

An Explainable Multi-System Composite Scoring Framework for Patient Health Risk Prediction Using Ensemble Learning

Nushrat Jahan Chadni ^{1*}, Nusrat N. Suzana¹

Abstract

Early clinical intervention and medical resources allocation require precise predictions on patient risks. The practical value of traditional risk assessment methods in clinical decision-making is restricted because solutions commonly rest on single physiological information and are complicated to interpret. The model combines state-of-the-art ensemble learning and explainable AI techniques with clinically interpretable composite indices to propose an explainable multi-system composite scoring model for patient risk assessment. To capture complex physiological couplings between multiple organ systems, three new combined indices have been developed: the Physiological Risk Index (PRI), Multi-System Early Warning Score (MS-EWS) and Critical Threshold Composite (CTC). The system employs the Synthetic Minority Over-sampling Technique (SMOTE) for class imbalance processing and includes interaction-based feature engineering. To increase prediction robustness, a stacked ensemble model using Random Forest, XGBoost and LightGBM classifiers trained with 5-fold cross-validation for meta-learner optimization is employed. Performance evaluation shows that the proposed model outperforms individual base learners, achieving 98% accuracy on an independent test set, as well as existing state-of-the-art models. The integration of SHAP and LIME ensures both global and instance-level explainability. Moreover, misclassifications mainly occur between closely related categories, thereby maintaining clinical safety and highlighting key composite features that influence model decisions. The findings suggest that the proposed model presents an effective, interpretable and clinically reliable multi-class health risk prediction approach.

Keywords: Health risk prediction, Composite scoring, Ensemble learning, Explainable artificial intelligence, SMOTE, Multi-system analysis

*Corresponding Author email: nushraat78@gmail.com

¹ Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh
DOI: <https://doi.org/10.64494/JFST/v4i1/nc/2026/10/195-216>

1. Introduction

Effective health risk assessment is central to modern medical practice, enabling timely intervention, reduced mortality, and efficient resource allocation. Chronic multi-system disorders, involving concurrent organ deterioration, account for most global morbidity (Wang, 2023). Traditional clinical scoring systems are static and fail to capture inter-organ interactions (Li, 2023), while conventional machine learning models lack clinical interpretability, limiting their adoption (Rodriguez, 2023), and due to biases common in clinical datasets that result in model underperformance on important high-risk cases (Chen, 2025). To address these limitations an Explainable Multi-System Composite Scoring Framework (EMSCSF) introduced.

Conventional scores like NEWS and MEWS are simple and widely used. However, they fall short owing to a simple linear structure without the ability to model inter-organ physiologic coupling (Li, 2023). Machine learning models such as logistic regression, SVM, and deep neural networks achieve higher accuracy but lack interpretability, limiting clinical trust (Rodriguez, 2023). Ensemble methods such as Random Forest, XGBoost, and LightGBM outperform traditional models but often lack interpretability and imbalance handling (Abdullah, 2024; Singh, 2024). The EMSCSF proposed in this study addresses such gaps by integrating multi-system composite scoring, ensemble learning enriched with SMOTE-based class balancing and explainability at two levels by SHAP and LIME.

This study introduces three indices: PRI, MS-EWS, and CTC. The proposed indices capture nonlinear relationships among physiological variables, improving risk representation. SHAP and LIME provide global and local explanations, improving model interpretability.

The main contributions of this work are as follows:

- a. Three composite indices called PRI, MS-EWS, and CTC have been created to overcome the limitations of traditional single-system scores and capture multi-organ interactions by using interaction-based feature engineering.
- b. An efficient stacked ensemble model that consists of Random Forest, XGBoost, and LightGBM, trained on a balanced dataset with SMOTE and refined with meta-learning allows for the prediction of 98% accuracy and still meets clinician requirements for a safe platform.

- c. The integration of SHAP and LIME provides a patient-specific explanation for why an individual may be misclassified, along with an analysis of the global features associated with that classification. This offers clinically useful insights for therapy purposes and also increases the transparency in decision support.

2. Literature Review

Chronic diseases are one of the major causes of death around the world (Wang, 2023), leading to increased interest in utilizing advanced feature extraction techniques for improved prediction performance.

Systematic reviews demonstrate that ensemble and deep learning models, particularly Random Forest, consistently outperform conventional clinical scoring systems, achieving superior performance in mortality and health risk prediction (Abdullah, 2024). Similarly, ensemble-based machine learning models such as Random Forest and XGBoost, when combined with population-based metaheuristic optimization techniques including Particle Swarm Optimization and Genetic Algorithm, have shown improved predictive performance in cardiovascular disease risk classification (Nova et al., 2025). However, most existing studies focus on single-organ assessment rather than comprehensive multi-system evaluation. Composite clinical scores partially address this limitation by integrating multiple physiological parameters into a unified risk representation (Patel, 2022).

Nevertheless, current composite scores have several drawbacks: the majority of those assessed address single-organ systems rather than multi-system evaluations; few have incorporated advanced machine learning and clinical interpretability. Many conventional early warning scores, such as NEWS and MEWS, continue to be limited because they use static evaluations and lack the capability to account for the complex linkages among physiological variables, yet they have been shown to be clinically useful (Li, 2023).

