Methodologies for Systematic Evaluation and Targeted Mitigation of Deficiencies in Critical Machine Learning Models

Tanmoy Sarkar Pias

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Daphne (Danfeng) Yao, Chair
T. M. Murali
Pearl Chiu
Ismini Lourentzou
Shalmali Joshi

July 28, 2025
Blacksburg, Virginia

Keywords: AI Trustworthiness, Responsiveness, Knowledge guided ML, Custom Loss, Healthcare

Methodologies for Systematic Evaluation and Targeted Mitigation of Deficiencies in Critical Machine Learning Models

Tanmoy Sarkar Pias

(ABSTRACT)

Despite the growing use of machine learning in healthcare, critical challenges remain unaddressed—models often fail to respond appropriately to life-threatening conditions, exhibit poor generalizability in real-world clinical settings, and show unequal performance across patient subgroups. These limitations compromise the reliability, safety, and equity of AI-driven decision-making, especially in high-stakes environments like intensive care. In this work, we outline a comprehensive evaluation and mitigation strategy to address both responsiveness and fairness shortcomings.. We develop testing approaches to systematically assess models' ability to respond to serious medical emergencies. Using generated test cases, we found that statistical machine-learning models trained solely from patient data are grossly insufficient and have many dangerous blind spots. Specifically, we identified serious deficiencies in the models' responsiveness, i.e., the inability to recognize severely impaired medical conditions or rapidly deteriorating health. For in-hospital mortality prediction, the models tested using our synthesized cases fail to recognize 66% of the test cases involving injuries. In some instances, the models fail to generate adequate mortality risk scores for all test cases. We also applied our testing methods to assess the responsiveness of 5-year breast and lung cancer prediction models and identified similar kinds of deficiencies. To address the low responsiveness of machine learning models to critical health conditions, we integrated domain knowledge into the modeling framework using two complementary strategies: (i) a custom loss function that penalizes violations of medical constraints, and (ii) a rule-based decision

tree derived from clinical knowledge, aggregated with a data-driven model. The resulting knowledge-guided models demonstrated notable improvements in performance, particularly under critical scenarios. For instance, recall improved by 7% on the full glucose test set and by 27% for critically high glucose cases, achieving 94–99% accuracy in detecting patients with severely abnormal glucose levels. Similar trends were observed for other vital signs. Moreover, the decision tree-based hybrid model enhanced early sepsis detection accuracy by 4%, underscoring the benefit of combining clinical knowledge with statistical learning for high-stakes medical applications. In addition, we address a bias problem we identified in models predicting type 2 diabetes, which disproportionately impacts younger adults—a growing segment of diabetes patients. In this research, we identify this deficiency in traditional machine learning models and propose an algorithm to mitigate the bias towards the young population when predicting diabetes. Deviating from the traditional concept of one-model-fits-all, we train customized machine-learning models for each age group. Our proposed solution consistently improves recall of diabetes class by 26% to 40% in the young age group (30-44). Moreover, our technique outperforms 7 commonly used whole-group sampling techniques such as random oversampling, SMOTE, and AdaSyns techniques by at least 36% in terms of diabetes recall in the young age group.

Methodologies for Systematic Evaluation and Targeted Mitigation of Deficiencies in Critical Machine Learning Models

Tanmoy Sarkar Pias

(GENERAL AUDIENCE ABSTRACT)

Despite the growing use of machine learning in healthcare, critical challenges remain unaddressed—models often fail to respond appropriately to life-threatening conditions, exhibit poor generalizability in real-world clinical settings, and show unequal performance across patient subgroups. These limitations compromise the reliability, safety, and equity of AI-driven decision-making, especially in high-stakes environments like intensive care. In this work, we outline a comprehensive evaluation and mitigation strategy to address both responsiveness and fairness shortcomings. In this research, we develop new methods to test ML models under critical health scenarios. Our findings reveal that current models, trained solely on patient data, have significant blind spots; many fail to recognize severe conditions, accurately predict mortality, or assess injury-related risks. For instance, in our tests, the models missed 66% of injury-related cases and often provided inadequate risk scores for patients who were actually at high risk. We observed similar limitations in models predicting long-term survival for breast and lung cancer, highlighting widespread responsiveness issues in current healthcare ML tools. To address the low responsiveness of machine learning models to critical health conditions, we integrated domain knowledge into the modeling framework using two complementary strategies: (i) a custom loss function that penalizes violations of medical constraints, and (ii) a rule-based decision tree derived from clinical knowledge, aggregated with a data-driven model. The resulting knowledge-guided models demonstrated notable improvements in performance, particularly under critical scenarios. For instance,

recall improved by 7% on the full glucose test set and by 27% for critically high glucose cases, achieving 94–99% accuracy in detecting patients with severely abnormal glucose levels. Similar trends were observed for other vital signs. Moreover, the decision tree-based hybrid model enhanced early sepsis detection accuracy by 4%, underscoring the benefit of combining clinical knowledge with statistical learning for high-stakes medical applications. In addition, we address a bias problem we identified in models predicting type 2 diabetes, which disproportionately impacts younger adults—a growing segment of diabetes patients. Many traditional ML models exhibit "digital ageism," or a tendency to overlook diabetes risk in younger populations. To counteract this, we designed age-specific models that improved detection accuracy by 26-40% for younger adults (ages 30-44) compared to conventional methods. Our approach also outperformed common techniques by at least 36% in recall for young adults, providing a more equitable solution for diabetes risk prediction.

Dedication

To my parents (Pran Bandhu Sarkar and Krishna Rani Mondal), brother (Mrinmoy Sarkar Anto), and my wife (Anindita Datta Prothoma) .

Acknowledgments

Pursuing my Ph.D. has been an intellectually transformative and personally humbling journey. It would not have been possible without the support, encouragement, and guidance of many individuals to whom I am deeply grateful.

First and foremost, I would like to express my deepest appreciation to my advisor and committee chair, Dr. Danfeng (Daphne) Yao. Her unwavering support, intellectual generosity, and thoughtful mentorship shaped my growth as a researcher and helped me develop the confidence to tackle meaningful and challenging problems. Her balanced and inspiring leadership has been a guiding light throughout my Ph.D. journey.

I would also like to sincerely thank my committee members, Dr. T. M. Murali, Dr. Pearl Chiu, Dr. Ismini Lourentzou, and Dr. Shalmali Joshi, for their valuable feedback, encouragement, and the diverse perspectives they brought to my research. Their insights enriched this work and contributed to its depth and rigor. I am grateful to Dr. Charles Nemeroff, Dr. Xinwei Deng, Dr. Sharmin Afrose, Moon Das Tuli, and Ipsita Hamid Trisha for their contributions and support in shaping the ideas that led to Chapter 1. Their interdisciplinary expertise and clinical perspectives helped ground my research in real-world impact.

This journey would not have been possible without the unwavering love and sacrifices of my family. Leaving my parents and home country to embark on this long academic path was one of the hardest decisions of my life. I am immensely thankful to my parents for trusting my decisions, standing by me through every phase, and constantly uplifting me with their words of encouragement. I am also thankful to my beloved wife for her continuous support. To my dearest friends, thank you for filling this journey with warmth, laughter, and moments of joy that helped me endure the most difficult times. Your companionship and empathy have

made a lasting difference in my life.

Above all, I thank God for guiding me throughout this journey, giving me strength during times of uncertainty, and helping me become wiser, more observant, and resilient.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The growing adoption of machine learning (ML) and artificial intelligence (AI) in healthcare holds transformative potential, offering predictive capabilities that can aid clinicians in managing complex conditions and improving patient outcomes. For instance, ML-based models in intensive care units (ICUs) can alert physicians to rapidly deteriorating conditions, thereby enhancing early intervention for patients in critical states. However, this technology comes with substantial challenges, particularly in ensuring that ML models accurately recognize high-risk conditions and do not inadvertently introduce or exacerbate biases. Moreover, the model should be fair Comprehensive model testing and performance assessment are essential, as mispredictions, especially false negatives, could have life-threatening consequences. Despite the promise of ML in healthcare, current models trained solely on patient data often show deficiencies, such as an inability to recognize severely impaired health conditions or rapidly deteriorating patients, leading to blind spots in mortality predictions.

