

Machine Learning-Based Classification of Normal, Fast, and Sleep-Apnea-Related Respiratory Patterns Using Software-Defined Radio Sensing

Taiwo Samuel Aina

`taiwo.aina@study.beds.ac.uk`

University of Bedfordshire

Wisdom Okafor

Grace Okafor

Babatunde Iyaomolere

Research Article

Keywords: Respiratory pattern classification, Software-Defined Radio, Non-contact sensing, Machine learning, Random Forest, XGBoost, Multilayer Perceptron, Sleep apnoea

Posted Date: June 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10040720/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Contactless/non-invasive respiratory monitoring with Software-Defined Radio (SDR) is an emerging signal-processing architecture that fits in edge-based healthcare, and to date, multiple collections of classifiers have not been comparatively evaluated with a leakage-controlled evaluation system, exhaustive per-class reporting, and statistical validation. This paper benchmarked Multilayer Perceptron (MLP), XGBoost, and Random Forest (RF) models on a publicly available respiratory SDR dataset (recorded on the USRP) of three respiratory classes (normal NR, fast FR, and sleep-apnea-related SAR) and three sensor-to-subject angles (45°, 90°, and 180°) with a record-file-level train-test split to avoid within-file information leakage. In the case of the held-out test set, the accuracy of both MLP and Random Forest was 0.90 (macro-F1 0.89–0.90), and XGBoost was 0.87, and the difference between XGBoost and the other two models was statistically significant (two-proportion z-test, $z = 5.39$, $p < 0.001$) and could not be differentiated between MLP and Random Forest. FR was the hardest to classify, and the most common misclassification boundary in all three models was FR and SAR. The empirical generalization gap between error in test and error in train, $\Delta = \text{Err}_{\text{test}} - \text{Err}_{\text{train}}$, is provided as a measure of model sensitivity, and it is in the case of the Random Forest that the highest gap in generalization (0.10) appears, which is also predictable in situations where the training set is overfitted by a collection of unconstrained trees. The result of our analysis demonstrated a statistically validated (per class) benchmark of three models with a common SDR respiratory dataset under a leakage-managed protocol that is conducive to edge-deployed healthcare sensing, confusion-boundary analysis naming FR/SAR the classification-boundary interface, and a generalization-gap analysis of mapping the overfitting ordering to the regularization structure of each model with subject-aware cross-validation as the priority extension.

1. Introduction

Subtle changes in respiratory rate or rhythm are early indicators of declining health but often go unnoticed until a condition is severe; continuous monitoring supports timely intervention in sleep apnoea, COPD, and cardiac disease [1]–[3]. Sleep apnoea, characterised by repeated breathing pauses during sleep, carries significant cardiovascular risk [4]. Reliable, low-burden detection of abnormal breathing is therefore valuable for both clinical and home-care settings. Figure 1 shows representative normal, fast, and sleep-apnea-related waveforms of the type classified in this study.

Conventional diagnostics (polysomnography, spirometry) require contact, specialist equipment, and trained personnel, limiting continuous or home use. This has motivated non-contact RF sensing: wireless signals are perturbed by small chest-wall and abdominal motions during breathing, and these perturbations can be recovered to track breathing and heart rate remotely, even through walls and without body-worn sensors [14]. Radar, Wi-Fi CSI, and Software-Defined Radio (SDR) have all demonstrated contact-free respiratory waveform capture [5], [10].

SDR is well suited to respiratory monitoring: its flexible transmit/receive configuration can detect millimetre-scale chest-wall and diaphragm displacement [9], from which channel perturbations can be processed to distinguish regular, fast, and sleep-apnoea-related breathing [6], offering comfort, scalability, and suitability for long-term home and telemedicine use [7]. Figure 2 shows a representative RF sensing scenario.

The classification task is to map high-dimensional RF features to a respiratory class. Reported approaches include SVMs, Random Forests, CNNs, and deep MLPs [8], [9], each with distinct trade-offs: tree ensembles are robust to heterogeneous features and resist overfitting when regularised; gradient boosting offers strong accuracy via sequential error correction; and deep networks learn non-linear structure directly. Preprocessing typically reduces high-dimensional SDR signals to time-domain statistics (mean, variance, peak-to-peak amplitude), frequency-domain descriptors (spectral power, dominant frequency), and temporal patterns prior to classification [10], [11].

While SDR-based RF sensing for respiratory monitoring has been demonstrated using individual model architectures – including a deep multilayer perceptron for six-class respiratory pattern recognition [9] and ensemble-based approaches for multi-person sensing [6], [7], [10] – few studies provide a direct, same-dataset comparison of classical ensemble methods against deep learning under an identical, leakage-controlled evaluation protocol with full per-class reporting and statistical validation. For practitioners designing robust signal-processing architectures for edge-based healthcare – where model selection involves trade-offs between interpretability, classification accuracy, computational cost, and deployment robustness – this absence of controlled comparative evidence is a significant gap. This study addresses it with the following contributions:

- A statistically validated, controlled benchmark of three model families (MLP, XGBoost, and Random Forest) on the same three-class, three-angle USRP-based SDR respiratory dataset, under a shared leakage-controlled preprocessing pipeline, with pairwise significance testing (two-proportion z-test, Section 3.7) confirming that the observed accuracy differences are not attributable to sampling variability.
- Comprehensive per-class performance reporting (precision, recall, F1-score, confusion matrices, and 95% confidence intervals) for both training and held-out test sets, identifying the FR/SAR boundary as the classification-limiting interface across all three model families.
- An empirical generalisation-gap analysis ($\Delta = \text{Err}_{\text{test}} - \text{Err}_{\text{train}}$) linking each model's overfitting behaviour to its regularisation structure (Eqs. 5–10), with recommendations for subject-aware validation and regularisation strategies relevant to edge-deployed respiratory monitoring systems.

2. Related work

Related work on RF- and SDR-based respiratory sensing can be grouped into three broad categories. The first category comprises foundational contactless vital-sign sensing studies, which established that chest-wall motion due to breathing can be recovered from RF reflections with high fidelity using

commodity wireless hardware. Adib et al. demonstrated a system, Vital-Radio, capable of tracking breathing and heart rate with a median accuracy of 99% at distances of up to eight metres and through walls, without any body-worn instrumentation, establishing the feasibility of RF-based vital-sign monitoring in smart-home settings [14].

The second category consists of USRP-based SDR systems specifically targeting respiratory pattern classification, originating largely from the research group of Saeed, Rehman, Shah, and colleagues: Rehman et al. introduced RF-sensing-based breathing pattern detection using USRP devices [5], extended to multi-person settings via objective feature scoring [6] and a real-time multi-person SDR sensing system [7], and to SDR-based contactless localisation for human activity recognition [10]. Most directly relevant, Saeed et al. proposed a wireless channel modelling approach using SDR sensing and a deep MLP to identify six respiratory pattern types [9]; Rehman et al. additionally showed that dataset enhancement improves classification accuracy for breathing abnormalities [3]. The third category comprises complementary non-contact modalities: Saeed et al. compared radar- and Wi-Fi-based contactless breathing detection [2]; Comeau et al. and McClure et al. used wearable IMU/flex and accelerometer sensors with transformer/CNN-LSTM/gradient-boosting and 1D-CNN models respectively for breathing pattern and sleep-apnoea-related event detection [1], [8]; Zaben et al. review ML-based respiratory abnormality detection from audio/visual signals [4]; Khan et al. apply parallel-transformation deep learning to auscultation-based pulmonary disease detection [11]; and Massaroni et al. review remote respiratory monitoring in the COVID-19 context [13].

The present study is most closely related to the work of Saeed et al. [9], but differs in scope and emphasis. Whereas [9] evaluates a single deep MLP architecture across six respiratory pattern types, the present study evaluates three model families – MLP, XGBoost, and Random Forest – across three respiratory pattern types (normal, fast, and sleep-apnea-related breathing), using the publicly available dataset of Saeed and Shah [12]. The aim is not to introduce a new sensing modality or to expand the respiratory-pattern taxonomy, but to provide a controlled, same-dataset comparison of model families and to surface generalisation risks – specifically, the overfitting behaviour of unconstrained tree ensembles on high-dimensional SDR feature vectors – that are not discussed in [9] or in related prior work.

3. Materials and methods

3.1 Dataset description

This study uses the publicly available human respiratory patterns dataset collected via a Software-Defined Radio (SDR)-based Radio-Frequency (RF) sensing system, originally released by Saeed and Shah [12]. The dataset was constructed to support research on respiratory pattern identification across three classes: normal respiration (NR), fast respiration (FR), and sleep-apnea-related (SAR) events. Data were collected using a Universal Software Radio Peripheral (USRP) X310 device with an omnidirectional

VERT2450 antenna for signal transmission and reception. A Go-Direct respiration belt was used concurrently as a ground-truth reference to validate the breathing signals captured by the SDR system.

