A Hybrid Explainable Ensemble Learning Framework for Health Risk Prediction

Hussain AlSalman
Department of Computer Science-College of Computer and Information Sciences,
King Saud University, Riyadh 11543, Saudi Arabia

***Abstract*—Early prediction of patients' health risk is a crucial component of the safe and effective implementation of clinical triage and timely intervention. Health risk data in the real world often tends to be small in size and limited, with class imbalance, making the overall accuracy and transparency of Machine Learning (ML) models insufficient and more difficult to achieve. This paper proposes a hybrid explainable ensemble learning framework for multi-class health risk prediction, which is built based on a stacking architecture with three strong base learners, namely Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). The XGBoost is chosen as the meta-learner due to its ability to learn complex non-linear probability mapping from the base models and provide complementary signals for the prediction. A complete preprocessing pipeline is implemented, covering missing data handling, systematic encoding of categorical variables, and strict separation between training and test sets to ensure unbiased assessment. Experimental results show that the proposed framework achieved an accuracy of 97.5%, which exceeds the accuracy results of individual models, which are 96.0%, 95.5%, and 96.0% for RF, XGBoost, and LightGBM, respectively. Additionally, the proposed framework integrates the predictive performance with the interpretable clinical decision support and transparency using SHapley Additive exPlanations (SHAP) method. The SHAP values are used to provide global and local explanations for revealing the most influential features that drive each prediction.**

***Keywords*—Health risk prediction; hybrid explainable; ensemble learning; meta-learner; LightGBM; SHAP**

I. Introduction

Accurate health risk prediction is an essential prerequisite for safe and efficient clinical decision-making [1]. In time-sensitive decision-making settings, clinicians faced the task of prioritizing patients' cases based on routinely available physiological measurements [2]. Conventional methodologies mostly rely on rule-based thresholds for integrating vital sign parameters and triggering escalation protocols [3]. A striking example is the National Early Warning Score 2 (NEWS2), which is widely implemented and standardized in clinical practice [3]. It explicitly uses oxygen supplementation and alternative calculation schemes to better incorporate respiratory risk into the normal clinical process of patient care [3]. Despite the fact that the Early Warning Score (EWS) is defined by simplicity, transparency, and ease of operationalization, its predictive efficacy differs from patient to patient and clinical environment to clinical environment [4]. Large-scale empirical studies have shown that NEWS2 is not better than its predecessor [4]. The NEWS, across the board, and, in some subgroups or in relation to some clinical endpoints, may in fact be shown to produce inferior predictive performance. Such constraints are the impetus for data-driven methodologies with the ability to unravel complex interrelationships of vital signs and clinical indicators [5].

In parallel, healthcare has seen a rapid growth in the number of machine learning (ML) methods with the goal of triage and deterioration prediction using routinely collected data during the medical examination and diagnosis phase [6]. Evidence from emergency departments' research shows that ML models are capable of exceeding conventional triage systems in predicting clinical outcomes [7]. This evidence presents the standard triage inputs, such as vital signs and basic clinical features, and provides support for the hypothesis that the process of algorithmic learners reduces mis-triage [7].

The reduction is achieved by identifying the interaction between features and non-linearity, which poses difficulties of representation within a deterministic scoring framework [7]. However, the effective use of models requires more than incremental increases in predictive accuracy. Clinical ML models need to be evaluated in realistic clinical operational conditions and should be designed with an adequate degree of interpretability. This process enables auditing, error analysis, and builds clinician trust. Hence, in addition to predictive accuracy, explainability is considered to be a very important criterion for therapeutic artificial intelligence [8]. Numerous high-performing models act as "black boxes" and interpretability layer takes on a particular importance in case of risk prediction [9].

A. Research Goals and Motivation

This study aims to develop an accurate and interpretable framework for multi-class health-risk level prediction from routinely available clinical indicators. The motivation for this work is two-fold. First, score-based systems, like NEWS2, are limited in the extent to which they use hand-crafted thresholds, additive rules, and are interpretable. This limitation may give rise to inconsistent discrimination across settings and endpoints [4].

Second, although traditional ML and single tree-based models provide powerful baseline models for structured prediction tasks [10], its application is sensitive to the sub-groups characteristics and may not be able to leverage the complementary strengths among learners fully. To overcome these shortcomings, an explainable hybrid ensemble framework is proposed to combine multiple base learners namely Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Light

Gradient Boosting Machine (LightGBM). The framework is able to learn an optimized combination policy by training a meta-learning, as is consistent with the theoretical underpinnings of stacked generalization. An important practical goal is to guarantee that performance improvements apply in realistic scenarios with limited and small sized imbalanced datasets. Accordingly, the developed framework focuses not only on overall accuracy but also macro-averaged performance measures.

Finally, to promote clinical transparency and decision-making, the framework is equipped with the SHapley Additive exPlanations (SHAP) explanation method. It can outline the features with the greatest impact at each risk assessment level using computed SHAP values [11]. Collectively, these goals align with a growing consensus that clinical machine learning applications should be both highly performant and interpretable to be credible and effective. In clinical triage and monitoring workflows, rapid risk stratification supports prioritization, escalation, and resource allocation decisions. Accordingly, the proposed model is designed as a decision-support tool that provides a risk-level estimate (Normal/Low/Medium/High) to assist clinicians rather than replacing clinical judgment.

B. Research Contributions

In the existing knowledge of data-driven clinical triage and vital-sign–based health risk classification, this work offers four main contributions, summarized as follows:

- An end-to-end, stacking-based framework is proposed that stacks multiple robust tree-based learners into a unified predictor for generalized, robust multi-class health risk level classification, addressing the problem of limited, small-sized datasets.
- Rather than a simple linear combiner, XGBoost is used as the meta-learner to learn a non-linear combination of base model probability outputs, thereby enabling recognition of cohort-specific integration patterns.
- Both general accuracy metrics and imbalance-aware metrics, such as macro-F1 scores and confusion matrices, are reported to ensure that improvements extend beyond the majority classes and remain meaningful across all risk classes.
- Global and instance-level explainability via SHAP is incorporated into the proposed framework for clinical transparency, audit, and feature-level reasoning.

The remaining part of the paper is organized into five Sections: Section II presents the related work on the considered problem and the analysis methods. Section III explains the research methodology, including the problem definition, the study dataset, building base learners, stacking the hybrid model, and model training and evaluation. Section IV presents the results and discusses the findings to justify the novelty and performance of the proposed framework's model. Section V illustrates the conclusions and future directions of current work.

II. Related Work

Health Risk classification using the rule-based Early Warning Score (EWS) remains in common use because of its easy formulation based on regularly documented vital signs measurements [3]. However, extensive research has shown that these scoring systems do not always improve discriminative performance across a wide range of outcomes and patient subgroups [4]. This limitation drives the movement toward more adaptable, data-driven approaches. Recent studies have continued to explore the use of ML methods for modeling and forecasting health risks, with a strong emphasis on the predictive ability of these systems [12]. These predictive systems have been used to enable intervention at early stages and preemptive decision-making, overcoming the limitations inherent in a conventional thresholding paradigm.