In clinical outcome prediction, machine learning models often struggle to generate reliable responses in clinically critical regions of feature space, where even small deviations in physiological variables can imply significant changes in patient risk. To address this issue, we incorporate domain knowledge into both the loss function and the input feature space to guide model learning toward more clinically meaningful behavior. Our approach introduces knowledge-guided (KG) loss functions that enforce domain-consistent risk gradients around known critical thresholds for vital signs such as glucose, respiratory rate, and systolic blood

pressure. We further explore the integration of domain knowledge as input features and evaluate its interaction with KG loss functions. Experimental results demonstrate that the KG loss significantly improves model responsiveness, particularly in high-risk regions, while varying the KG loss coefficient allows flexible control over the strength of domain enforcement. Although the integration of domain knowledge features alone has limited effect, combining them with the KG loss leads to further performance gains. Both LSTM and Transformer models show substantial increases in recall for severe cases, confirming the effectiveness of our approach in enhancing clinical sensitivity without compromising general performance.

The critical need for fairness and accuracy extends beyond mortality prediction. With diseases like type 2 diabetes increasingly affecting younger populations, traditional ML models face further limitations in identifying high-risk individuals within minority subgroups. Most existing models struggle to diagnose diabetes effectively in young adults, reflecting a broader issue of "digital ageism." This bias is evident in several machine learning models, where younger populations are frequently overlooked in favor of older, more commonly represented age groups. Additionally, imbalances in healthcare data can compound these biases, leading to disparities in predictive accuracy across demographics, including age, gender, and ethnicity. For diabetes prediction, generalized metrics such as AUROC may mask performance gaps, concealing poor recall rates in high-risk, minority age groups.

This work addresses these critical gaps by focusing on two main objectives. First, we develop and implement systematic testing methods for mortality prediction models to evaluate their responsiveness to critical health conditions, ensuring models can identify rapidly deteriorating patients in ICUs and handle complex, time-series data. Also, we test the fairness of the machine learning model. Second, we propose targeted solutions to mitigate the responsiveness issues as well as the fairness issues. These two approaches offer a comprehensive framework to enhance the trustworthiness and fairness of ML in clinical environments, ad-

dressing both model responsiveness and demographic inclusivity. This work thereby contributes essential methodologies and insights to the field, promoting AI-driven healthcare solutions that are both effective and equitable.

1.1 Contribution

Our work makes three key contributions toward improving the safety and reliability of machine learning models in healthcare.

1.1.1 Contribution to evaluating the responsiveness of ML models to critical patient condition

First, we present a rigorous testing framework designed to evaluate how well models respond to critical patient conditions. This framework goes beyond traditional accuracy metrics by simulating high-stakes medical scenarios, revealing important blind spots, deficiencies, and failure modes that would otherwise remain hidden. For instance, we find that many models trained solely on data tend to underreact or completely miss severe health events, which could have serious consequences in clinical practice.

- Development of Systematic Test Case Generation: We create systematic approaches to generate new, high-risk test cases that extend beyond the original dataset, improving the assessment of machine learning model responsiveness in clinical scenarios. These generated cases are either absent or underrepresented in training datasets, exposing potential blind spots in ML model predictions.
- Incorporation of Domain Expertise: Our test case generation process is informed by domain knowledge and input from medical experts, ensuring that the generated cases

reflect realistic clinical risks and conditions. Through interviews with medical experts, we validate estimated risks for some generated test cases, ensuring the relevance and accuracy of our testing approach.

- Evaluation of Model Responsiveness: We conduct experiments involving binary classification tasks, such as in-hospital mortality prediction and 5-year cancer survivability prognosis, to evaluate ML models' responsiveness to time-series data and critical health events. Our work identifies critical deficiencies in current machine learning models, particularly their limited responsiveness to severe health conditions, underscoring the need for improved trustworthiness in digital health applications.

1.1.2 Contribution to Improve Responsiveness of ML Models for Critical Conditions

Second, to address these responsiveness gaps, we propose a novel approach that combines domain knowledge with data-driven learning. We incorporate medical expertise into the model in two complementary ways: through a custom loss function that penalizes clinically implausible predictions and through decision trees that encode medical rules and are integrated with statistical models. These knowledge-guided methods significantly improve the model's ability to detect and respond to critical conditions, particularly for patients with extreme vital signs or early signs of deterioration.

- We integrate domain knowledge into the model to improve its responsiveness in clinically critical ranges. This integration allows the model to better reflect medical expectations, particularly under extreme physiological conditions.

- We introduce a knowledge-guided loss function that explicitly aligns model sensitivity with domain-informed critical zones. This approach is systematically evaluated on an

in-hospital mortality prediction task, revealing improved risk responsiveness without compromising baseline accuracy.

- We design a hybrid system by embedding a domain knowledge–based decision tree into a data-driven machine learning model. This setup is tested on the sepsis prediction task and shows improved alignment with clinical thresholds while maintaining competitive performance.

1.1.3 Contribution to fair and accurate diabetes prediction model for young population

Third, we examine the fairness of these models, particularly their performance across different age groups. We find that many conventional models disproportionately fail to identify early signs of disease in younger adults, an issue we refer to as digital ageism. To address this, we develop age-aware models that offer more accurate and equitable predictions for underrepresented and at-risk subpopulations.

- We address a critical gap in the literature by developing a machine learning approach specifically designed for precision type 2 diabetes (T2D) diagnosis in a young adult population (30-44 years old). This is a new study to target this specific age group for T2D diagnosis using machine learning.
- We show that multiple machine learning models and a number of sampling techniques (SMOTE, random sampling, etc.) fail to achieve fair performance in terms of detecting T2D in young adults.
- We propose a bias correction technique specifically for improving T2D diagnosis in young adults. This demonstrates the effectiveness of subgroup-focused bias correction,

promoting fairer and more accurate machine learning models in healthcare settings.

Together, these three contributions—rigorous responsiveness testing, domain knowledge integration, and fairness-aware modeling—offer a comprehensive path toward more trustworthy and clinically useful machine learning systems.

1.2 Organization

This document is organized as follows. Chapter 2 Reviews the related literature, providing a comprehensive overview of current methodologies and challenges in machine learning applications within healthcare. Chapter 3 Demonstrates our proposed framework for evaluating the responsiveness of machine learning models, focusing on their ability to recognize and respond to critical health conditions. Chapter 4 Outlines our mitigation techniques aimed at improving model responsiveness for critical cases using domain knowledge. Chapter 5 Discusses biases present in models predicting type 2 diabetes among young people, highlighting the impact of digital ageism and evaluating approaches to mitigate this issue. Finally we conclude this dissertation in chapter 6.

Chapter 2

Review of Literature

In this chapter, we explore the literature to find out existing solutions and their limitations. Here we also attempt to position our work within the literature space filling the gap.

2.1 Evaluating Responsiveness of Machine Learning

Most current machine learning models are tested using datasets drawn from the same distribution as their training data, resulting in overly optimistic performance metrics that may not reflect real-world conditions. Yuchi et al. [1] introduced two techniques for evaluating neural networks in the context of automated vehicles: (i) estimating neuron coverage, or the proportion of neurons activated during predictions, and (ii) augmenting input images with predefined modifications to evaluate model robustness. Neuron coverage aims to assess if the network is utilizing its full capacity or if certain neurons remain inactive. However, the hypothesis that unactivated neurons could lead to mispredictions lacks both mathematical and empirical validation. Additionally, augmenting images using handcrafted modifications—such as adding blur, rain effects, or adjusting lighting—can produce varied samples, but this approach has significant limitations. Handcrafted augmentations cannot account for the full range of potential input scenarios, making comprehensive testing impractical and insufficient.