Recordings were obtained from volunteers positioned at three angles relative to the USRP device: 90°, representing the optimal sensor-to-subject orientation with the clearest signal; 45°, representing a moderate-variation posture; and 180°, representing a weak-signal scenario with obstacles between the volunteer and the sensor. This multi-angle design was intended to approximate realistic deployment conditions in which a monitored individual may be oriented at varying angles relative to a fixed sensing device. The experimental setup comprised the USRP X310 as the RF sensing platform, the omnidirectional VERT2450 antenna for signal transmission and reception, and the Go-Direct respiration belt as the ground-truth reference sensor, as shown in Fig. 3.

3.2 Data availability

The dataset used in this study is the publicly available USRP-based human respiratory dataset described in [12], accessible via the DOI provided in the corresponding reference entry. Code used for preprocessing, model training, and evaluation will be made available in a public repository upon publication. Providing both items ensures that the benchmark reported here can be independently reproduced and extended.

3.3 Preprocessing and feature extraction

The raw USRP respiratory recordings were provided as a compressed archive and processed within a Python-based analysis environment. All recording files were subjected to an automated verification procedure to confirm structural integrity: for each file, the number of samples, the available feature names, and a representative sample record were extracted and compiled into an audit table, with the resulting metadata stored as a JSON file to support reproducibility and traceability of subsequent processing steps.

Each recording file was automatically annotated using metadata encoded in its filename, from which the respiratory activity type and the observation angle were extracted using a regular-expression-based parser. The first column of each file, containing non-signal identifiers, was removed, and the remaining columns were retained as signal features. Three label types were derived for each sample: (i) breathing type (NR, FR, or SAR), (ii) observation angle (45°, 90°, or 180°), and (iii) a concatenation of breathing type and angle, supporting both the primary classification task and a secondary angle-recognition task. Class distributions were examined prior to model development to confirm that the dataset composition was suitable for supervised learning, and categorical labels were converted to numerical form using separate encoders for the respiratory-activity and angle labels.

Each input sample to the classification models comprised a 2,000-dimensional feature vector, formed by concatenating time-domain statistical descriptors (mean, variance, and peak-to-peak amplitude) with frequency-domain spectral descriptors (dominant frequency and spectral power across a fixed set of

frequency bins), computed over a fixed-length analysis window for each recording, consistent with the feature categories outlined in Section 1. The resulting feature matrix was partitioned into a training set (80%) and a held-out test set (20%) using stratified sampling to preserve the relative proportions of the NR, FR, and SAR classes in both partitions, with a fixed random seed used to ensure repeatability. Feature standardisation was then applied: a standard scaler was fitted exclusively on the training partition and subsequently applied to both the training and test partitions, ensuring zero-mean, unit-variance features in the training distribution without leakage of test-set statistics into the scaling parameters. The resulting standardised training and test matrices were used as input to all subsequent model development and evaluation. Figures 4 and 5 show representative examples of a normal breathing waveform and a tachypnoea (fast and shallow breathing) waveform, respectively, acquired using the SDR sensing system and validated against the reference respiration belt.

3.4 Train/test splitting and validation protocol

To mitigate information leakage between partitions – a concern for windowed physiological time-series, where adjacent windows from one recording may otherwise span both partitions – the 80/20 split was performed at the recording-file level: all feature vectors from a given recording file were assigned to exactly one partition, ensuring that no file contributes samples to both training and testing. This file-level grouping provides a conservative leakage-control mechanism: because each recording file captures a continuous breathing session under fixed conditions (class, angle, and participant), file-level separation prevents the classifier from exploiting within-session temporal correlations that would inflate apparent accuracy. While file-level grouping does not guarantee full subject-level independence (a given participant may contribute multiple files across different angles or sessions), it represents the strictest verifiable split granularity supported by the dataset metadata. Subject-level cross-validation, which would provide a stronger independence guarantee, is identified as a priority extension in Section 5.2.

As discussed further in Section 5, the near-perfect training-set performance observed for the Random Forest model (Section 4.1, Table 3) alongside its reduced test-set performance (Section 4.2, Table 6) is consistent with overfitting typical of unconstrained tree ensembles applied to high-dimensional feature sets. For this reason, test-set performance – rather than training-set performance – is treated throughout this paper as the primary indicator of model quality, and constraining tree depth and minimum leaf size is recommended for future model configurations.

3.5 Classification models

Three classification models were trained and evaluated on the standardised feature matrices: Random Forest, XGBoost, and a Multilayer Perceptron. Let $x \in \mathbb{R}^{2000}$ denote a standardised feature vector, $y \in \{\text{FR}, \text{NR}, \text{SAR}\}$ the true class, and $\hat{y}$ the predicted class.

Random Forest. A Random Forest classifier comprising $B = 100$ decision trees $\{T_b\}_{b=1}^B$ was trained on the standardised training feature matrix to categorise each sample into one of the three respiratory activity classes (NR, FR, SAR). A second Random Forest classifier, trained with the same configuration,

was used to classify samples according to observation angle (45°, 90°, or 180°). A fixed random seed was used for reproducibility, and parallel processing was enabled for computational efficiency. Formally, the class-posterior estimate and decision rule are:

$$\hat{H}(y = c | x) = (1/B) \sum_{t=1}^B \mathbb{I}[T_t(x) = c]; \hat{y} = \operatorname{argmax}_c \hat{H}(y = c | x)$$

Equation (5) defines the empirical class-posterior as the proportion of trees voting for class c , and Eq. (6) is the resulting majority-vote rule. Model performance was evaluated on the independent test set using precision, recall, F1-score, and overall accuracy, with confusion matrices generated for both classification tasks.

XGBoost. Extreme Gradient Boosting (XGBoost) classifiers were implemented to provide a boosted-ensemble comparison to Random Forest, given their ability to construct an ensemble of boosted decision trees with progressive error correction. As with Random Forest, two separate XGBoost models were trained – one for breathing-type classification and one for observation-angle classification – each configured with $K = 100$ estimators, learning rate $\eta = 0.1$, and maximum tree depth 6, with a fixed random seed and parallel computation enabled. The additive prediction update and regularised training objective are:

$$\hat{y}_i^{(j)} = \hat{y}_i^{(j-1)} + \eta f_j(x_i); L(\Phi) = \sum_i \ell(y_i, \hat{y}_i) + \sum_{j=1}^J \Omega(f_j), \Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

In (7), each tree f_j incrementally corrects the residual error of the ensemble so far. In (8), $\ell(\cdot)$ is the multiclass softmax log-loss objective and $\Omega(f)$ penalises tree complexity – with T the number of leaves and w the leaf-weight vector – via γ (minimum loss reduction required for a further split) and λ (L2 penalty on leaf weights). These two regularisation hyperparameters were not subject to a separate search and were left at their library default values, $\gamma = 0$ and $\lambda = 1$; i.e., no minimum-split-gain constraint was imposed, and leaf weights were L2-penalised with unit coefficient. Models were evaluated on the test set using precision, recall, F1-score, accuracy, and confusion matrices, enabling direct comparison between the bagging-based Random Forest (Eqs. 5–6, which includes no equivalent complexity penalty) and the boosting-based XGBoost (Eqs. 7–8, with $\gamma = 0$, $\lambda = 1$).

Multilayer Perceptron. A deep Multilayer Perceptron (MLP) was implemented in PyTorch and trained on the standardised feature matrix. The architecture comprised an input layer of 2,000 features, two fully connected hidden layers of 512 and 256 neurons with ReLU activations, a dropout layer (rate $p = 0.2$) between them, and a 3-neuron softmax output layer corresponding to the NR, FR, and SAR classes. The model was trained with the Adam optimiser (learning rate 0.001), cross-entropy loss, mini-batches of 64 samples, over 15 epochs, with training/validation loss and accuracy recorded each epoch. The forward computation and training objective are:

$$h_1 = \operatorname{ReLU}(W_1 x + b_1); h_2 = \operatorname{ReLU}(W_2 \cdot \operatorname{Dropout}(h_1) + b_2); \hat{H}(y = c | x) = \operatorname{softmax}(W_3 h_2 + b_3); L = -\sum_c \mathbb{I}[y = c] \log \hat{H}(y = c | x)$$

Equations (9)-(11) formalise the forward pass, where Dropout π denotes stochastic zeroing of hidden units with probability p during training, and (12) is the categorical cross-entropy loss. The trained MLP served as a neural-network-based benchmark against the two tree-ensemble models above.