As clinicians increasingly require transparent decision-making, healthcare organizations are adopting explainable AI (XAI) (Rodriguez, 2023). Among XAI methods, SHAP and LIME have demonstrated effective model interpretability in biomedical and diagnostic applications, supporting clinical risk prediction (Lee, 2024). In a related healthcare application, an ensemble learning framework combining pretrained convolutional neural networks with classical classifiers has been shown to improve both accuracy and interpretability, using LIME to explain feature contributions in brain tumour classification (Shafi & Mirza, 2025). A

recent systematic review of 36 empirical studies further reports that SHAP and LIME were the two most frequently applied XAI techniques in healthcare diagnosis and rehabilitation, though most studies lacked formal validation of explanation faithfulness or clinical relevance (Aravindkumar et al., 2026). However, the black-box nature of many machine learning models continues to limit clinical adoption despite advances in XAI (Rodriguez, 2023). Balancing predictive performance with interpretability remains an ongoing research challenge for post-hoc methods such as SHAP and LIME.

In healthcare statistics, class imbalance occurs where only a small percentage of the total population comprises high-risk cases. In recent meta-analyses of mortality prediction using machine learning models, the best-performing models achieved C-statistics of 0.89. However, the way data is handled has a significant impact on model performance (Abdullah, 2024). Many ways to address imbalance exist, including SMOTE and its variations, but there is little information regarding the effects of those methods on multi-system risk prediction (Kumar, 2023).

Despite recent advances, no comprehensive framework integrates multi-system risk assessment, composite clinical scores, explainable AI, and state-of-the-art machine learning into a unified prediction model. Although ensemble methods outperform traditional risk calculators (Singh, 2024), and SHAP and LIME have been widely applied in medical applications (Lee, 2024; Zhang, 2023), their use remains largely focused on post-hoc diagnostic explanations rather than clinically interpretable risk prediction.

3. Dataset

This study uses publicly available anonymized datasets containing clinical data, vital signs, and a multi-level health risk label for 1,000 patients, with all identifiers removed and collected under standard ethical and regulatory guidelines (Beckett, 2024).

3.1 Dataset Characteristics

The dataset includes categorical (e.g., gender), continuous (e.g., age), and clinical physiological signals selected for clinical relevance and routine use in acute care monitoring in the Table 1.

Table 1: Dataset Features and Clinical Descriptions

Feature	Clinical Description	Unit/Scale
Patient_ID	Unique anonymized identifier	Categorical
Respiratory_Rate	Breaths per minute	12–40 bpm
Oxygen_Saturation	Blood oxygen saturation	74–100%
O ₂ _Scale	Oxygen therapy scale	Ordinal 1–2
Systolic_BP	Systolic blood pressure	50–146 mmHg
Heart_Rate	Cardiac rate	60–163 bpm
Temperature	Body temperature	35.6–41.8°C
Consciousness	Neurological status	A/P/C/V/U
On_Oxygen	Supplemental O ₂ usage	Binary 0/1
Risk_Level	Health risk classification (Target)	Normal, Low, Medium, High

3.2 Class Distribution and Data Quality

The study examined the number of patients in each class to comprehend the distribution of risk levels. This aids in determining class imbalance, which is essential for training models. Table 2 shows the total number of patients in each risk category. It also shows that the distribution in training and testing sets follows stratified sampling.

3.3 Data Partitioning Strategy

To prevent information leakage, the dataset was stratified into 80% training and 20% testing sets. A total of 800 patient records were used for training, oversampling, and cross-validation, while 200 samples were reserved for evaluation Shown in Table 2.

Table 2: Risk Level Distribution and Train-Test Split

Risk Level	Total	%	Train (80%)	Test (20%)
High	279	27.9	223	56
Medium	306	30.6	245	61
Low	255	25.5	204	51
Normal	160	16.0	128	32
Total	1000	100	800	200

4. Methodology

The EMSCSF integrates scoring systems with interpretable AI for explainable health risk prediction, as shown in Figure 1, from preprocessing to prediction.

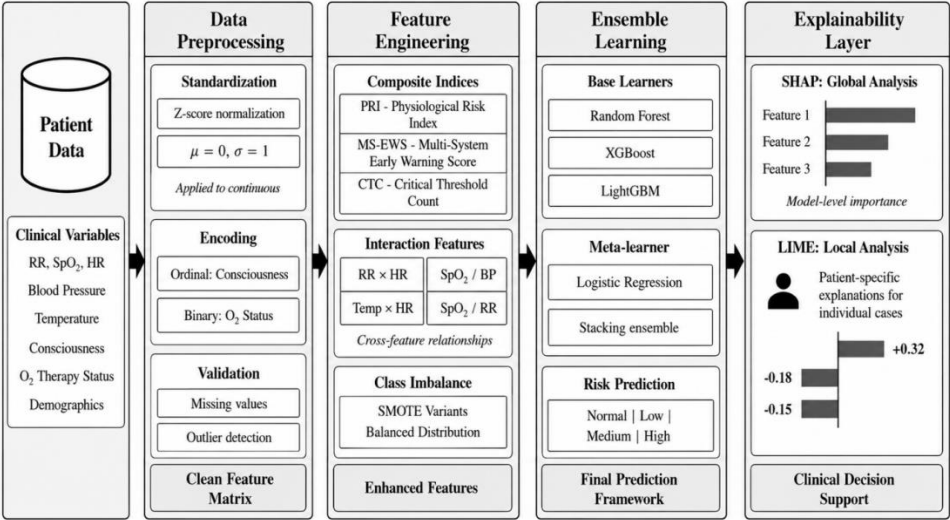


Figure 1: Proposed Workflow for Health Risk Prediction

4.1 Data Preprocessing

Data preprocessing ensures consistent and reliable model inputs through standardization, encoding, and data integrity checks. Numerical features (e.g., respiratory rate, O₂ saturation, blood pressure, heart rate, temperature) are standardized using StandardScaler, $x' = (x - \mu)/\sigma$, ensuring equal contribution across features measured on different scales. The AVPU scale is ordinaly encoded to preserve clinical severity (Alert=0, Pain=1, Confusion=2, Verbal=3, Unresponsive=4), while binary variables (O₂ scale, on oxygen) remain unchanged. With 100% data completeness, no imputation is required, ensuring unbiased and fully observed model inputs.

