Exploring Advanced Machine Learning Techniques for Efficient Prediction of Preterm Birth

Capstone II Report

Anelya Alimgozhina

Nazarbayev University
Department of Electrical and Computer Engineering
School of Engineering and Digital Sciences

This project report was created on TexStudio editing platform using LaTeX. All the figures were drawn using draw.io online software tool.

NAZARBAYEV UNIVERSITY

Electrical and Computer Engineering
Nazarbayev University
http://www.nu.edu.kz

Title:
Exploring Advanced Machine Learning Techniques for Efficient Prediction of Preterm Birth

Theme:
Biomedical Signal Processing

Project Period:
Fall 2024 - Spring 2025

Project Group:
Applications of Signal Processing Lab

Participant(s):
Anelya Alimgozhina

Supervisor(s):
Muhammad Tahir Akhtar

Copies: 1

Page Numbers: 42

Date of Completion:
April 25, 2025

Abstract:

This work's purpose is to contribute to the field of premature birth prediction with the help of Electrohysterogram (EHG) signals. Since preterm birth is critical public and health challenge, the project aims to find a reliable way to deal with the issue of class imbalance - one of the problems in prediction - by comparing various oversampling, undersampling, and ensemble techniques, with and without Cross-Validation (CV). Additionaly, exploration of different features and their effect on the model performance has been conducted. The results have shown that using stratified 5-fold CV in conjunction with resampling approaches might be a useful strategy for managing unbalanced datasets and producing more accurate performance measures. Resampling techniques can aid in "balancing" the dataset, but their effectiveness is best evaluated with CV, which guarantees that the model's performance accurately represents its generalization. Moreover, the model's feature type has a big impact on performance: categorical features showed superior results without CV, but when CV has been implemented, performance became more consistent across feature sets.

Contents

List of Figures

List of Tables

Preface

The topic of preterm birth prediction is of big significance in the field of maternal and neonatal health. As our understanding of this complex phenomenon evolves, the urgency to develop reliable predictive models becomes increasingly critical. This work aims to explore the phenomenon of overly positive results of preterm birth prediction, bringing together different research findings, methodologies, and innovative approaches that can enhance clinical practice and improve outcomes for mothers and their children. We mention why such cases occur and investigate the best way to come up with realistic prediction results.

The structure of this report is as follows. Chapter 1 is the Introduction, which provides general information on the field of machine learning, feature selection, and preterm birth. Chapter 2 is the Background, offering an overview of the basics of preterm birth prediction. Chapter 3 is the Methodology, detailing the experimental setup of the project. Chapter 4 is Results and Discussion, where the findings of the experiments are presented and analyzed. Finally, Chapter 5 is the Conclusion, which summarizes the work done and outlines future plans.

As a member of the Applications of Signal Processing Lab, I have been fortunate to witness firsthand the collaborative spirit that drives our research efforts. I would like to extend my heartfelt gratitude to Professor Muhammad Tahir Akhtar, the ASP Lab's and this project's supervisor, whose vision and dedication have been instrumental in organizing the seminars that enriched our lab's knowledge base. I would like to acknowledge that our lab meetings have become the catalyst for this project. I also wish to thank my fellow lab members, whose diverse experiences and insights enriched our discussions and inspired me. The collective knowledge shared during these seminars has been instrumental in shaping the ideas presented here.

Nazarbayev University, April 25, 2025

Anelya Alimgozhina
<anelya.alimgozhina@nu.edu.kz>

Chapter 1

Introduction

1.1 Background

Machine learning is a branch of computer science and AI (Artificial Intelligence), whose primary objective is to imitate the human learning process using algorithms [1]. There are three main classes of machine learning techniques [2]:

- Supervised Learning. The training process utilizes labeled inputs paired with the outputs. This way algorithms recognize the patterns and relationships. An example of the application is an electronic mail classification (spam/not spam) [2].
- Unsupervised Learning. The training set consists of unlabeled inputs; the algorithm does not require human intervention or guidance. An example of the application is categorizing different documents depending on their topic [2].
- Reinforcement Learning. The training process is a "hybrid" of the previous types: algorithms receive human feedback, but only after making a connection between an input and output. It can be considered a "trial and error" process.

Using the definition of World Health Organization (WHO), premature birth is described as the births before the end of 37 weeks of pregnancy (or less than 259 days after the first day of the woman's last period) [3, 4]. Unfortunately, preterm birth cases are quite common globally; nearly 1 in 10 babies are born too early, and approximately 1 000 000 infants died from health complications related to premature birth in 2020 [5]. Moreover, there are many long-term consequences that make child's life more difficult: respiratory problems (chronic respiratory morbidity, bronchopulmonary dysplasia) [6], neurological disorders (cerebral palsy),

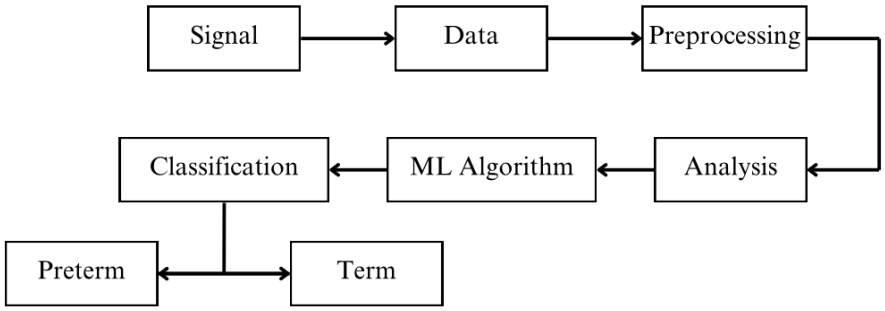

Figure 1.1: The classical premature birth prediction model, showing the main steps of the procedure [15].

delay in development [7], higher risk of enteric and invasive infections [8, 9], sensory impairments (ex. hearing and eyesight problems) [10]. Concequently, it may negatively affect academic performance and result in higher need of medical attention later in life [11]. Therefore, prediction of preterm birth using machine learning techniques has become a growing research field.

Feature selection (FS) is a process that focuses on locating a subset of "useful" features from the original set [12]. FS process contains four key steps [12]:

- Subset generation: creating subsets by adding or removing single or multiple features.
- Subset evaluation: measuring the "relevance" of the present features.
- Finding stopping criteria: to avoid an exhaustive search stopping criteria are defined to find the optimal subset.
- Result: getting the resulting subset.

The metrics that assess the performance of the FS result are accuracy rate, error rate, sensitivity, precision, recall, specificity, and geometric mean [12]. Four main types of features exist: Time Domain (ex. variance, mean absolute value), Frequency Domain (ex. peak frequency, spectral moments), Time-Frequency Domain (ex. spectrum envelope), and Deep Features [13]. Moreover, there are bio-inspired algorithms (BIA) that are based on natural phenomena and processes (e.g. animal behavior, physical processes, etc.) used for feature selection [14]. For example, Ant Colony Algorithm, Artificial Immune System, Evolution Strategy, Artificial Flora algorithm, and Lion algorithm. They are considered to be effective when dealing with non-linear and high-dimensional data [14].

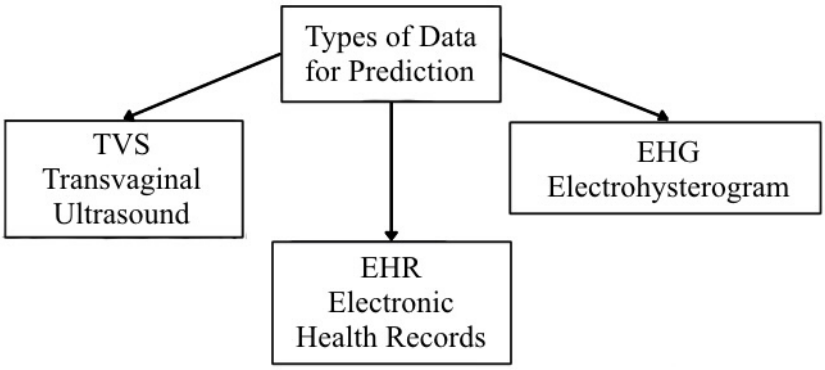

Figure 1.2: Types and examples of signals used in preterm birth prediction: TVS, EHR, EHG.

Figure 1.1 illustrates a common model for predicting preterm birth [15]. The unprocessed signal serves as the input for a quantifier block. Then, the preprocessing phase involves signal processing techniques (e.g. digital and adaptive filters [16]), feature extraction and selection, and data normalization. After analyzing the extracted data, machine learning algorithms are used to categorize (classify) the signal as either preterm or term.

Figure 1.2 shows the three types of signals that are commonly applied for preterm birth prediction: Electronic Health Records (EHR), Transvaginal Ultrasound (TVS), and Electrohysterogram (EHG) (or Uterine Electromyography (EMG)). [17, 18, 19]. EHR is made up from the data that is included in medical history of the patient (weight, past hospital visits, smoking habits, etc.). Unfortunately, there is no one standard template used in hospitals (be it in different cities or internationally). Therefore, it is hard to compare how different authors evaluate and predict outcomes since they use various features [15]. TVS is acquired using sound waves, creating an image of the pelvic organs [20]. Unfortunately, this is an invasive technique. This procedure is inconvenient and painful for both mother and fetus, and also there is a possibility of infection and it can put both lives in danger [21]. Moreover, it is performed only once - in the first trimester. EHG is measured through a non-invasive method that depicts contractile activity of uterus with the help of electrodes [22]. Using various prediction methods demonstrated that EHG records are able to provide adequate data for preterm labor prediction, and help diagnosing labor more accurately than other clinical methods [23].

Two methods are typically used to forecast preterm birth based on EHG uterine records: either to differentiate between preterm and term delivery, or to differentiate between pregnancy and labor in preterm or term cases [23]. The two methods can be separated into two groups: methods that deal with the complete EHG records or signals, and methods that deal with specific contraction bursts that correspond to uterine contractions. The uterus is considered a complex, non-linear

dynamic system since it is made up of billions of intricately coupled cells with non-linear reactions.

1.2 Literature Review

Serious consequences of premature birth highlight the critical need to enhance existing preterm labor prognosis technologies. However, several challenges persist in prediction efforts [15]:

- **Difficult Timely Detection:** Due to the spontaneous nature of pregnancy data, it can be challenging to identify signs of preterm labor promptly. Numerous factors can contribute to premature birth, including previous health issues, smoking, alcohol consumption, drug use, maternal age, and multiple pregnancies (such as twins or triplets). Unfortunately, not all causes have been fully identified, complicating timely detection [24]. Most preterm deliveries happen suddenly ("spontaneous preterm birth (sPTB)") without the mentioned known factors, making it difficult to get the right referrals and treatment for specialized care [25].

- **Class Imbalance Issue:** The challenge arises when the number of samples in the class of interst is significantly lower than other classes', leading to unrealistically high predictive outcomes.

One of the most popular methods to deal with data imbalance are oversampling techniques, especially the creation of synthetic samples, like Synthetic Minority Over-Sampling Technique [26] (SMOTE) and Adaptive Synthetic sampling approach [27] (ADASYN) [28, 29]. Despite the frequent usage in the research, there is a flaw to these methods: it is essentially creates samples from "thin air" [29].

Another resampling method is undersampling. Although they are not as frequently used as oversampling techniques, new papers are being published every year, exploring new possibilities [30]. Unfortunately, their use is not recommended for small datasets. Specifically, undersampling can greatly decrease the number of samples available for training the model, which may result in an underestimation of its performance [31].

According to [29], it is encouraged to use ensemble methods to mitigate the problem of class imbalance. They bring together multiple techniques to create a more complete and robust solution. Some of them are Random Data Partitioning, EasyEnsemble, SMOTEBagging, etc. [32, 33, 34]. Such methods have consistently outperformed other algorithms when dealing with datasets that have class imbalance issues [35].

Although extensive research on preterm birth has been conducted, there is no definite answer to the questions like ”What is the “best” frequency content of the

signals to extract features for classification task?", nor ”Which features could be considered the “best” ones to achieve a classification accuracy as high as possible?".

According to [23], various features have been used to classify individual pregnancy and labor contraction bursts. These include the Root Mean Square (RMS) value; amplitude and area under the contraction curve; contraction power; and spectral features such as peak frequency, mean, peak, and median frequencies of the power spectrum, mean power frequency. Other approaches have considered Electrohysterogram (EHG) propagation velocity [36], wavelet-based features, and time-domain metrics (like time reversibility, sample entropy, and variance entropy) [37].

1.3 Related Works

Predicting preterm birth is a complex challenge. As noted earlier, various issues make the process quite difficult. For instance, some studies ([38, 39]) have pointed out that current methods often claim unrealistically high accuracy rates, often due to wrong handling of the class imbalance issue. Therefore, it’s essential to consider the ratio of samples in the dataset and prepare it accordingly, as well as correctly.

It has been found that a number of papers have used oversampling methods before dividing the dataset into testing and training sets [40]. The mentioned strategy may cause data leakage. It happens because created training samples are closely related to the original instances that will be in the testing (validation) set. As a result, some of these artificially generated samples may also end up in the testing set, which can lead to overly optimistic results that mainly show how well the model can remember the training samples instead of demonstrating its true predictive performance [40].