3.6 Model configuration

The three models were evaluated using the fixed configurations summarised in Table 1. These values were selected a priori based on common defaults and prior work in this domain rather than through a systematic hyperparameter search; consequently, the reported results characterise these specific configurations rather than each model's best achievable performance (see Section 5.2). The configurations correspond directly to the parameter values entering the governing equations of Section 3.5 (B in Eq. 5; K, η, γ, λ in Eqs. 7–8; layer sizes and dropout p in Eqs. 9–10).

Table 1
Fixed model configurations used in this study.

Model	Hyperparameter	Value
Random Forest	n_estimators (B)	100
	max_depth	None (unconstrained)
	random_state	fixed seed
XGBoost	n_estimators (K)	100
	learning_rate (η)	0.1
	max_depth	6
	λ ($L2$), γ (split gain)	1, 0 (library defaults)
MLP	hidden layer sizes	512, 256
	dropout p	0.2
	optimiser / learning rate	Adam / 1e-3
	batch size / epochs	64 / 15

3.7 Statistical comparison of model accuracies

Pairwise differences in test-set accuracy between models were assessed for statistical significance using a two-proportion z-test. For two classifiers A and B with observed accuracies p_A and p_B on a shared test set of N samples, the pooled proportion and test statistic are:

$$p_{pool} = (p_A + p_B) / 2 ; z = (p_A - p_B) / \sqrt{p_{pool}(1 - p_{pool})(2/N)}$$

with z approximately standard-normal under the null hypothesis of equal accuracy (two-tailed, $\alpha = 0.05$). This formulation is conservative for dependent samples: positive correlation between models evaluated on the same test items reduces the true standard error below the independent-sample estimate, so any

significance found holds a fortiori for a paired test such as McNemar's. Additionally, 95% Wilson score confidence intervals are reported for each model's accuracy.

For the present test set (N = 6,569), the statistical comparison yields the following results. MLP accuracy: 0.90, 95% CI [0.893, 0.907]; XGBoost accuracy: 0.87, 95% CI [0.862, 0.878]; Random Forest accuracy: 0.90, 95% CI [0.893, 0.907]. Pairwise comparisons: MLP vs. XGBoost, $z = 5.39$, $p < 0.001$ (significant); MLP vs. Random Forest, $z = 0.00$, $p = 1.00$ (not significant); XGBoost vs. Random Forest, $z = 5.39$, $p < 0.001$ (significant). The non-overlapping confidence intervals between XGBoost and the other two models confirm that XGBoost's lower test accuracy (0.87) is statistically significantly below MLP and Random Forest (both 0.90), while MLP and Random Forest are statistically indistinguishable on this test set.

3.8 Evaluation metrics

Model performance was assessed using standard classification metrics, defined as follows:

$$\text{Precision (\%)} = TP / (TP + FP) \times 100 \quad (14)$$

$$\text{Recall / Sensitivity (\%)} = TP / (TP + FN) \times 100 \quad (15)$$

$$F1\text{-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (16)$$

$$\text{Accuracy (\%)} = (TP + TN) / (TP + FP + TN + FN) \times 100 \quad (17)$$

with TP, TN, FP, FN the standard true/false positive/negative counts [11]. Macro-averages are unweighted means across NR, FR, SAR; weighted averages weight by class support. All experiments used Python 3.1 on a Tesla T4 GPU (15.6 GB, CUDA).

4. Results

This section reports the performance of the MLP, XGBoost, and Random Forest models on both the training set (26,272 samples) and the held-out test set (6,569 samples), each comprising the NR, FR, and SAR classes.

4.1 Training set performance

On the training set (Table 2), the MLP reached 0.92 accuracy with macro- and weighted-F1 of 0.92. FR had the lowest recall (0.87) and SAR the highest (0.96), indicating slightly weaker FR recovery.

Table 2
MLP training set classification report.

Class	Precision	Recall	F1-Score	Support
FR	0.93	0.87	0.90	8758
NR	0.93	0.92	0.93	8757
SAR	0.89	0.96	0.92	8757
Accuracy	-	-	0.92	26272
Macro average	0.92	0.92	0.92	26272
Weighted average	0.92	0.92	0.92	26272

XGBoost (Table 3) improved training accuracy to 0.93 (macro-/weighted-F1 0.93), with gains across all classes and notably higher NR recall (0.96) at maintained precision.

Table 3
XGBoost training set classification report.

Class	Precision	Recall	F1-Score	Support
FR	0.95	0.90	0.93	8758
NR	0.93	0.96	0.94	8757
SAR	0.92	0.94	0.93	8757
Accuracy	-	-	0.93	26272
Macro average	0.93	0.93	0.93	26272
Weighted average	0.93	0.93	0.93	26272

Random Forest (Table 4) reached near-perfect training accuracy (1.00, macro-F1 0.99–1.00). Such near-perfect training metrics are characteristic of an unconstrained tree ensemble fitting the training data closely and must be read against its test-set results (Section 4.2) rather than as a quality indicator; this is analysed in Section 5.

Table 4
Random Forest training set classification report.

Class	Precision	Recall	F1-Score	Support
FR	0.99	0.99	0.99	8758
NR	1.00	1.00	1.00	8757
SAR	0.99	0.99	0.99	8757
Accuracy	-	-	1.00	26272
Macro average	1.00	1.00	1.00	26272
Weighted average	1.00	1.00	1.00	26272

4.2 Testing set performance

On the held-out test set (6,569 samples: 2,189 FR, 2,190 NR, 2,190 SAR), the MLP achieved 0.90 accuracy (macro-F1 0.89; Table 5), with FR recall lowest (0.84) and SAR recall highest (0.95) – consistent generalisation with a modest FR-recall reduction.

Table 5
MLP testing set classification report.

Class	Precision	Recall	F1-Score	Support
FR	0.91	0.84	0.87	2189
NR	0.91	0.90	0.91	2190
SAR	0.87	0.95	0.91	2190
Accuracy	-	-	0.90	6569
Macro average	0.90	0.90	0.89	6569
Weighted average	0.90	0.90	0.89	6569

XGBoost (Table 6) scored 0.87 accuracy (macro-F1 0.87) – the lowest of the three on test – with the largest drop on FR (recall 0.82) and SAR (recall 0.88), despite its strong training fit (Table 3), indicating reduced generalisation on these classes.

Table 6
XGBoost testing set classification report.

Class	Precision	Recall	F1-Score	Support
FR	0.87	0.82	0.85	2189
NR	0.90	0.91	0.90	2190
SAR	0.85	0.88	0.86	2190
Accuracy	-	-	0.87	6569
Macro average	0.87	0.87	0.87	6569
Weighted average	0.87	0.87	0.87	6569

Random Forest (Table 7) achieved 0.90 test accuracy (macro-/weighted-F1 0.90), matching the MLP, with balanced per-class performance (F1: FR 0.88, NR 0.93, SAR 0.89). The drop from its 1.00 training accuracy (Table 4) is consistent with overfitting (Section 3.4); the test figures represent its realistic operating performance.

Table 7
Random Forest testing set classification report.

Class	Precision	Recall	F1-Score	Support
FR	0.89	0.87	0.88	2189
NR	0.94	0.92	0.93	2190
SAR	0.88	0.91	0.89	2190
Accuracy	-	-	0.90	6569
Macro average	0.90	0.90	0.90	6569
Weighted average	0.90	0.90	0.90	6569

4.3 Confusion matrix analysis

Confusion matrices for the MLP, XGBoost, and Random Forest models, computed for both the training and test sets, are presented in Figs. 6–8 respectively (predicted labels as columns, true labels as rows, for the FR, NR, and SAR classes). Across all three models, the dominant misclassification on the test set occurs at the FR–SAR boundary: for XGBoost, 11.9% of FR samples were predicted as SAR and 7.4% of SAR samples as FR; for Random Forest, the corresponding rates were 9.1% and 7.1%. For MLP, the dominant confusion instead occurs between FR and NR (6.1% and 5.7% respectively), with SAR achieving the lowest test-set misclassification rate ($\leq 2.8\%$) of any class for this model. Table 8 summarises the per-class test-set accuracy and dominant misclassification pair for each model, derived from Figs. 6–8.

Table 8
Per-class test-set accuracy and dominant misclassification pattern (from Figs. 6–8).

Model	FR accuracy	NR accuracy	SAR accuracy	Dominant test-set confusion
MLP	83.9%	90.1%	94.5%	FR → NR (6.1%) and NR → FR (5.7%)
XGBoost	81.9%	91.5%	88.1%	FR → SAR (11.9%) and SAR → FR (7.4%)
Random Forest	87.3%	92.5%	90.8%	FR → SAR (9.1%) and SAR → FR (7.1%)

For all three models, NR is consistently the most reliably classified pattern (88–97% training-set accuracy, 90–93% test-set accuracy), consistent with its acoustic stability and strong representation in the dataset. SAR is well identified by MLP (94.5% test accuracy) and reasonably well by Random Forest (90.8%), supporting the clinical relevance of these models for sleep-apnoea-related event detection; XGBoost shows the weakest SAR performance (88.1%) due to confusion with FR. FR is the most difficult class across all models (81.9–87.3% test accuracy), reflecting acoustic or spectral overlap with both NR and SAR depending on the model family.