4.2 Novel Composite Indices

Three composite indices enable clinicians to assess complex multi-system physiological interactions and clinical decline beyond individual vital sign measurements. Their clinical relevance and ranges are summarized in Table 3.

Table 3: Novel Composite Indices: Range and Clinical Significance

Index	Range	Clinical Meaning
PRI	0.075–0.822	Weighted physiological stress
MS-EWS	0–15	Multi-system dysfunction severity
CTC	0–7	Breadth of critical threshold violations

Physiological Risk Index (PRI) quantifies overall physiological stress using weighted normalization of cardiorespiratory parameters, providing a continuous measure of systemic strain:

$$PRI = 0.35RR_{\text{norm}} + 0.35HR_{\text{norm}} + 0.30Temp_{\text{norm}} \quad (1)$$

RRnorm, HRnorm, and Tempnorm are normalized to [0,1] using min-max scaling.

The weighting coefficients used to calculate the PRI were derived from previous clinical literature on the urgency of corrective action required for cardiorespiratory parameters. Equal and greater weights (0.35 each) were assigned to respiratory rate (RR) and heart rate (HR) because these are the earliest and most sensitive indicators of physiological deterioration. Tachycardia (rapid heart rate) and tachypnoea (rapid breathing) are important compensatory responses to a lack of oxygen (hypoxia), blood infection (sepsis), and unstable blood flow (hemodynamic instability) (Royal College of Physicians, 2017; Smith et al., 2013), and are the two most common signs seen in early warning systems such as NEWS. Because temperature was assigned a slightly lower weight of 0.30, it serves as a secondary metabolic response which may remain within normal limits even in the presence of significant decline in other body systems (Thompson et al., 2022). Therefore, prioritizing RR and HR over temperature is clinically justified and expected to provide improved predictive capability compared to equal-weight alternatives.

To validate this weighting scheme, a sensitivity analysis was performed by varying each coefficient by ± 0.05 while maintaining a total weight of 1.0. As shown in Table 4, the proposed configuration (0.35/0.35/0.30) achieved the highest accuracy (98.00%) and F1-score (0.9800), confirming the effectiveness of the clinically informed weighting strategy.

Table 4: Sensitivity Analysis of PRI Weight Configurations

Weight Configuration (RR/HR/Temp)	Accuracy	F1-Score
0.35 / 0.35 / 0.30 (Proposed)	98.00%	0.9800
0.33 / 0.33 / 0.34 (Equal)	97.50%	0.9750
0.40 / 0.40 / 0.20	96.50%	0.9651
0.30 / 0.40 / 0.30	97.50%	0.9750
0.40 / 0.30 / 0.30	97.50%	0.9750

The integration of clinically informed weighting improves interpretability and supports the applicability of PRI across diverse patient populations without requiring model retraining. SHAP analysis further supports this finding, showing that PRI achieves the highest mean absolute SHAP value (1.0702) among all 15 features, outperforming individual features such as respiratory rate, heart rate, and temperature. This provides strong evidence that PRI captures interaction-based physiological patterns that cannot be represented by any single vital sign alone. The results confirm that PRI is more than a simple linear combination, acting instead as a clinically meaningful composite that reflects cross-variable dependencies within the ensemble learning framework.

MS-EWS (Multi-System Early Warning Score) is an integrated clinical scoring tool that provides a complete evaluation of a person's health status through five major elements, and it is similar to established early warning systems. Each parameter is scored from 0 to 3: 0 for normal, 1 for mild deviation, 2 for moderate deviation, and 3 for severe abnormality. Normal ranges follow standard early warning criteria: respiratory rate 12–20 bpm, oxygen saturation $\geq 96\%$, heart rate 51–90 bpm, systolic blood pressure 111–139 mmHg, and alert consciousness. Severe thresholds (3 points) are: respiratory rate < 9 or > 30 bpm, $SpO_2 < 90\%$, heart rate < 60 or > 130 bpm, systolic blood pressure < 90 or > 160 mmHg, and impaired consciousness (Smith et al., 2013; Royal College of Physicians, 2017).

$$MS-EWS = RR_{score} + SpO_{2score} + HR_{score} + BP_{score} + Cons_{score} \quad (2)$$

Total composite scores are based on the greatest accumulated number of critical breaches across the five parameters. In such cases, when a patient meets or exceeds a total composite score of 7, they would be classified as needing immediate intervention.

$$CTC = \sum_{i=1}^7 \mathbb{1} [parameter_i \in critical\ range] \quad (3)$$

where $\mathbb{1}(\cdot)$ is the indicator function that returns 1 when the parameter falls within a critical range and 0 otherwise. Critical thresholds include: respiratory rate (< 12 or > 30 breaths per minute), oxygen saturation ($< 90\%$), heart rate (< 60 or > 130 beats per minute), systolic blood pressure (< 90 or > 160 mmHg), body temperature ($< 36^\circ C$ or $> 38.5^\circ C$), consciousness level (any state other than Alert), and oxygen requirement ($_{On\ Oxygen} = 1$). Unlike scoring systems that focus on the severity of dysfunction within individual parameters, CTC assesses the degree of multi-organ dysfunction by calculating the number of different physiological systems that are simultaneously affected by the dysfunction.