It has been noticed that most of the time researches use either a resampling method or K-fold Cross-Validation (CV). However, only a small percentage seems to utilize both techniques and show the potential problems in the prediction results, such as high bias [41]. To the author’s best knowledge, not a lot of papers exist that make a full comparative analysis of an extensive list of oversampling techniques with CV (during and after them) and without one as a baseline like in [42]. Figure 1.3 demonstrates the difference between approaches. In "CV after Resampling" approach, oversampling is performed on the original datasets before CV and performance evaluation, and complexity measures are calculated for the oversampled sets. In "Resampling during CV" approach, oversampling, naturally, occurs during CV, but only the training folds are oversampled. Classifiers are trained on these oversampled training sets and tested on the original test folds. It should be noted that the CV after oversampling leads to unrealistic results, making the second approach more reliable.

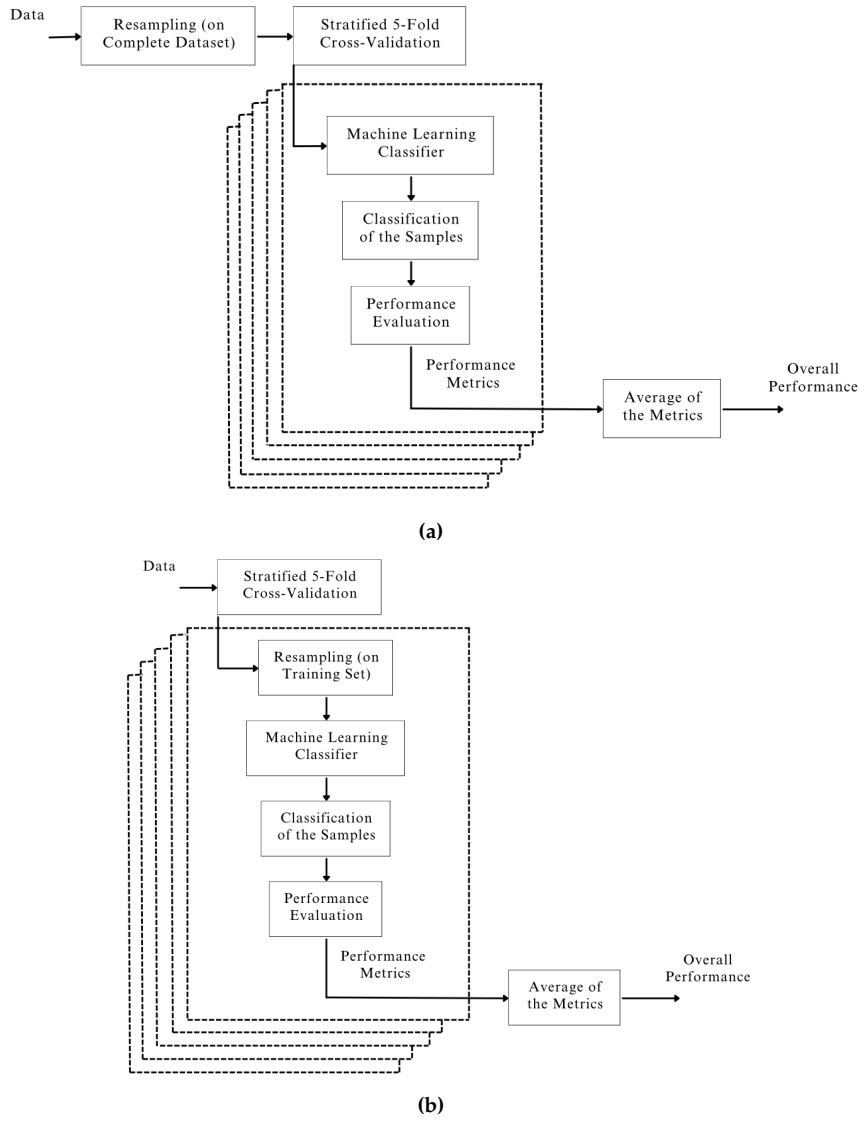

Figure 1.3: Experimental setup architecture of [42], illustrating the work's 2 main approaches: a) the incorrect approach - CV after resampling on the whole dataset; b) the correct approach - resampling during each fold of CV.

1.4 Motivation and Problem Statement

The project's main objective is to enhance current methods for detecting cases of premature birth. Preterm delivery is an important issue that affects millions of people across the world. By improving the accuracy of preterm birth predictions, this technology can potentially save lives and better prepare both medical personnel and mothers for the challenges that may arise. The robust method of handling

the highly imbalanced data contributes to the ongoing research on EHG analysis and may accelerate the process of creating a useful tool for obstetricians and neonatologists.

The main aim of this work is to make a comprehensive comparative analysis of using CV with not only the oversampling methods but also undersampling and ensemble ones to create a helpful guide for other researchers to have reliable and realistic prediction models.

1.5 Ethical and Professional Responsibilities

1.5.1 Ethical Responsibility

Working with a predictive model for preterm birth involves several ethical considerations that must be addressed to ensure responsible implementation. One of the main concerns is the privacy and confidentiality of sensitive health data. In this study, we will use several established open-source databases from a reputable source. This way, the privacy and confidentiality, as well as the informed consent, are not the issues for this project. Nevertheless, it is crucial to ensure that data does not contain any personally identifiable information (PII) that could compromise privacy of the people involved. The anonymity of the used data should be thoroughly checked.

Incorrect handling of data, especially regarding resampling strategies before training, can lead to overly optimistic performance. This can mislead healthcare workers and patients about the predictions' reliability. Therefore, this project will implement CV techniques and study the performance to ensure that accuracy metrics reflect realistic results.

Finally, there is a risk of overly relying on the predictions. Predictive tools can provide important insights, but they should not completely replace clinical judgment. It should be stated that predictive model is just a support tool for professionals.

1.5.2 Informed Judgments

Ensuring that decisions made during the development of the project are well-informed requires a multi-dimensional approach. We will start with a thorough analysis of the dataset, including understanding its structure, potential biases, and limitations. Moreover, we will make sure to communicate with the supervisor in order to get the needed feedback from the knowledgeable expert. Throughout the project, we will implement an iterative feedback loop where we continuously evaluate our approaches based on findings and the supervisor's input. This will allow us to adapt our strategies as needed.

If possible, we will consult with experts from various fields, including data science, machine learning, obstetrics, and public health. This collaboration will help us integrate diverse perspectives and ensure that technical decisions align with real-world healthcare and societal needs.

In addition, we will continuously update our literature review in order to have the most recent and up-to-date information and publications from reputable sources, such as peer-reviewed journals with good impact factors. Using the mentioned tactics, informed judgments can be made.

1.5.3 Global Context

The project's goal is to improve predictive accuracy by using effective methods for handling imbalanced data (such as oversampling, undersampling and ensemble techniques). This is especially important in regions with fewer resources, where the ability to accurately identify preterm pregnancies can lead to timely interventions and potentially save lives. Dependable predictions can help utilize medical resources more effectively, ensuring that "at-risk" individuals receive necessary care.

The findings from this project can theoretically contribute to international efforts that seek to reduce preterm birth rates, aligning with the goals of organizations like the World Health Organization. By sharing insights on best practices for managing imbalanced data, this project can provide a valuable framework for similar studies globally. This way, we will make a contribution to the scientific world.

In conclusion, this project can make a meaningful impact on maternal and infant health worldwide. Not only it will contribute to the academic community but also potentially help families around the world.

1.5.4 Economic Impact

Reliable prediction in the short term can improve healthcare efficiency by allowing for more precise risk management. Identifying high-risk pregnancies early on could lower unnecessary hospitalizations and emergency interventions, resulting in immediate cost savings for healthcare providers. Nonetheless, incorporating the predictive tool into clinical practice will require training healthcare professionals. While essential for successful implementation, these training sessions may initially draw resources and time away from regular clinical duties.

Preterm birth prediction using machine learning has the potential to significantly lower healthcare costs associated with premature delivery. By minimizing complications and hospital stays through early interventions, families and healthcare systems can save significant amounts on treatment costs. Moreover, better

health outcomes for mothers and infants can lead to a healthier workforce, positively affecting productivity levels. Healthier children are likely to have lower rates of long-term health issues, reducing future healthcare expenses.

Additionally, by providing a reliable prediction method, the project can help make quality prenatal care accessible worldwide, particularly in underserved communities. This improvement can lead to broader economic benefits for those populations.

Limited budgets and resources can make it hard to use predictive tools, especially if there isn't enough money or infrastructure. It's important to make sure the model is affordable and can grow as needed. If the results are too good (overly optimistic), it could lead to financial losses, which would defeat the purpose of using automatic predictions.

1.5.5 Environmental Impact

This project does not create any waste or byproducts. However, we still need to take into account its environmental impact.

By utilizing open-source datasets, we significantly reduce the need for extensive data collection and associated resource consumption. This approach minimizes the environmental footprint linked to data generation since we use existing data rather than initiating new resource-intensive studies.

In terms of computational resources, we will prioritize using the least amount of electricity. We will use optimized algorithms for efficiency, reducing the computational power required for training and testing. This efficiency results in lower energy consumption, making the entire predictive modeling process more eco-friendly.

One of the aims is to utilize existing hardware whenever possible, avoiding buying new equipment that contributes to electronic waste. If new hardware will be required, it would be used to its fullest potential and bought from responsible manufacturers that follow environmentally friendly practices, including recycling and energy-efficient production methods. By implementing these sustainability measures into the project design, we try to advance the prediction of preterm births and set an example for future research initiatives.

1.5.6 Societal Impact

One of the benefits of the project is the improvement in maternal and infant health research. One day, by employing a dependable predictive approach, healthcare professionals can improve their detection rates of high-risk pregnancies (at an early stage). This early identification will help with timely medical interventions, which can significantly decrease complications of premature births. Improved prenatal

care can lead to reduced rates of infant mortality, directly benefiting both mothers and babies and fostering healthier family environments.

In addition to direct health advantages, lowering the number of premature births can result in substantial cost savings for families and healthcare systems. Decreased complications and hospitalizations lead to reduced medical expenses (ranging from medications and hospital visits to specialized equipment like inhalers and oxygen concentrators), enabling families to allocate resources to education and other essential needs.

By making predictive tools reliable, accessible and relevant, we can help to ensure that vulnerable groups receive the care they need, promoting equality in healthcare access and outcomes. This can also educate people by raising awareness about maternal health issues and starting discussions that encourage better health practices and support systems.

Chapter 2

Investigated Methodology

2.1 Experimental Setup

In this study, a comparative analysis has been conducted of three different resampling techniques — oversampling, undersampling, and ensemble methods — and their impact on the classification performance of a Random Forest model. For each sampling method, five different techniques have been tested. The analysis is performed both with and without the use of CV to assess the effectiveness of these techniques in addressing class imbalance and improving model generalization.

To assess the generalization ability of the models, Stratified 5-Fold CV has been used. Stratified K-Fold ensures that each fold has the same class distribution as the original dataset, providing a more reliable estimate of model performance when dealing with imbalanced datasets. The model is trained and evaluated on each fold.

Figures 2.1-2.3 illustrate the flow chart of the experiment's process. For each resampling technique, the following experiments have been performed:

- **Oversampling**: Oversampling techniques are applied to the minority class to generate new samples, followed by training a Random Forest classifier.
- **Undersampling**: Undersampling techniques are applied to the majority class to balance the dataset, and then a Random Forest model is trained.
- **Ensemble Method**: Ensemble classifiers are trained using balanced bootstrapped samples.

Each of these experiments is conducted both with and without Stratified K-Fold CV, resulting in 6 different experimental conditions (without including baseline cases). Figure 2.2 shows the framework of performing classification without CV. Resampling is used only on training set. It should be noted that in the Figure

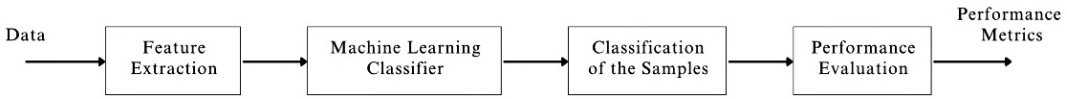

Figure 2.1: Classification without any resampling techniques, marked as Baseline (with and without CV).

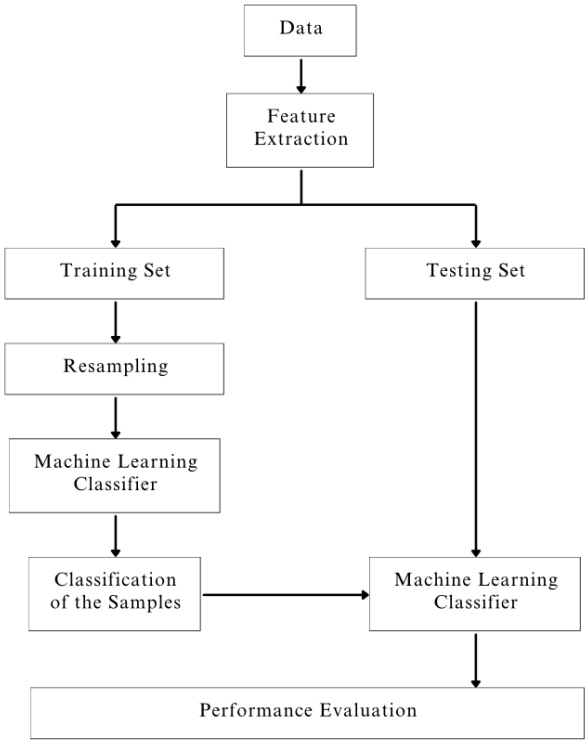

Figure 2.2: Classification with resampling techniques and without CV.