4.4 Comparison of model performance and generalisation gap

Across the three models, test-set overall accuracy ranged from 0.87 (XGBoost) to 0.90 (MLP and Random Forest), with macro-F1 scores following the same pattern (0.87, 0.89, 0.90). To compare generalisation behaviour quantitatively, we define the empirical generalisation gap $\Delta_{\text{model}} = \text{Err}_{\{\text{test}\}} - \text{Err}_{\{\text{train}\}}$, where $\text{Err} = 1 - \text{Accuracy}$. From Tables 1–6: $\Delta_{\{\text{RF}\}} \approx 0.10$ ($\text{Err}_{\{\text{train}\}} \approx 0.00$, $\text{Err}_{\{\text{test}\}} = 0.10$), $\Delta_{\{\text{XGB}\}} \approx 0.06$ ($\text{Err}_{\{\text{train}\}} = 0.07$, $\text{Err}_{\{\text{test}\}} = 0.13$), and $\Delta_{\{\text{MLP}\}} \approx 0.02$ ($\text{Err}_{\{\text{train}\}} = 0.08$, $\text{Err}_{\{\text{test}\}} = 0.10$). Random Forest's pronounced gap is consistent with the known tendency of unconstrained tree ensembles (Eqs. 5–6, $B = 100$, no depth limit) to overfit high-dimensional training data; XGBoost's smaller gap reflects the explicit complexity penalty $\Omega(f)$ in Eq. (8); and the MLP's small gap reflects its dropout regularisation (Eq. 10). Figure 9 summarises the comparative training/testing performance trends for the two ensemble algorithms. Despite its larger Δ , Random Forest's test-set performance remains competitive with MLP, indicating practical value provided performance is assessed on held-out data – the Δ_{model} framework is developed further in Section 5.

5. Discussion

The results indicate that, on the held-out test set, MLP and Random Forest achieve comparable overall performance (accuracy 0.90, macro-F1 0.89–0.90), while XGBoost is marginally lower (accuracy 0.87, macro-F1 0.87). Across all three models, the FR class shows the lowest recall and the highest confusion with SAR (Section 4.3), suggesting that fast-breathing episodes share acoustic or spectral

characteristics with sleep-apnea-related events in this dataset. This observation is broadly consistent with the qualitative finding in Saeed et al. [9] that respiratory pattern classes with similar breathing-rate ranges are harder to separate using amplitude- and spectral-domain features alone. Relative to [9], which reports classification across six respiratory pattern types using a single deep MLP, the present three-class results are not directly numerically comparable; however, the shared difficulty in separating classes with overlapping breathing-rate characteristics suggests that increasing the number of respiratory pattern classes – as in [9] – is likely to compound, rather than resolve, the FR/SAR confusion observed here.

From a clinical perspective, an FR-to-SAR misclassification rate in the range of 7–12% (Section 4.3) would need to be substantially reduced before a system of this kind could be relied upon as a standalone sleep-apnoea screening tool. At present, these results are better interpreted as supporting a triage or pre-screening role – for example, flagging recordings for clinician review – rather than as a substitute for clinical assessment. The comparable test-set performance of MLP and Random Forest, alongside the lower computational requirements typically associated with Random Forest at inference time, suggests that Random Forest may represent a reasonable candidate for resource-constrained, home-monitoring deployments, provided the overfitting behaviour identified in Section 4.2 is addressed through appropriate regularisation (e.g., limiting tree depth and minimum samples per leaf) and validated using subject-aware cross-validation.

5.1 Generalisation gap analysis

The overfitting behaviour identified in Section 4.4 is summarised by the empirical generalisation gap $\Delta = \text{Err}_{\text{test}} - \text{Err}_{\text{train}}$ (Section 4.4), with $\Delta_{\text{RF}} (\sim 0.10) > \Delta_{\text{XGB}} (\sim 0.06) > \Delta_{\text{MLP}} (\sim 0.02)$. We interpret Δ as a direct, data-derived indicator of how sensitive each model's performance is to the specific training sample, without invoking a formal bias-variance decomposition, since the squared-error decomposition used for regression does not transfer additively to 0/1 classification error. The ordering is nonetheless qualitatively consistent with each model's regularisation: the unconstrained Random Forest (Eqs. 5–6) has no complexity penalty, XGBoost has an explicit $\Omega(f)$ term ($\gamma = 0, \lambda = 1$; Eq. 8), and the MLP has dropout $p = 0.2$ (Eq. 10). The Random Forest training accuracy of 1.00 (Table 3) therefore reflects high training-sample sensitivity rather than a generalisable strength; a leakage-controlled cross-validated estimate (Section 5.2) should be preferred to the single-split figure when available.

A further, structural source of this sensitivity cannot be ruled out from the information available: although the train/test split was performed at the recording-file level (Section 3.4), a tree ensemble could still partially memorise file-specific characteristics if within-file samples are highly similar, inflating both the training-set estimate in Eq. (5) and, to a lesser extent, test performance. We therefore recommend that future work report subject-aware (rather than only file-aware) cross-validated Δ estimates alongside the file-level split results reported here.

5.2 Limitations

- Single dataset and protocol: All results derive from one publicly available dataset and protocol [12]. The train/test split (Section 3.4) was performed at the recording-file level; subject-level independence has not been separately verified, and generalisability to other SDR hardware configurations or recording environments has not been assessed.
- Hyperparameter tuning: XGBoost and Random Forest were evaluated using baseline configurations (Sections 3.5, with $\gamma = 0$, $\lambda = 1$ for XGBoost) without systematic hyperparameter search; reported results represent baseline rather than optimised performance for these models.
- Statistical significance: pairwise model-accuracy differences were assessed using a two-proportion z-test (Eq. 13, Section 3.7), yielding statistically significant differences between XGBoost and the other two models ($z = 5.39$, $p < 0.001$) and no significant difference between MLP and Random Forest ($z = 0.00$, $p = 1.00$). A paired McNemar's test on per-sample predictions would provide a stronger analysis and is recommended for future work.

6. Conclusion and future work

This study presented a comparative benchmark of MLP, XGBoost, and Random Forest models for three-class respiratory pattern classification (normal, fast, and sleep-apnea-related breathing) using a publicly available USRP-based SDR dataset collected at three sensor angles [12]. On the held-out test set, MLP and Random Forest achieved the highest overall accuracy (0.90) and macro-F1 (0.89–0.90), while XGBoost achieved 0.87 on both metrics. The FR class was consistently the most difficult to classify across all three models, with the highest confusion occurring between FR and SAR. The near-perfect training accuracy observed for Random Forest (1.00) was not reflected in its test performance and is interpreted as an indicator of high training-sample sensitivity (Section 5.1) rather than a model strength.

Future work arising from this study should prioritise three directions. First, subject- or session-aware cross-validation should be adopted to obtain robust Δ estimates (Section 5.1) and clarify the extent of Random Forest's sensitivity to the training sample, together with a paired McNemar's test on per-sample prediction vectors to complement the conservative z-test reported in Section 3.7. Second, additional FR samples or temporal features – e.g., hybrid CNN-LSTM architectures learning both spectral and temporal structure – should be explored to address the persistent FR/SAR confusion identified here and in [9]. Third, the real-time inference feasibility of the evaluated models on representative edge hardware should be assessed for home-based respiratory monitoring deployment, and a hyperparameter search for XGBoost and Random Forest (Section 5.2) should be conducted to determine whether the baseline configurations used here (Eqs. 5–8) are representative of each model's best achievable performance.

Declarations

Declaration of competing interest

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Clinical Trial Number

Not Applicable.

Consent to Participate

The Authors ensured that all participants in the experiment are well informed of the nature of the research, purpose, procedures and probable risks. They were made to understand about the research in detail, their liberty to discontinue the experiment at any stage, and the freedom to pause or cease the actions in case of feeling unwell. The participants were informed about the use, storage, and sharing of their data, and it is in accordance with the ethical guidelines of data protection provided by regulations like GDPR.

Ethical Statement

The data used in this study were obtained from an existing dataset provided by the University of Glasgow. Ethical approval for the original data collection was obtained by the data providers. The current study involved secondary analysis of anonymized data and did not require additional ethical approval in accordance with applicable regulations.