Kone et al. [13] introduced a study of health-risk prediction and prevention using ML methods integrated into protective healthcare pipelines. Multi-modal risk analysis has gained focus, with heterogeneous signals integrated to augment personalization and early alerting in a real-world context [14]. With the same goal and in the context of emergency department settings, triage models based on ML have been shown to perform better in terms of discrimination against clinically important outcomes [15]. Raita et al. [7] built ML models to predict Emergency Department (ED) outcome and compared them with classical triage methods and found that ML models can demonstrate better discrimination using routinely available prediction variables such as demographic data, vital signs, and presenting complaints. Likewise, Goto et al. [16] found a result of improved discrimination and reductions in both under- and over-triage among pediatric ED populations using ML-based prediction models compared to traditional triage tiers. Ensemble learning is one of the current approaches to increasing robustness in structured health data sets, especially when different constituent models exhibit complementary error dynamics [17].

Alternative representations to conventional tabular modeling have been analyzed to enhance usability for predicting health risks. Knowledge-graph-based methodologies have been used to aid risk classification by mining relationships among health concepts to facilitate an organized understanding of interconnected health entities [18]. Variants of deep learning architectures based on multi-layered perceptrons have also been explored for predicting health risk, hence the potential of representation learning in identifying complex patterns in medical data [19]. Workflow and infrastructure are huge factors in the real-world adoption of risk prediction systems. According to empirical investigations, adoption might be limited by insufficient integration with electronic health records (EHRs), which would result in deployment limitations that go beyond model correctness [20].

Explainability has become critical in the deployment of healthcare systems due to the need to explain complex ensemble predictions in risk-stratification situations [21]. Mukilan et al. [22] proposed a hybrid SVM-Naive Bayes ensemble with a combination of LIME and SHAP to achieve transparent air quality and health risk prediction. The authors showed that explanatory layers can be coupled with ensemble learners synergistically in order to increase interpretability [22]. Jayakarthik et al. [23] proposed an approach, called EnvHealthNet, a multimodal model for predicting environmental health risks. The approach follows the general

trend for integrating multiple sources of information. In maternal and antenatal contexts, ML-based risk-level prediction has also been investigated, and more recently, explicit modeling of explainability levels using techniques such as SHAP and LIME has been applied to explain maternal risk prediction. [24, 25]. Across studies and systematic reviews of ML in emergency department triage, the potential of clinical data and techniques to improve both consistency and accuracy in triage decisions is repeatedly highlighted, alongside existing issues such as a lack of datasets and a need for transparent evaluation and reporting [26, 27].

III. Methodology

This section presents the research methodology for developing the proposed framework for health risk prediction. Fig. 1 summarizes the complete workflow of the methodology. In the beginning, the study's dataset is described in detail. Then, the problem statement is formally defined to provide a foundation for the proposed solution.

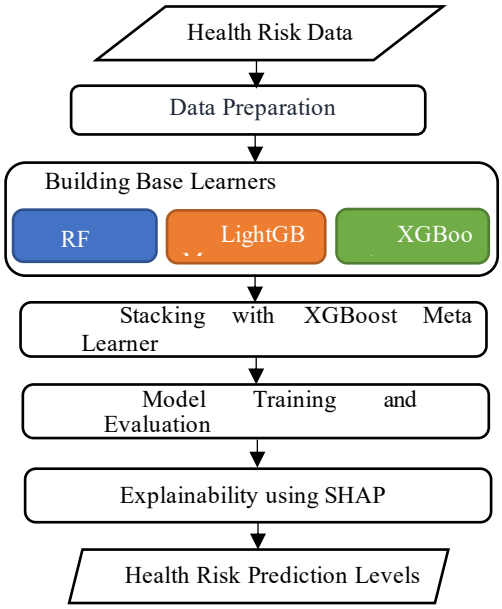

Fig. 1. Flowchart of research methodology.

Afterward, the preprocessing pipeline is described to handle missing values and categorical attributes consistently across both the training and inference phases. The base learners are then introduced, and the stacking ensemble is elaborated using XGBoost as the meta-learner to merge the probabilistic outputs from the base models. Next, the model training and evaluation process is applied to multi-class health risk levels. Finally, an interpretability layer based on SHAP is introduced to provide both global and instance-level explanations of the model decisions.

A. Problem Definition

The problem definition formulates the health risk prediction as a supervised multi-class classification task. Let $X \in R^{n \times d}$ denotes a feature matrix constructed from routinely recorded physiological indicators and clinical attributes. In addition, let $y \in \{1, \ldots, K\}$ represent the ground-truth risk class, in which $K = 4$ corresponds to {High, Low, Medium, Normal}. The goal is to learn a mapping $f: R^d \rightarrow \{1, \ldots, K\}$ that assigns the correct risk level to unseen cases. The risk classes represent ordered levels of severity relevant to real-world triage and monitoring, and the predictive performance should be effective with limited, small datasets, uniform across all classes, and not skewed towards the majority class.

B. Dataset of Study Design

The study uses an anonymized health-risk dataset consisting of variables based on vital signs and related clinical indicators, including metrics for oxygenation and a categorical variable for state of consciousness. The dataset consists of real-world patient data collected from Central Medical College and Hospital, with identifiers removed for privacy protection. It is publicly available on the KAGGLE platform under the name "Health Risk Prediction (Anonymized Real Data)" for conducting studies in health risk multi-class classification [28]. It contains 1,000 anonymized patient records with 9 features and a multi-class target with 4 risk labels: High, Low, Medium, and Normal. The dataset description and variable definitions are provided in Table I, while the class distribution is reported in Table II. We note that Patient_ID is used only for record indexing and is excluded from model training.

TABLE I. Description of Dataset Features

#	Column	Type (Role)	Description
1	Patient_ID	object (Identifier)	An anonymized patient record identifier (used for indexing; not a predictive clinical feature).
2	Respiratory_Rate	Integer (Feature)	Breathing rate (typically breaths per minute); reflects respiratory distress.
3	Oxygen_Saturation	Integer (Feature)	Peripheral oxygen saturation (SpO_2, typically %); indicates oxygenation status.
4	O2_Scale	Integer (Feature)	Oxygen scale/category used for scoring oxygenation (often aligns with clinical early-warning scoring scales).
5	Systolic_BP	Integer (Feature)	Systolic blood pressure (mmHg). It reflects the hemodynamic stability.
6	Heart_Rate	Integer (Feature)	Heart rate (beats per minute); indicates cardiovascular response/stress.
7	Temperature	float (Feature)	Body temperature (°C); indicates fever/hypothermia risk.
8	Consciousness	Object (Feature)	Level of consciousness (e.g., AVPU categories such as A/V/P/U); proxy for neurological status.
9	On_Oxygen	Integer (Feature)	Whether the patient is receiving supplemental oxygen (binary indicator, e.g., 0/1).
10	Risk_Level	Object (Target)	Multi-class health-risk label (Normal / Low / Medium / High) used as the prediction outcome.