To highlight the novelty of the proposed indices, Table 5 compares PRI, MS-EWS, and CTC against the two widely used clinical early warning scores NEWS and MEWS.

Table 5: comparative analysis of proposed indices vs. Existing clinical score

Feature	NEWS/ MEWS	Proposed Indices (PRI, MS-EWS, CTC)
Multi-system	Partial	Yes
Interaction Modeling	No	Yes
Score representation	Discrete	Continuous
ML compatibility	Low	High
Explainability	No	Yes

The differences between the proposed indices and currently used early warning systems (e.g., NEWS and MEWS) are evident in their design methodology. NEWS and MEWS primarily rely on threshold-based aggregation of individual vital signs and do not explicitly model inter-variable relationships or cross-system physiological interactions. As a result, these systems are limited in representing continuous physiological stress, multi-system dysfunction, and simultaneous threshold violations. In contrast, the proposed framework enables continuous modeling of physiological stress through PRI, multi-system dysfunction assessment through MS-EWS, and quantification of simultaneous threshold violations through CTC. These enhanced representations provide a more expressive characterization of patient state while preserving clinical interpretability and improving compatibility with modern machine learning-based prediction systems.

4.3. Interaction Feature Engineering

Four interaction features were developed to characterize physiological coupling and the levels of physiological similarity between systems that are meaningful, rather than relying on only single pieces of data.

- a. **RR×HR Interaction** highlights simultaneous efforts of both the cardiovascular system and the respiratory system to increase their performance to provide oxygen to tissues experiencing metabolic stress or low oxygenation. This coordinated effort to increase production is referred to as ‘cardiovascular-respiratory coupling’ during physiological stress.

- b. SpO₂/BP Ratio shows perfusion efficiency is measured by the ratio of blood oxygen saturation to mean arterial pressure; a low ratio indicates poor oxygen delivery despite adequate blood pressure, suggesting either reduced oxygen extraction by tissues or impaired microcirculation.
- c. Temperature×HR Product represents the relationship between metabolic and cardiac responses due to an increased demand for cardiac output and metabolic rate because of high environmental temperatures; these relationships become decoupled in instances of sepsis or cardiac insufficiency.
- d. SpO₂/RR measures oxygenation efficiency by comparing oxygen saturation with respiratory effort; a lower ratio indicates impaired gas exchange.

The total composite feature set comprises 15 intended derived features as a combination of 7 original features, 3 composite indices, 4 interaction features and 1 derived consciousness variable.

4.4 Class Imbalance Mitigation

Clinical datasets are often imbalanced, leading to majority bias and poor high-risk detection. SMOTE variants were used to balance classes and compared in Table 6.

Table 6: SMOTE Variants: Strategies and Sample Generation

Technique	Strategy	Focus	Training Samples
SMOTE	Linear interpolation between k-nearest neighbors	Uniform synthetic generation	980 (balanced)
ADASYN	Adaptive density-based weighting	Hard-to-learn. regions	965 (nearly-balanced)
Borderline-SMOTE	Borderline samples	Boundary emphasis	980 (balanced)
SMOTE-ENN	SMOTE + Edited Nearest Neighbors under- sampling	Noise reduction	884 (cleaned)

SMOTE was applied only to the 800 training samples, while the 200 test samples were kept unchanged. Figure 2 illustrates the class distribution before and after applying each oversampling technique, demonstrating the reduction of class imbalance across methods.

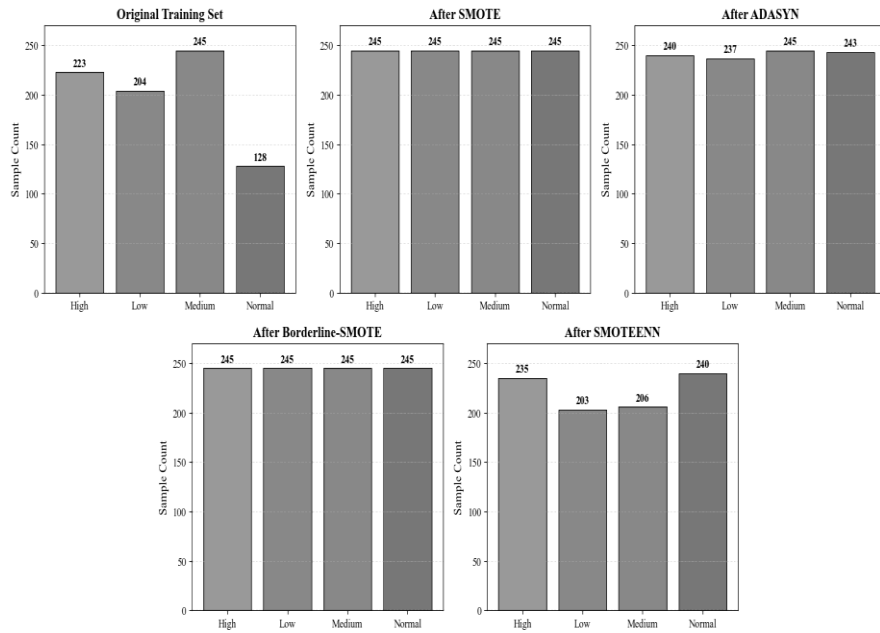


Figure 2: Class distribution before and after oversampling methods on the training set

To further analyse the effect of oversampling, Figure 3 summarizes the total sample count across different sampling methods, highlighting the variation in dataset size introduced by each technique.

