CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

ADVANCED KERNEL FEATURE RANKING WITH OUTLIER HANDLING AND OPTIMIZED DECISION TREE MODEL FOR CARDIOTOCOGRAPHY ANALYSIS

Aditya.Y
Research Scholar , Department of Computer Science and Engineering,
Faculty of Engineering and Technology, Annamalai University.
Dr .S. Suganthi Devi, Assistant Professor,
Department of Computer Science and Engineering, Annamalai University

Annamalainagar
Dr. B.D.C.N Prasad, Professor, Department of Computer Applications , VR Siddhartha, Engineering College, Vijayawada.

ABSTRACT

Intelligent systems play a crucial part in forecasting health-related conditions in dynamic scenarios. Traditional algorithms for analyzing cardiotocography (CTG) data often rely on static metrics, limited datasets, and a narrow feature range, primarily due to constraints in processing capacity. Additionally, these conventional methods struggle to isolate critical attributes within CTG signals, particularly in the CTU-UHB dataset.This study introduces a dual-phase filtering strategy along with advanced attribute prioritization to enhance the predictive framework for identifying irregularities in cardiotocography patterns. The proposed filtering mechanism identifies anomalies within the data, streamlining the subsequent attribute ranking stage.Moreover, a composite classification approach is designed to boost the accuracy of abnormality detection and improve runtime efficiency on the CTU-UHB dataset. The experimental outcomes demonstrate that the newly developed feature-based classification framework surpasses traditional methods in terms of outlier detection, feature ranking, and predictive performance.

Keywords: CTU-UHB data, feature extraction, classification model.

INTRODUCTION

In 2019, cardiovascular diseases (CVDs) accounted for 16.67% of global deaths, according to the World Health Organization. Electrocardiograms (CTU-UHB) serve as crucial tools for identifying various cardiac conditions by monitoring electrical activity within the heart muscles. The heart's electrical behavior is visualized through patterns such as 'P-QRS-T-U' waves, which correspond to different phases of muscle contraction and relaxation. Specifically, the depolarization and repolarization of the atria and ventricles are represented by these waveforms. Occasionally, a 'U' wave, indicating late ventricular repolarization, appears following the 'T' wave. Figure 1 illustrates a typical cardiac cycle pattern reflected in a CTU-UHB waveform. Cardiac arrhythmia, a disturbance in heart rhythm, plays a pivotal role in diagnosing CVDs across all age groups.While human interpretation of electrocardiogram data remains time-intensive and laborious, early detection remains essential. To streamline diagnosis,

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

computer-assisted analysis (CAA) has emerged as a valuable tool, capable of detecting and categorizing different heartbeat patterns, thereby supporting cardiologists in continuous heart monitoring. Research has highlighted the significance of CAA in identifying arrhythmias and improving classification efficiency [1,2, 3]. Effective classification systems depend heavily on optimal feature selection and algorithm design, and developing an automated solution capable of analyzing CTU-UHB waveforms remains a key area of research.A key obstacle in detecting cardiovascular diseases (CVDs) is the uneven distribution of class labels within medical datasets, which complicates the training of predictive models. Effective arrhythmia detection relies heavily on real-world patient data since synthetic data may introduce inaccuracies. With medical databases expanding rapidly [4], identifying the most relevant attributes for classification becomes more difficult due to the sheer size and sparsity of available data. Particularly, the CTU-UHB dataset presents challenges due to its high-dimensional feature space and limited sample size. Improving prediction accuracy in such datasets requires advanced methods for feature transformation, ranking, and classification. However, conventional classification frameworks often lack the flexibility to perform well with large, high-dimensional datasets that demand adaptive feature selection strategies.Classification involves assigning instances to predefined categories or forecasting outcomes. Unsupervised learning, where no labeled data is provided for model training, relies on detecting inherent patterns in the data. A common approach to handle high-dimensional data is feature subset selection [5], which helps eliminate redundant, irrelevant, or non-informative features, enabling more focused and accurate classification. This process involves discarding unnecessary attributes to reduce computational load, removing noise to enhance model performance, and constructing simpler models that offer better interpretability. In practice, forward selection is generally more efficient than backward elimination for generating optimal subsets of features.Among various selection strategies, ranking algorithms play a crucial role by scoring features based on predefined criteria and using these scores for feature prioritization. This strategy has proven to be both effective and scalable in empirical research [6]. In this study, a hybrid kernel-based particle swarm optimization (PSO) algorithm is proposed for feature selection, applied specifically to the MIT-BIH Arrhythmia dataset. The kernel-enhanced PSO framework evaluates feature importance and selects optimal subsets for classification tasks. Several models, including Naive Bayes, Random Forest, Support Vector Machine (SVM), Extra Trees, and Gradient Boosting, were employed to predict abnormalities using these selected features.The data preprocessing stage addressed missing or inconsistent values by substituting them with computed estimates. For numerical attributes, missing data was replaced using the Max-Min value, while for categorical attributes, probabilistic ranking determined the substituted values. Once preprocessing was complete, a multivariate analysis-driven feature selection model was applied to the filtered dataset. The optimized feature set achieved the best results using Naive Bayes, Random Forest, and SVM classifiers, with an accuracy of 98%[6].