TABLE II. DISTRIBUTION OF RISK LEVEL CLASSES

Class Label	Number of Instances
High	279
Low	255
Medium	306
Normal	160

C. Data Splitting and Preprocessing

A hold-out splitting strategy is used to randomly split the dataset into 80:20 training and test subsets to obtain a reliable estimate of generalization performance. The split is performed with a fixed random seed of 42 for reproducibility. The test subset is only for final evaluation and is not used to calibrate preprocessing or train meta-learners. Stratified partitioning is applied to ensure that imbalanced classes can be preserved across splits. This strategy is similar to real-world deployment scenarios in which models must handle unexpected data. Data preprocessing is then implemented as a standard pipeline to apply the same transformations during training and testing.

First, all features are checked for missing values. If numerical features contain missing values, median values are used to replace them. Also, missing values in categorical features are imputed using the mode. After that, categorical variables are converted to numerical values using one-hot encoding. It enables discrete clinical states to be represented without forcing a spurious ordinal relationship. The resulting transformation, z = T(x), is fit only on the training set and then used on both the training and test sets, ensuring that the evaluation is unbiased and that the deployed model will be able to handle new cases with the same preprocessing.

D. Building Base Learners

To capture heterogeneous patterns in structured clinical datasets, the proposed framework trains a diverse set of tree-based base learners at the first hierarchical level. Each base learner $m \in \{1, \dots, M\}$ induces a probability distribution over the K risk classes for a given transformed input $x \in \mathbb{R}^d$, given by $p_m(x) = [p_m(y = 1 \mid x), \dots, p_m(y = K \mid x)] \in \mathbb{R}^d$. In this study, powerful ensemble learners are built for tabular data type, especially bagging tree (RF), as well as gradient boosted decision trees (LightGBM and XGBoost). These ensemble learners can capture such nonlinear effects and feature interactions, which often appear in physiological measurements.

E. Stacking with XGBoost Meta-Learner

The proposed framework combines the base learners through stacked generalization implemented using a stacking classifier. In this configuration, the ensemble operates at the probability level, meaning that the inputs to the second-stage model are not raw features but the class-probability outputs produced by the first-stage learners. For an input instance x, each base learner returns a K-dimensional probability vector over the risk classes. Stacking mechanism constructs a meta-feature vector by concatenating these probability vectors across all base learners:

$$g(x) = concat(p_1(x), p_2(x), \dots, p_M(x)) \in \mathbb{R}^{MK} \quad (1)$$

Where M denotes the number of base learners, which is 3 in our framework. This representation captures complementary confidence patterns across models and provides a compact, information-rich input for second-stage fusion.

For the second stage, the framework uses XGBoost as the meta-learner, specified in the code as the final estimator. The meta-learner learns a nonlinear mapping from the stacked probability space g(x) to the final class probabilities:

$$p_{stack}(y \mid x) = h(g(x)), \hat{y} = arg \max_{k \in \{1,\dots,K\}} p_{stack}(y = k \mid x) \quad (2)$$

Using predicted probability of the stacked method, the meta-learner operates on a concatenation of base-model probability vectors $g(x) = concat(p1(x), \dots, pM(x))$. The XGBoost meta-learner learns a nonlinear fusion, $p_{stack}(y \mid x) = h(g(x))$, with $c\ v = 5$ used to generate unbiased meta-features during training.

To mitigate optimistic bias in meta-learning, the stacking construction specifies that the base-level probabilities used to train the meta-learner are generated via 5-fold cross-validation, which trains the framework to generate meta-features using cross-validated base predictions during fitting. We adopt stacking rather than fixed-weight aggregation because it learns a data-driven fusion rule from base-model probability outputs, enabling adaptive weighting across patient profiles. To validate this choice empirically, the proposed approach will be benchmarked against strong individual baselines using the same split, and both accuracy and macro-averaged metrics will be reported.

F. Model Training and Evaluation

The training phase is performed exclusively on the training subset. During fitting, the base learners are first trained on the training data, and the stacking mechanism internally constructs the meta-learning dataset as specified by the 5-fold cross-validation. Precisely, the training set is split into five folds; for each fold, the base learners are trained on the remaining folds and used to generate out-of-fold class probabilities for the held-out fold. These out-of-fold probabilities become the meta-features for training the XGBoost meta-learner, which learns the final fusion rule from base-model probability signals. After the meta-learner is trained, the base learners are refit on the full training set, enabling the final stacked model to use all available training data when producing predictions for unseen cases.

Evaluation is conducted strictly on the held-out test partition, which is never used during model fitting or meta-feature construction. After training, predicted labels are obtained using the trained pipeline of the base learners to produce $\hat{y}$. Performance is quantified using accuracy to summarize overall correctness and macro-F1-score to provide an imbalance-aware assessment that weights all classes equally. Accuracy, Precision, Recall, and F1- score are computed as:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (3)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (4)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (5)$$

$$F1\text{-}score = \frac{2\times(Precision\times Recall)}{(Precision+Recall)} \quad (6)$$

where, $TP, TN, FP,$ and FN are the number of true positives, true negatives, false positives, and false negatives of all classes.

No resampling like SMOTE, manual threshold shifting, or cost-sensitive reweighting is applied. To ensure fair assessment under imbalance, macro-precision, macro-recall, macro-F1-score, and class-wise results are reported in addition to accuracy. In other words, to confirm that performance is not dominated by the majority class, the macro-F1-score is computed by first calculating precision and recall for each class k, then averaging the per-class F1-scores as:

$$Macro\ F1\text{-}score = \frac{1}{K}\sum_{k=1}^{K} F1\text{-}score_k \quad (7)$$

In addition to scalar metrics, the confusion matrix is computed to inspect class-wise error patterns and identify systematic confusions between adjacent severity levels (e.g., Medium vs. High). The confusion matrix $C \in N^{K\times K}$ is defined as:

$$C_{i,j} = |\{x: y = i \wedge \hat{y} = j\}| \quad (8)$$

The confusion matrix counts how many instances of true class i are predicted as class j. Together, Accuracy, Macro-F1-score, and the confusion matrix provide a comprehensive evaluation: Accuracy captures overall correctness, Macro-F1-score emphasizes balanced performance under class imbalance, and the confusion matrix supports detailed clinical error analysis. In addition, the receiver operating characteristic (ROC) curve will be visualized to show the relationship between the true positive rate and the false positive rate at different decision thresholds.