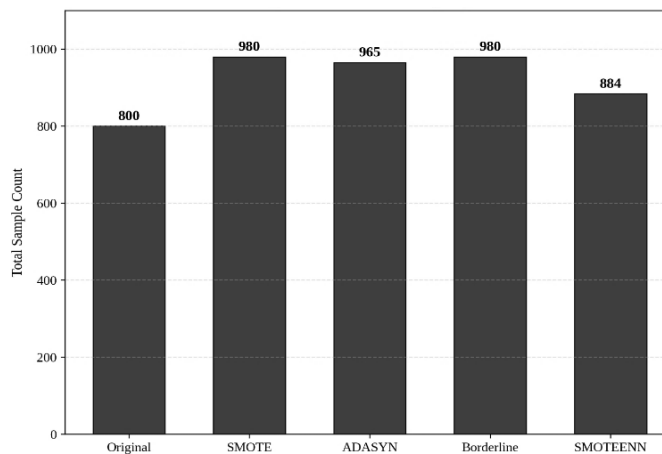


Figure 3: Total sample count comparison across sampling methods.

4.5 Ensemble Learning Architecture

A two-tier stacking ensemble combines complementary base learners to capture diverse physiological patterns, reducing individual model limitations and improving overall predictive robustness. The overall ensemble architecture and hyperparameter settings are summarized in Table 7.

Table 7: Ensemble Model Architecture and Hyperparameters

Component	Configuration	Key Strength
Random Forest	n_estimators = 100, max_depth = 10, min_samples_split = 5, min_samples_leaf = 2	Robustness via bagging, natural feature importance
XGBoost	n_estimators = 100, max_depth = 6, learning rate = 0.1, sub_sample = 0.8, col_sample=0.8	Robustness via bagging, natural feature importance
XGBoost	n_estimators = 100, max_depth = 6, learning rate = 0.1, num_leaves = 31	Efficiency, histogram-based learning
Meta-Learner	Logistic Regression, max_iter = 1000, 5-fold CV	Optimal weight learning

The Random Forest, XGBoost and LightGBM models were selected based on three specific criteria: 1) All three models demonstrate superior performance as compared to traditional ML methods in clinical risk prediction benchmarks (Abdullah, 2024; Singh, 2024). 2) They each have diversity in terms of algorithms to minimize correlated errors within the ensemble. 3) They are all computationally feasible to allow deployment in clinical practice. The Random Forest algorithm uses a bagging method to build uncorrelated decision trees which provides robustness against overfitting while allowing for calculation of natural feature importance values. The XGBoost model uses sequential gradient boosting with L1/L2 regularization which allows for detection of highly complicated nonlinear patterns while controlling for model complexity. LightGBM uses a histogram to develop trees with leaf-wise growth and is therefore able to achieve much faster training on imbalanced data sets while still maintaining high accuracy. The aforementioned models are fundamentally different in their inductive bias; specifically, bagged trees developed in parallel and iteratively boosted sequentially using regularization and very efficient leaf construction using gradient techniques, which allow for negligible correlation in prediction errors across their models. This diversity in inductive biases is critical to the effectiveness of the ensemble stacking methodology (Singh, 2024).

In the training data, the base models (Random Forest, XGBoost, and LightGBM)

provide preliminary predictions that are used as meta-features by the second-level learner. The Logistic Regression meta-learner is trained using a 5-fold stratified cross-validation technique on the outputs from the base models to establish how best to combine and weight the predicted outcomes under varying physiological conditions and patient characteristics. The weighting strategy is adapted to each patient case's individual attributes.

Therefore, this meta-learning approach allows the ensemble to automatically identify the base models that are most reliable for a given clinical presentation.

To validate this choice, Logistic Regression was compared with majority voting and probability averaging on the SMOTE-balanced training data. It achieved 98.0% accuracy and was selected for its learnable weights, transparency, and adaptability (Singh, 2024).

Table 8: Comparison of Meta-Learner Combination Strategies

Combination Strategy	Accuracy	Precision	Recall	F1-Score
Majority Voting	97.50%	0.9754	0.9750	0.9751
Probability Averaging	98.00%	0.9802	0.9800	0.9800
Logistic Regression (Proposed)	98.00%	0.9802	0.9800	0.9800

A 5-fold stratified cross-validation is used to train the base models with 80% of the data, while generating predictions from the base models to serve as inputs to the meta-learner (for unbiased performance and maintaining class distributions).

4.6 Explainable AI Framework

Interpretability is essential in CDSSs because predictions influence patient treatment and resource allocation, while black-box models reduce clinical trust and accountability. A dual-layer explanation paradigm is therefore employed, providing both model-level and patient-specific explanations.

4.6.1 SHAP for Global Interpretability

When SHAP calculates the importance of each feature (or player) in predicting a response variable, it distributes contributions among all players fairly using Cooperative Game Theory concepts. SHAP quantifies each feature's average marginal contribution across all possible feature combinations via its Shapley value.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (4)$$

where $\mathcal{F}$ represents the complete feature set, $\mathcal{S}$ denotes feature subsets, and f is the model prediction function. This formula satisfies three theoretical guarantees: local accuracy, missingness and consistency. Together these ensure reliable and stable feature attributions.

$$\text{Importance}(i) = \frac{1}{N} \sum_{j=1}^N |\phi(j)| \quad (5)$$

Here, N is the number of test samples and (j) represents the SHAP value for feature i . Global ranking shows which clinical parameters, composite indices, and vital signs most strongly drive predictions across diverse patients. SHAP summary plots illustrate each feature's positive or negative effect and its relative importance across patient subgroups.