Pei et al. proposed DeepXplore [2], which leverages gradient from trained networks to

augment input images, creating new test samples by altering lighting or adding occlusions. While DeepXplore introduces gradient-guided augmentation, these adjustments cover only limited types of variations in the input space. Importantly, both Yuchi et al. [1] and Pei et al.'s [2] techniques are designed for image data and cannot be easily generalized to other data types, such as tabular or time-series data, which are prevalent in healthcare settings. This gap underscores the need for more adaptable and comprehensive testing methodologies applicable across various data types in clinical machine learning models.

d'Eon et al. proposed Spotlight [3], a clustering algorithm designed to identify semantically related misprediction regions within the embedded input space. However, this method has notable limitations. First, Spotlight does not automatically determine the number of misprediction segments, requiring manual input to identify the most meaningful clusters. Second, the method relies on manual inspection to assign semantic values to the misprediction regions, which may work well in straightforward cases (e.g., recognizing gender from facial features of an Asian child) but poses challenges for more complex datasets. While the approach could theoretically identify misprediction zones in other data types, such as time-series data, assigning semantic values to these clusters would be difficult without substantial domain expertise.

A significant body of work has focused on detecting distribution shifts [4, 5, 6, 7, 8] and addressing performance disparities across demographic subgroups [9, 10, 11, 12]. However, despite these advancements, none have proposed a systematic framework for evaluating a machine learning model's responsiveness to critical conditions—a key factor in ensuring model reliability and fairness, particularly in high-stakes domains like healthcare. This gap underscores the need for methodologies that not only detect shifts or disparities but also assess the model's ability to respond appropriately under varied clinical scenarios.

2.2 Disparity in healthcare models

Disparities in models refer to the unequal performance of machine learning algorithms across different demographic groups, often leading to biased predictions that can negatively impact minority or underrepresented populations. Data imbalance is a well-documented driver of biased predictions in machine learning models, particularly when distributions across target classes and demographic groups are uneven, potentially leading to serious disparities in outcomes[13]. High-profile cases have illustrated the harmful effects of such biases across different domains. For example, the widely used criminal risk assessment tool COMPAS has been shown to exhibit lower accuracy and higher false positive rates for Black defendants, raising concerns over racial discrimination in judicial settings [14]. Similar issues have been observed in online advertising, where public record advertisements appear more frequently for Black-associated names than for white-associated names, suggesting algorithmic biases in demographic representation [15]. Gender-based biases have also been identified in facial recognition systems that misclassified darker-skinned females more frequently than other groups [16].

There have been a lot of studies identifying the prevalence of disparity in machine learning models in healthcare [9, 10, 11, 12]. Biases emerge due to the inherent imbalances in medical datasets, where certain conditions or demographic groups are underrepresented or overrepresented. For instance, a recent study demonstrated that algorithms used to enroll high-risk patients in health programs favor white patients, despite Black patients showing a 26.3% higher rate of chronic health conditions within the same risk category [17]. Similarly, racial disparities are evident in algorithmic predictions of osteoarthritis pain, with studies indicating a 43% bias against certain racial groups [18]. Moreover, the structure of case-control studies frequently introduces temporal biases, which can diminish predictive accuracy and perpetuate inequities in clinical decision-making [19]. We identify that the tra-

ditional machine learning models trained on imbalance BRFSS (with only 15% representing the diabetes population) dataset [20], tend to misdiagnose diabetes more frequently in the younger population (30-44 years) compared to other subgroups. However, this issue hasn't been reported in the literature. We identified this issue and proposed a mitigation technique.

2.3 Domain knowledge integration

Traditional machine learning (ML) models have achieved success in pattern recognition tasks but often underperform in domains where the dataset fails to capture the true characteristics of different attributes and the relationship with the dependent variable. It's mostly observed in domains such as healthcare, biology, and environmental science. These limitations have led to the emergence of *knowledge-guided machine learning* (KGML), a paradigm that explicitly incorporates domain knowledge—including clinical rules, physical laws, or expert annotations—into the learning process to improve generalizability, interpretability, and trustworthiness.

2.3.1 Benefits of Knowledge-Guided Learning

Studies across domains report consistent benefits of KGML:

- **Improved Generalization:** Enhanced performance in low-data regimes.
- **Interpretability:** Transparent features and rule-based splits enable human understanding.
- **Domain Trust:** Alignment with known rules fosters clinical and scientific acceptance.

- **Scientific Validity:** Constraint-driven models avoid physically or biologically implausible outputs.

2.3.2 Forms of Domain Knowledge and Integration Techniques

Domain knowledge can be encoded in various forms: rules, ontologies, graphs, physical equations, or simulation outputs. These are injected at multiple stages of the ML pipeline:

- **Preprocessing:** Using domain-based rules for imputing missing values or filtering noise.
- **Feature Engineering:** Constructing domain-relevant indices or discarding clinically redundant features.
- **Model Learning:** Modifying the loss function to penalize violations of clinical or physical constraints.
- **Post-Processing:** Applying rule-based filters or ensembling with expert systems.

Sirocchi et al. [21] introduced a modular framework for integrating clinical knowledge into machine learning models across all stages of the pipeline. In scientific domains, KGML often incorporates physical constraints. Karpatne et al. [22] introduced the Theory-Guided Data Science (TGDS) framework, embedding physical knowledge into neural architectures. This KGML was used as a loss component in a neural network [23, 24] and RNN model [25] to estimate lake temperature [26, 27, 28] using the well-established relationship between water depth and temperature. Another study introduces physics-guided deep learning model to estimate blood pressure [29].

However, existing approaches often suffer from limited generalizability. Most are tailored to specific tasks or constrained to particular model architectures, and they are typically not designed to handle time-series data. Furthermore, these methods usually incorporate only a single knowledge constraint into the loss function, limiting their expressiveness. To address these shortcomings, we introduce a novel loss function that can integrate multiple domain-informed constraints across diverse feature types—including both tabular and time-series modalities—and is compatible with a wide range of deep learning architectures, such as LSTM and Transformer models.

Chapter 3

Low Responsiveness of Machine Learning Models to Critical or Deteriorating Health Conditions

3.1 Introduction

Artificial intelligence (AI) machine learning technologies are rapidly made available for incorporation into clinical workflows. The Food Drug Administration (FDA) authorized the first autonomous AI diagnostic system in 2018[30], which is for detecting diabetic retinopathy. Reports also showed that hospitals have adopted machine learning models for clinical uses[31], e.g., Sepsis Watch at Duke University Hospital reduced inpatient mortality from 9.6% to 8.1% [32]. AI models are also deployed to predict surgery time in hospitals [33].

For medical applications, mispredictions of machine learning (ML) models may have serious consequences. For example, missed detection (i.e., false negatives) in-hospital mortality prediction or 5-year cancer prognosis[9] may cause death or underestimate patients' risks. A widely deployed early warning system for sepsis in U.S. hospitals, the Epic Sepsis Model, was found to give poor prediction performance (AUC-ROC 0.63) in a study involving 27,697 patients[34]. Many factors contribute to mispredictions. It is widely recognized that biased

training data cause prediction errors, especially for minority patients in the minority disease class, e.g., Black patients who die in the hospital[9, 35]. The sole use of whole population-wide metrics may also mask the deficiencies in the prediction accuracy of minority-class patients, further exacerbating the problem [9]. Noisy training images may also interfere with the model's learning process, as shown in a skin cancer application [36]. Similar accuracy concerns also exist in other critical AI applications, such as self-driving vehicles, which have resulted in fatalities [37] and severe injuries [38].

For mortality prediction, it is important to measure whether or not ML models can promptly respond to deteriorating patients' conditions. Systematically testing medical AI machine learning models before clinical adoption can help reveal the prediction deficiencies, motivate model corrections, and improve their trustworthiness [39]. Exhaustive testing is both unnecessary and impossible in most medical AI applications, due to the immense complexity of the problem space. Thus, it is crucial to develop strategic testing approaches focusing on the most critical conditions.

The current ML testing practice is very limited in terms of i) the coverage of minority class cases, ii) responsiveness to disease conditions, and iii) testing time-series cases. First, existing testing is largely restricted to a small percentage (10-15%) of the existing dataset, i.e., test set, as the majority of the data is reserved for training. Because data imbalance in medicine is common, a typical test set likely has a low coverage of various critical medical conditions and minority prediction class cases. For in-hospital mortality prediction, the minority prediction class is the death cases, e.g., 13.5% are death cases and the majority (86.5%) of the patients do not die after staying in hospital [40, 41]. Even with cross-validation and bootstrapping, the test set is largely limited to the original data. During clinical deployment, new patient conditions may occur out of the distribution of the test set. Second, a medical AI model's responsiveness to specific disease conditions needs to be evaluated.