G. Explainability Using SHAP

The methodology integrates SHAP explanations into the final stacked model, thereby enabling feature-level reasoning and clinical transparency. By decomposing a prediction into a baseline and contributions from various input variables, SHAP provides additive feature attributions. It is used to produce both local explanations and global explanations. Local explanations support the individual risk assignments by indicating the features that increase or decrease the probability of each class. For global explanations, mean absolute SHAP values specify the most important predictors in the ensemble learners group. The SHAP methodology gives coherent and computationally tractable explanations of tree-based models. By making the links between predictions and clinically significant signs, this interpretability layer aids in building trust, detecting unexpected model behavior, and simplifying the clinical review process.

IV. Experimental Results

In this section, the experiments of the proposed framework are conducted and discussed to provide in-depth interpretations of quantitative results as well as explanations for explainability analyses. The evaluation and findings are reported through three experiments. The first experiment presents the confusion matrix, ROC curves, class-specific measurements, and analysis to validate the predictive performance of the developed framework on a held-out test set. The second experiment compares the framework to individual base models to quantify the value of stacking and confirm that improvements stay consistent despite class imbalance. Finally, the third experiment demonstrates the results of interpretability integrated into the proposed framework. The SHAP values of global feature attributions are mapped to the interdependencies between data-level features and provide, at the same time, instance-level explanations for a representative test case. In the first experiment, the proposed hybrid ensemble is evaluated on the test set to demonstrate its performance in multi-class health risk classification. Fig. 2 shows the confusion matrix, which takes a class-by-class breakdown of correct predictions and misclassifications of the four risk classes.

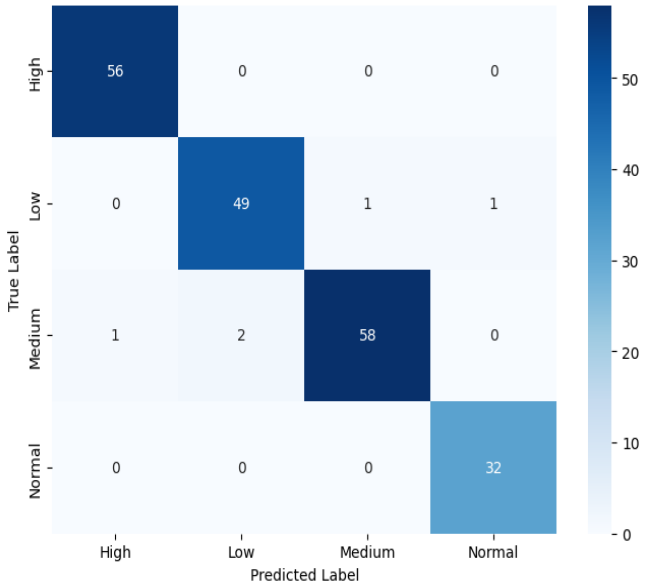

Fig. 2. Confusion matrix of the hybrid ensemble model.

As shown in Fig. 2, the main diagonal of the confusion matrix suggests that the model has a majority fit the observations to the correct risk categories. The corrected classified instances provide evidence of worthy overall discriminative performance. The model maintains high sensitivity for severe cases, which is an essential requirement for triage-oriented prediction. Following the confusion matrix, the results of other evaluation quantities are given in Table III, including the per-class precision, recall, F1-score, and the macro-averaged metrics.

TABLE III. Evaluation Results of the Hybrid Ensemble Model

Class Label	Precision	Recall	F1-score
High	0.9825	1.0000	0.9912
Low	0.9608	0.9608	0.9608
Medium	0.9831	0.9508	0.9667
Normal	0.9697	1.0000	0.9846
Accuracy	0.9750		
Macro avg.	0.9740	0.9779	0.9758
Weighted avg.	0.9751	0.9750	0.9749

As given in Table III, the obtained results support the confusion matrix findings. The model has an accuracy of 0.9750, along with a strong macro-average precision of 0.9740, macro-average recall of 0.9779, and macro-average F1-score of 0.9758. These results shaped the performance to be balanced in all classes. Furthermore, the high-risk class has a recall of 1.0000 and an F1-score of 0.9912, which implies that no severe cases are missed. The medium class shows a slightly lower recall of 0.9508 compared with the other classes, which is consistent with the low number of medium risk, classified as low or high risk, as shown in Fig. 2. However, the model's F1-score for the medium class is relatively high (0.9667), underlining good performance even for the most overlapped class. Fig. 3 shows the AUC-ROC plot of the hybrid ensemble model for each of the classes. The one-versus-rest ROC curve is obtained by treating each of the classes as the positive class, compared to the other classes at different thresholds of decision and the AUC is a threshold independent measure of discriminative capacity.

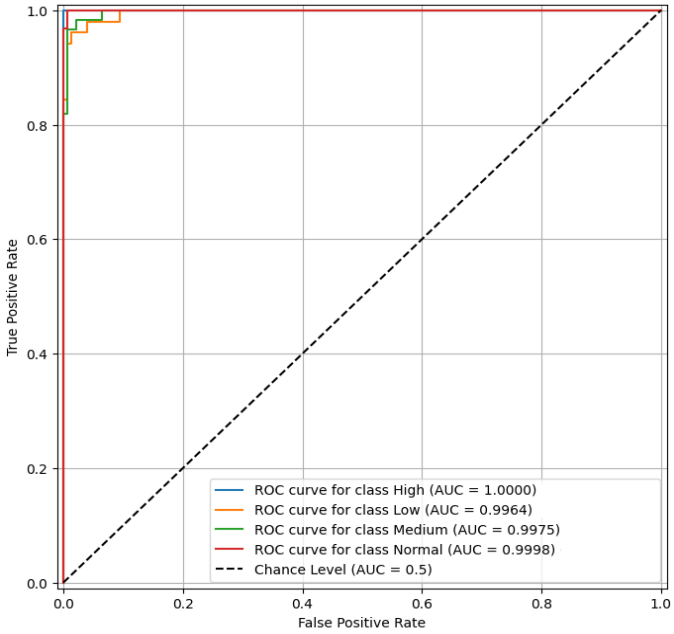

Fig. 3. One-versus-rest ROC curve of the hybrid ensemble model.

As shown in Fig. 3, the ROC curves for all classes in one-versus-rest classification are clustered in the upper-left quadrant, indicating that the classifier's discriminatory ability is respectable across all tested threshold values. The AUC scores are extremely high (1.0000 for High class, 0.9964 for Low class, 0.9975 for Medium class, and 0.9998 for Normal class). These results show that the ensemble model's probabilistic outputs achieve high separability between the risk categories and the remaining classes, supporting reliable threshold-based implementation when different operating points are required.

In the second experiment, the performance comparison between the hybrid ensemble model and single learners is documented in Fig. 4 and Table IV. Fig. 3 shows a comparison of the test set accuracy, and Table IV lists the macro-averaged precision, recall, and F1-score for each model. Fig. 4 shows that the hybrid ensemble model achieves the highest accuracy (0.9750), outperforming RF (0.9600), LightGBM (0.9600), and XGBoost (0.9550). This evidence identifies the framework as an approach for effectively utilizing the complementary strengths between the constituent base learners. The hybrid ensemble model improves generalization beyond that expected from any single robust model.