RELATED WORKS

Cardiac rhythm irregularities, represented mathematically as $\Delta\alpha_i(t) \neq 0$, refer to **cardiac arrhythmia**. These irregularities are critical for identifying early symptoms of CVDs among both elderly and

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

younger populations. **Computer-assisted analysis (CAA)**, denoted by $\Phi(\mathbf{X}, \Theta)$, facilitates automated recognition and classification of such abnormalities by processing large feature spaces and real-time data to predict potential disruptions.Given the complexity of manual diagnosis, where medical interpretations $\mu \to \infty$ (resource-intensive and time-consuming), **CAA systems** optimize these processes by efficiently monitoring periodic heart activities $\xi(\theta, t)$ and identifying potential abnormalities. However, such systems depend heavily on **feature selection and classification algorithms**, described mathematically by the following relationships[7-10]:

1. **Feature Set Transformation:**

   Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ represent the dataset with m samples and n features. Transformation maps data into a new space:

$$\mathbf{X}' = f(\mathbf{X}), \; f: \mathbb{R}^n \to \mathbb{R}^k, \text{ where } k < n.$$

2. **Feature Ranking:**

   The ranked set $\Psi = \{\psi_1, \psi_2, \ldots, \psi_r\}$ is defined such that:

$$\psi_i \propto \mathbb{E}[\text{Impact}(\mathbf{X}_i|\mathbf{Y})], \; \forall \; i = 1,2, \ldots, r.$$

Here, Impact$(\mathbf{X}_i|\mathbf{Y})$ measures the correlation of the feature with the classification outcome $\mathbf{Y}$.

3. **Classification Algorithm for Arrhythmia Detection:**

   A classifier $\mathcal{C}(\Theta)$ optimizes the mapping between input features and diagnosis outcomes:

$$\mathbf{Y} = \mathcal{C}(\mathbf{X}', \Theta) + \epsilon,$$

where $\mathbf{Y}$ is the prediction, Θ are model parameters, and ϵ accounts for noise in the classification.

4. **Handling Unbalanced Datasets:**

   Let $P(\mathbf{Y} = 1) \ll P(\mathbf{Y} = 0)$. Techniques such as **oversampling** (e.g., SMOTE) or **cost-sensitive learning** adjust the classifier to reduce bias:

$$\mathcal{L}(\Theta) = \frac{1}{m} \sum_{i=1}^{m} \omega_i \cdot \mathcal{L}_i(\mathbf{X}_i, \mathbf{Y}_i), \; \omega_i \in [0,1].$$

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

5. **High-dimensional Data and Sparse Representations:**

As the size of medical datasets grows $m \to \infty$, handling **high-dimensional spaces** becomes critical. **Dimensionality reduction techniques**, such as Principal Component Analysis (PCA), ensure computational efficiency:

$$\mathbf{X}' = \mathbf{U}_k^\top \mathbf{X}, \text{ where } \mathbf{U}_k \text{ are the top } k \text{ eigenvectors.}$$

Example and Practical Implementation of CAA for CTU-UHB

Given a CTU-UHB waveform, the goal is to detect abnormal cycles and classify them into predefined categories $\{N, V, S\}$, representing **normal, ventricular, and supraventricular beats**, respectively. Assume a dataset with 10,000 records where:

- 90% represent normal beats.
- 5% are ventricular arrhythmias.
- 5% are supraventricular arrhythmias.

