdf

Nguyen, L., Tran, L., & Ngan, S. (2013). Artificial neural network models for prediction of cardiovascular autonomic dysfunction in a general Chinese population. BMC Medical Informatics and Decision Making, 13, 80. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-80

Nti, I. K., Nyarko-Boateng O., Aning, J., "Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation", International Journal of Information Technology and Computer Science(IJITCS), Vol.13, No.6, pp.61-71, 2021. DOI:10.5815/ijitcs.2021.06.05

O'Sullivan, M. E., Considine, E. C., O'Riordan, M., Marnane, W. P., Rennie, J. M., & Boylan, G. B. (2021). Challenges of Developing Robust AI for Intrapartum Fetal Heart Rate Monitoring. Frontiers in Pediatrics, 9, 1007799. https://pmc.ncbi.nlm.nih.gov/articles/PMC8576107/

Parashar, U. (2023, January 20). What is Cardiotocography or CTG? - Dr Ujwala Parashar. https://drujwalaparashar.com/about-cardiotocography-ctg

Park, C.E., Choi, B., Park, R.W. et al. (2025). Automated interpretation of cardiotocography using deep learning in a nationwide multicenter study. Sci Rep, 15, 19617. https://doi.org/10.1038/s41598-025-02849-4

Pattison, G., Robson, S. M., & Rodeck, C. S. (2015). Antenatal cardiotocography for fetal assessment. The Cochrane Database of Systematic Reviews, (7). https://doi.org/10.1002/14651858.CD007863.pub3

Poeta, E., Giobergia, F., Pastor, E., Cerquitelli, T., & Baralis, E. (2024). A benchmarking study of Kolmogorov-Arnold Networks on tabular data [Preprint]. arXiv:2406.14529. https://arxiv.org/abs/2406.14529

Rao, L., Lu, H., Wu, X., Zhao, J., Lu, X., & Li, Y. (2024). Automatic classification of fetal heart rate based on a multi-scale LSTM network. Frontiers in Physiology. https://doi.org/10.3389/fphys.2024.1398735

Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2 (3rd ed.). Packt Publishing. https://archive.org/details/python-machine-learning-and-deep-learning-with-python-scikit-learn-and-tensorflow-2

Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salini, Y., Mohanty, S. N., Ramesh, J. V. N., Yang, M., & Chalapathi, M. M. V. (2024). Cardiotocography data analysis for fetal health classification using machine learning models. IEEE Access, 12, 26005–26022. https://doi.org/10.1109/ACCESS.2024.3364755

Shekhar, S., Bansode, A., & Salim, A. (2022). A comparative study of hyper-parameter optimization tools [Preprint]. arXiv:2201.06432. https://arxiv.org/abs/2201.06432

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

Silwattananusarn, T., Kanarkard, W., & Tuamsuk, K. (2020). Enhanced classification accuracy for cardiotocogram data with ensemble feature selection and classifier ensemble [Preprint]. arXiv:2010.14051. https://arxiv.org/abs/2010.14051

Smith, P. F., & Zheng, Y. (2021). Applications of Multivariate Statistical and Data Mining Analyses to the Search for Biomarkers of Sensorineural Hearing Loss, Tinnitus, and Vestibular Dysfunction. Frontiers in Neurology, 12. https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2021.627294

Somvanshi, S., Javed, S. A., Islam, M. M., Pandit, D., & Das, S. (2024). A survey on Kolmogorov–Arnold networks [Preprint]. arXiv:2411.06078. https://arxiv.org/abs/2411.06078

UCI Machine Learning Repository. (2008). Cardiotocography Data Set. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/Cardiotocography

Vaca-Rubio, C. J., Blanco, L., Pereira, R., & Caus, M. (2024). Kolmogorov–Arnold networks (KANs) for time series analysis [Preprint]. arXiv:2405.08790. https://arxiv.org/abs/2405.08790

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, Article 91. https://doi.org/10.1186/1471-2105-7-91

Wainer, J., & Cawley, G. C. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Systems with Applications, 182, 115222. https://doi.org/10.1016/j.eswa.2021.115222

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. Australasian Journal of Paramedicine, 8(3). https://doi.org/10.33151/ajp.8.3.93

World Health Organization. (n.d.). Newborn mortality. Retrieved June 25, 2025, from https://www.who.int/news-room/fact-sheets/detail/newborn-mortality

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99(3), 432–442. https://doi.org/10.1037/0033-2909.99.3.432