A typical machine learning model, prioritizing majority class samples, may underestimate mortality risks and fail to produce high enough mortality risk scores for critically ill patients [9]. For example, we found that a model trained on the MIMIC-III benchmark [41] gives a recall of 88.7% for non-death cases (majority prediction class), but 57% recall for death cases (minority prediction class). Thus, testing needs to consider this disparity and prioritize underrepresented disease conditions for evaluation. Third, time series data is pervasive in medicine. However, methods for systematically assessing ML models for time series are missing. Generative technologies have been proposed to produce curated manmade images for self-driving vehicles [1, 2]. However, image-based solutions do not address the unique temporal challenge in time series applications.

In this work, we develop systematic approaches for generating new test cases beyond the original dataset to assess the responsiveness of machine learning models to critical health conditions that may occur in clinical settings. Our test case generation is guided by domain knowledge and medical experts. Our experiments involve binary classification tasks, including time-series-based in-hospital mortality prediction and 5-year breast and lung cancer survivability prognosis (Figure 3.1). We develop multiple methods for generating high-risk test cases that do not exist in the training data or are underrepresented in the training set. We also conduct interviews with medical experts to obtain their estimated risks on some of the generated test cases. Our work reveals alarming prediction deficiencies of machine learning models and points out that ML responsiveness is an important aspect of trustworthiness in digital health.

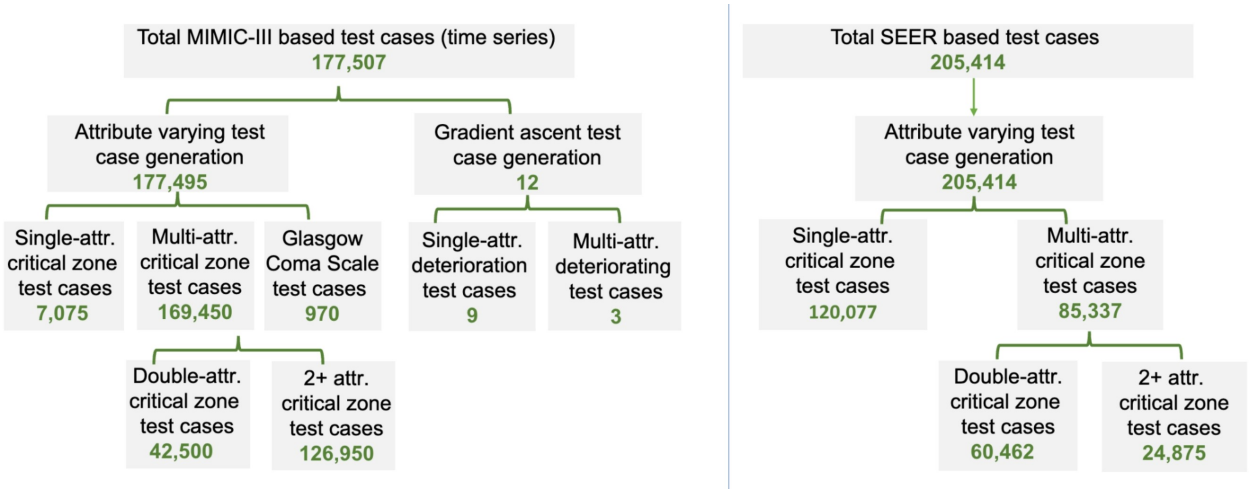


Figure 3.1: The Enhanced-DP approach in contrast to traditional approaches, enriches the minority age groups and creates new training sets by replicating diabetic samples (1 to n times) from a minority age group. n machine learning models are trained on each of the n versions of the training sets. The best model is selected based on performance metric balance accuracy (Bal_Acc) and area under precision and recall curve (AUPRC). The top left bar chart represents age distribution (histogram) in the original dataset.

3.2 Methods

3.2.1 Prediction tasks, datasets, and model selection

Our work aims to test medical machine learning (ML) models for their binary classification accuracy under serious disease conditions. We focus on three binary prediction tasks, namely i) 48-hour in-hospital mortality (IHM) risk prediction, ii) 5-year breast cancer survivability (BCS) prediction, and iii) 5-year lung cancer survivability (LCS) prediction.