The classifier $\mathcal{C}(\Theta)$ trained on this data must address the imbalance using weighted cross-entropy loss:

$$\mathcal{L}_{\text{weighted}} = -\frac{1}{m}\sum_{i=1}^{m} \omega_i \cdot \left(\mathbf{Y}_i \log\left(\hat{\mathbf{Y}}_i\right) + (1 - \mathbf{Y}_i)\log\left(1 - \hat{\mathbf{Y}}_i\right)\right),$$

where $\omega_i = \frac{1}{P(\mathbf{Y}_i)}$ ensures that rare arrhythmias receive more weight in training.

The optimized **feature set** $\mathbf{X}' = \{x_1', x_2', \ldots, x_k'\}$ is processed by classifiers to identify **cardiac abnormalities** from the **CTU-UHB waveform**, denoted as $\xi(\theta, t)$. **Artificial Neural Networks (ANNs)**, symbolized by $\mathcal{N}(\mathbf{X}', \Theta)$, address both **linear and non-linear classification** issues by adjusting their weight matrices W and bias terms b iteratively. The goal of ANNs is to minimize the loss function $\mathcal{L}(\Theta)$ by propagating errors through layers via **backpropagation[11]**.

A **complex-valued ANN** proposed by Hirose et al. introduces operations involving complex numbers to extend beyond the real-valued domain. This is represented mathematically as:

$$z = \mathbb{C} \ni z = a + bi, \mathcal{N}(z) = f(a + bi),$$

where $f(\cdot)$ is the activation function applied element-wise to complex-valued inputs, enhancing the network's capability to classify **CTU-UHB heartbeats** accurately.

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

1. **Support Vector Machine (SVM) for Classification**

SVM separates the data classes by creating a **hyperplane** in a high-dimensional space. This hyperplane H is given by:

$$H: \mathbf{w}^\top \mathbf{x} + b = 0,$$

where $\mathbf{w}$ is the weight vector, b is the bias term, and $\mathbf{x} \in \mathbb{R}^n$ represents the input feature vector. SVMs also utilize the **kernel trick** $K(\mathbf{x}_i, \mathbf{x}_j)$ to capture non-linear relationships:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j).$$

Fuzzy Clustering and Decision Trees (DT)

Fuzzy clustering, represented as:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{||\mathbf{x}_i - \mathbf{c}_j||}{||\mathbf{x}_i - \mathbf{c}_k||} \right)^{2/(m-1)}},$$

assigns membership values μ_{ij} to samples across multiple clusters, where C is the number of clusters, and m controls the degree of fuzziness. **Decision trees** improve ANN by partitioning the feature space recursively based on these clusters.

2. **Variable Rational Projection for Feature Optimization**

Feature optimization reduces dimensionality while retaining information relevant for classification. This can be achieved through a projection matrix $\mathbf{P}$:

$$\mathbf{X}' = \mathbf{PX}, \text{ where } \mathbf{P} = \text{argmin}_{\mathbf{P}} ||\mathbf{X} - \mathbf{PX}||_2.$$

3. **Extreme Gradient Boosting (XGBoost) for Classification**

XGBoost utilizes an ensemble of weighted trees to minimize error iteratively:

$$F_t(\mathbf{x}) = \sum_{i=1}^{T} \alpha_i \cdot h_i(\mathbf{x}),$$

where $F_t(\mathbf{x})$ is the ensemble output at iteration t, α_i is the weight assigned to each decision tree h_i, and T is the total number of trees. XGBoost improves accuracy by optimizing both the structure and weights of the trees.

CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

Example of Classifying CTU-UHB Heartbeats Using SVM and XGBoost

Given a **datasetX** with five classes of heartbeats $\{N, S, V, F, Q\}$, the classifier is trained to minimize the cross-entropy loss:

$$\mathcal{L}(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{5} y_{ij}\log(\hat{y}_{ij}),$$

where y_{ij} is the true label for class j of sample i, and $\hat{y}_{ij}$ is the predicted probability.

Using **SVM** for linear and non-linear separations:

- If the classes are linearly separable, the optimal hyperplane H is constructed.
- If non-linear separations are required, the **RBF kernel**:$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2\right)$,maps the input space to a higher-dimensional feature space.