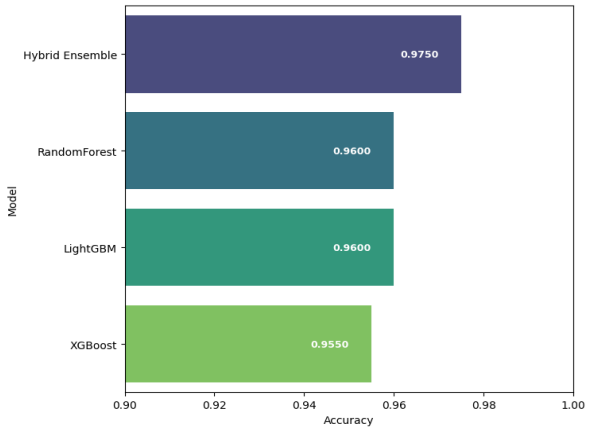

Fig. 4. Comparison of test accuracy between the hybrid ensemble model and individual baseline models.

TABLE IV. COMPARISON OF MACRO-AVERAGED EVALUATION METRICS BETWEEN THE HYBRID ENSEMBLE MODEL AND INDIVIDUAL BASELINE MODELS

Model	Macro F1-Score	Macro Precision	Macro Recall
RF	0.963	0.967	0.960
XGBoost	0.959	0.961	0.956
LightGBM	0.963	0.967	0.960
Hybrid Ensemble	0.976	0.974	0.978

To confirm that the majority class does not affect the improvement, Table IV presents the macro-averaged performance values. The hybrid mode achieved the best results, with a macro F1-score of 0.976, macro recall of 0.978, and macro precision of 0.974. In contrast, the individual models have lower scores. The RF and LightGBM have a macro F1-score of 0.963, whereas the XGBoost has a macro F1-score of 0.959. These results highlight that the hybrid model improved the balanced multi-class performance.

Furthermore, Table IV shows greater reliability for minority or critical classes than for overall accuracy alone. To validate the proposed model's results against the state-of-the-art method, a direct comparison with previously published methods on the same dataset is currently not possible, as no peer-reviewed benchmarks are available. Therefore, the results of the proposed model are compared against strong individual baselines on the same split and report macro-averaged metrics to account for class imbalance.

The third experiment evaluates the transparency of the framework using SHAP analysis. To provide global explanations of the model's predictions, SHAP summary plots are shown for each risk label, as shown in Fig. 5 to 8. These visualizations give each feature its relative importance in the aggregate model output. They show how feature values, whether high or low, influence the predicted probability towards or away from each class. Presentation of such summary plots by the risk

label is especially informative in multi-class settings since the same feature may contribute to divergent risk outcomes. This depends on the range in which its value falls, as well as its interactions with other variables. For example, a low oxygen saturation level may increase the probability of High-risk classification but also decrease the probability of Normal-risk classification.

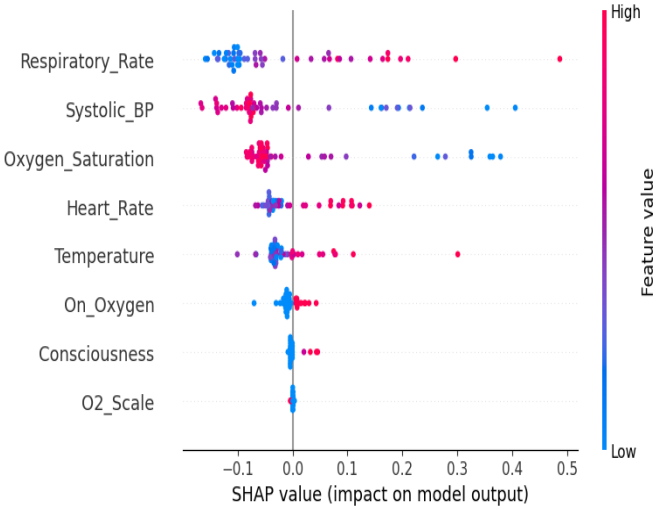

Fig. 5. SHAP summary plot for the high-risk class.

Fig. 5 gives the SHAP summary plot for the High-risk class. The most influential predictors, based on the selected variables, are respiratory rate and systolic blood pressure, followed in importance by oxygen saturation, heart rate, and temperature. This ordering agrees with known clinical expectations, in which respiratory decompensation and hemodynamic instability are the major indicators of high risk. The chromatic distribution further biases the model's output toward the High-risk category by applying pronounced values for features such as a high respiratory rate, low systolic blood pressure, and low oxygen saturation.

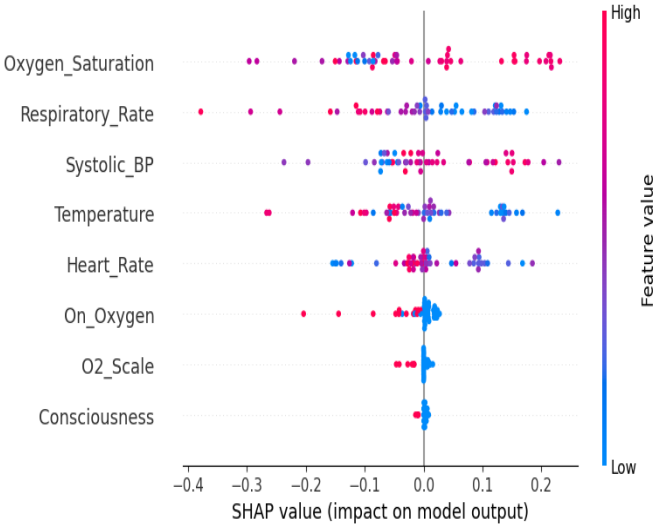

Fig. 6. SHAP summary plot for the low-risk class.

In Fig. 6, the SHAP summary plot for the Low-risk class shows the boundary of the features that influence the model's decision on Low risk, and how those feature value ranges modify the decision. The principal contributors mirror the clinically salient vital signs variables identified for each of the other classes, such as oxygen saturation, respiratory rate, systolic blood pressure and heart rate. However, their effects correspond with an intermediate spectrum of severity. In particular, the feature values reflecting mild deviations from physiology are associated with a higher probability of Low-risk. Those representing stable physiology will move the model's predictions toward Normal, and more obvious abnormal values will steer the model's predictions away from Low-risk toward Medium-risk or High-risk.