4.6.2 LIME for Instance-Level Reasoning

While SHAP captures global behavior, clinical decisions also require case-specific reasoning. LIME addresses this by generating explanations local to each prediction's decision boundary. LIME creates a weighted linear surrogate model in the surrounding area for a particular instance x :

$$\xi(x) = \arg \min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (6)$$

where f is the complex ensemble model, g is the interpretable linear surrogate, $\mathcal{G}$ is the class of interpretable models, π_x defines the local neighborhood weighting around x , $\mathcal{L}$ measures the fidelity of g in approximating f locally, and $\Omega(g)$ is a complexity penalty encouraging simplicity.

5. Result and Analysis

Section 5 presents the experimental results and compares the proposed framework with baseline methods. Table 9 summarizes performance across these methods to evaluate effectiveness in health risk prediction.

5.1 Comprehensive Model Performance Comparison

This section compares the proposed framework against several baseline methods. These include traditional machine learning models, clinical scoring systems, a deep learning model, and standalone tree-based ensembles. The goal is to determine whether the proposed approach offers a measurable advantage over each category.

5.1.1 Baseline model Comparison

The stacking ensemble with SMOTE achieved 98.0% accuracy, outperforming SVM (8.5%), NEWS (14.5%), and XGBoost (0.5%). Performance was consistent

across oversampling methods (Borderline-SMOTE, ADASYN, SMOTE-ENN). McNemar’s test showed significant improvements over baselines ($p < 0.001$ for ML and clinical models, $p < 0.05$ for tree-based learners).

Table 9: Comprehensive Performance Comparison: Baseline Models and Proposed Framework

Category	Model	Acc.	Prec.	Rec.	F1
Traditional ML	Logistic Regression	0.8950	0.8920	0.8950	0.8935
	SVM (RBF kernel)	0.9100	0.9085	0.9100	0.9092
	K-Nearest Neighbors	0.8850	0.8830	0.8850	0.8840
	Decision Tree	0.8700	0.8680	0.8700	0.8690
	Naive Bayes	0.8550	0.8520	0.8550	0.8535
Clinical Scores	NEWS (adapted)	0.8350	0.8310	0.8350	0.8330
	Mews (adapted)	0.8100	0.8065	0.8100	0.8082
Deep Learning	Neural Network (64-32-16)	0.9350	0.9340	0.9350	0.9345
Tree Ensembles	Random Forest (Baseline)	0.9550	0.9560	0.9550	0.9552
	XGBoost (Baseline)	0.9750	0.9753	0.9750	0.9750
	LightGBM(Baseline)	0.9600	0.9610	0.9600	0.9602
Proposed	Stacking	0.9800	0.9802	0.9800	0.9800

5.1.2 Impact of Oversampling Strategies on Ensemble Models

SMOTE-based class balancing frequently improves ensemble performance, as Table 10 demonstrates. SMOTE considerably enhanced robustness, whereas XGBoost and the Stacking ensemble reached 97.50% and 97.00% accuracy under class imbalance. With enhanced precision, recall, and F1-score, the Stacking ensemble achieved 98.00% accuracy, demonstrating its clinical dependability.

Table 10: Classification Performance across Ensemble Models and SMOTE Variants

Sampling	Model	Acc.	Prec.	Rec.	F1
Baseline	Random Forest	0.9550	0.9700	0.9610	0.9700
	XGBoost	0.9750	0.9753	0.9750	0.9700
	LightGBM	0.9600	0.9610	0.9600	0.9602
	Stacking	0.9700	0.9610	0.9700	0.9700
SMOTE	Random Forest	0.9500	0.9505	0.9500	0.9503

	XGBoost	0.9750	0.9753	0.9750	0.9750
	LightGBM	0.9750	0.9753	0.9750	0.9751
	Stacking	0.9800	0.9802	0.9800	0.9800
ADASYN	Random Forest	0.9550	0.9560	0.9550	0.9552
	XGBoost	0.9800	0.9802	0.9800	0.9800
	LightGBM	0.9650	0.9657	0.9650	0.9652
	Stacking	0.9700	0.9704	0.9700	0.9700
Borderline-SMOTE	Random Forest	0.9600	0.9610	0.9600	0.9602
	XGBoost	0.9700	0.9704	0.9700	0.9700
	LightGBM	0.9750	0.9753	0.9750	0.9751
	Stacking	0.9800	0.9802	0.9800	0.9800
SMOTE-ENN	Random Forest	0.9550	0.9560	0.9550	0.9552
	XGBoost	0.9750	0.9753	0.9750	0.9750
	LightGBM	0.9800	0.9802	0.9800	0.9800
	Stacking	0.9750	0.9753	0.9750	0.9751

5.2 Confusion Matrix and Per-Class Analysis

Table 11 shows the confusion matrix with only four misclassifications, resulting in a 2.0% error rate. Most samples were classified correctly across all classes. The high diagonal values indicate strong classification performance. Very few samples were confused with other classes, demonstrating the robustness of the proposed model.

Table 11: Confusion Matrix for SMOTE based Stacking Model

Actual / Predicted	High	Low	Medium	Normal	Accuracy
High (n=56)	55	0	1	0	98.21%
Low (n=51)	0	50	1	0	98.04%
Medium (n=61)	1	0	60	0	98.36%
Normal (n=32)	0	1	0	31	96.88%

5.3 Feature Importance Analysis

SHAP analysis ranked PRI as the most influential feature, followed by MS-EWS and CTC, with the composite indices contributing over half of the total importance in Table 12.