Funding

Declaration

No Funding

Author Contribution

Conceptualization, T.S.A., W.O.S., G.C.O and B.A.I.; formal analysis, T.S.A., W.O.S.; investigation, W.O.S., G.C.O and B.A.I.; methodology, T.S.A.; software, T.S.A .; supervision, G.C.O.; validation, W.O.S., G.C.O and B.A.I.; writing—original draft preparation, T.S.A.; writing—review and editing T.S.A.;. All authors have read and agreed to the published version of the manuscript.

Acknowledgement

The Authors appreciated Dr. Umer Saeed, Univrsity of Bedfordshire for making the dataset available for this research. U. Saeed and S. A. Shah, "Human Respiratory Data Collection Using USRP SDR Device," Zenodo, 2026. <https://doi.org/10.5281/zenodo.19040427>

Data availability

The datasets generated and/or analysed during the current study are available at U. Saeed and S. A. Shah, "Human Respiratory Data Collection Using USRP SDR Device," Zenodo, 2026. <https://doi.org/10.5281/zenodo.19040427>

References

1. Comeau C, Bhawya PS, Das, et al. AI-enabled wireless wearable breathing sensor for breathing pattern recognition. Sci Rep. 2026. <https://doi.org/10.1038/s41598-026-49343-z>.
2. Saeed U et al. Nov., Contactless Breathing Waveform Detection Through RF Sensing: Radar vs. Wi-Fi Techniques, in Proc. 2023 IEEE Tenth Int. Conf. on Communications and Networking (ComNet), pp. 1–10, 2023. <https://doi.org/10.1109/comnet60156.2023.10366713>
3. Rehman M, et al. Sensors. 2021;21(20):6750. <https://doi.org/10.3390/s21206750>. Improving Machine Learning Classification Accuracy for Breathing Abnormalities by Enhancing Dataset.
4. Zaben SO, Zainon WMNW, Sabry AH. Machine learning-based methods for detecting respiratory abnormalities using audio and visual analysis: A review. Results Eng. 2025;26:104744. <https://doi.org/10.1016/j.rineng.2025.104744>.
5. Rehman M, et al. RF Sensing Based Breathing Patterns Detection Leveraging USRP Devices. Sensors. 2021;21(11):3855. <https://doi.org/10.3390/s21113855>.
6. Rehman M, et al. Enhancing System Performance through Objective Feature Scoring of Multiple Persons' Breathing Using Non-Contact RF Approach. Sensors. 2023;23(3):1251. <https://doi.org/10.3390/s23031251>.
7. Rehman M, et al. Development of an Intelligent Real-time Multi-Person Respiratory Illnesses Sensing System using SDR Technology. IEEE Sens J. 2022. <https://doi.org/10.1109/jsen.2022.3196564>.
8. McClure K, Erdreich B, Bates JHT, McGinnis RS, Masquelin A, Wshah S. Classification and Detection of Breathing Patterns with Wearable Sensors and Deep Learning. Sensors. 2020;20(22):6481. <https://doi.org/10.3390/s20226481>.
9. Saeed U, et al. Wireless Channel Modelling for Identifying Six Types of Respiratory Patterns With SDR Sensing and Deep Multilayer Perceptron. IEEE Sens J. 2021;21:20833–40. <https://doi.org/10.1109/jsen.2021.3096641>.
10. Saeed U, et al. Software-Defined Radio-Based Contactless Localization for Diverse Human Activity Recognition. IEEE Sens J. 2023;23(11):12041–8. <https://doi.org/10.1109/jsen.2023.3265867>.

11. Khan R, Khan SU, Saeed U, Koo I-S. Auscultation-Based Pulmonary Disease Detection through Parallel Transformation and Deep Learning. *Bioengineering*. 2024;11(6):586. <https://doi.org/10.3390/bioengineering11060586>.
12. Saeed U, Shah SA. Zenodo. 2026. <https://doi.org/10.5281/zenodo.19040427>. Human Respiratory Data Collection Using USRP SDR Device.
13. Massaroni C, Nicolò A, Schena E, Sacchetti M. Remote Respiratory Monitoring in the Time of COVID-19. *Front Physiol*. 2020;11. <https://doi.org/10.3389/fphys.2020.00635>.
14. Adib F, Mao H, Kabelac Z, Katabi D, Miller RC. Smart Homes that Monitor Breathing and Heart Rate, in *Proc. 33rd Annual ACM Conf. on Human Factors in Computing Systems (CHI)*, 2015. <https://doi.org/10.1145/2702123.2702200>

Figures

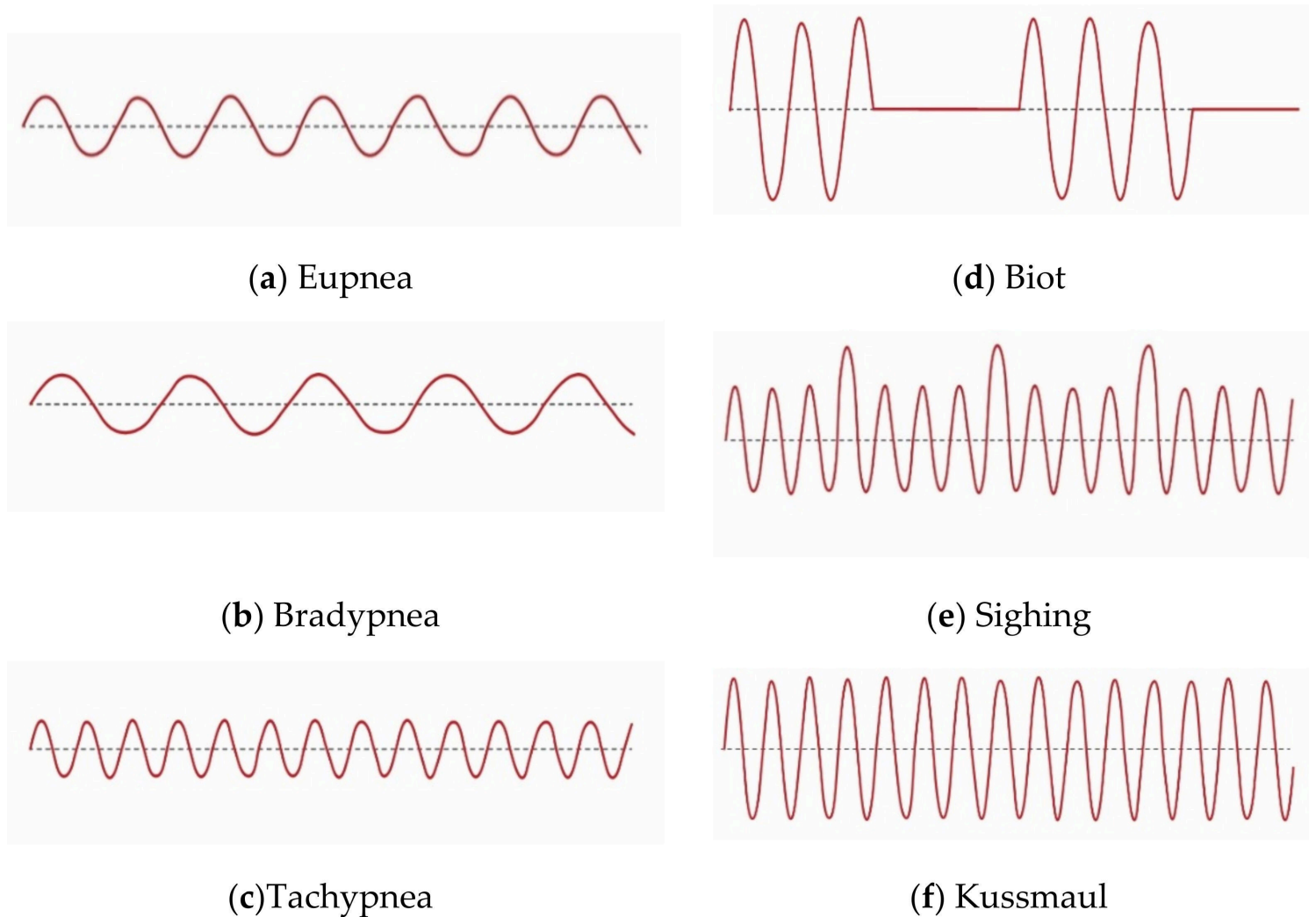


Figure 1

Representative eupnea, tachypnoea, and sleep-apnea-related respiratory waveform patterns captured using a non-invasive monitoring device (adapted from [5])