2.3 oversampling is performed during CV: after splitting the dataset into folds, on the training set only. Finally, the performance metrics are averaged across all folds to find mean values and assess the models. Baseline results (Figure 2.1) are found, meaning no resampling technique is applied to the dataset. By evaluating these experimental conditions, the study's aim is to identify the combined effects of resampling techniques and CV, and how different sampling methods affect the results.

In order to investigate the effect of features on the prediction results, this experimental setup has been utilized on 3 sets of different features:

- The categorical features that can also be referred as EHR (ex. mother's age);
- The features extracted from the EHG signal;

- Combined set of the aforementioned features.

2.2 Resampling Techniques

Three resampling types have been evaluated to address the class imbalance present in the dataset. For each type, five different techniques have been implemented to assess their impact on classification performance, that are publicly available Python scikit-learn library "Imbalanced-learn" (X is original set of data):

- **Oversampling**:
  - **Random Oversampling**: A technique that duplicates samples from the minority class randomly to balance the class distribution.

$$X_{\text{oversampled}} = X_{\text{minority}} + \{X_{\text{minority}}\} \tag{2.1}$$

  - **SMOTE**: A technique that creates synthetic samples for the minority class by interpolating between existing instances of minority class.

$$X_{\text{synthetic}} = X_{\text{minority}} + \alpha(X_{\text{minority,1}} - X_{\text{minority,2}}) \tag{2.2}$$

    where $\alpha \in [0, 1]$ is a random number and $X_{\text{minority,1}}, X_{\text{minority,2}}$ are randomly selected minority class instances.
  - **ADASYN**: A technique that generates synthetic samples for the minority class, focusing more on instances that are harder to classify (near the decision boundary).

$$X_{\text{synthetic}} = X_{\text{minority}} + \alpha(X_{\text{minority,1}} - X_{\text{minority,2}}) \tag{2.3}$$

    where $X_{\text{minority,1}}, X_{\text{minority,2}}$ are instances near the decision boundary.
  - **SVMSMOTE**: A technique that combines SMOTE with Support Vector Machine (SVM) to generate synthetic samples along the decision boundary.

$$X_{\text{synthetic}} = X_{\text{minority}} + \alpha(X_{\text{SVMLine}} - X_{\text{minority}}) \tag{2.4}$$

    where X_{SVMLine} represents the decision boundary defined by the SVM.
  - **BorderlineSMOTE**: A variant of SMOTE that generates synthetic samples near the decision boundary.

$$X_{\text{synthetic}} = X_{\text{minority}} + \alpha(X_{\text{boundary,1}} - X_{\text{boundary,2}}) \tag{2.5}$$

    where $X_{\text{boundary,1}}, X_{\text{boundary,2}}$ are instances near the decision boundary.
- **Undersampling**:

- **Random Undersampling**: A technique that randomly selects and discards samples from the majority class.

$$X_{\text{undersampled}} = X_{\text{majority}} - \{X_{\text{majority,discarded}}\} \tag{2.6}$$

- **Cluster-based Undersampling**: A technique that reduces the majority class by clustering its instances and selectively removing redundant or less representative ones.

$$X_{\text{clustered}} = \text{Cluster}(X_{\text{majority}}), \quad X_{\text{undersampled}} = X_{\text{majority}} - \text{RedundantClusters}. \tag{2.7}$$

- **Tomek Links Undersampling**: A technique that identifies pairs of nearest neighbors where one is from the majority class and the other from the minority class, aiming to clean the decision boundary.

$$X_{\text{tomek}} = \{X_{\text{majority}}, X_{\text{minority}}\}, \quad \text{such that } \text{dist}(X_{\text{majority}}, X_{\text{minority}}) < \epsilon, \tag{2.8}$$

  where ϵ is small distance threshold below which the majority class sample is considered near the decision boundary and is thus removed

- **NearMiss Undersampling**: A technique that uses distance to select which majority class samples to keep or discard.

$$X_{\text{nearMiss}} = \{X_{\text{majority}} : \text{dist}(X_{\text{majority}}, X_{\text{minority}}) \leq \epsilon\} \tag{2.9}$$

  where ϵ is a user-defined threshold distance that determines how close the majority class samples must be to the minority class samples to be selected for inclusion.

- **ENN Undersampling**: A technique that uses nearest neighbors (NN) to remove noisy and misclassified samples from the majority class.

$$X_{\text{ENN}} = \{X_{\text{majority}} : \text{NN}(X_{\text{majority}}) \notin \text{class}\} \tag{2.10}$$

- **Ensemble Methods**:

  - **Easy Ensemble**: An ensemble method that combines multiple models trained on balanced subsets of the majority class via random undersampling.

$$X_{\text{ensemble}} = \{\text{RandomUndersampling}(X_{\text{majority}})\} \tag{2.11}$$

  - **SMOTEBoost**: A technique that integrates SMOTE with boosting, generating synthetic minority class samples during the boosting process.

$$X_{\text{SMOTEBoost}} = \text{SMOTE}(X_{\text{minority}}) + \text{Boosting}(X_{\text{minority}}) \tag{2.12}$$

– **RUSBoost**: Similar to SMOTEBoost, but it combines random undersampling of the majority class with boosting.

$$X_{\text{RUSBoost}} = \text{RandomUndersampling}(X_{\text{majority}}) + \text{Boosting}(X_{\text{minority}}) \tag{2.13}$$

– **RAMOBoost**: An extension of RUSBoost that uses a more advanced sampling technique to adaptively undersample the majority class.

$$X_{\text{RAMOBoost}} = \text{AdaptiveSampling}(X_{\text{majority}}) + \text{Boosting}(X_{\text{minority}}) \tag{2.14}$$

– **Balanced Random Forest**: A technique that modifies the traditional random forest by applying random undersampling to each bootstrap sample of the majority class.

$$X_{\text{BalancedRF}} = \text{Bootstrap}(X_{\text{majority}}) + \text{Undersample}(X_{\text{majority}}) \tag{2.15}$$

2.3 Model: Random Forest

It has been decided to use Random Forest classifier for classification for this project based on the results of the Preliminary Study shown in the Experimental Outcomes section of Results and Discussion chapter. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the majority class from the individual trees. It is fairly common classifier used in preterm birth prediction. Hyperparameters of the Random Forest model (for example, maximum depth, minimum samples split) are optimized using grid search based on the training data (for each technique separately). The Random Forest classifier can be expressed as follows:

$$f(x) = \text{MajorityVote}\left(T_1(x), T_2(x), \ldots, T_M(x)\right), \tag{2.16}$$

where M is the number of decision trees in the forest, $T_m(x)$ is the prediction made by the m-th tree for input x. MajorityVote takes the predictions from all trees and returns the most frequent result.

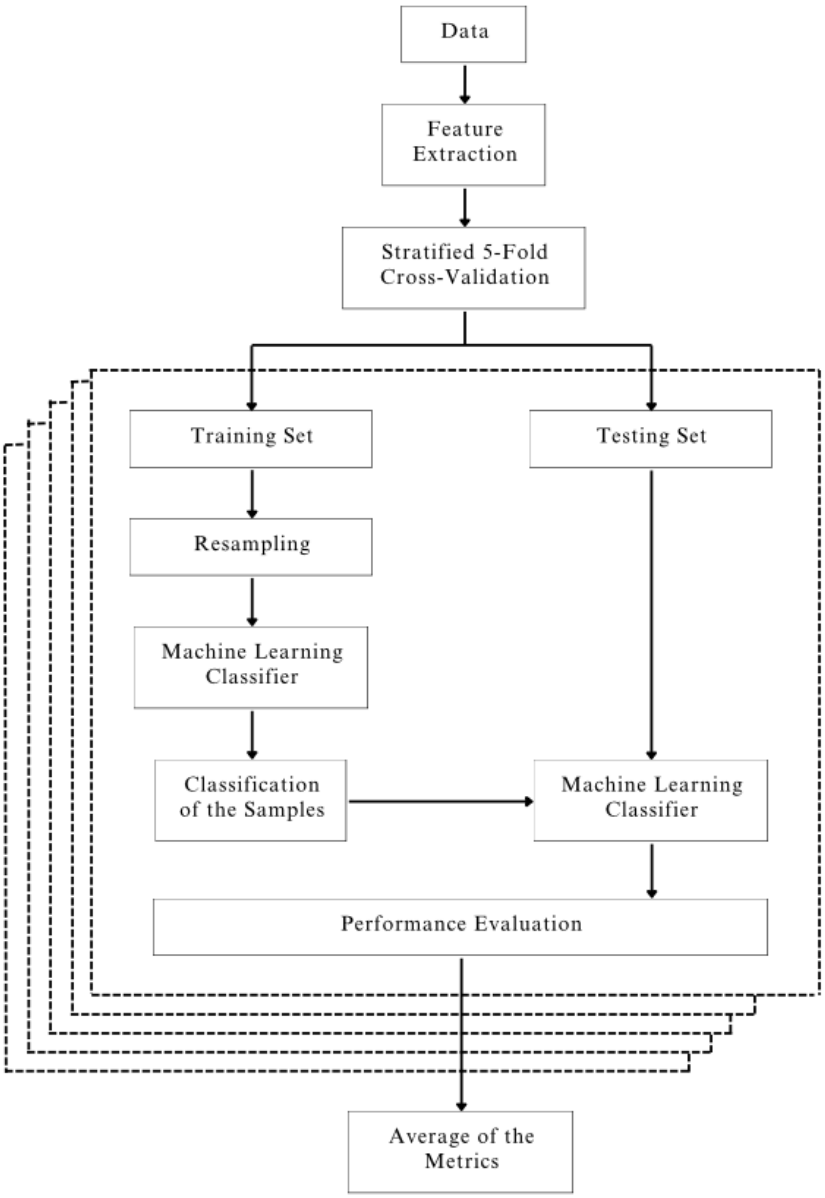

Figure 2.3: Classification with resampling techniques and CV.

Chapter 3

Results and Discussions

3.1 Datasets

3.1.1 Dataset for Preliminary Study

A small dataset of 58 samples has been used. It has contained already extracted features, presented numerically [43]. The included features are: Count Contraction - number of contractions, Length of Contraction (in relation to the womb), STD (standard deviation), Entropy (randomness of molecules), and Contraction's Time Interval. From Figure 3.1, it is evident that this dataset is also imbalanced: only 32.8 percent of all samples are marked as preterm. In order to see how features correlate to the prematurity, correlation color-coded plot (or also known as the heatmap) has been created as seen in Figure 3.2. The higher correlation indicator, the more values are linearly related. Therefore, this matrix could be used to identify and choose the features that are "strongly" correlated with the target value.

3.1.2 Term-Preterm Electrohysterogram Database

The data used in this study comes from the Term-Preterm Electrohysterogram (EHG) Database (TPEHG DB) available on Physionet [44]. This dataset contains EHG recordings from 300 pregnancies, with the goal of classifying pregnancies as either term or preterm based on uterine electrical activity. The data includes both physiological signals and medical history of the patients.

The dataset consists of 300 EHG recordings, of which 262 correspond to term pregnancies and 38 correspond to preterm pregnancies. The data has been collected from 300 different patients, with each subject having recorded uterine electrical signals during their pregnancy.

When the records from patients with incomplete or missing medical histories have been removed, the dataset has been reduced to 169 subjects, including 19

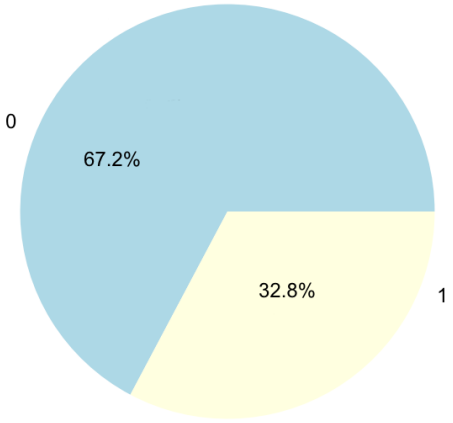

Figure 3.1: Class distribution of the dataset [43] used in preliminary study, where "1" - preterm birth, "0" - term birth.