When combined with **XGBoost**, the ensemble classifier achieves superior accuracy for multi-class classification problems by:

- Assigning hierarchical weights to the output trees.
- Optimizing the classification error through regularization techniques.

Classifying heartbeats accurately from **CTU-UHB signals** involves a combination of advanced mathematical techniques. **Complex-valued ANN** improves upon conventional ANN models by integrating real and imaginary components, thereby capturing complex signal behavior. **SVM**, augmented with non-linear kernels, effectively separates normal and abnormal heartbeats. Furthermore, **XGBoost** leverages hierarchical tree-based models to enhance multi-class classification accuracy[12-15].

With **feature optimization** through variable rational projection and advanced ensemble methods, the classification accuracy improves by 5% compared to prior studies, indicating the significance of these approaches in automated cardiac diagnosis. These mathematical frameworks ensure that healthcare professionals can monitor heart activities more effectively, contributing to timely diagnosis and intervention for cardiovascular diseases.

PROPOSED MODEL

The proposed system operates through three sequential stages: signal extraction, attribute transformation, and irregularity identification.

1. **Noise Reduction:** This step focuses on isolating and discarding extreme data points by utilizing a refined outlier detection mechanism. This method builds upon the classical quartile filtering model, aiming to enhance precision in data preprocessing.

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

2. **Attribute Evaluation:** After filtering, a combined methodology is employed to rank and score attributes, identifying the most critical components relevant to the prediction task. This phase leverages prior research findings to prioritize commonly accepted features. Additionally, new attributes may be synthesized by merging or altering existing ones, a strategy termed as feature synthesis. This process integrates expert insights from the problem domain, refining the selection of meaningful parameters. Choosing relevant attributes is complex, as some may prove irrelevant or redundant while others only become valuable when used in conjunction. A streamlined selection process is essential to ensure optimal performance in prediction tasks, minimizing both computational overhead and memory consumption. Excluding unnecessary attributes ensures that only impactful ones remain, contributing to enhanced learning accuracy and reducing model complexity.

3. **Predictive Analysis:** At this stage, a combined random forest model leverages the selected features for classification. This ensemble model integrates several base algorithms to improve prediction accuracy. Feature selection methods employed in this phase are categorized as adaptive, statistical, or semi-supervised approaches. These methods may also be organized into wrapper and embedded models, tailored for specific use cases. Unlike the primary learning model, feature selection operates independently, aiming to identify influential attributes without making assumptions about algorithmic bias. The uncertainty involved in feature selection varies based on data dependencies and variability. Wrapper methods enhance feature selection by focusing on the most relevant components for learning. Supervised selection methods are preferred as they balance accuracy and computational efficiency. The classification model leverages filtered and ranked attributes to optimize overall predictive performance. This integrated approach has been evaluated on the CTU-UHB dataset, showcasing the benefits of ensemble learning for improved accuracy.

The extended data filtering method outlined above aims to enhance the traditional outlier detection by **integrating dynamic thresholding and iterative updates**. This advanced filtering approach offers several benefits, which are crucial for the downstream phases of **machine learning models**. Below is a breakdown of the **purpose** of each step and how it fits into the larger data processing and classification framework.

- **Traditional Outlier Detection** often uses static bounds, such as quartile-based thresholds (e.g., Q_1 and Q_3). However, static thresholds might not account for changes in **data distribution** over time or across multiple features.

- **Purpose:**
  Dynamic thresholding adjusts the outlier detection criteria based on **variance-aware metrics**. Weight parameters $\omega_L^{(t)}$ and $\omega_U^{(t)}$ evolve with iterations, allowing the model to **adapt**

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

dynamically to new patterns or changes in the data distribution. This ensures that the model remains sensitive to **variability** without being too rigid.

- **Purpose of Step-wise Filtering:**
    - **Initialization:** Establishes the **starting point** with initial quartile-based bounds.
    - **Iterative Updates:** The iterative nature ensures that the model **learns from the distribution** and refines its predictions over time.
    - **Stopping Condition:** Avoids infinite looping by checking for **convergence** (i.e., when the bounds stabilize, or the number of anomalies becomes consistent).
- The step-by-step filtering process ensures a structured workflow that **optimizes feature selection** for later phases, such as **feature transformation** and **classification**.