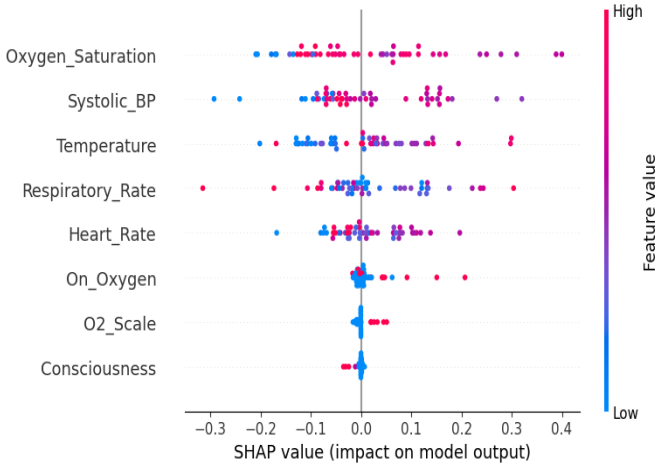

Fig. 7. SHAP summary plot for the Medium risk class.

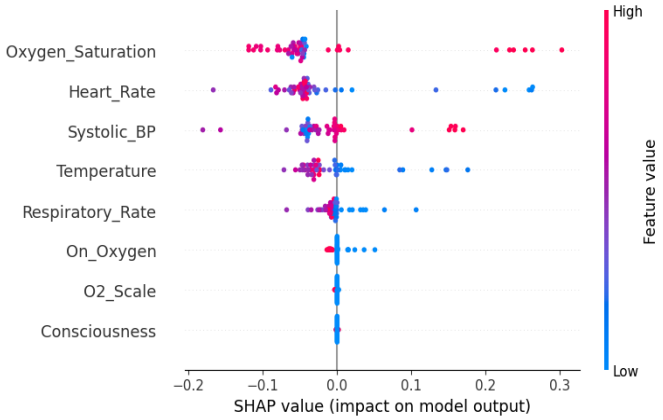

Fig. 8. SHAP summary plot for the Normal risk class.

Similarly, Fig. 7 and 8 represent the Medium-risk and Normal-risk classes, respectively. They show that oxygen saturation continues to have a dominant effect, whereas the contributions of heart rate, temperature, and respiratory rate vary across classes. The global explanations are consistent with the behavior of the confusion matrix. This consistency shows that misclassifications mostly concentrate near the Low and Medium risk border region, where physiological patterns can overlap. Together with Fig. 5, 7, and 8. Fig. 6 completes the class-wise global interpretability analysis and provides evidence that the model's Low-risk decisions are driven by clinically coherent

signals rather than spurious artifacts. For the dominant features, identified by SHAP, Fig. 9 visualizes their relationships using a pairwise scatter distribution plot. The points of the scatter distribution plot are colored by risk label to show class separation and overlap in feature space.

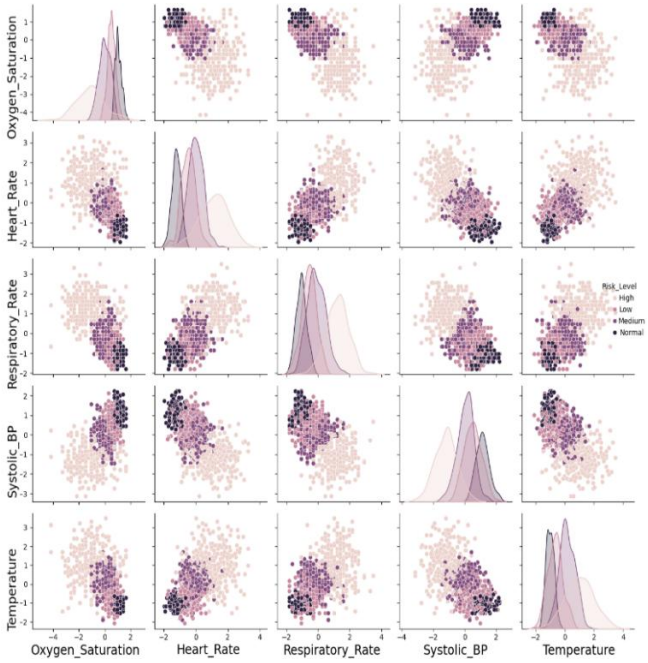

Fig. 9. Visualization of pairplot for the relationships between dominant features.

From Fig. 9, the scatter distribution of risk levels provides an empirical rationale for SHAP-based feature rankings. The visualization shows that the risk classes are in partially-separable regions of the combined feature space. For example, the separation between classes is increased once oxygen saturation is plotted side-by-side with respiratory rate and systolic blood pressure. This separation provides direct support for the prominence of these variables among the leading SHAP contributors. The overlap between classes, Low and Medium risk, provides a visual explanation for the constrained number of misclassifications described in Fig. 2, and the slightly reduced recall for the Medium class described in Table III.

In addition to global SHAP summaries, instance-level explanations are provided for a representative test case. The waterfall plots report feature contributions for each class (Normal, Low, Medium, and High), demonstrating not only why the model supports the Medium decision but also why competing classes are rejected. The local SHAP waterfall explanations for a specific test instance with true label Medium are given in Fig. 10 to 13, reporting class-wise contributions. Fig. 10 indicates that key features, notably systolic blood pressure, heart rate, temperature, and oxygen saturation, contribute negatively to the High class, preventing escalation to High-risk for this case. Fig. 11 highlights that respiratory rate strongly decreases the Low-risk probability, consistent with a physiological profile more concerning than the low class. Fig. 12 shows strong positive contributions from respiratory rate, systolic blood pressure, heart rate, temperature, and oxygen saturation, collectively increasing the probability toward the Medium class, which explains the final predicted outcome. Fig. 13 shows negative contributions from multiple vital-sign features, reducing the Normal class probability and justifying why the model rejects Normal for this instance.

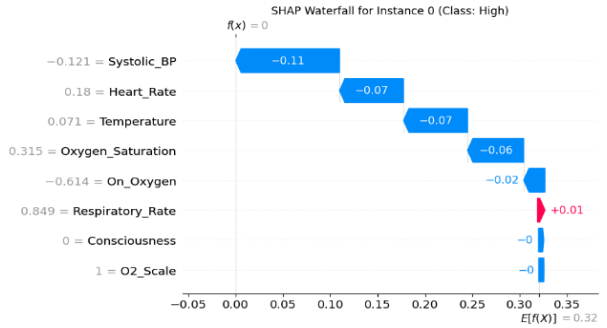

Fig. 10. SHAP waterfall plot for a specific test instance, with features influencing its probability of being classified as High risk.

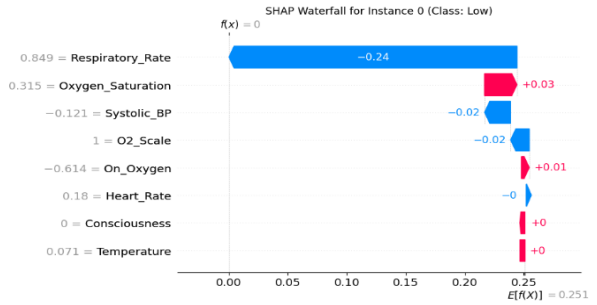

Fig. 11. SHAP waterfall plot for a specific test instance, with features influencing its probability of being classified as Low risk.