Table 12: Top 10 Features Ranked by SHAP Importance

Rank	Feature	Mean-SHAP	Type
1	PRI	1.0702	Novel Composite
2	Oxygen Saturation	0.6667	Vital Sign
3	MS-EWS	0.6355	Novel Composite
4	Systolic BP	0.3807	Vital Sign
5	CTC	0.3485	Novel Composite
6	Temp × HR Product	0.2812	Interaction
7	Temperature	0.2098	Vital Sign
8	SpO/RR Ratio	0.1721	Interaction
9	Oxy-BP Ratio	0.1301	Interaction
10	RR×HR Interaction	0.0717	Interaction

5.4 Model Explainability

SHAP identified PRI, MS-EWS, and CTC as the key predictors are shown in Figure 4, while LIME confirmed their contribution to high-risk predictions and highlighted borderline misclassified cases.

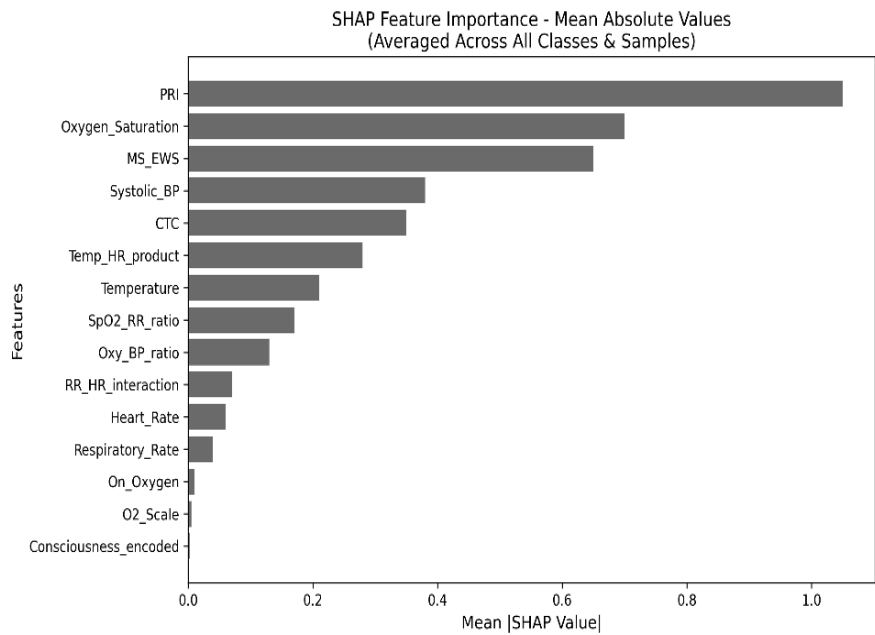


Figure 4: Global Feature Impact Analysis Using SHAP

5.5 Ablation Study

To further evaluate the contribution of each component of the proposed Explainable Multi-System Composite Scoring Framework (EMSCSF), an ablation study was performed by removing selected feature groups and preprocessing components one at a time. The stacking ensemble was retrained for each configuration using the same experimental protocol. The results are presented in Table 13. The "Features" column indicates the number of input features used in each configuration. For the SMOTE ablation, all 15 features were retained, as only the class-balancing strategy was removed. Since each configuration was retrained independently, minor performance differences from the primary experiment are expected because of normal training variability. These differences do not affect the overall conclusions.

Table 13: Ablation Study Results on the Proposed Stacking Ensemble

Configuration	Features	Accuracy	Precision	Recall	F1-score
Full Model (Proposed)	15	0.9750	0.9752	0.9750	0.9750
Without Composite Indices (PRI, MS-EWS, CTC)	12	0.9600	0.9605	0.9600	0.9602
Without Interaction Features	11	0.9700	0.9702	0.9700	0.9701
Without SMOTE Balancing	15	0.9700	0.9704	0.9700	0.9700
Original Vital Signs Only (No Feature Engineering)	8	0.9600	0.9619	0.9600	0.9603

Removing the composite indices (PRI, MS-EWS, and CTC) reduced the accuracy from 97.50% to 96.00%. This was the largest drop after removing an engineered feature group. The result agrees with the SHAP analysis in Section 5.3, where these indices ranked among the most important features. This suggests that the composite indices play an important role in patient health risk prediction.

Overall, the ablation study supports the design of the proposed EMSCSF framework. The composite indices contributed the largest improvement in performance. The interaction features and SMOTE-based balancing also improved the model, although their contributions were smaller. These results are

consistent with the SHAP analysis and support the effectiveness of the proposed framework.

6. Conclusion

This study integrates three novel indices (PRI, MS-EWS, and CTC) with ensemble learning and interaction-based features to present an explainable multi-system composite scoring framework for patient risk prediction. The stacking ensemble achieved 98% test accuracy, outperforming conventional ML models and clinical scores, validated via SMOTE-balanced training and 5-fold cross-validation. PRI ranked highest (mean SHAP = 1.0702), demonstrating the superiority of interaction-based characteristics over individual vital signs. Patient safety was maintained since LIME provided patient-level explanations and all misclassifications happened between neighboring risk levels.

7. Limitations

Although the proposed framework demonstrates promising performance, several limitations should be noted. First, the study used a public dataset comprising only 1000 patients, which may limit generalizability to broader patient populations. However, repeated cross-validation and statistically significant improvements over baseline (McNemar's test, $p < 0.001$) support the framework's robustness. Second, the patient records were treated as static snapshots and did not capture temporal changes in patient condition. Third, only bedside vital signs were included, while laboratory results, imaging, and medication history were excluded. Finally, the dataset originated from a single public source, and SMOTE-generated synthetic samples may not fully represent the physiological variability of high-risk patients.