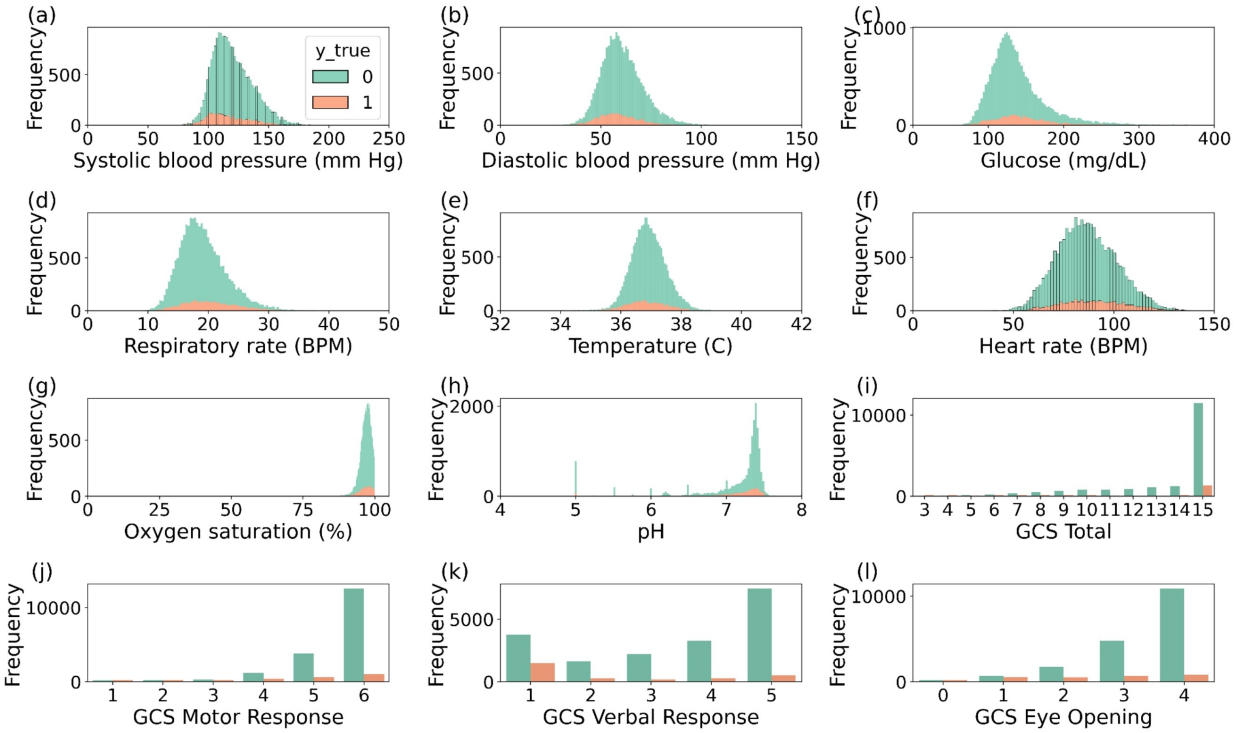


Figure 3.2: Distributions of different attributes (vitals) of the MIMIC-III

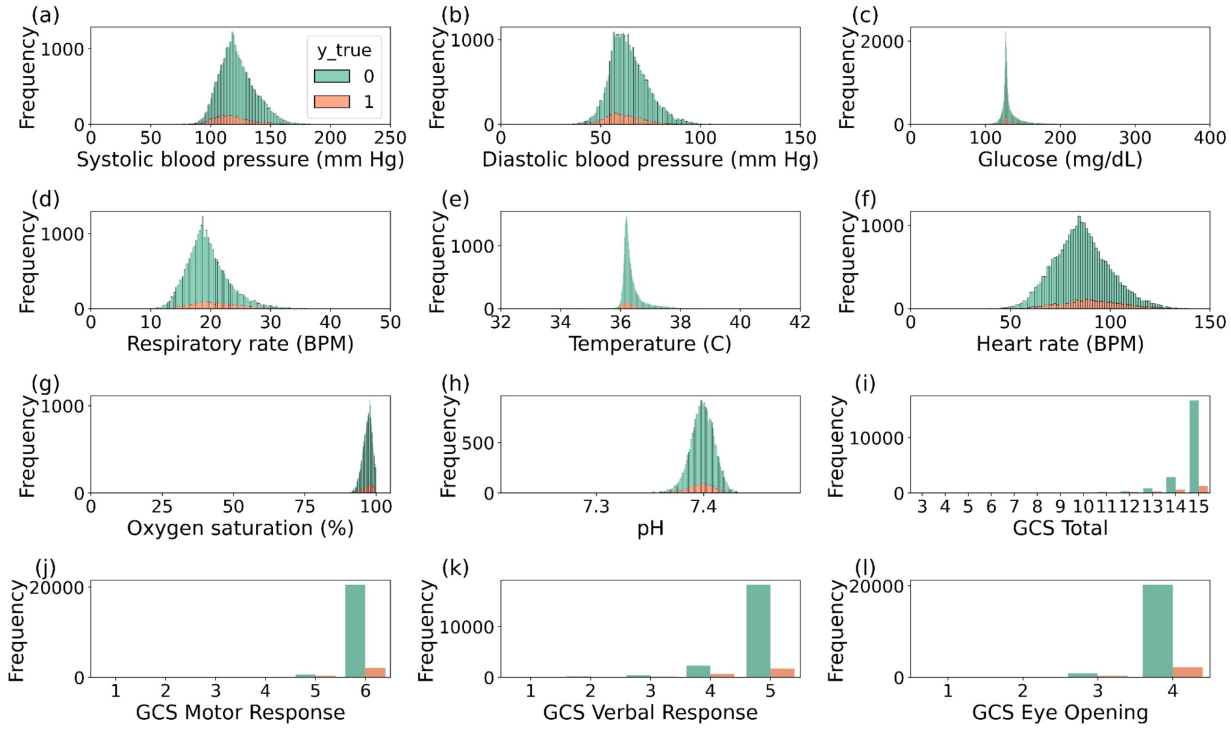


Figure 3.3: Distributions of different attributes (vitals) of the eICU 48 hours ICU mortality dataset.

The datasets in our study include i) a 2019 benchmark [41] based on the MIMIC III dataset [42, 43], ii) a 2020 benchmark [44] based on the eICU [45] dataset, and iii) a 2018 reproducibility benchmark [46] based on the SEER (5-years breast and lung cancer) dataset [47]. The first two datasets contain patients' 48-hour time series data in critical care units (ICU). Our study excludes clinical free text notes. As with many medical datasets, the MIMIC-III dataset for IHM, containing 21,139 samples, is imbalanced, with only 13.2% death cases (Class 1), and 86.8% non-death cases (Class 0). The eICU IHM benchmark dataset contains a total of 30,681 (88.5% for Class 0 and 11.5% for Class 1) samples with similar attributes and time lengths to the MIMIC III benchmark [44]. Figure 3.2 and 3.3 shows the distributions of key attributes of both MIMIC III and eICU datasets respectively. The SEER BCS dataset contains 248,751 patient cases with 56 attributes (7 numerical and continuous

features and 49 categorical). In the SEER BCS dataset, only 12.7% of cases are death cases (Class 0); the rest are survived cases (Class 1). Figure 3.4 shows the distributions of key attributes of SEER BCS and LCS dataset. The SEER LCS dataset contains 205,555 cases with 47 features (7 numerical and continuous features and 40 categorical). 84% of patients died in the LCS dataset.

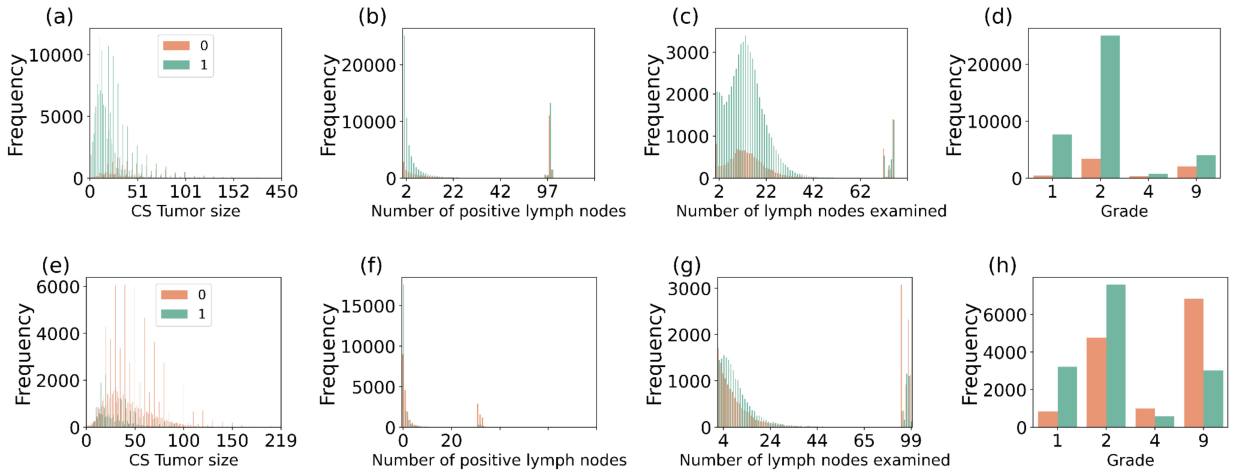


Figure 3.4: Distributions of different attributes of the (a)-(d) SEER BCS and (e)-(h) SEER LCS dataset.

We select ML models that are commonly used for these prediction tasks or commonly used in medical literature. Specifically, we select Long Short Term Memory (LSTM) as it is commonly used for predicting mortality risk in a 48-hour ICU time series dataset—used in recent literature [9, 48, 49, 50]. Similarly, for cancer survivability prediction, we selected multi-layer perceptron (MLP), which was commonly used in analyzing SEER datasets, e.g., [9, 46, 51]. The performance of LSTM and MLP reported in these existing studies is comparable to other top models. In addition, we also evaluated general-purpose ML models commonly seen in medical literature, including XGBoost, AdaBoost, Random Forest, Gaussian Naive Bayes, and K-nearest Neighbor (KNN). For mortality prediction, we also include CW-LSTM and linear logistic regression models from the benchmark study [41] and a Transformer model.

3.2.2 Dataset preprocessing

We train machine learning models with benchmark datasets of MIMIC-III [41], eICU [44], and SEER breast and lung cancer survivability studies [46], following the conventional pre-training processing (e.g., encoding, standardization). As MIMIC-III and eICU benchmark datasets contain missing values, we imputed the values that are missing using the most recent observation (within 48 hours) if it exists, otherwise, a value from the normal range of corresponding vitals is mentioned in [41]. Masking was used to indicate whether the vital value was original or imputed. The categorical variables, including binary ones, were encoded using a one-hot vector. The numerical features, such as diastolic blood pressure and glucose level, were converted to their standardized form. After preprocessing, each time-series data point became a 76-by-48 matrix (76 computed features and 48 hours). The processed dataset was used for training and testing neural network-based models such as LSTM and CW-LSTM models. For non-neural network models that cannot directly process time series, we extracted 6 statistical features (mean, min, max, standard deviation, skew, and number of measurements) from various sub-periods (first/last 10%, 25%, 50%, and full 100%). We did not encode the categorical variables, as they contain values with a meaningful scale. The missing values were replaced with mean values computed on the training set and numerical variables were standardized. In total, we obtained 714 features from each 48-hour time series with 17 vitals. The continuous variables were standardized before training. After encoding, the feature vector length of the BCS and LCS datasets became 1,418 and 1,314, respectively.