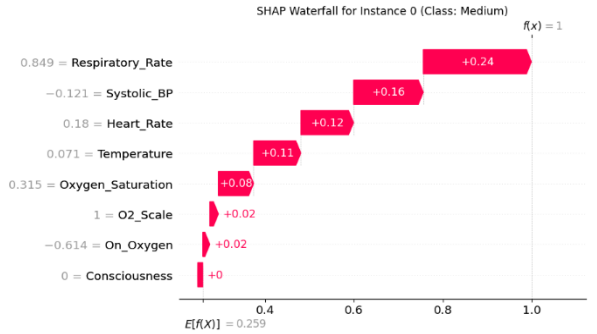

Fig. 12. SHAP waterfall plot for a specific test instance, with features influencing its probability of being classified as Medium risk.

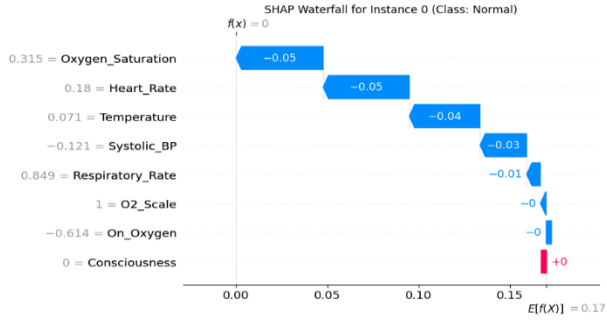

Fig. 13. SHAP waterfall plot for a specific test instance, with features influencing its probability of being classified as Normal risk.

Overall, these local explanations support the fact that the decision of the model is not only accurate but also traceable. As the same clinically relevant features identified at the global level in Fig. 5 to 8, Fig. 10 to 13 offer case-specific reasoning for providing cohesive transparency for clinical interpretation. These case-level attributions can support clinician review by identifying whether respiratory compromise (e.g., elevated respiratory rate, reduced oxygen saturation) or hemodynamic signals (e.g., lower systolic blood pressure) are the primary drivers of risk escalation. Across all experiments, the framework demonstrates the following advantages:

- Strong multi-class predictive performance with accuracy equal to 97.50%, and high per-class discrimination as demonstrated in Fig. 2, Fig. 3, and Table III.
- Consistent improvement over individual models in both accuracy and imbalance-aware macro metrics, as shown in Fig. 4 and Table IV.
- Clinically successful explanations through SHAP at both global and local levels, as visualized in Fig. 5 to 13.

These advantages support the proposed framework as an accurate and explainable framework for practical health-risk classification. For deployment, integration with EHR/HIS requires consistent data availability at the point of care, robust handling of missingness, latency-aware execution, and monitoring for dataset shifts over time. From an ethics and accountability perspective, the proposed model can be used as human-in-the-loop decision support, with auditability and explanation logs, subgroup performance checks, clinician override, and prospective validation before clinical adoption.

V. Conclusions and Future Work

This study proposed a hybrid explainable ensemble learning framework for multi-class health risk classification. The proposed framework is based on routinely collected clinical indicators, such as respiratory rate, oxygen saturation, blood pressure, heart rate, temperature, oxygen-related variables, and consciousness status. The hybrid classification model in the framework is built using the stacking ensemble concept to address a small, limited dataset and a class-imbalanced problem. The hybrid model generates synthesized complementary decision signals from multiple tree-based decision learners. It produces a singular, robust health risk classifier across four clinically significant classes: High, Low, Medium, and Normal.

Experimental results have shown that the framework demonstrates respectable predictive performance on a held-out test set, achieving 97.50% accuracy. More importantly, the model maintains high, balanced performance despite class imbalance, as shown by a high macro-level F1-score and valuable macro precision and recall values. The confusion matrix also shows that misclassifications are rare and mainly occur between adjacent severity levels, such as Low versus Medium. This outcome facilitates expected clinical overlap in borderline physiological presentations and is generally preferable to large severity jumps.

Beyond the predictive performance, the investigation highlighted the value of transparency and clinical interpretability. The SHAP-based explainability assessment provided both global and local insights into how the model operates. Global SHAP summary plots showed that variables of clinical importance, and particularly, variables reflecting respiratory compromise and physiological instability, such as respiratory rate, oxygen saturation, and systolic blood pressure, are predominant in risk discrimination. This interpretability provides evidence of the clinically plausible nature of the inferred decision function. The pairwise scatter distribution is a useful complement to these global explanations that revealed feature space configurations consistent with class separation patterns. The local SHAP waterfall visualizations provided patient-level reasoning for a representative, medium-risk case that captures all competing labels. Collectively, this body of evidence suggests that the framework has not only accuracy but also auditability. These interpretable justifications can be used to build clinician confidence and to provide an analysis of error in authentic decision-support contexts.

Because no peer-reviewed studies have reported machine-learning benchmarks on the dataset used in this study, this work provides the first academic baseline and a reproducible reference for future comparisons. A limitation of this work is that results are reported under a single hold-out split. Future work will include repeated runs with multiple random seeds and bootstrap confidence intervals to quantify performance uncertainty. Although this study adopts a stacking-based ensemble with an XGBoost meta-learner as the final configuration, other ensemble variants may be considered depending on the target operating requirements. For example, soft voting or weighted averaging can provide simpler aggregation. At the same time, alternative meta-learners, such as logistic regression or LightGBM, and different base-learner combinations may yield different trade-offs between performance, stability, and computational cost. A systematic comparison of these alternative configurations under the same data split and evaluation protocol is left for future work.

Moreover, future work will focus on validating the framework in external hospitals and care environments to quantify robustness to dataset shift and investigate subgroup performance across datasets. In addition, the probability calibration and the optimization of operating thresholds will be investigated to better align model outputs with the requirements of clinical decision-making. A second direction is to move the framework beyond its role as a snapshot prediction system towards use as more actionable decision support. This includes incorporating temporal trends from repeated vital sign measurements to enable earlier detection of deterioration and enrich features with other EHR signals. Finally, the SHAP layer will be improved with stability checks and clinician-facing explanation summaries, and then pilot or prospective evaluation to explore usability and impact within actual clinical workflows.

Acknowledgment

The author is grateful to the Deanship of Scientific Research at King Saud University's College of Computer and Information Sciences (CCIS) for funding this research.

References

[1] N. L. Edoh, V. M. Chigboh, S. J. C. Zouo, J. Olamijuwon, and Dentistry, "Improving healthcare decision-making with predictive analytics: A conceptual approach to patient risk assessment and care optimization,"

International Journal of Scholarly Research in Medicine, vol. 3, no. 2, pp. 1-10, 2024.