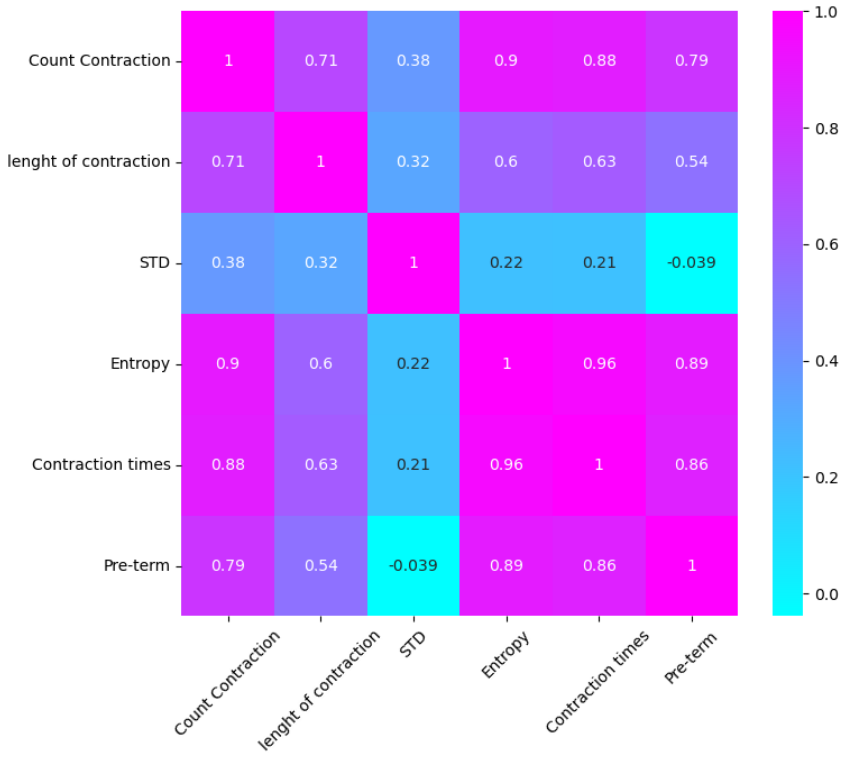

Figure 3.2: Correlation heatmap based on the data from [43].

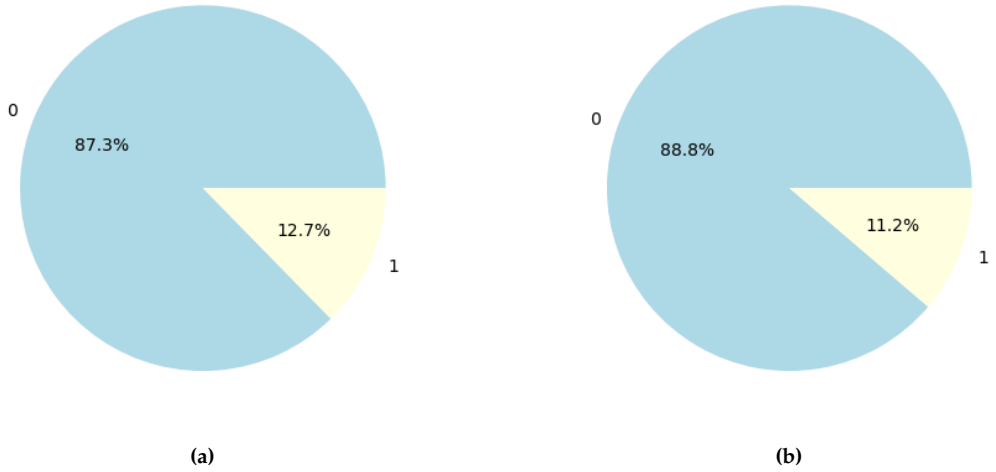

Figure 3.3: Comparison of TPEHG Dataset before and after dropping incomplete entries: a) original distribution of classes, b) the distribution after cleaning the data. "1" - preterm birth, "0" - term birth

preterm pregnancies and 150 term pregnancies. As seen from Figures 3.3, the distribution of the classes does not suffer significantly.

The authors of the dataset have already filtered the data using filtered with different bandwidth filters. It has been selected to work with the data processed by a 0.08-4.0 Hz bandwidth filter, since it has been recognized that the uterine EMG content ranges from 0 to 5 Hz [44].

The patient data includes both EHG recordings and associated medical history, stored in the form of a header file. The medical history (or EHR) contains several key attributes, such as pregnancy duration, gestation duration at the time of recording, maternal age, the number of previous deliveries, number of previous abortions, weight at the time of recording, whether hypertension has occurred, whether the patient is diabetic, placental position, whether there has been bleeding during the first trimester, whether there has been bleeding during the second trimester, whether funneling has occurred, and whether the patient is a smoker. Since some categorical features have not been numerical, ordinal encoding has been performed. In addition, several key features have been extracted from the EHG signals. These features have been shown to be useful in distinguishing between term and preterm pregnancies [45]. The most commonly used features include:

- Median Frequency – statistical measure of the central tendency of the frequency spectrum.
- Peak Frequency – frequency at which the signal showcases its maximum power.

- Sample Entropy – measure of signal regularity and complexity, useful for detecting changes in uterine activity.
- Root Mean Square (RMS) – measure of the signal's amplitude that is commonly used in signal processing.

The combination of EHG signal features and maternal medical history provides a robust dataset for developing predictive models for preterm birth risk.

3.2 Performance Measures

To evaluate the performance of the model, several metrics have been used, providing a comprehensive understanding of its strengths and weaknesses across different aspects:

- **Accuracy:** Accuracy is the proportion of correctly predicted instances out of the total instances. It is calculated as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3.1}$$

  where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. Although widely used, accuracy can be misleading in imbalanced datasets, being biased towards the majority class [42]. Therefore, it has been decided to focus on the following metrics.

- **Precision:** Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is defined as

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{3.2}$$

  High precision indicates that the model has a low rate of false positives.

- **Recall:** Recall, also known as sensitivity or true positive rate, is the proportion of actual positive instances that have been correctly identified by the model. It is given by

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3.3}$$

  High recall indicates that the model is successful inpredicting most positive cases, but it may also have a higher number of false positives.

- **F1 Score:** The F1 Score is the harmonic mean of precision and recall, providing a single measure that balances both metrics. It is particularly useful when the class distribution is imbalanced. It is calculated as

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3.4}$$

  The F1 score ranges between 0 and 1, with 1 being the best possible score.

- **ROC AUC:** The Receiver Operating Characteristic (ROC) curve plots the true positive rate (Recall) against the false positive rate (FPR). The Area Under the Curve (AUC) provides an aggregate measure of performance across all possible classification thresholds. An AUC value closer to 1 indicates a better performing model. AUC is insensitive to class distribution and is a useful metric in imbalanced datasets.

3.3 Experimental Outcomes

3.3.1 Preliminary Study

It has been decided to test the CV technique's efficiency. For this purpose, the dataset mentioned earlier has been used. The performance of 4 different classifiers has been compared: Decision Tree (DT) Classifier, Random Forest (RF) Classifier, Support Vector Classifier (SVC), and Stacked Classifier (combination of RF and SVC) - with and without CV. For this experiment, Stratified K-fold CV has been used (K=20, which is the optimal parameter according to [46]).

After completing the code and running it, the results have been presented in Table 3.1 and Table 3.2. It is evident that the results without CV are highly unrealistic, with 3 out of 4 classifiers reporting ideal performance. On the other hand, prediction models with CV demonstrated slightly more realistic metrics. As it can be seen, RF has shown great results. Therefore, it has been decided to use this classifier for the next experiments.

Table 3.1: Experiment results' metrics without CV

Classifier	Accuracy(%)	Precision(%)	F1 Score(%)	Recall(%)
Decision Tree	93.3	80	88.9	100
Random Forest	100	100	100	100
Support Vector Classifier	100	100	100	100
Stacked Classifier	100	100	100	100

Table 3.2: Experiment results' metrics with CV

Classifier	Accuracy(%)	Precision(%)	F1 Score(%)	Recall(%)
Decision Tree	98	90	90	90
Random Forest	100	95	95	95
Support Vector Classifier	98	92.5	93.3	95
Stacked Classifier	100	95	95	95

3.3.2 Key Findings

This subsection represents the combined effect of resampling techniques and CV. Each bar chart represents the performance metric for more clear view on the differences between techniques of the same sampling method. Although multiple performance metrics - Precision, Recall, and F1 Score - have been evaluated, only the ROC AUC is presented and discussed in the main body of this work. This is because ROC AUC provides a comprehensive measure of classification performance across different experiments and is sufficient for drawing overall conclusions in this context. The remaining metrics are provided in the Appendix A for reference.

Figure 3.4 and Figure 3.5 present the values of ROC AUC metric for the oversampling techniques without and with the presence of the CV respectively. Baseline values have also been included for comparison. It should be noted that most techniques have outperformed the baseline values except ROS. Without CV, models seem to be overfitting, showing high ROC AUC scores especially in case of using categorical dataset. However, other metrics have varied significantly and remain inconsistent (for example, SMOTE and ADASYN). With CV, the performance across all metrics becomes more stable and reliable. The values of the conducted 3 experiments have become more comparable.

Figure 3.6 and Figure 3.7 show the performance metrics for the undersampling techniques. For better comparison, baseline values have also been included. Most of the techniques have outperformed baseline values. However, Tomek Links method has not presented any improvement aside from the ROC AUC value no matter which features have been utilized. It could be explained by the fact that this method did not significantly improve the class distribution. Unlike oversampling techniques case, the set of categorical features did not result in highest performance measures. After applying CV, all metrics have shown mixed effect: half of the techniques reported improvement while the remaining ones decreased considerably. For most techniques, the set of combined features has worked better.

Figure 3.8 and Figure 3.9 present the values for the ensemble techniques without and with CV respectively. Similarly with the previous figures, baseline values have also been included for comparison. All techniques have outperformed baseline values. The values of metrics, overall, are greater than in the case of undersampling. After applying CV, all metrics have shown mixed effect: 3 out of

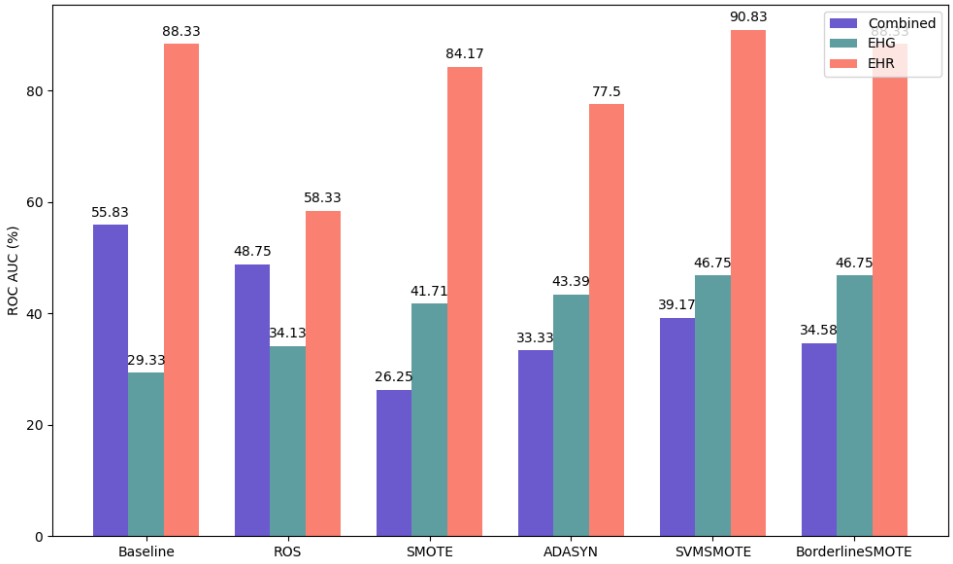

Figure 3.4: ROC AUC comparison of different oversampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

5 techniques reported improvement while the remaining ones decreased considerably. However, the range of values across different methods have decreased: improving the ”weakest" models and reduced the performance of the best ones. Moreover, there is no definite answer to which feature set has performed best since the results are mixed in both cases (with and without CV). However, it is worth mentioning that the values of the metrics have become relatively similar in all 3 cases.

Based on the results of the ”combined" feature set, ROC AUC generally improved across different resampling techniques. It indicates that the model learned to better discriminate between classes and generalize, possibly due to improved class balance from resampling and the more robust evaluation from CV. On the other hand, Precision, Recall, and F1 score are more sensitive to class imbalances because they focus on the positive class performance. While CV stabilizes performance and prevents overfitting on a single training split, the effect on Precision, Recall, and F1 score might not be as uniform. It seems this instability stems from the resampling technique, which can be sensitive to fluctuations in data distribution, especially with imbalanced classes.