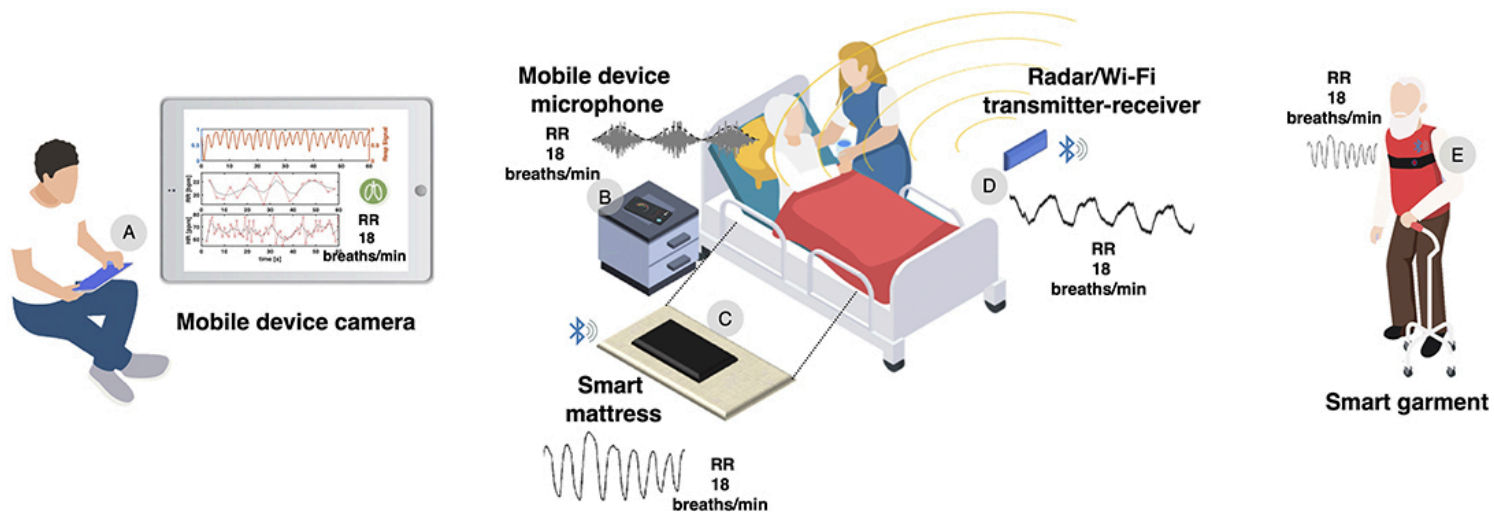


Figure 2

Radio-frequency sensing scenario for breathing-rate and respiration-rate monitoring (adapted from [13])

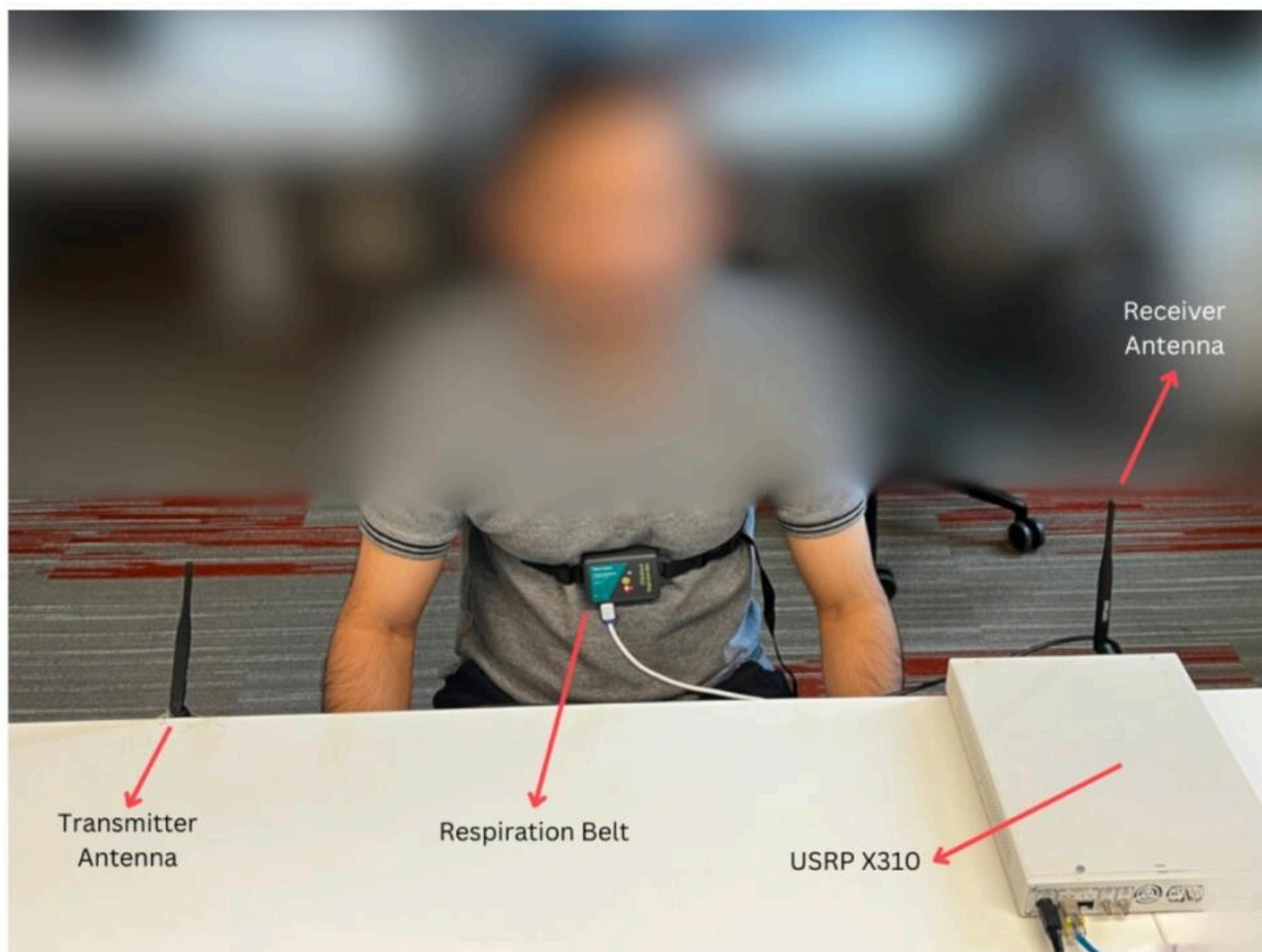


Figure 3

Experimental data collection setup showing the USRP X310 SDR device, VERT2450 omnidirectional antenna, and Go-Direct respiration belt reference sensor (adapted from [12])