3.2.3 Configurations of machine learning models

For in-hospital mortality risk prediction, we utilized the Long Short Term Memory (LSTM) model, Channel-wise Long Short Term Memory (CW-LSTM) model, Transformer, Logistic Regression (LR), AdaBoost, XGBoost, and Random Forest (RF) models. For 5-year breast cancer survivability (BCS) prediction, we used Multi-layer perceptron (MLP), AdaBoost, XGBoost, and random forest models. We utilized the optimal settings of neural network models (i.e., layers, activation, hyperparameters) for each of the tasks from corresponding benchmarks [41, 46]. The LSTM model consisted of an input layer (76 dimensions), a masking layer (76 dimensions), a bidirectional LSTM layer (16 dimensions), an LSTM layer (16 dimensions), a dropout layer, and finally a dense layer (1 dimension). In total, the LSTM had 7,569 trainable parameters. The CW-LSTM layer consisted of an input layer (76 dimensions), masking layer (76 dimensions), 17 channel layers (for each 17 input features), 17 bidirectional layers (connected to one of the 17 channels layers), another set of 17 bidirectional layers, a concatenation layer connecting all 17 bidirectional layers, bidirectional layer (64 dimensions), LSTM layer (36 dimensions), dropout layer (64 dimensions), and finally a dense layer (1 dimension). In total, the CW-LSTM model had 153,025 parameters. The size of CW-LSTM's parameters was 20 times that of LSTM's. The CW-LSTM model allows independent pre-processing of each variable before combining them. For both LSTM-based models, the optimal hyperparameters are selected using grid search [41] [Harutyunyan 2019]. For example, the batch size, dropout, and time-step are set to 8, 0.3, and 1, respectively. The transformer model consisted of an input layer (76 dimensions), a masking layer (76 dimensions), a positional encoding layer (76 dimensions), 2-3 transformer encoder blocks, a global average pooling layer, a batch normalization layer, a dropout layer (0.3 or 05), a dense layer (32 or 64 units), and finally a dense layer (1 dimension). Each transformer encoder block included a multi-head attention layer with 4 heads (key dimension 76), followed by

layer normalization and residual connections. The feed-forward dense layers within each encoder block contained a hidden dimension of 16. In total, the transformer model contains a total of 881,677 parameters (trainable parameters: 293,841, optimizer parameters: 587,684, and non-trainable parameters: 152), larger than both the LSTM and CW-LSTM models. For hyperparameter tuning, grid search was employed to select the best hyperparameters (Table 3.1 and 3.2). The logistic regression (LR) model was from the sklearn library, utilizing the L2 regularization penalty. To prevent overfitting and to enhance the generalization capability of the model, the parameter C is 0.001. This choice of a small C value effectively controls the amount of regularization applied during training. The remaining hyperparameters were left at their default values, following the standard implementation provided by the Python Sklearn library. This model was trained with the standardized training set. The MLP model used for BCS survivability prediction consists of 2 hidden layers, where each hidden layer contains 20 neurons. The hidden layer used Relu as an activation function. Dropout rate of 0.1 after each hidden layer was used to avoid overfitting. The last layer predicted binary labels using the sigmoid activation function. The MLP model contained 28,831 trainable parameters. MLP hyperparameter is empirically selected using grid-search from a list of predefined values such as the number of hidden layers (1, 2, 3, and 4), number of nodes in each layer (20, 50, 100, and 200), and dropout (0, 0.1, 0.2, 0.3, 0.4, and 0.5) [46]. The other models are implemented using Python's Sklearn library and hyperparameters are tuned using grid search (Table 3.1 and 3.2).

3.2.4 Model training, threshold tuning, and imbalance correction methods

Model training, threshold tuning, and imbalance correction methods For in-hospital mortality prediction, LSTM models and transformer models were trained for 100 epochs using the

MIMIC-III or eICU dataset. For 5-year cancer survivability prediction, MLP models were trained for 25 epochs with the SEER BCS or LCS dataset with optimal hyperparameter settings mentioned in [46]. Other models, including XGBoost, AdaBoost, and Random forest, are trained using the best hyperparameters obtained from grid search (Table 3.1 and 3.2). The models were trained using binary cross-entropy loss. An epoch was selected based on the threshold-agnostic validation area under the precision-recall curve (AUPRC) and validation loss to avoid overfitting. Specifically, we first selected the top 3 epochs with the highest validation AUPRC and then selected the epoch with the minimum validation loss (Tables A.1 and A.2). We monitored the validation loss and training loss difference to prevent overfitting. In all experiments, the chosen machine learning model demonstrated a small loss difference (Tables A.1 and A.2).

Besides evaluating models trained on the original training sets, we also experimented with resampling and reweighting techniques and measured how well the resulting bias-corrected machine learning models performed in our critical zone tests. The reweighting technique has demonstrated superior performance in healthcare datasets, as evidenced by prior studies [52]. For resampling, we tested two generative resampling approaches, SMOTE (Synthetic Minority Oversampling Technique) and AdaSyn (Adaptive Synthetic Sampling). We employed Python's Imblearn library to apply SMOTE and AdaSyn oversampling techniques, generating balanced training sets by increasing samples from the minority class (sizes shown in Table A.3). For reweighting, we utilized Python's Sklearn library to compute balanced class weights based on the training sets (Table A.4). These methods are applied to the LSTM model for mortality prediction and to the MLP model for cancer survivability prediction.

The training, validation, and test set breakdown for MIMIC-III and eICU datasets is 70%, 15%, and 15% and 80%, 10%, and 10% for the BCS and LCS datasets. After model calibration, a threshold-tuning process is conducted on the validation set, and an optimal threshold

is selected based on balanced accuracy and F1 score for the minority class. Specifically, after training, we first conducted model calibration by applying Isotonic Regression using the validation set. Model calibration mapped the predicted probabilities to actual probabilities. Then, we performed threshold tuning to determine the optimal threshold. The minority F1 score and balanced accuracy were computed on the validation set for each threshold ranging from 0.0 to 1.0 with a step size of 0.01. Subsequently, the top three thresholds yielding the highest minority F1 scores were identified, and the optimal threshold maximizing balanced accuracy across all validation samples was selected. This process was repeated for 3 independently trained models of each type, and the average threshold was calculated from these independent trials. Thresholds are shown in Table A.5. The tasks were executed on a machine with Ubuntu 18.04 operating system, x8664 corei9 architecture, 8 physical cores (16 virtual cores), and 32 GB RAM. The experimental code and models were written using Python 3.7, TensorFlow 1.15, and Keras 2.1.2. The cancer survivability prediction MLP model was trained on a machine with x86_64 Intel(R) Xeon(R) CPU 2.40GHz (40 cores) and 125 GB RAM. The experimental code and model were written using Python 3.6, TensorFlow 2.9.0, and Keras 2.9.0.

3.2.5 Mapping neuron activations

We visualized the activated neurons in a neural network model for a particular input. The Keras backend was used to capture the neuron outputs from the bidirectional layer output and LSTM layer output for the mortality risk prediction model. Sigmoid activation was applied to obtain neuron output values in the range of [0, 1]. To quantify changes in neuron activation, we defined and computed Neural Zone Activation (NZA) and average zone difference NAZ. A zone is defined by the attribute range bounded by two values. NZA calculates the average neurons' activations within a zone, where a zone can be a critically low, critically

Table 3.1: Hyperparameter tuning set for grid search

Model	**Hyperparameter search set**
XGBoost	n_estimators: [50, 100, 150, 200, 300]
	max_depth: [None, 3, 5, 7, 10]
	learning_rate: [0.01, 0.1, 0.2, 0.5, 1]
	subsample: [0.1, 0.5, 0.8, 1.0]
	colsample_bytree: [0.1, 0.5, 0.8, 1.0]
AdaBoost	n_estimators: [50, 100, 150, 200, 300]
	learning_rate: [0.01, 0.1, 0.2, 0.5, 1]
Random Forest	n_estimators: [50, 100, 150, 200, 300]
	max_depth: [None, 3, 5, 7, 10]
	min_samples_split: [2, 5, 10]
	min_samples_leaf: [1, 2, 4]
K Nearest Neighbors	n_neighbors: [3, 5, 7, 9, 11]
	weights: ['uniform', 'distance']
	metric: ['euclidean', 'manhattan', 'minkowski']
	p: [1, 2]
Naive Bayes	var_smoothing: [1e-9, 1e-8, 1e-7, 1e-6]

Table 3.2: Selected best hyperparameters through grid search

Model	Hyperparameter	Best hyperparameters			
		eICU	MIMIC-III	SEER-BCS	SEER-LCS
XGBoost	n_estimators	200	300	300	300
	max_depth	5	3	5	5
	learning_rate	0.1	0.1	0.1	0.1
	subsample	1	1	0.8	0.8
	colsample_bytree	1	1	0.8	1
AdaBoost	n_estimators	200	200	150	150
	learning_rate	0.1	0.1	1	1
Random Forest	n_estimators	300	300	300	300
	max_depth	None	None	None	None
	min_samples_split	10	10	2	10
	n_neighbors	11	11	11	11
K Nearest Neighbors	weights	distance	distance	distance	distance
	metric	euclidean	manhattan	euclidean	euclidean
	p	1	1	1	1
	var_smoothing	1e-7	1e-6	1e-9	1e-9

high, or normal range (Equation 3.1). NAZ computes the average NZA difference between two zones (Equation 3.2), such as normal and critically high zones, indicating how much neurons react to zone changes. There is no standard value for NAZ. A relatively higher value indicates a good response.