Categorical features are often easier for a model to work with, especially when there is a clear relationship between the feature and the target variable. When CV has not been applied, the model have performed best with categorical features, likely because it could quickly ”learn" these relationships from the oversampled data. Without CV, oversampling may also lead to overfitting to these easily learned

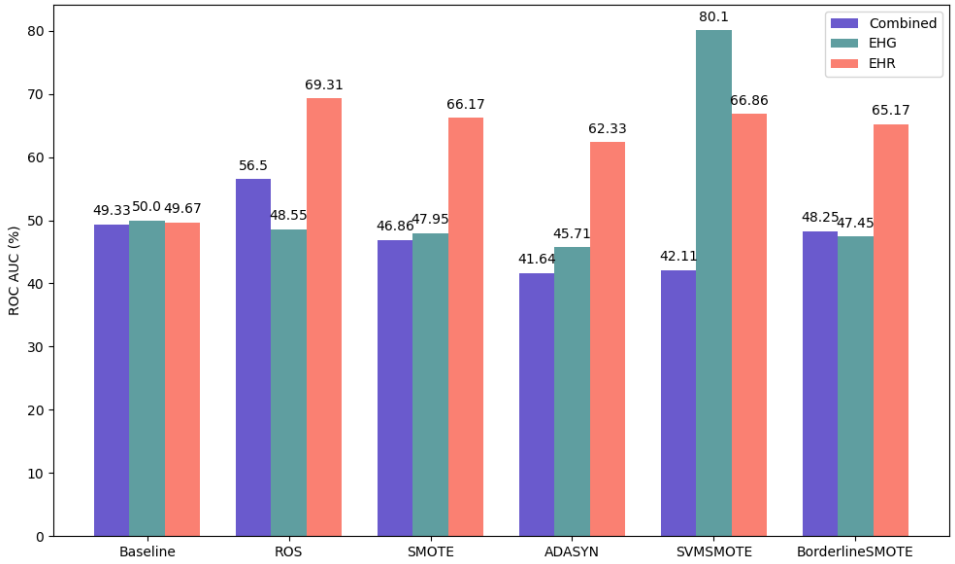

Figure 3.5: ROC AUC comparison of different oversampling techniques using 3 different feature sets - ”’EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with Stratified 5-fold CV.

features, inflating performance.

On the other hand, EHG Signal Feature are more complex and may contain noise, but they often have a richer representation of the underlying patterns. When CV has been applied, however, the performance across all feature sets has become more similar. This suggests that the added complexity of EHG signal features, combined with the more rigorous evaluation under CV, resulted in a more stable performance across different features. The model may have been forced to focus on the generalizable patterns in the data rather than memorizing the intricacies of the EHR or EHG features.

When using both categorical and signal features, the model has been exposed to a more diverse set of information. The combination of these features has provided a richer representation of the data, but with CV the performance across all feature sets converged. This convergence could be attributed to the fact that the resampling techniques and CV mitigated the potential overfitting that might have occurred when dealing with more complex feature sets.

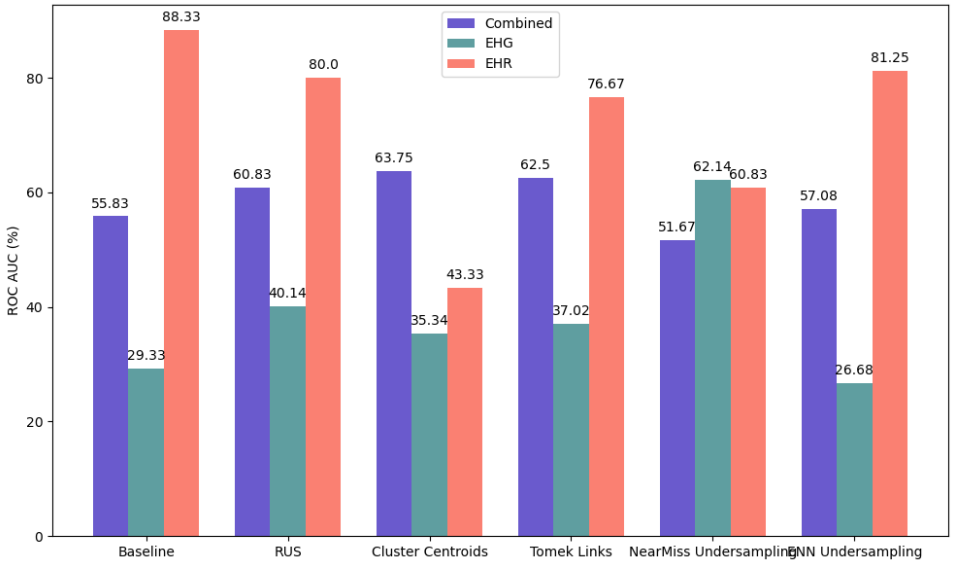

Figure 3.6: ROC AUC comparison of different undersampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

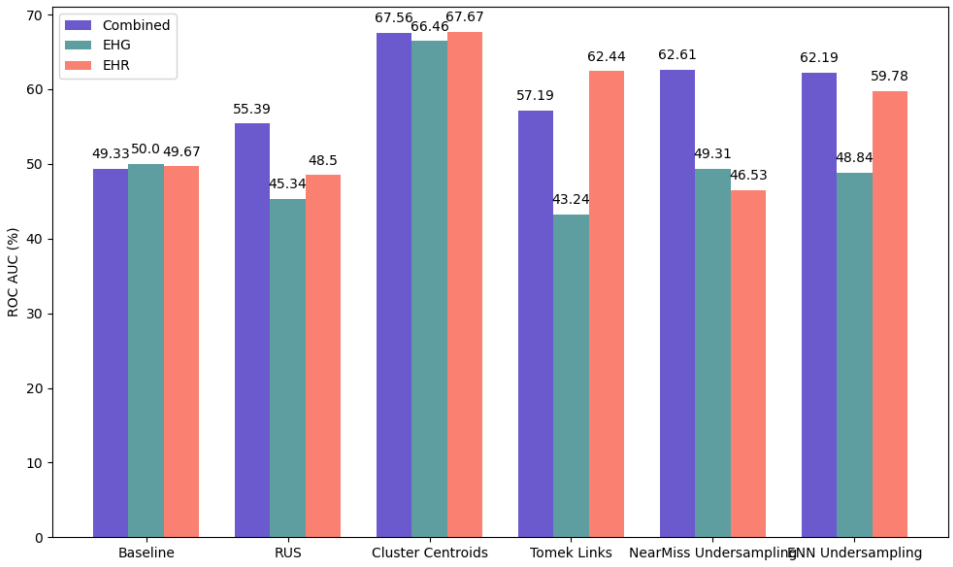

Figure 3.7: ROC AUC comparison of different undersampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with Stratified 5-fold CV.

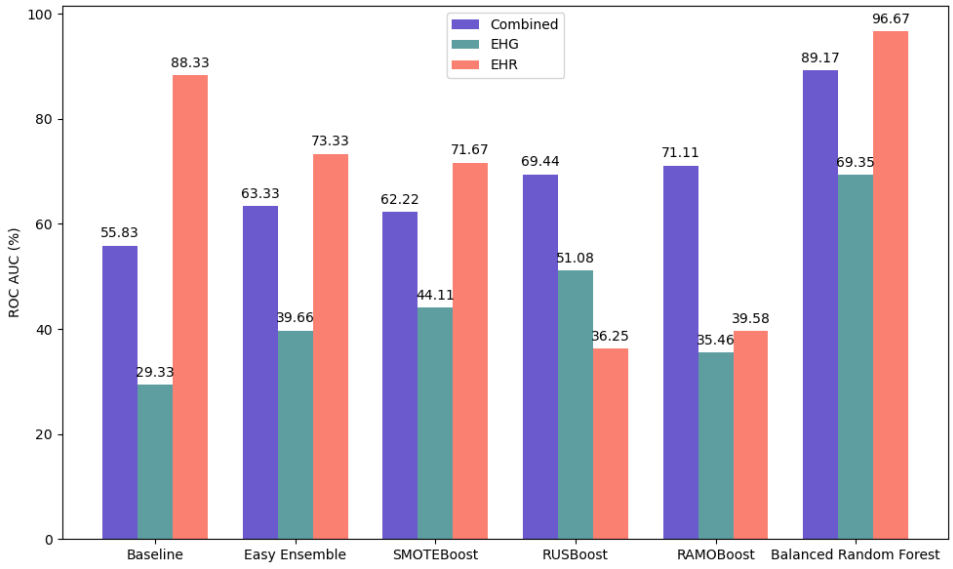

Figure 3.8: ROC AUC comparison of different ensemble techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

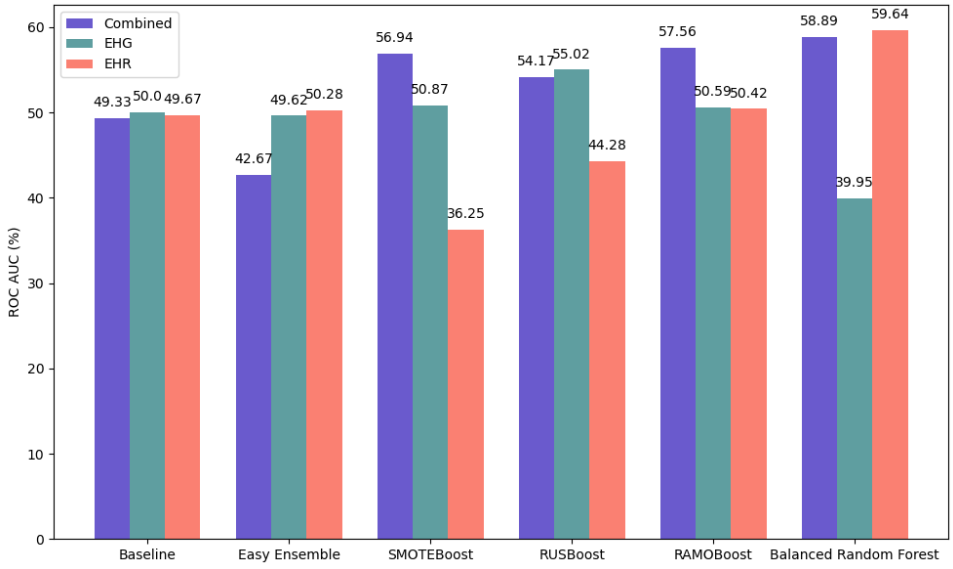

Figure 3.9: ROC AUC comparison of different ensemble techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with Stratified 5-fold CV.

Chapter 4

Conclusion

4.1 Summary of Work Done

During the Capstone project, various experiments on EHG signals have been carried out. Literature Review has been updated and revised. More importantly, in Capstone II feature extraction has been studied in more detail. The experimental setup outlined in Capstone I has been used in order to investigate the effect of different types of features on the prediction and its performance metrics. Experiments have been carried out using Python. The results show that the combination of resampling techniques and Stratified K-Fold CV can be an effective way to handle imbalanced datasets and provide more reliable performance metrics. While resampling methods help balancing the dataset, their impact is best assessed when coupled with CV, which ensures that the model's performance reflects its ability to generalize. In addition, the type of features that are used in the model also plays a significant role in final model performance, with categorical features showing better results without CV, while the performance became more consistent across feature sets when CV has been applied.

4.2 Future Work

In the future, it would be beneficial to implement more machine learning classifiers, like SVM and Decision Trees, and compare their performance. In addition, hybrid approaches could be used: like SMOTE+Tomek Links (SMOTETomek) or SMOTE+ENN Undersampling, which combine oversampling and undersampling to reduce the chance of introducing noise and overfitting while maintaining a good balance between the classes.

List of Publications

1. A. Alimgozhina and M. T. Akhtar, "Exploring the joint effect of resampling techniques and cross-validation for preterm birth prediction," in *Proc. 47th Annual International Conference IEEE Engineering Medicine Biology Society (EMBC 2025)*, July 14–18, 2025, Copenhagen, Denmark, 2025 (accepted).

2. A. Alimgozhina and M. T. Akhtar, "The joint performance of resampling techniques with cross-validation and its effects on prediction of preterm birth," *IEEE Access*, 2025 (in preparation).

Bibliography

[1] International Business Machines (IBM). *Machine Learning*. https://research.ibm.com/topics/machine-learning. Accessed: 2024-02-20.

[2] Osvaldo Simeone. "A very brief introduction to machine learning with applications to communication systems". In: *IEEE Transactions on Cognitive Communications and Networking* 4.4 (2018), pp. 648–664.

[3] World Health Organization. *Preterm Birth*. https://www.who.int/news-room/fact-sheets/detail/preterm-birth. Accessed: 2024-09-18.

[4] Joshua P Vogel et al. "The global epidemiology of preterm birth". In: *Best practice & research Clinical obstetrics & gynaecology* 52 (2018), pp. 3–12.

[5] World Health Organization. *Born too soon: decade of action on preterm birth*. World Health Organization, 2023.

[6] Anne Greenough. "Long-term respiratory consequences of premature birth at less than 32 weeks of gestation". In: *Early human development* 89 (2013), S25–S27.

[7] Daniela Morniroli et al. "Beyond survival: the lasting effects of premature birth". In: *Frontiers in Pediatrics* 11 (2023), p. 1213243.

[8] Magdalena Wolska, Tomasz Piotr Wypych, and Pilar Rodríguez-Viso. "The Influence of Premature Birth on the Development of Pulmonary Diseases: Focus on the Microbiome". In: *Metabolites* 14.7 (2024), p. 382.

[9] Tobias Strunk et al. "Innate immunity in human newborn infants: prematurity means more than immaturity". In: *The journal of maternal-fetal & neonatal medicine* 24.1 (2011), pp. 25–31.

[10] Monique Rijken et al. "Mortality and neurologic, mental, and psychomotor development at 2 years in infants born less than 27 weeks' gestation: the Leiden follow-up project on prematurity". In: *Pediatrics* 112.2 (2003), pp. 351–358.

[11] David Drummond et al. "Educational and health outcomes associated with bronchopulmonary dysplasia in 15-year-olds born preterm". In: *PLoS One* 14.9 (2019), e0222286.