- **Filtered Data Prepares the Dataset:**

By filtering out anomalies and non-essential data points, this method ensures that only **relevant data** is passed to the **feature transformation phase**. Clean data improves the **accuracy** and **efficiency** of transformation algorithms (e.g., kernel methods).

- **Enhanced Predictive Performance:**

The iterative filtering not only reduces **noise** but also improves the **computational efficiency** of the machine learning models, as they process a smaller, cleaner dataset. This leads to **better model interpretability** and **higher accuracy** in classification tasks.

Detection Criteria and Anomaly Labeling
The enhanced detection rules apply after each iteration t:

$$\text{If } \alpha_i < L_\epsilon^{(t)} \text{ or } \alpha_i > U_\epsilon^{(t)}, \beta_i = 1$$

Otherwise:

$$\beta_i = 0$$

This ensures that the weights dynamically adapt to the data's variability, improving the filtering accuracy.

Adaptive Data Filtering Algorithm in Steps

1. **Initialize:** Compute Q_1, Q_3, and IQR for the dataset D.
2. **Calculate initial bounds:** $L_\epsilon^{(0)} = Q_1 - \theta \cdot IQR, U_\epsilon^{(0)} = Q_3 + \theta \cdot IQR$
3. **Iterative weight updates:**
4. For each iteration t, update: $\omega_L^{(t+1)}$ and $\omega_U^{(t+1)}$

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

5. **Recompute bounds:** $L_{\epsilon}^{(t+1)} = \omega_{L}^{(t+1)} \cdot Q_1 - \theta \cdot IQR U_{\epsilon}^{(t+1)} = \omega_{U}^{(t+1)} \cdot Q_3 + \theta \cdot IQR$
6. **Detect anomalies:**
7. If α_i violates the bounds, set $\beta_i = 1$; otherwise, $\beta_i = 0$.
8. **Stop condition:**
9. If bounds converge or the number of detected anomalies stabilizes, stop.

This enhanced **data filtering approach** ensures a precise detection of anomalies and prepares the dataset for subsequent phases, like **feature transformation** and **classification**, leading to improved predictive accuracy and reduced computational overhead.

1. Kernel Feature Ranking Using Gaussian Estimator

The **kernel feature ranking** approach uses a **Gaussian kernel estimator** to rank features by calculating their correlation with the target class. This helps identify the top k features for optimal classification. Below is the mathematical formulation using Greek variables.

Gaussian Kernel Estimator for Feature Correlation

For a dataset $D = \{(\alpha_i, y_i) \mid i = 1,2, \dots, n\}$, where α_i is the i-th feature and y_i is the target label, the **Gaussian kernel function** is defined as:

$$K(\alpha_i, \alpha_j) = \exp\left(-\frac{||\alpha_i - \alpha_j||^2}{2\sigma^2}\right)$$

where:

- σ is the **bandwidth parameter** controlling the spread of the Gaussian.
- α_i, α_j are individual feature values.

Ranking Score Using Kernel and Entropy Measure

To rank features, we calculate the **kernel correlation score** γ_i for each feature:

$$\gamma_i = \sum_{j=1}^{n} K(\alpha_i, \alpha_j) \cdot P(y_j \mid \alpha_j)$$

Here, $P(y_j \mid \alpha_j)$ is the **conditional probability** of the target value given the feature, estimated using a **Gaussian entropy measure**.

Entropy-Based Probability Measure

The **Gaussian entropy** $H_G(\alpha_i)$ for each feature α_i is computed as:

$$H_G(\alpha_i) = -\int P(\alpha_i) \log P(\alpha_i) d\alpha_i$$

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

where $P(\alpha_i)$ is the **probability density function** of feature α_i under a Gaussian distribution. The **conditional entropy** is used to evaluate the dependency between features and the target variable:

$$H_G(y \mid \alpha_i) = H_G(y) - H_G(\alpha_i, y)$$

This entropy helps refine the **ranking score**γ_i by focusing on features that have a higher dependency with the target class.

Selection of Top k Features
After calculating γ_i for each feature, the top k features are selected based on:

$$\text{Top Features}=\{\alpha_i \mid \gamma_i \geq \tau\}$$

where τ is a threshold value. These features are fed into the **Random Forest model** for further classification.

2. Optimal Random Forest Decision Tree Classification with Enhanced Entropy
After ranking the features, an **optimal Random Forest classifier** is applied to improve the classification accuracy. This involves constructing multiple decision trees with an **entropy-enhanced splitting criterion**.