Neural activation

Given neuron i, its activation value $\text{act}(\text{neuron}_i(k))$ for input k, and a vital zone bounded by n_1 and n_2 (representing critically low, critically high, or normal vital range), we define the average neural zone activation $\text{NZA}_i(n_1, n_2)$ in Equation (1). Given two vital zones z_1 and z_2, we further define the average neural zone activation difference $\Delta\text{NZA}_{\text{avg}}(z_1, z_2)$ in Equation (2). In Equation (2), z_{1s} and z_{1e} are the zone starting and ending values, respectively. b is the number of neurons in the activation layer (16 in our case).

$$\text{NZA}_i(n_1, n_2) = \frac{1}{|n_1 - n_2|} \sum_{k=n_1}^{n_2} \text{act}(\text{neuron}_i(k)) \tag{3.1}$$

$$\Delta\text{NZA}_{\text{avg}}(z_1, z_2) = \frac{1}{b} \sum_{i=1}^{b} |\text{NZA}_i(z_{1s}, z_{1e}) - \text{NZA}_i(z_{2s}, z_{2e})| \tag{3.2}$$

3.2.6 Statistical methods

Model performance is reported using the average and standard deviations, which are calculated using 9 or 15 trials. The trials were performed using 3 model instances that have identical architecture and were trained on the same training set with random model parameter initialization. Each of the 3 model instances is evaluated with 3-5 test sets. The distribution shift of the synthesized test dataset from the original training sets was quantified by Wasserstein distance (WD) [53, 54]. We used an implementation from the Python library

called *scipy.stats.wasserstein_distance*. First, the WD was calculated between the same features from the whole original dataset and the synthesized test set. Then, the feature-specific WD was averaged to obtain the mean WD for quantifying the distribution shift.

3.2.7 Attribute-based test case generation for in-hospital mortality risk prediction

We created new cases by increasing or decreasing one or multiple vital health parameters in the seeding records. To reduce computing complexity, we prioritized by focusing on the most influential features. Relevant medical terminologies are explained in the Appendix notes A.1

.

In the single-attribute variation, we generated new test cases by varying a single attribute at a time while keeping other attributes unchanged. We then evaluated how the model reacts to these changes and its ability to recognize associated risks (e.g., hypoglycemia). Specifically, given an attribute A, single-attribute variation for time series involved the following operations. First, we identified A's minimum and maximum values in the MIMIC-III or eICU datasets, which defined the observed range. Then, the mean and the variance of attribute A were computed from the entire dataset. Using the variance and the observed range, we generated a series of random values for every value from that range, one value for each of the 48 hours. Then, the new test case was formed by having these generated values for attribute A and other attribute values directly inherited from the seed. We repeat this process for every possible attribute value from the observed range with step 1.

Algorithm for Single-Attribute Variation Test Case Generation for Time Series Data

Input:

- Dataset D (e.g., MIMIC-III or eICU)
- Attribute $A \in$ {Diastolic bp, Glucose, Respiratory rate, ...}
- Time-series length $T = 48$ hours
- Seed S

Output: Set of test cases with varied single attribute A

Procedure:

1: Compute minimum $A_{\min}$, maximum $A_{\max}$, mean μ_A, and standard deviation σ_A values of attribute A in dataset D.
2: For each value μ in the range $[A_{\min}, A_{\max}]$ with step size α:

1. Generate a vector of length T by sampling random values from the normal distribution $N(\mu, \sigma_A^2)$.
2. Replace the feature value of A with the new vector x_A in seed S.
3. Save the new sample, S_μ, which is a representative of attribute value μ.

Multi-attribute variation generated new test cases by modifying two or more attributes, aiming to represent medical conditions that were characterized by variations in multiple related attributes. We further differentiated two scenarios: i) a single set of medically correlated attributes driven by one underlying disease condition, e.g., high diastolic and systolic blood pressure due to hypertension, and ii) medically correlated attributes due to multiple underlying conditions, e.g., hypertension and diabetes. These test cases were used to assess the machine learning model's ability to respond to the risks of multiple disease conditions in patients. One of the test sets was created by changing multiple vitals such as systolic blood pressure, diastolic blood pressure, blood glucose level, respiratory rate, heart rate, and body temperature at the same time. A test case was assigned a ground truth label using existing literature or under the guidance of medical doctors. 6 multi-attribute test

cases and 12 deteriorating test cases were directly labeled by the medical doctor (Tables A.6 and A.7).

3.2.8 Deteriorating test case generation for MIMIC-III

We leveraged the gradients of LSTM to guide the generation of new test cases. This method is automatic and does not require the specification of attributes to change – aiming to generate new test cases that are challenging for machine learning models to classify correctly. Such cases typically occur at the decision boundary of the classifier. Our method started from a healthy patient's record (i.e., a seed with low or zero mortality risk). The seed is a time-series record far away from the classifier's decision boundary. We incrementally adjusted the attribute values of the seed by following the steepest direction (i.e., gradient) that can maximize the loss (i.e., prediction errors of the machine learning model). This process explores the local hyperspace and iteratively produces new cases that are closer and closer to the machine learning model's decision boundary. Computationally, given a trained machine learning model and a healthy patient's time series record as the seed, we computed the derivative of the model's loss function, i.e., gradient (Equation 3.3). The gradient is a vector of partial derivatives describing the direction and rate of changes of the loss function. Then, we changed the test case in the direction of increasing gradient. Our algorithm is described in Equation 3.3. The step size or learning rate to control the magnitude of the change was set to 0.001-0.2 in our experiment depending on the attribute (Table A.8 and A.9). Our method focuses on generating special samples; it differs from the straightforward gradient ascent process, which adjusts model weights to minimize loss. We have two ways of creating gradient-based test cases from the MIMIC-III dataset – single attribute gradient approach and multi-attribute gradient approach. In the former, we focus on a single attribute and apply gradient ascent solely to modify that specific attribute. This

approach allows one to observe the individual impact of each attribute on the mortality risk. In the latter, we simultaneously change values of multiple attributes using gradient ascent. Gradient approaches create test cases that represent deteriorating health conditions in time series. Table A.10 shows the various categories of test sets and their sizes.

Gradient ascent

Given a sample x, we add the gradient G_i calculated from the trained model with the seed sample x_i to generate a new sample x_i^{new}, where i represents the feature to be changed by the gradient ascent process.

$$x_i^{\text{new}} = x_i + \alpha G_i \tag{3.3}$$

Where:

- i is the selected feature index from $[0, 1, \ldots, 17]$
- x_i^{new} is the sample with updated feature i
- x_i is the current feature value
- α is the step size or learning rate that controls the magnitude of change in the input feature
- $G_i = \frac{\delta y}{\delta x_i}$ is the gradient of the output y with respect to feature x_i

3.2.9 Glasgow coma scale test case generation

The Glasgow Coma Scale (GCS) is a neurological scale that assesses a patient's level of consciousness. It evaluates responses in three categories: eye-opening (E), verbal response

(V), and motor response (M), adding up to a score ranging from 3 to 15 [Jain 2018]. A lower score indicates a more severe impairment of consciousness. Definitions of values in each category are given in Table A.11. A GCS score can be representative of multiple sets. For example, GCS total 10 can be the outcome of (E, V, M) = (3, 3, 4), or (4, 4, 2), etc. The GCS total test set contains all the possible combinations of (E, V, M) for each particular GCS total value. The double attribute-based GCS cases were also created by varying both attributes and keeping the other constants to healthy values.