[12] Amina Benkessirat and Nadjia Benblidia. "Fundamentals of feature selection: An overview and comparison". In: *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2019, pp. 1–6.

[13] Garima Sharma, Kartikeyan Umapathy, and Sridhar Krishnan. "Trends in audio signal feature extraction methods". In: *Applied Acoustics* 158 (2020), p. 107020.

[14] Tin H Pham and Bijan Raahemi. "Bio-inspired feature selection algorithms with their applications: a systematic literature review". In: *IEEE Access* (2023).

[15] Tomasz Włodarczyk et al. "Machine learning methods for preterm birth prediction: a review". In: *Electronics* 10.5 (2021), p. 586.

[16] Kamalraj Subramaniam, Nisheena V Iqbal, et al. "A review of significant researches on prediction of preterm birth using uterine electromyogram signal". In: *Future Generation Computer Systems* 98 (2019), pp. 135–143.

[17] Jinshan Xu et al. "Review on EHG signal analysis and its application in preterm diagnosis". In: *Biomedical Signal Processing and Control* 71 (2022), p. 103231.

[18] Cheng Gao et al. "Deep learning predicts extreme preterm birth from electronic health records". In: *Journal of biomedical informatics* 100 (2019), p. 103334.

[19] Pihla Kuusela et al. "Second-trimester transvaginal ultrasound measurement of cervical length for prediction of preterm birth: a blinded prospective multicentre diagnostic accuracy study". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 128.2 (2021), pp. 195–206.

[20] Cleveland Clinic. *Transvaginal Ultrasound*. https://my.clevelandclinic.org/health/diagnostics/4993-transvaginal-ultrasound. Accessed: 2024-09-18.

[21] R Parameshwari and S Shenbaga Devi. "Acquisition and analysis of electrohysterogram signal". In: *Journal of Medical Systems* 44.3 (2020).

[22] Martim Almeida et al. "Electrohysterography extracted features dependency on anthropometric and pregnancy factors". In: *Biomedical Signal Processing and Control* 75 (2022), p. 103556.

[23] Franc Jager, Sonja Libenšek, and Ksenija Geršak. "Characterization and automatic classification of preterm and term uterine records". In: *PLoS One* 13.8 (2018), e0202125.

[24] P. Moglia. "Premature birth". In: *Magill's Medical Guide (Online Edition)* (2023).

[25] Marc Hershey et al. "Predicting the risk of spontaneous premature births using clinical data and machine learning". In: *Informatics in Medicine Unlocked* 32 (2022), p. 101053.

[26] Nitesh V Chawla et al. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[27] Haibo He et al. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee. 2008, pp. 1322–1328.

[28] Ibraheem M Alkhawaldeh, Ibrahem Albalkhi, and Abdulqadir Jeprel Naswhan. “Challenges and limitations of synthetic minority oversampling techniques in machine learning”. In: *World Journal of Methodology* 13.5 (2023), p. 373.

[29] Ahmad S Tarawneh et al. “Stop oversampling for class imbalance learning: A review”. In: *IEEE Access* 10 (2022), pp. 47643–47660.

[30] Zhongqiang Sun et al. “Undersampling method based on minority class density for imbalanced data”. In: *Expert Systems with Applications* 249 (2024), p. 123328.

[31] Tarid Wongvorachan, Surina He, and Okan Bulut. “A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining”. In: *Information* 14.1 (2023), p. 54.

[32] Ahmad B Hassanat et al. “Rdpvr: Random data partitioning with voting rule for machine learning from class-imbalanced datasets”. In: *Electronics* 11.2 (2022), p. 228.

[33] Xu-Ying Liu and Zhi-Hua Zhou. “Ensemble methods for class imbalance learning”. In: *Imbalanced learning: Foundations, algorithms, and applications* (2013), pp. 61–82.

[34] Zhang Yongqing et al. “Improved SMOTEBagging and its application in imbalanced data classification”. In: *IEEE Conference Anthology*. IEEE. 2013, pp. 1–5.

[35] Azal Ahmad Khan, Omkar Chaudhari, and Rohitash Chandra. “A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation”. In: *Expert Systems with Applications* (2023), p. 122778.

[36] Lasse Lange et al. “Velocity and directionality of the electrohysterographic signal propagation”. In: *PloS one* 9.1 (2014), e86775.

[37] Dima Alamedine, Mohamad Khalil, and Catherine Marque. “Comparison of different EHG feature selection methods for the detection of preterm labor”. In: *Computational and mathematical methods in medicine* 2013.1 (2013), p. 485684.

[38] Franc Jager et al. “Assessing velocity and directionality of uterine electrical activity for preterm birth prediction using EHG surface records”. In: *Sensors* 20.24 (2020), p. 7328.

[39] Jinshan Xu et al. “Realistic preterm prediction based on optimized synthetic sampling of EHG signal”. In: *Computers in Biology and Medicine* 136 (2021), p. 104644.

[40] Gilles Vandewiele et al. “Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling”. In: *Artificial Intelligence in Medicine* 111 (2021), p. 101987.

[41] Aydin Demircioğlu. “Applying oversampling before cross-validation will lead to high bias in radiomics”. In: *Scientific Reports* 14.1 (2024), p. 11563.

[42] Miriam Seoane Santos et al. “Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]”. In: *ieee ComputatioNal iNtelligeNCe magaziNe* 13.4 (2018), pp. 59–76.

[43] Nithish Paidimarri. *Preterm Birth Prediction using Machine Learning*. https://github.com/Nithish-456/Pre-Term-Birth-Prediction/tree/main. Accessed: 2024-08-01.

[44] Gašper Fele-Žorž et al. “A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and pre-term delivery groups”. In: *Medical & biological engineering & computing* 46 (2008), pp. 911–922.

[45] Gilles Vandewiele et al. “A critical look at studies applying over-sampling on the tpehgdb dataset”. In: *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*. Springer. 2019, pp. 355–364.

[46] Uri Goldsztejn and Arye Nehorai. “Predicting preterm births from electrohysterogram recordings via deep learning”. In: *Plos one* 18.5 (2023), e0285219.

Appendix A

Supplementary Performance Metrics

In this section, the results for Precision, Recall, and F1 score are presented.

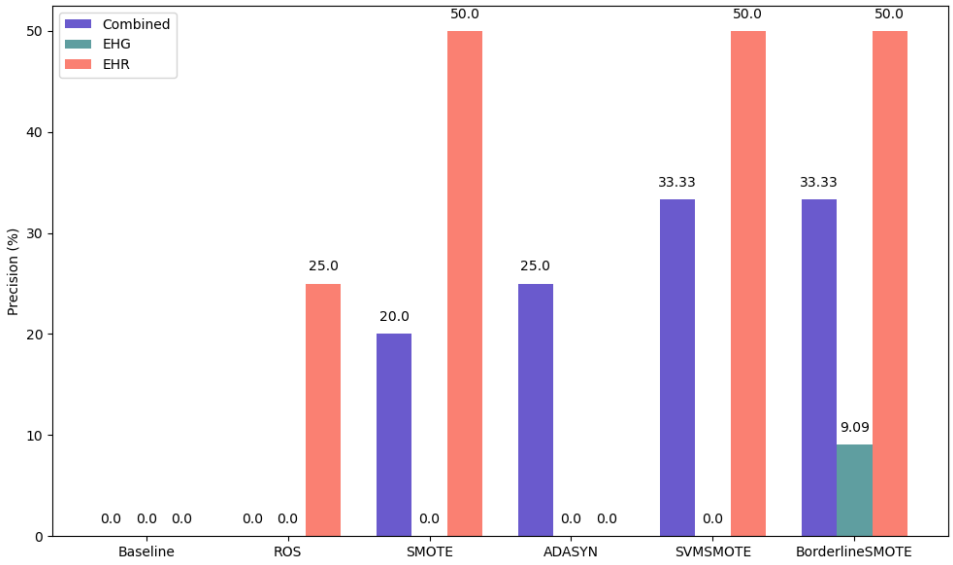

Figure A.1: Precision comparison of different oversampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

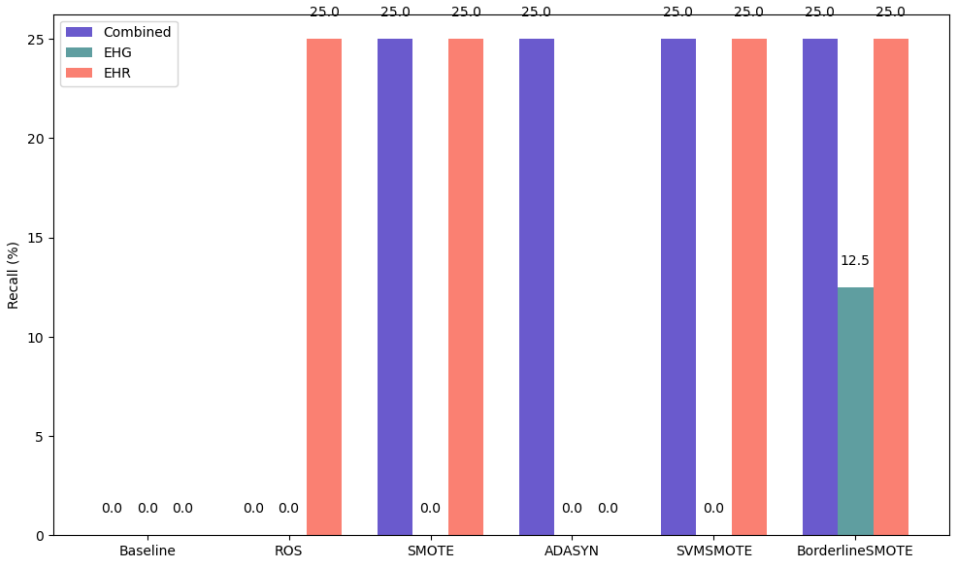

Figure A.2: Recall comparison of different oversampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

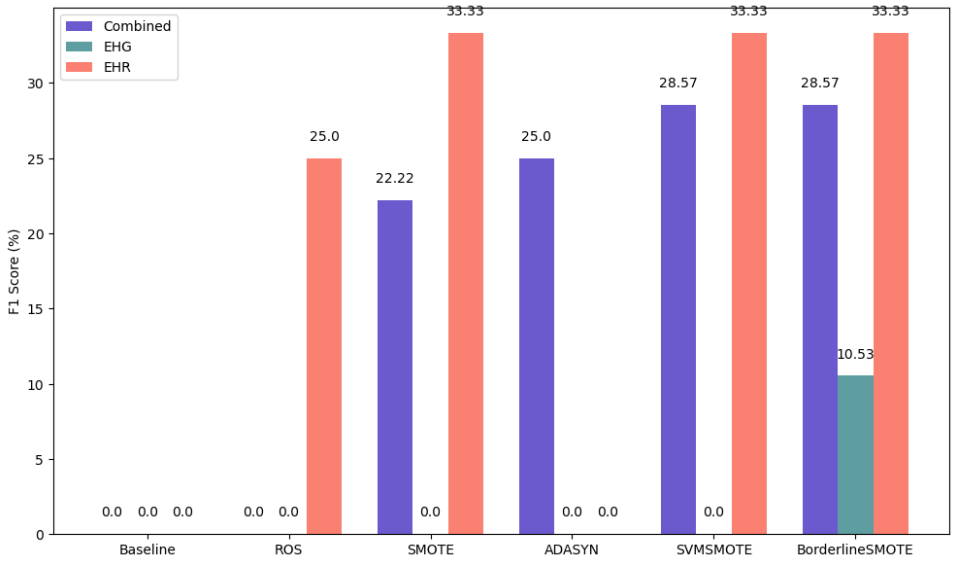

Figure A.3: F1 Score comparison of different oversampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

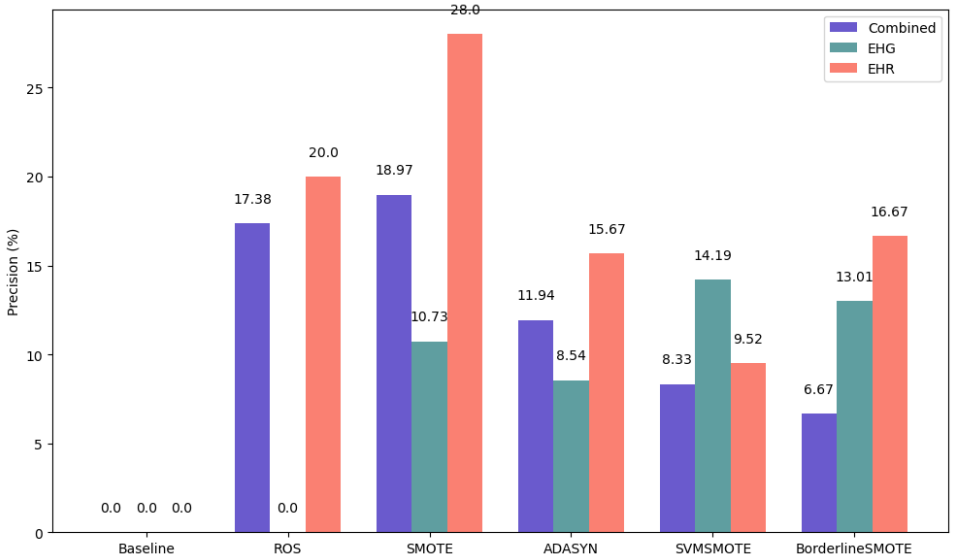

Figure A.4: Precision comparison of different oversampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with CV.