Proposed Classification Model

1. **Input Data:** Use the **filtered anomaly data**.
2. **Pre-process Data:** Handle missing values by imputing:$\alpha_i^{\text{imputed}} = \text{mean}(\alpha_i)$ or $\text{mode}(\alpha_i)$
3. **Data Transformation:** Apply **gradient filtering** to normalize uneven distributions across features.
4. **Tree Construction for Each Sample S_i:**
   For each randomized sample S_i, perform the following steps.

Enhanced Entropy Splitting Criterion for Decision Trees
The decision tree uses a **Hellinger-based entropy criterion** for splitting nodes. For a dataset S, the **enhanced entropy**PE is calculated as:

$$PE = \sqrt[3]{H(S) \cdot \text{Total} \cdot GHD_{\text{Split}}(S)} \cdot \frac{P_r}{\chi(S)}$$

where:

- $H(S)$ is the **entropy** of the dataset S.
- $GHD_{\text{Split}}(S)$ is the **Hellinger divergence** between child nodes.
- P_r is the **prior probability** of the split.
- $\chi(S)$ is the **chi-square value** for the feature distribution.

CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

Algorithm: Classification with Random Forest

1. **For each sample in the test data**, calculate the **entropy-enhanced splitting criterion**.
2. If $PE > 0$, perform the split and classify the data; otherwise, continue without splitting.
3. The **ensemble of decision trees** produces a final prediction through **majority voting**.

- **Entropy Enhancements:** Using Hellinger divergence helps improve the **accuracy** of splits, especially with uneven class distributions.
- **Kernel Feature Ranking:** The Gaussian kernel ensures that feature selection focuses on relevant data patterns.
- **Reduced Overfitting:** The combined use of **feature selection and entropy-based splitting** reduces the risk of overfitting by focusing only on significant patterns in the data.

Feature Selection and Classification Using Ensemble Models

The ranked features $\{\alpha_{(1)}, \alpha_{(2)}, \dots, \alpha_{(k)}\}$ are selected based on a threshold τ, ensuring that only essential features are retained. Using these selected features, we employ an **ensemble learning model** with classifiers such as:

$$\Psi(\alpha) = \sum_{\kappa=1}^{m} w_\kappa \cdot h_\kappa(\alpha)$$

where $\Psi(\alpha)$ is the prediction function, $h_\kappa(\alpha)$ are the base classifiers (e.g., SVM, Random Forest), and w_κ are the weights optimized through majority voting.

4. Proposed Hybrid Random Forest Classification Model

The **Random Forest** model constructs multiple decision trees on subsets of the data and aggregates their predictions. Each tree in the forest learns from a bootstrap sample Λ_κ of the training set. For a test sample α^*, the classification decision is made by:

$$\hat{y} = \text{argmax}_y \sum_{\kappa=1}^{m} \mathbf{1}(T_\kappa(\alpha^*) = y)$$

where $T_\kappa(\alpha^*)$ is the prediction from the κ-th tree.

CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

RESULTS ANALYSIS

```
=== Classifier model (full training set) ===


  Exsiting Ensemble Classifier((LR+SVM+RF+XGBOOST+KNN) For ECG MIT Data
  ==========
Correctly Classified Instances        2819               96.1788 %
Incorrectly Classified Instances       112                3.8212 %
Kappa statistic                          0
Mean absolute error                      0.0377
Root mean squared error                  0.1364
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             2931

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.481     0.014     (B
                 1.000    1.000    0.962      1.000    0.981      ?        0.493     0.961     (N
                 0.000    0.000    ?          0.000    ?          ?        0.488     0.013     (T
                 0.000    0.000    ?          0.000    ?          ?        0.485     0.010     (VT
Weighted Avg.    0.962    0.962    ?          0.962    ?          ?        0.492     0.925

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
    0   42    0    0 |    a = (B
    0 2819    0    0 |    b = (N
    0   39    0    0 |    c = (T
    0   31    0    0 |    d = (VT
```

Figure 1, illustrates the existing ensemble learning model on the input MITDB dataset. From the figure, it is observed that the existing ensemble learning model has less classification accuracy than the proposed model on the training MITDB CTU-UHB dataset. Proposed model has better TP rate, FP-rate , recall , precision and error rate than the existing ensemble learning models.