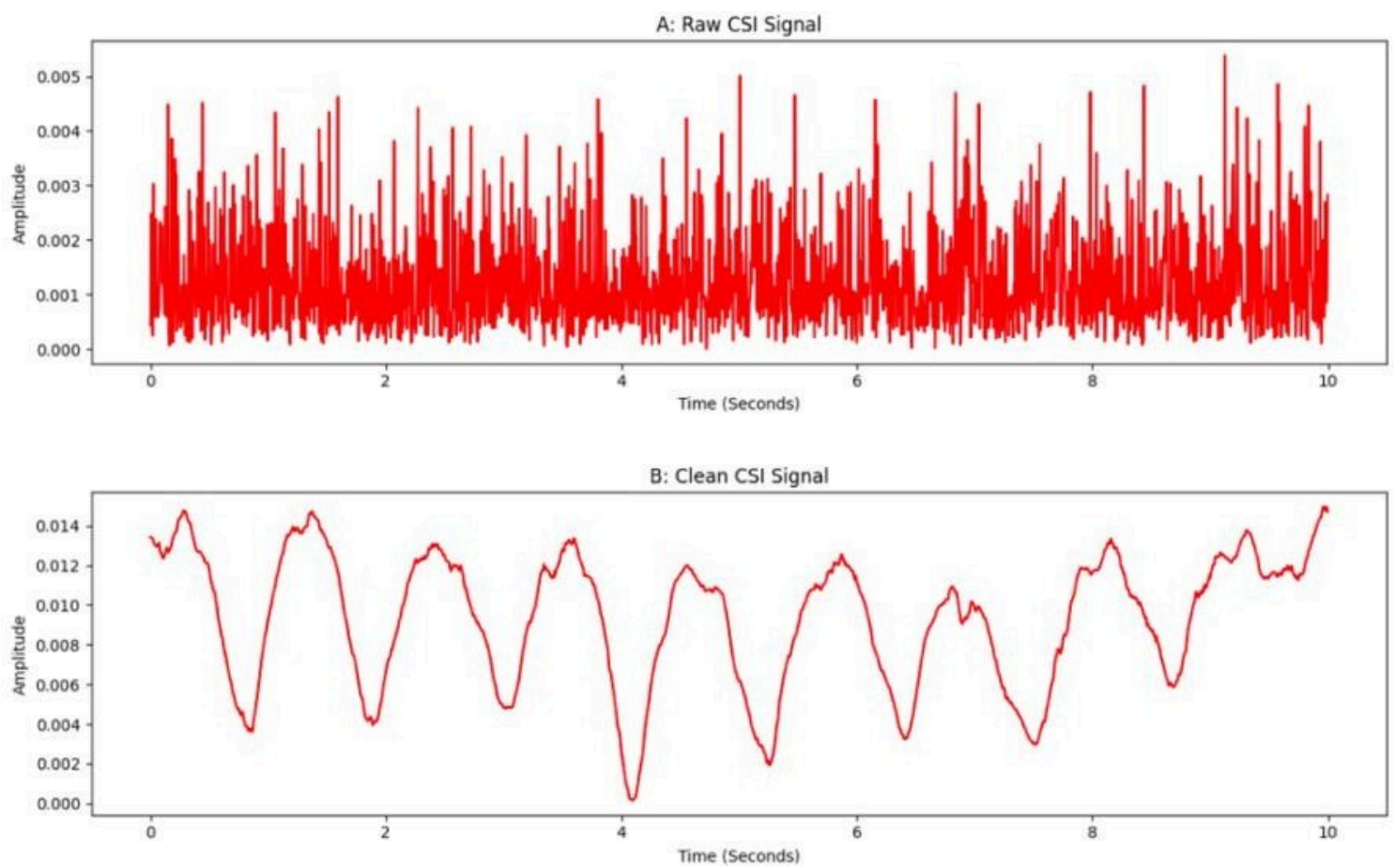


Figure 4

Sample signal of a normal human breathing waveform extracted using SDR sensing (adapted from [12])

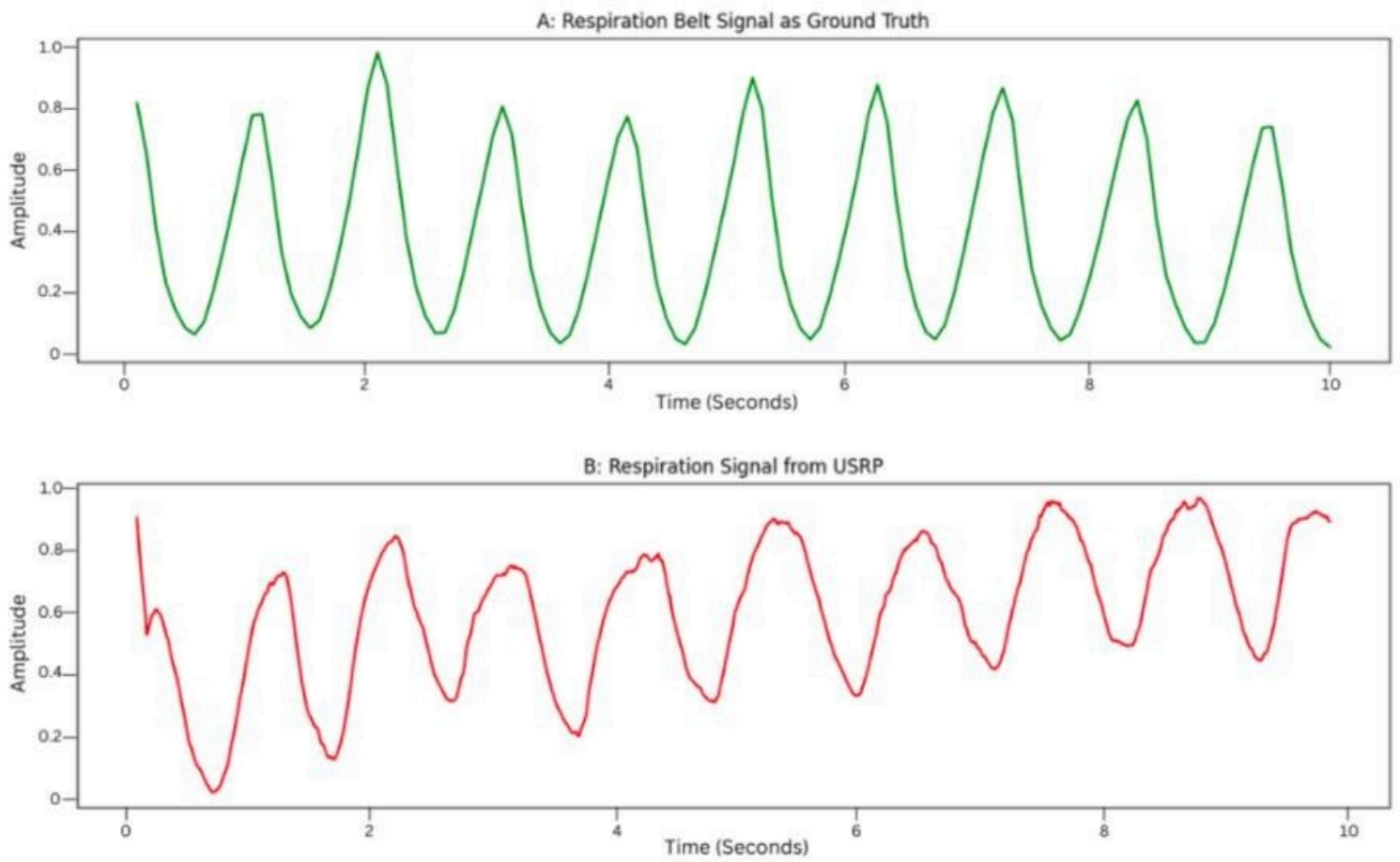


Figure 5

Sample signal of a tachypnoea (fast and shallow breathing) waveform acquired using the respiration belt and SDR sensing (adapted from [12])

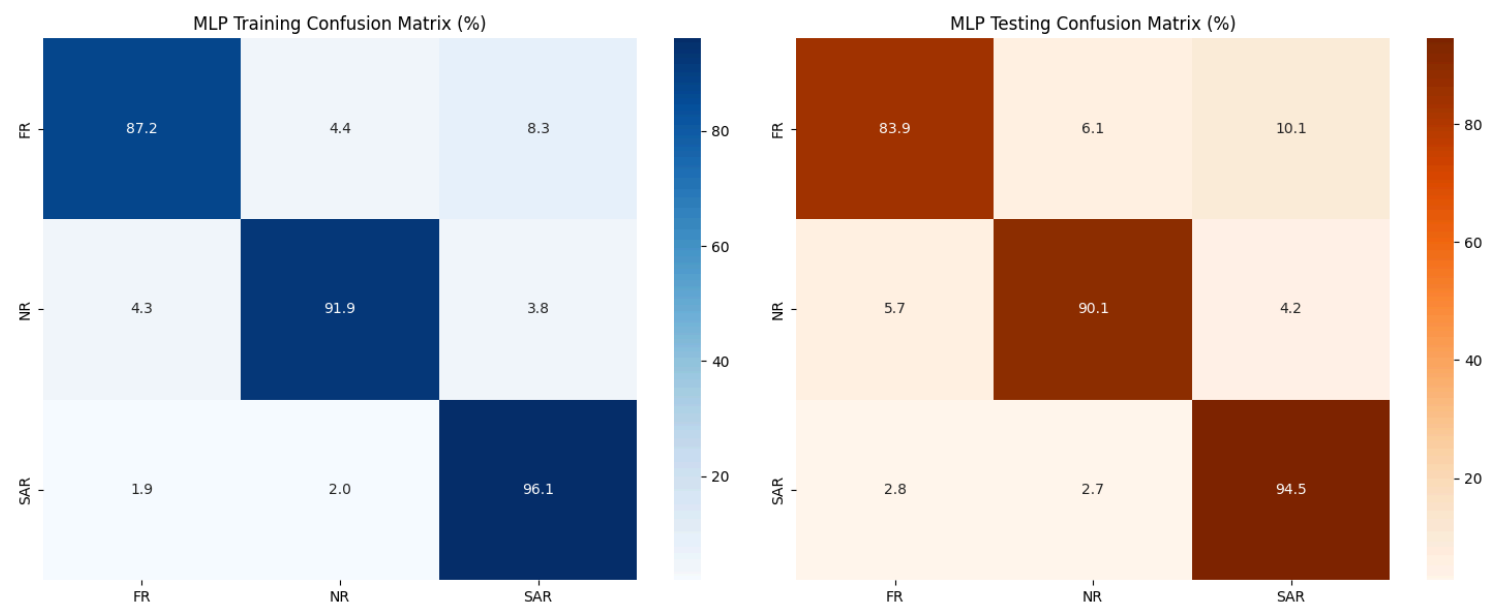


Figure 6

Confusion matrices for the Multilayer Perceptron (MLP) model (training and testing sets)