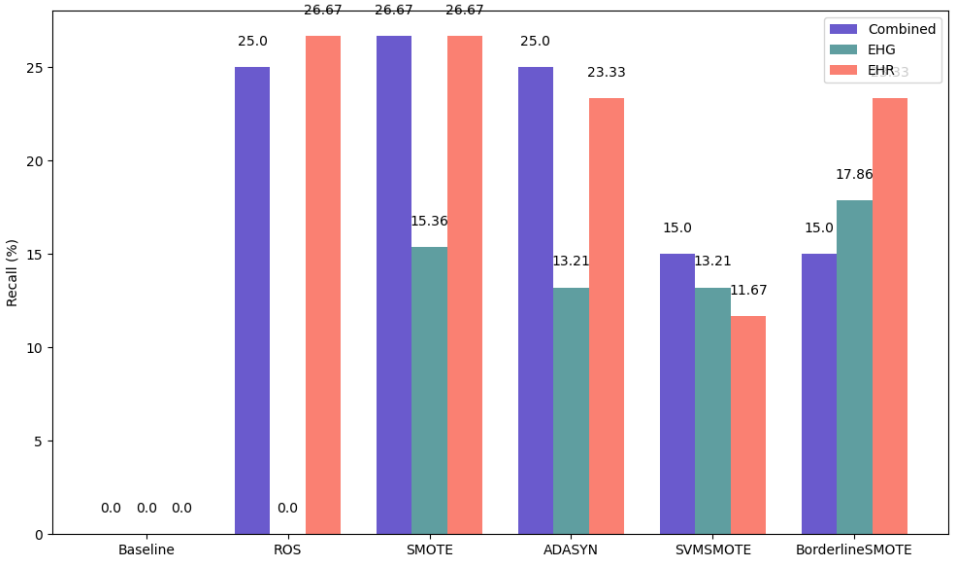

Figure A.5: Recall comparison of different oversampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with CV.

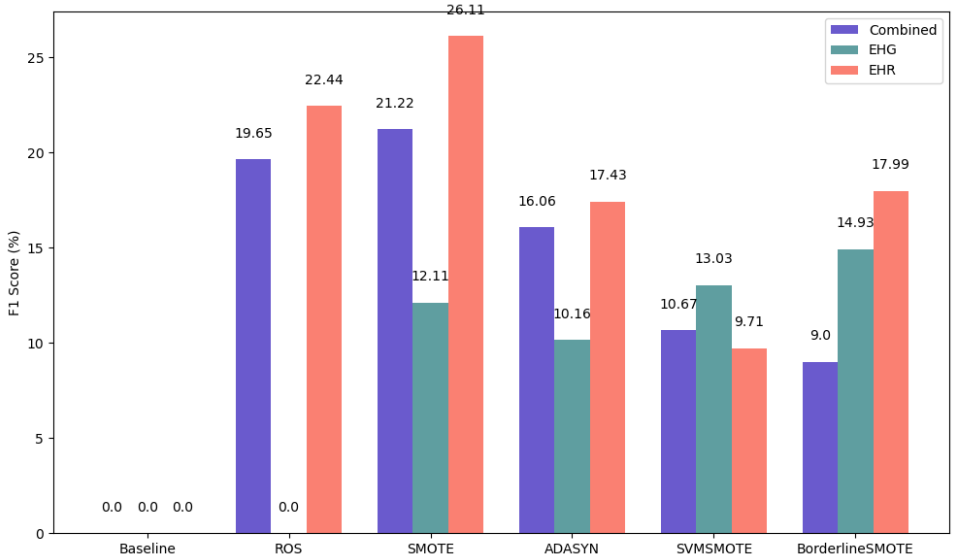

Figure A.6: F1 Score comparison of different oversampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with CV.

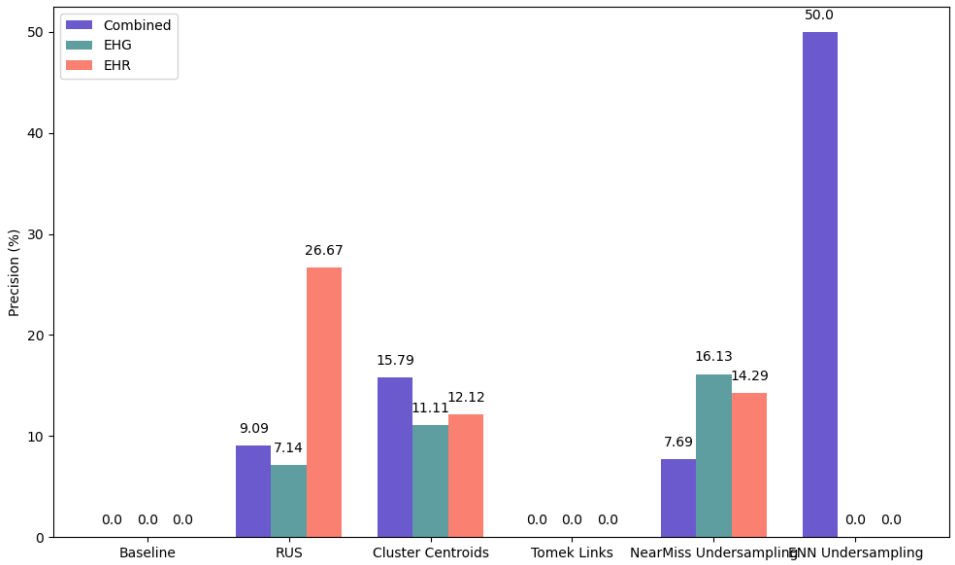

Figure A.7: Precision comparison of different undersampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

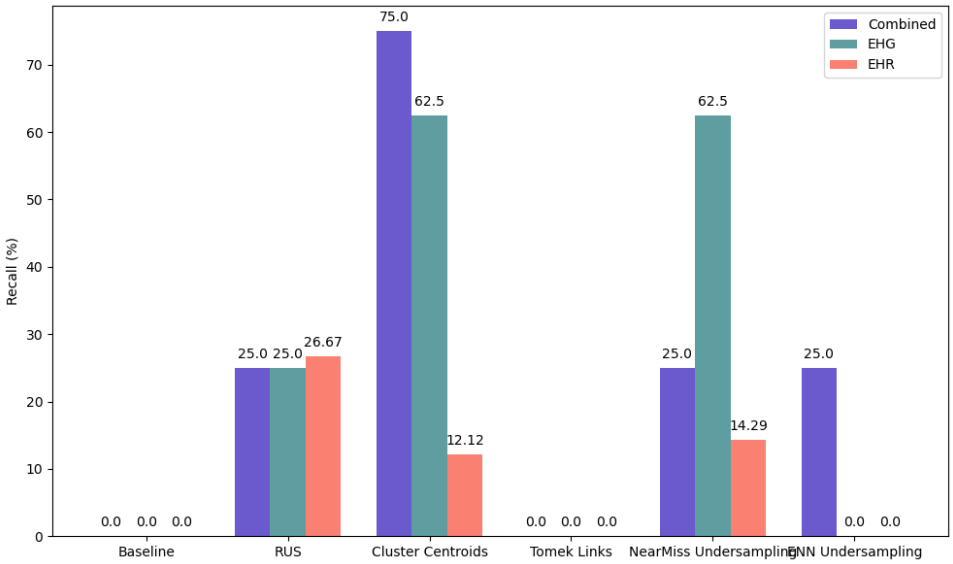

Figure A.8: Recall comparison of different undersampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

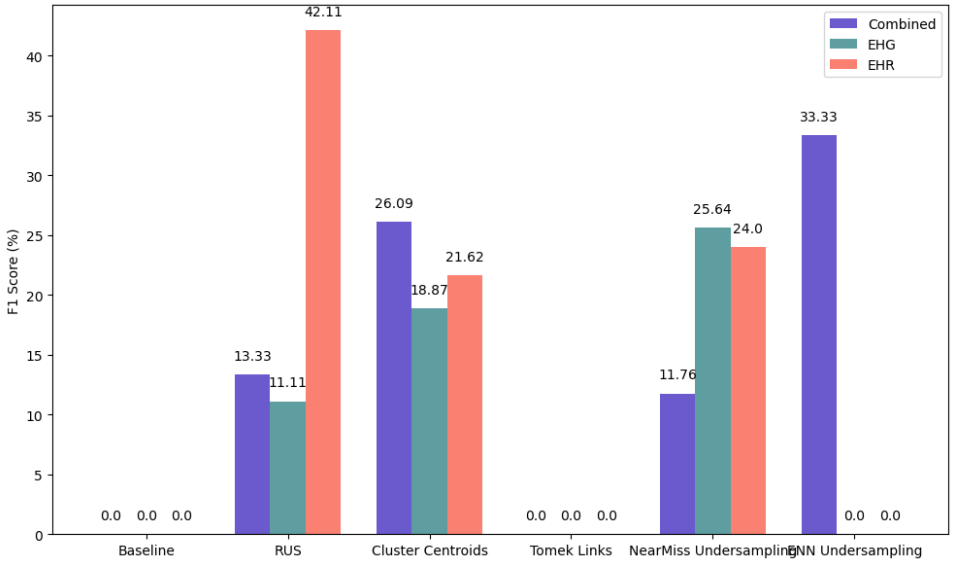

Figure A.9: F1 Score comparison of different undersampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

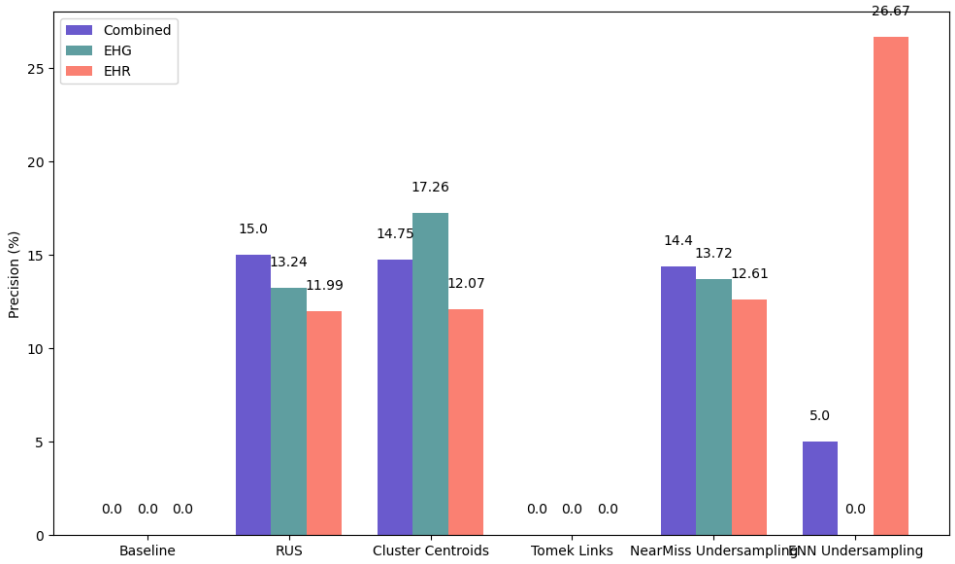

Figure A.10: Precision comparison of different undersampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with CV.

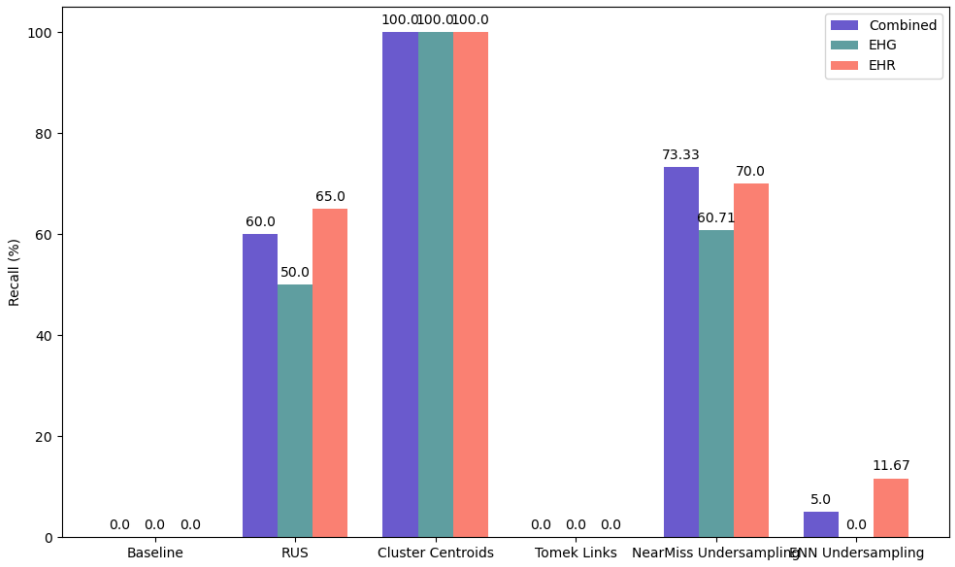

Figure A.11: Recall comparison of different undersampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with CV.