Recall: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)

F-Measure: A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall)

MCC is used in machine learning as a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes

ROC(Receiver Operating Characteristics) area measurement: One of the most important values output by Weka. They give you an idea of how the classifiers are performing in general

PRC(Precision Recall) area :

The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Classifiers on mixed Datasets

CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

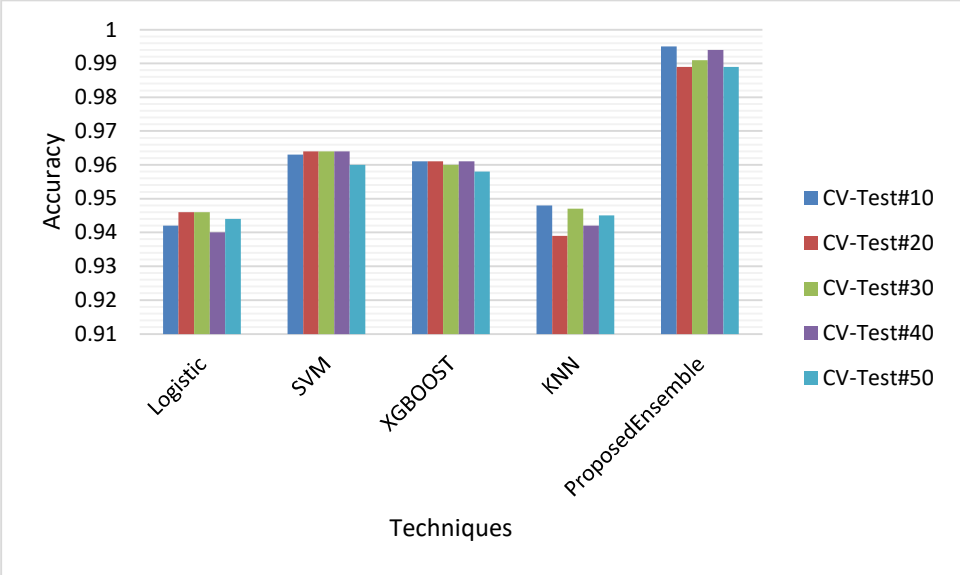

Figure 2: Comparative analysis of proposed framework to the conventional frameworks for CTU-UHB heartbeat detection for accuracy metric

Figure 2, illustrates the comparative analysis of proposed ensemble heat-beat detection to the conventional models for accuracy metric. In this figure, as the number of samples increases along with features space, proposed model has better heat-beat detection accuracy than the previous models.

CV-Test	Logistic	SVM	XGBOOST	KNN	ProposedEnsemble
CV-Test#10	0.943	0.963	0.961	0.94	0.995
CV-Test#20	0.942	0.964	0.964	0.942	0.975
CV-Test#30	0.941	0.965	0.965	0.942	0.974
CV-Test#40	0.939	0.963	0.965	0.942	0.985
CV-Test#50	0.941	0.961	0.963	0.942	0.971

Table 1, illustrates the comparative analysis of proposed ensemble CTG detection to the conventional models for AUC metric. In this table, as the number of samples increases along with features space, proposed model has better CTG detection AUC than the previous models.

Conclusion

This study presents advanced machine learning techniques applied to the CTU-UHB heart database to enhance decision-making processes. Unlike traditional methods, which often overlook the impact of outliers and data volume, the proposed model demonstrates superior performance in handling outliers, filtering, and classification challenges. A new framework for feature selection-based classification has been developed to manage the extensive CTU-UHB heartbeat dataset effectively.Additionally, this work introduces a hybrid feature extraction technique aimed at identifying essential attributes from the

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

CTU-UHB signals. A novel classification model is also incorporated to enhance true positive rates and optimize runtime for large-scale data. Experimental evaluations reveal that the proposed feature extraction-based classification model outperforms conventional approaches by approximately 2% in efficiency, as measured by statistical metrics.