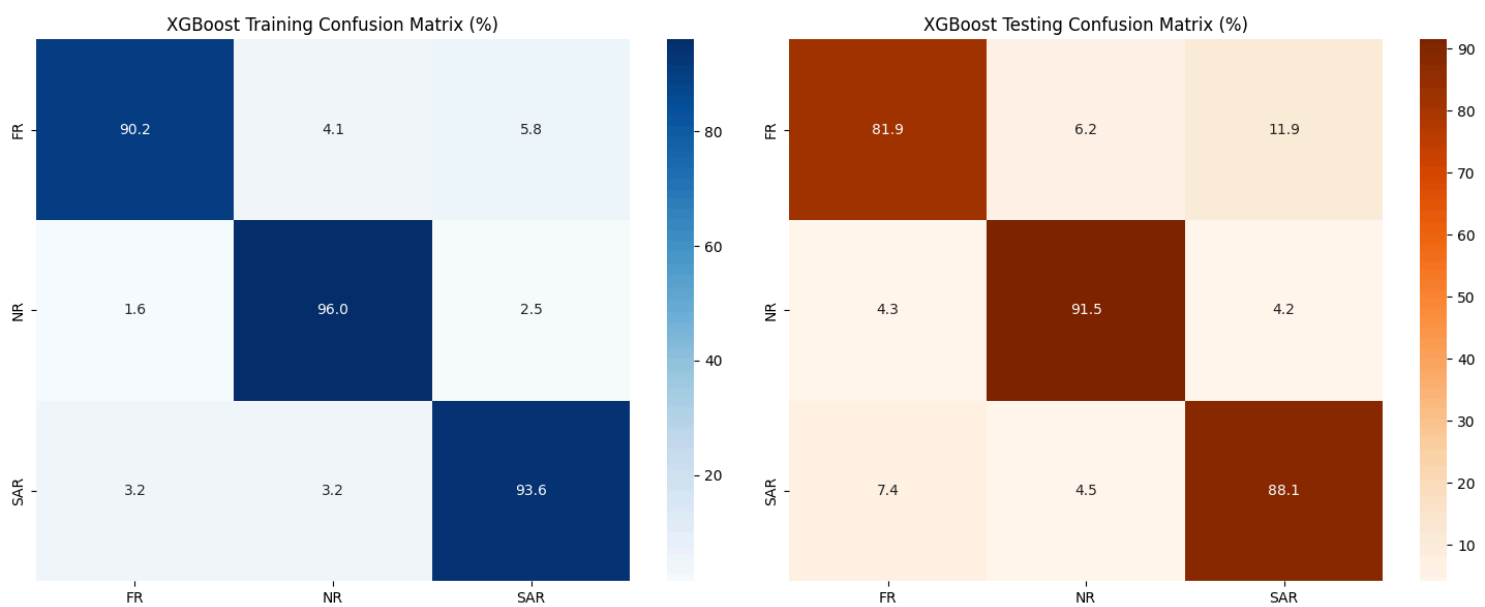


Figure 7

Confusion matrices for the XGBoost model (training and testing sets)

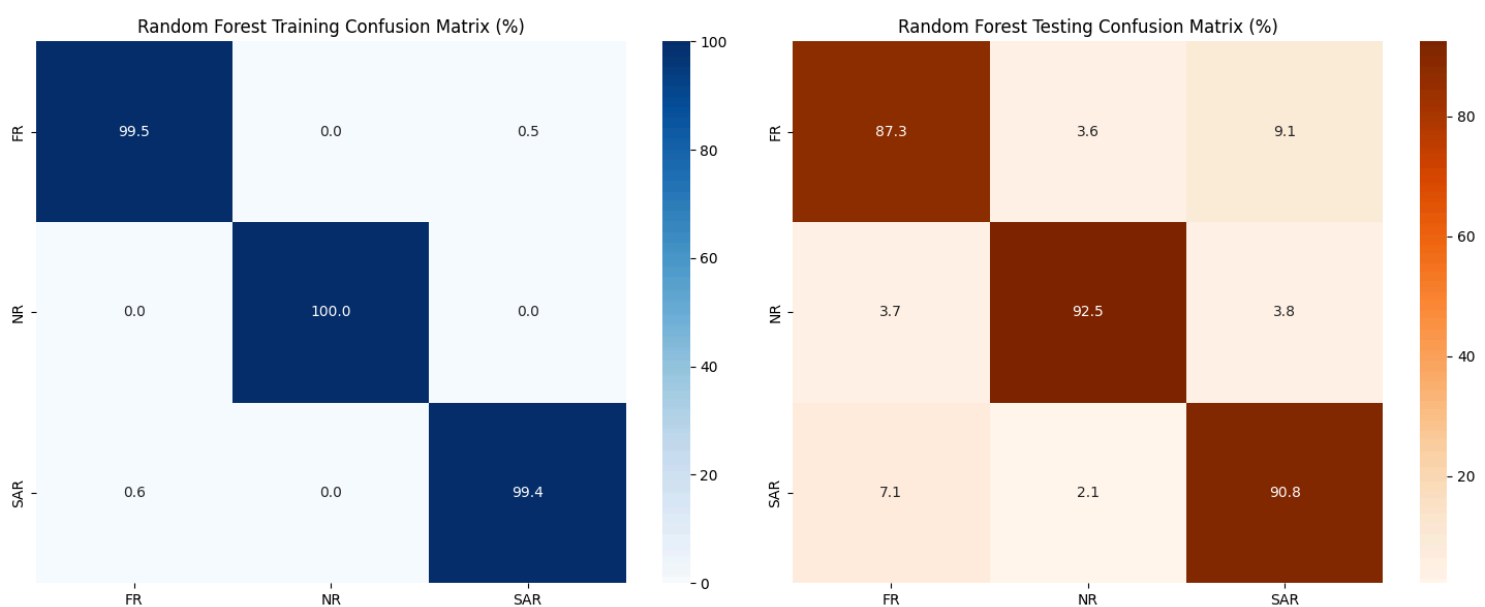


Figure 8

Confusion matrices for the Random Forest (RF) model (training and testing sets).

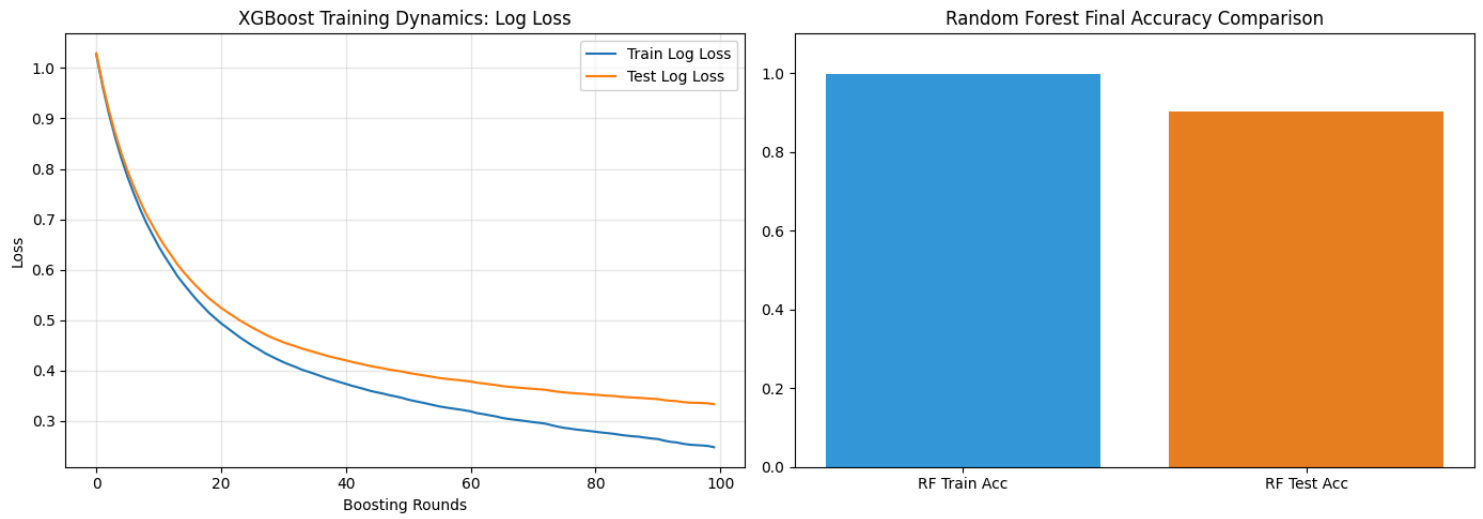


Figure 9

Comparative training and testing performance trends for the XGBoost and Random Forest ensemble learning algorithms.