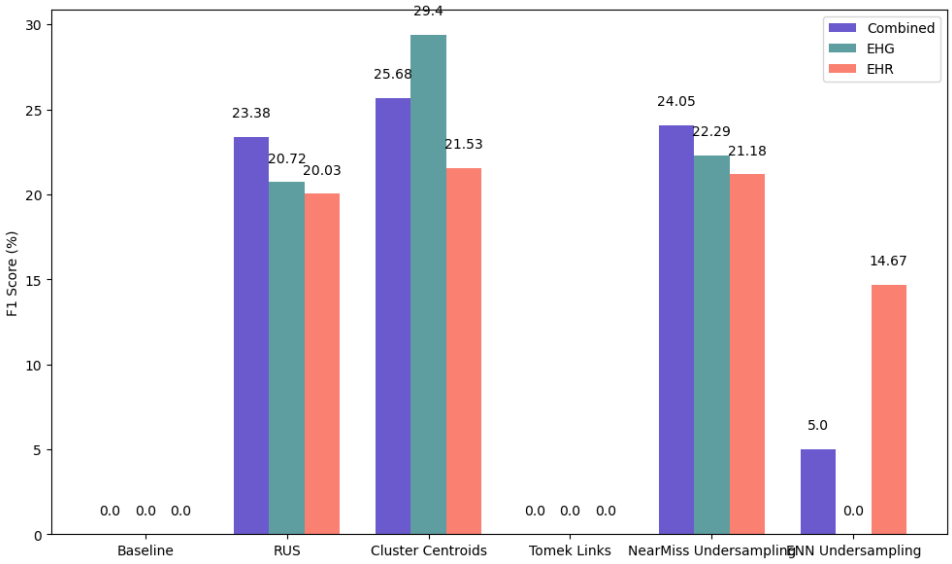

Figure A.12: F1 Score comparison of different undersampling techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with CV.

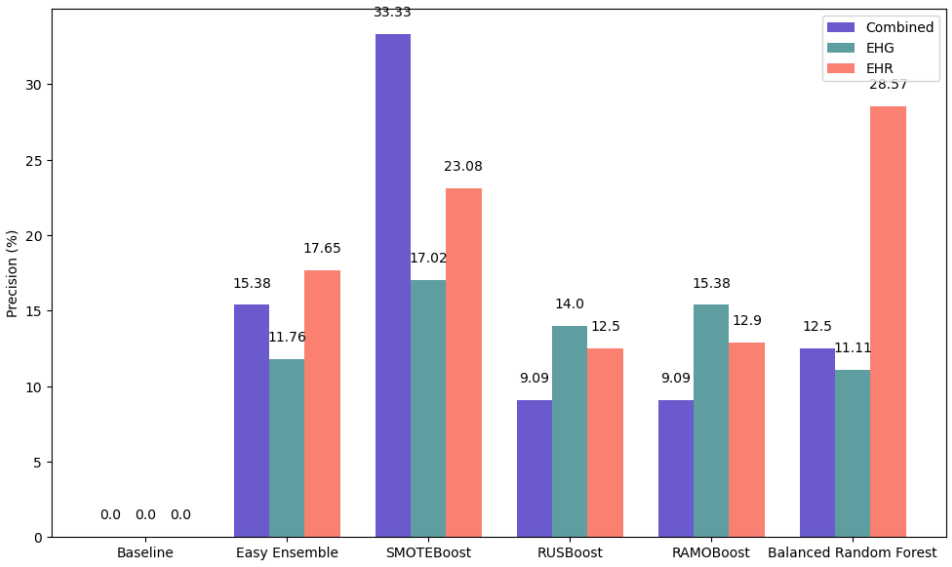

Figure A.13: Precision comparison of different ensemble techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

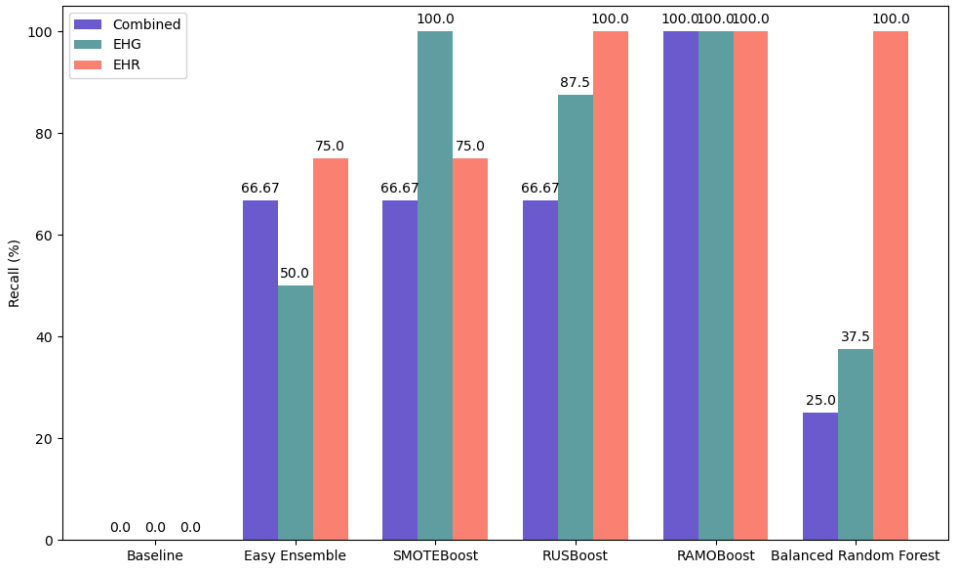

Figure A.14: Recall comparison of different ensemble techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

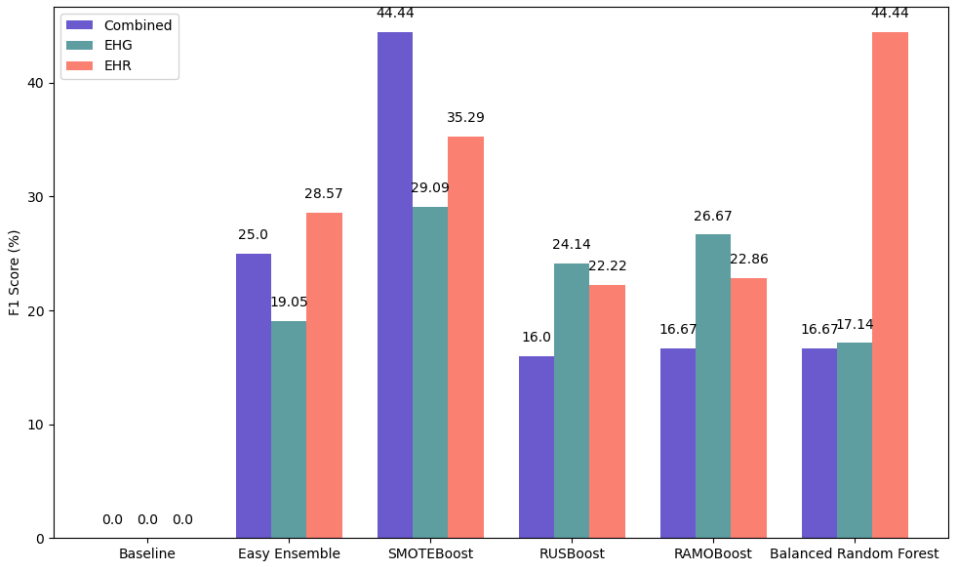

Figure A.15: F1 Score comparison of different ensemble techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - without using CV.

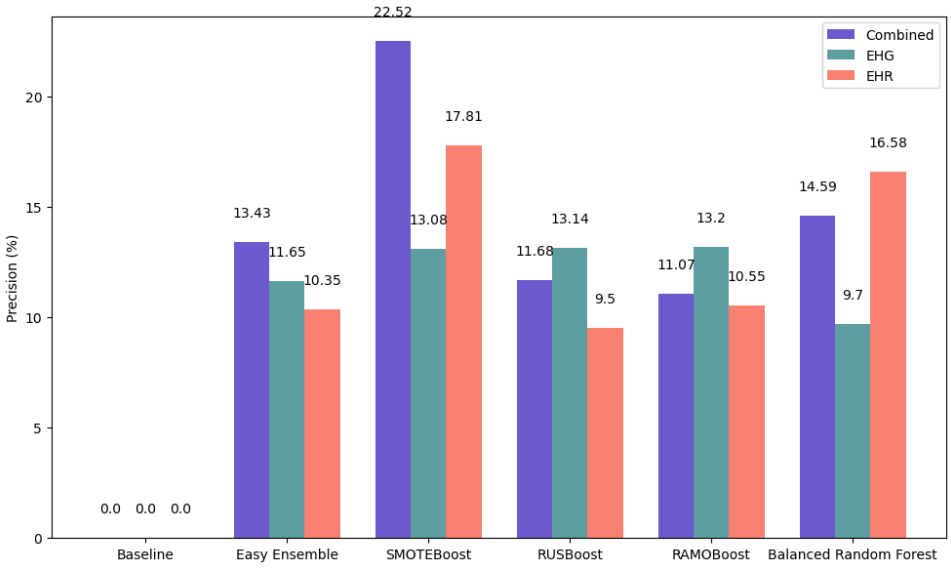

Figure A.16: Precision comparison of different ensemble techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with CV.

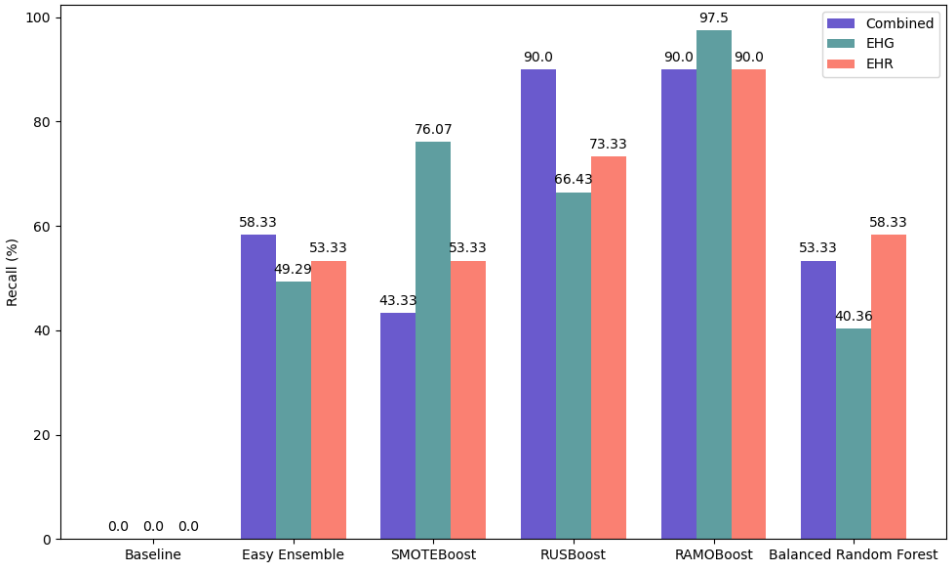

Figure A.17: Recall comparison of different ensemble techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with CV.

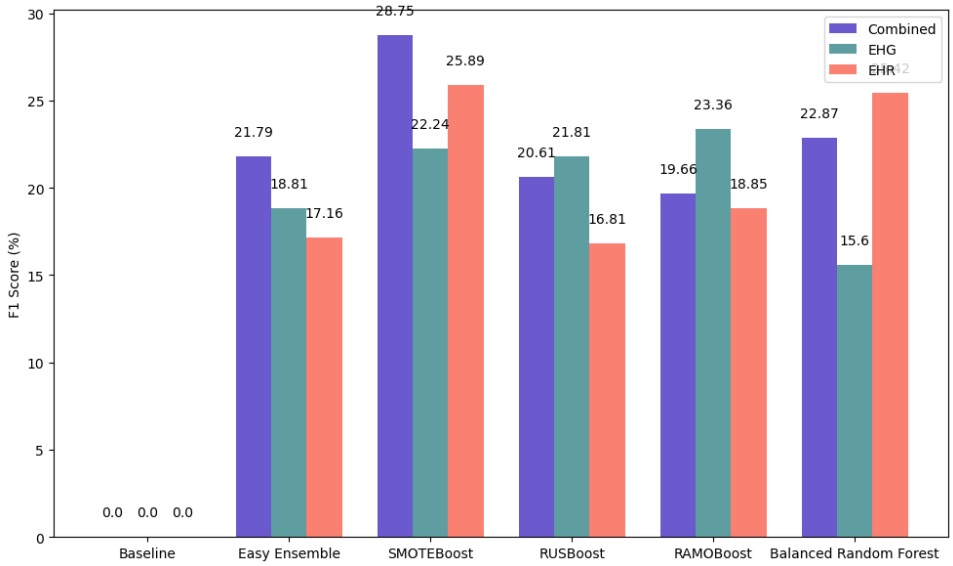

Figure A.18: F1 Score comparison of different ensemble techniques using 3 different feature sets - ”EHR" (contains categorical features), ”EHG" (contains features extracted from the EHG signal), ”combined" (combination of both feature sets) - with CV.