[2] L. Sandman and J. Liliemark, "Should severity assessments in healthcare priority setting be risk-and time-sensitive?," Health Care Analysis, vol. 31, no. 3, pp. 169-185, 2023.

[3] G. B. Smith, O. C. Redfern, M. A. Pimentel, S. Gerry, G. S. Collins, J. Malycha, D. Prytherch, P. E. Schmidt, and P. J. Watkinson, "The national early warning score 2 (NEWS2)," Clinical Medicine, vol. 19, no. 3, p. 260, 2019.

[4] M. A. Pimentel, O. C. Redfern, S. Gerry, G. S. Collins, J. Malycha, D. Prytherch, P. E. Schmidt, G. B. Smith, and P. J. Watkinson, "A comparison of the ability of the National Early Warning Score and the National Early Warning Score 2 to identify patients at risk of in-hospital mortality: A multi-centre database study," Resuscitation, vol. 134, pp. 147-156, 2019.

[5] N. S. Mosavi and M. F. Santos, "Enhancing clinical decision support for precision medicine: A data-driven approach," in Informatics, 2024, vol. 11, no. 3, p. 68: MDPI.

[6] R. Sánchez-Salmerón, J. L. Gómez-Urquiza, L. Albendín-García, M. Correa-Rodríguez, M. B. Martos-Cabrera, A. Velando-Soriano, and N. Suleiman-Martos, "Machine learning methods applied to triage in emergency services: A systematic review," International Emergency Nursing, vol. 60, p. 101109, 2022.

[7] Y. Raita, T. Goto, M. K. Faridi, D. F. Brown, C. A. Camargo Jr, and K. Hasegawa, "Emergency department triage prediction of clinical outcomes using machine learning models," Critical care, vol. 23, no. 1, p. 64, 2019.

[8] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and P. Q. Consortium, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," BMC medical informatics decision making, vol. 20, no. 1, p. 310, 2020.

[9] S. Niu, Q. Yin, J. Ma, Y. Song, Y. Xu, L. Bai, W. Pan, and X. Yang, "Enhancing healthcare decision support through explainable AI models for risk prediction," Decision Support Systems, vol. 181, p. 114228, 2024.

[10] M. Banerjee, E. Reynolds, H. B. Andersson, and B. K. Nallamothu, "Tree-based analysis: a practical approach to create clinical decision-making tools," Circulation: Cardiovascular Quality Outcomes, vol. 12, no. 5, p. e004879, 2019.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in neural information processing systems, vol. 30, 2017.

[12] F. Rahimian, G. Salimi-Khorshidi, A. H. Payberah, J. Tran, R. Ayala Solares, F. Raimondi, M. Nazarzadeh, D. Canoy, and K. Rahimi, "Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records," PLoS medicine, vol. 15, no. 11, p. e1002695, 2018.

[13] N. Kone, S. Singh, M. Noor, and N. Ranjan, "Health Risk Prediction and Prevention Using Machine Learning," in International Conference on AI Systems and Sustainable Technologies, 2025, pp. 101-113: Springer.

[14] A. B. Abraham, R. K. JS, N. V. Babu, and R. G. Nancy, "Multi-Modal AI Personal Health Risk Prediction and Early Alert System," in 2025 4th International Conference on Automation, Computing and Renewable Systems (ICACRS), 2025, pp. 775-782: IEEE.

[15] G. Feretzakis, A. Sakagianni, A. Anastasiou, I. Kapogianni, R. Tsoni, C. Koufopoulou, D. Karapiperis, V. Kaldis, D. Kalles, and V. S. Verykios, "Machine learning in medical triage: A predictive model for emergency department disposition," Applied Sciences, vol. 14, no. 15, p. 6623, 2024.

[16] T. Goto, C. A. Camargo Jr, M. K. Faridi, R. J. Freishtat, and K. Hasegawa, "Machine learning–based prediction of clinical outcomes for children during emergency department triage," JAMA network open, vol. 2, no. 1, p. e186937, 2019.

[17] L. R. Namamula and D. Chaytor, "Effective ensemble learning approach for large-scale medical data analytics," International Journal of System Assurance Engineering Management, vol. 15, no. 1, pp. 13-20, 2024.

[18] X. Tao, T. Pham, J. Zhang, J. Yong, W. P. Goh, W. Zhang, and Y. Cai, "Mining health knowledge graph for health risk prediction," World Wide Web, vol. 23, no. 4, pp. 2341-2362, 2020.

[19] T. Bikku, "Multi-layered deep learning perceptron approach for health risk prediction," Journal of Big Data, vol. 7, no. 1, p. 50, 2020.

[20] V. Sharma, I. Ali, S. van der Veer, G. Martin, J. Ainsworth, and T. Augustine, "Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records," BMJ health care informatics, vol. 28, no. 1, p. e100253, 2021.

[21] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, and R. Sharma, "Explainable AI for healthcare 5.0: opportunities and challenges," IEEE Access, vol. 10, pp. 84486-84517, 2022.

[22] A. Mukilan, K. Abirami, K. Keerthishree, M. Srivani, and V. Vanitha, "Hybrid SVM-Naive Bayes Ensemble with LIME and SHAP for Transparent Air Quality and Health Risk Prediction," in 2025 13th International Conference on Intelligent Systems and Embedded Design (ISED), 2025, pp. 364-370: IEEE.

[23] R. Jayakarthik, D. Gopinath, M. A. Begum, A. D. Fuladi, A. Balaram, and C. Raja, "EnvHealthNet: A Multi-Modal Machine Learning Model for Commercial Environmental Health Risk Prediction," in 2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN), 2025, pp. 1121-1128: IEEE.

[24] F. Ahamad, "Detection of Antenatal Health Risk Level Using Machine Learning," in 2024 International Conference on Control, Computing, Communication and Materials (ICCCCM), 2024, pp. 426-429: IEEE.

[25] A. Rahman and M. G. R. Alam, "Explainable AI based maternal health risk prediction using machine learning and deep learning," in 2023 IEEE World AI IoT Congress (AIIoT), 2023, pp. 0013-0018: IEEE.

[26] J. S. Hinson, D. A. Martinez, S. Cabral, K. George, M. Whalen, B. Hansoti, and S. Levin, "Triage performance in emergency medicine: a systematic review," Annals of emergency medicine, vol. 74, no. 1, pp. 140-152, 2019.

[27] A. Naemi, T. Schmidt, M. Mansourvar, M. Naghavi-Behzad, A. Ebrahimi, and U. K. Wiil, "Machine learning techniques for mortality prediction in emergency departments: a systematic review," BMJ open, vol. 11, no. 11, p. e052663, 2021.

[28] R. Alam, "Health Risk Prediction (Anonymized Real Data)," URL https://www.kaggle.com/datasets/ludocielbeckett/health-risk-prediction-anonymized-real-data/data, [Last Access: 25-01-2026