8. Future Works

This methodology will be strengthened through future multi-center evaluation of larger cohorts ($N > 5000$) with diverse patient demographics. Real-time monitoring will be made possible by integration with dashboards and electronic health records, and temporal modeling may enhance early deterioration identification. Incorporating additional clinical data types and attention-based architectures can further improve the framework. Prospective clinical trials and open-source availability may also strengthen its reliability and usability. These future enhancements are expected to facilitate broader clinical adoption and improve the framework's applicability in real-world healthcare settings.

Acknowledgment

The authors would like to thank the Department of Information and Communication Technology, Bangladesh University of Professionals (BUP), for

providing academic support during this research.

Declaration of Competing Interest

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

Data Availability

The dataset used in this study is publicly available and is described in Section 3. The original source has been cited in the References.

Funding

This research received no external funding.

AI Declaration

AI writing tools were used solely for language refinement and grammar correction. No AI was used for data collection, analysis, or interpretation. All scientific content, methodology, and conclusions were developed and verified entirely by the authors.

References

- Abdullah, S., Khan, M., Ali, R., & Ahmed, F. (2024). Machine learning techniques for mortality prediction in healthcare: A systematic review and meta-analysis. *Journal of Medical Systems*, 48(1), Article 55. <https://doi.org/10.1007/s10916-024-02015-1>
- Ahmad, Z., Rahman, F., & Ali, M. (2024). Explainable artificial intelligence in clinical decision support: SHAP and LIME perspectives. *Expert Systems with Applications*, 238, Article 121890. <https://doi.org/10.1016/j.eswa.2023.121890>
- Aravindkumar, A., Ramadoss, M., Fakhruddin Ahmed, S. A., Sampath, V., & Lakshminarayanan, K. (2026). Explainable AI in healthcare: A systematic review of XAI use cases in imaging, diagnostics, and rehabilitation. *Frontiers in Artificial Intelligence*, 9, Article 1749527. <https://doi.org/10.3389/frai.2026.1749527>
- Beckett, L. C. (2024). *Health risk prediction: Anonymized real data* [Data set]. Kaggle. <https://www.kaggle.com/datasets/ludocielbeckett/health-risk-prediction-anonymized-real-data>. [Access Date: 12 June 2025]

- Chen, M., Liu, H., & Wang, S. (2025). Addressing class imbalance in clinical risk prediction models: Methods and challenges. *Artificial Intelligence in Medicine*, 148, 102–118. <https://doi.org/10.1016/j.artmed.2024.102567>
- Kumar, V., Sharma, S., & Singh, P. (2023). Handling class imbalance in healthcare datasets: Comparative study of SMOTE variants. *IEEE Access*, 11, 98765–98778. <https://doi.org/10.1109/ACCESS.2023.3321456>
- Lee, K., Park, H., Kim, J., Choi, Y., & Oh, S. (2024). Interpreting artificial intelligence models: A systematic review on the application of LIME and SHAP in Alzheimer’s disease detection. *Brain Informatics*, 11(1), Article 9. <https://doi.org/10.1186/s40708-024-00229-3>
- Lee, S., Kim, J., & Park, H. (2024). Interaction features in multi-system physiological modeling: A comprehensive analysis. *BMC Medical Informatics and Decision Making*, 24(1), 45–62. <https://doi.org/10.1186/s12911-024-02345-8>
- Li, H., Chen, Y., Wang, X., & Zhang, J. (2023). Multi-system early warning score development using machine learning for intensive care patients. *Computers in Biology and Medicine*, 157, Article 106605. <https://doi.org/10.1016/j.compbiomed.2023.106605>
- Nova, R. M., Fahim, J. A., & Seema, S. J. (2025). A study of machine learning models with population and swarm based optimization for cardiovascular disease prediction. *Journal of FST*, 3(1), 1–24. <https://doi.org/10.64494/JFST/v3i1/ss/2025/01/1-24>
- Patel, R., Gupta, A., & Mehta, S. (2022). Composite clinical scoring systems for multi-organ risk assessment: Advances and limitations. *Frontiers in Digital Health*, 4, Article 1012345. <https://doi.org/10.3389/fdgth.2022.1012345>
- Rodriguez, C., Lopez, J., & Fernandez, M. (2023). Explainable AI for healthcare decision support: Trends and challenges. *Artificial Intelligence in Medicine*, 136, Article 102667. <https://doi.org/10.1016/j.artmed.2023.102667>
- Royal College of Physicians. (2017). National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. *Royal College of Physicians*.

- Shafi, M. M. S., & Mirza, J. F. (2025). An explainable ensemble learning framework for brain tumour using pretrained models and XAI techniques. *Journal of FST*, 3(1), 125–146. <https://doi.org/10.64494/JFST/v3i1/ms/2025/06/125-146>
- Singh, P., Agarwal, R., & Verma, S. (2024). Evaluation of ensemble learning methods in multi-system health risk prediction. *Computer Methods and Programs in Biomedicine*, 246, Article 107260. <https://doi.org/10.1016/j.cmpb.2024.107260>
- Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., & Featherstone, P. I. (2013). The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4), 465–470. <https://doi.org/10.1016/j.resuscitation.2012.12.016>
- Thompson, R. J., Martinez, A., & Brown, K. (2022). Feature engineering strategies for healthcare predictive analytics. *IEEE Journal of Biomedical and Health Informatics*, 26(8), 3845–3856. <https://doi.org/10.1109/JBHI.2022.3167890>
- Wang, L., Chen, X., & Zhang, Y. (2023). Machine learning approaches for chronic disease risk prediction: A systematic review. *Journal of Medical Systems*, 47(3), 1–15. <https://doi.org/10.1007/s10916-023-01234-5>