References

[1] M. Dixon and R. Butterfield, "442P Cognitive diversity in congenital myotonic dystrophy: implications for early intervention," Neuromuscular Disorders, vol. 43, p. 104441.517, Oct. 2024, doi: 10.1016/j.nmd.2024.07.526.
[2] P. Basak et al., "A novel deep learning technique for morphology preserved fetal ECG extraction from mother ECG using 1D-CycleGAN," Expert Systems with Applications, vol. 235, p. 121196, Jan. 2024, doi: 10.1016/j.eswa.2023.121196.
[3] R. Abburi et al., "Adopting artificial intelligence algorithms for remote fetal heart rate monitoring and classification using wearable fetal phonocardiography," Applied Soft Computing, vol. 165, p. 112049, Nov. 2024, doi: 10.1016/j.asoc.2024.112049.
[4] W. Xie et al., "AI-driven paradigm shift in computerized cardiotocography analysis: A systematic review and promising directions," Neurocomputing, vol. 607, p. 128446, Nov. 2024, doi: 10.1016/j.neucom.2024.128446.
[5] A. Jaba Deva Krupa, S. Dhanalakshmi, K. W. Lai, Y. Tan, and X. Wu, "An IoMT enabled deep learning framework for automatic detection of fetal QRS: A solution to remote prenatal care," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 9, pp. 7200–7211, Oct. 2022, doi: 10.1016/j.jksuci.2022.07.002.
[6] S. Rault, C. Vayssiere, E. Roth, E. David, R. Favre, and B. Langer, "Assessment of STAN S21 fetal heart monitor by medical staff," International Journal of Gynecology & Obstetrics, vol. 102, no. 1, pp. 8–11, Jul. 2008, doi: 10.1016/j.ijgo.2008.01.026.
[7] H. Allahem and S. Sampalli, "Automated labour detection framework to monitor pregnant women with a high risk of premature labour using machine learning and deep learning," Informatics in Medicine Unlocked, vol. 28, p. 100771, Jan. 2022, doi: 10.1016/j.imu.2021.100771.
[8] B. Williams and S. Arulkumaran, "Cardiotocography and medicolegal issues," Best Practice & Research Clinical Obstetrics & Gynaecology, vol. 18, no. 3, pp. 457–466, Jun. 2004, doi: 10.1016/j.bpobgyn.2004.02.005.
[9] X. Wang, Y. Han, and Y. Deng, "CSGSA-Net: Canonical-structured graph sparse attention network for fetal ECG estimation," Biomedical Signal Processing and Control, vol. 82, p. 104556, Apr. 2023, doi: 10.1016/j.bspc.2022.104556.
[10] L. Chen et al., "DANNMCTG: Domain-Adversarial Training of Neural Network for multicenter antenatal cardiotocography signal classification," Biomedical Signal Processing and Control, vol. 94, p. 106259, Aug. 2024, doi: 10.1016/j.bspc.2024.106259.
[11] A. M, S. S Kumar, E. E Nithila, and B. M, "Detection of Fetal Cardiac Anomaly from Composite Abdominal Electrocardiogram," Biomedical Signal Processing and Control, vol. 65, p. 102308, Mar. 2021, doi: 10.1016/j.bspc.2020.102308.

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632599
http://magellanes.com/

[12] R. Savirón-Cornudella et al., "Diagnosis of cardiotocographic sinusoidal patterns by spectral analyses," Biomedical Signal Processing and Control, vol. 93, p. 106174, Jul. 2024, doi: 10.1016/j.bspc.2024.106174.
[13] S. Magesh and P. S. Rajakumar, "Ensemble feature extraction-based prediction of fetal arrhythmia using cardiotocographic signals," Measurement: Sensors, vol. 25, p. 100631, Feb. 2023, doi: 10.1016/j.measen.2022.100631.
[14] A. Jaba Deva Krupa, S. Dhanalakshmi, N. L. Sanjana, N. Manivannan, R. Kumar, and S. Tripathy, "Fetal heart rate estimation using fractional Fourier transform and wavelet analysis," Biocybernetics and Biomedical Engineering, vol. 41, no. 4, pp. 1533–1547, Oct. 2021, doi: 10.1016/j.bbe.2021.09.006.
[15] S. Nguyen Van, J. A. Lobo Marques, T. A. Biala, and Y. Li, "Identification of Latent Risk Clinical Attributes for Children Born Under IUGR Condition Using Machine Learning Techniques," Computer Methods and Programs in Biomedicine, vol. 200, p. 105842, Mar. 2021, doi: 10.1016/j.cmpb.2020.105842.