3.2.10 Attribute-based test case generation for 5-year cancer prognosis

Single-attribute variation: Similarly, we engineered cancer test cases by varying one attribute of a seed record. The attribute may be the size of the tumor (T), the number of positive lymph nodes (N), the number of examined lymph nodes (ELNs), or the grade of the cancer cell. T and N are the two most important factors for determining cancer severity or stage [SBC 2024]. T has 4 categories based on the size. The tumor test set was created by varying the size of 3 seeds in the surviving class, using the range (0-986 mm) from the original SEER dataset. This BCS tumor size test set contains 12,891 cases, including 18 T0 cases, 243 T1 cases, 390 T2 cases, and the rest of 12,240 T3 cases. The LCS tumor size contains 8,367 cases, including 12 T0 cases, 171 T1 cases, 273 T2 cases, and 7,911 T3 cases. (T4 cases cannot be created, as it is not associated with a quantitative value). The number of positive lymph nodes (N) is divided into 4 categories. The positive lymph node test case was created similarly by changing the corresponding value from the same 3 seeds using the attribute range (0-84). For BCS, we generated 7,686 test cases, including 90 N0 cases, 270 N1 cases, 546 N2 cases, and 6,780 N3 cases. For LCS, we generated 24,264 test cases, including 333 N0 cases, 999 N1 cases, 1998 N2 cases, and 20,934 N3 cases. Appendix

Notes A.1 have more details of T and N category definitions.

The ELN test case was created similarly by varying the number of ELNs (range in [0, 86]) from 3 seeds and keeping other values the same as the seeds. The ELN test set contains a total of 3,510 cases for BCS and 1,835 cases for LCS. Although the number of examined lymph nodes (ELNs) is not directly related to the cancer staging, it is crucial for diagnosing cancer. Several studies proposed that there should be a standard (or a minimum) number of ELN cancer diagnoses [55, 56, 57, 58]. The grade of the cancer cell represents the spreading and growth intensity of the cancer cell [59]. The SEER dataset contains 1-4 grades where the higher grade represents faster growth and speed and another grade 9 for undetermined (not stated/applicable). For BCS, we created test sets for each of 1-4 grades, where each set contains 24,875, created from 21,723 cases from the majority Class 1 (survival) and 3,152 cases from the minority Class 0 (death). We utilized the entire validation set as the seed pool, allowing for a more comprehensive evaluation. In total, the 1-4 grade test set contains 99,500 cases (24,875 cases for each grade). To create each grade test set, we set the corresponding grade value to all data points in the validation set.

Double- and triple-attribute variations. Double-attribute variation generated new BCS test cases by changing a pair of attributes from the 3 continuous attributes, i) the size of the tumor (T), ii) the number of positive lymph nodes (N), and iii) the number of examined lymph nodes (ELNs). The grade attribute was excluded, as it is categorical. The tumor size and positive lymph node combination test set contains 18,531 test cases. The tumor size and number of examined lymph nodes combination test set also contains 18,531 test cases. The number of examined lymph node and positive lymph node combination test sets contain 23,400 test cases. The triple-attribute test set was created by setting three attributes simultaneously to represent serious disease conditions, e.g., tumor size to T4, number of positive lymph nodes to N3, and grade to 4. The validation set, consisting of

24,875 cases including 21,723 cases from Class 1 (survived) and 3,152 from Class 0 (death), was used as seeds. While the tumor size and number of positive lymph nodes are continuous variables, we treated them as categorical by selecting a value from the T4 and N3 range respectively. As a result, the triple-attribute test set contains 21,723 cases derived from Class 1 seeds and 3,152 cases derived from Class 0 seeds. Table A.12 and A.13 summarizes the various categories of test sets and their sizes. We performed double- and triple-attribute variation tests for BCS models, not on LCS models.

For labeling generated breast and lung cancer test cases, we used authoritative literature to assign labels. We labeled cases with Class 0 (indicating low survivability) if there was a strong presence of cancer (i.e., T 1-3, N 1-3, and grade 2-4). For ELNs, the previous studies using SEER datasets [Sun 2020, Chi 2017] suggested using at least 8-9 ELNs for stage T1 diagnosis, 37 ELNs for T2 diagnosis, and 87 ELNs for T3 diagnosis. As ELN is not directly responsible for the death, that attribute was not considered during labeling.

3.2.11 Selection of Seeds

We used existing patient records from the original dataset as seeds (i.e., starting points) to generate synthetic test sets. We selected seeds from the in-hospital mortality dataset that are real-world non-death patient cases that exhibit healthy attribute values. Seeds were chosen as follows. For attribute-based test case generation, we randomly selected seeds from MIMIC-III Class 0 (survival case) following two criteria. First, the mean (of 48 hours) attribute values are within the range of ideal health conditions defined in Table A.14. In addition, the standard deviation of each attribute needs to be less than or equal to the mean standard deviation (Table A.15) of the MIMIC-III dataset. Our evaluation of attribute-based test case generation involved 5 seeds and the statistics of these 5 seeds are given in Table A.16. The

deterioration test case generation involved another 3 seeds, which were selected randomly from Class 0 of MIMIC-III. Since the eICU dataset contains similar samples with identical features and a consistent 48-hour time duration, we utilized the same test set generated from MIMIC to evaluate models trained on the eICU dataset. Additionally, the selected seed attributes fall within the healthy (ideal) range, minimizing the out-of-distribution effects on models trained on the eICU dataset. For the cancer survivability prediction task, test cases involving changing a numerical variable were generated from 3 randomly selected seeds from the surviving class. Test sets are separately generated from each of the SEER BCS and LCS datasets. Test sets involving categorical variables, such as grades test and triple-attribute test sets, were generated using all validation data points from SEER.

3.3 Results

For in-hospital mortality prediction, we generated 177,507 new time-series test cases based on MIMIC-III to represent serious patient conditions and used them to evaluate the responsiveness of machine-learning models (Table A.10). 126,950 cases are generated by modifying multiple vital attribute values in 5 seed records, 42,500 cases by modifying double attributes in seed record, 7,075 cases by modifying a single attribute value in a seed record, 970 cases by modifying Glasgow coma scale (GCS), and 12 cases by gradient ascent. Modifications to vital attributes are bounded by the minimum and maximum values of the attribute in the in-hospital mortality (IHM) datasets and focus on critically high and critically low ranges of the 6 vitals. We carefully use literature [60, 61, 62, 63, 64] to identify these ranges (Table A.14). The test case generation also ensures the continuity of the time series. The 6 types of attributes include systolic blood pressure, diastolic blood pressure, blood glucose level, respiratory rate, heart rate, and body temperature. A seed record is a real-world patient

case selected from the MIMIC-III dataset that is a non-death case whose attributes are in the typical healthy ranges [61, 62, 63, 64]. The other 12 test cases are generated using a gradient-ascent approach, which modifies the seed by following the direction of the steepest increasing loss function.

Each synthetic test case is assigned a label, death (Class 1) or survival (Class 0) for in-hospital mortality (IHM) prediction. Labels are verified either by a medical doctor or confirmed by the literature. These labels are considered ground truth in our study. Two medical doctors reviewed 18 generated test cases (6 attribute-based cases and 12 gradient-based cases), where the test cases are time series data and the risk scores of the medical doctors' output are quantitative, between 0 and 1. The medical experts estimated risk values are in Tables A.6 and A.7. Labels of the other 177,489 test cases are inferred based on expected ranges of vital health parameters of healthy individuals extracted from medical literature. Synthetic test cases persistently containing vital values in critical zones represent patients in sustained critical health conditions, and thus are labeled Class 1. These cases should receive a high mortality risk prediction from machine learning models. We define machine learning responsiveness as the model's ability to react to significant changes in input values, e.g., by increasing the mortality risk score for IHM prediction.

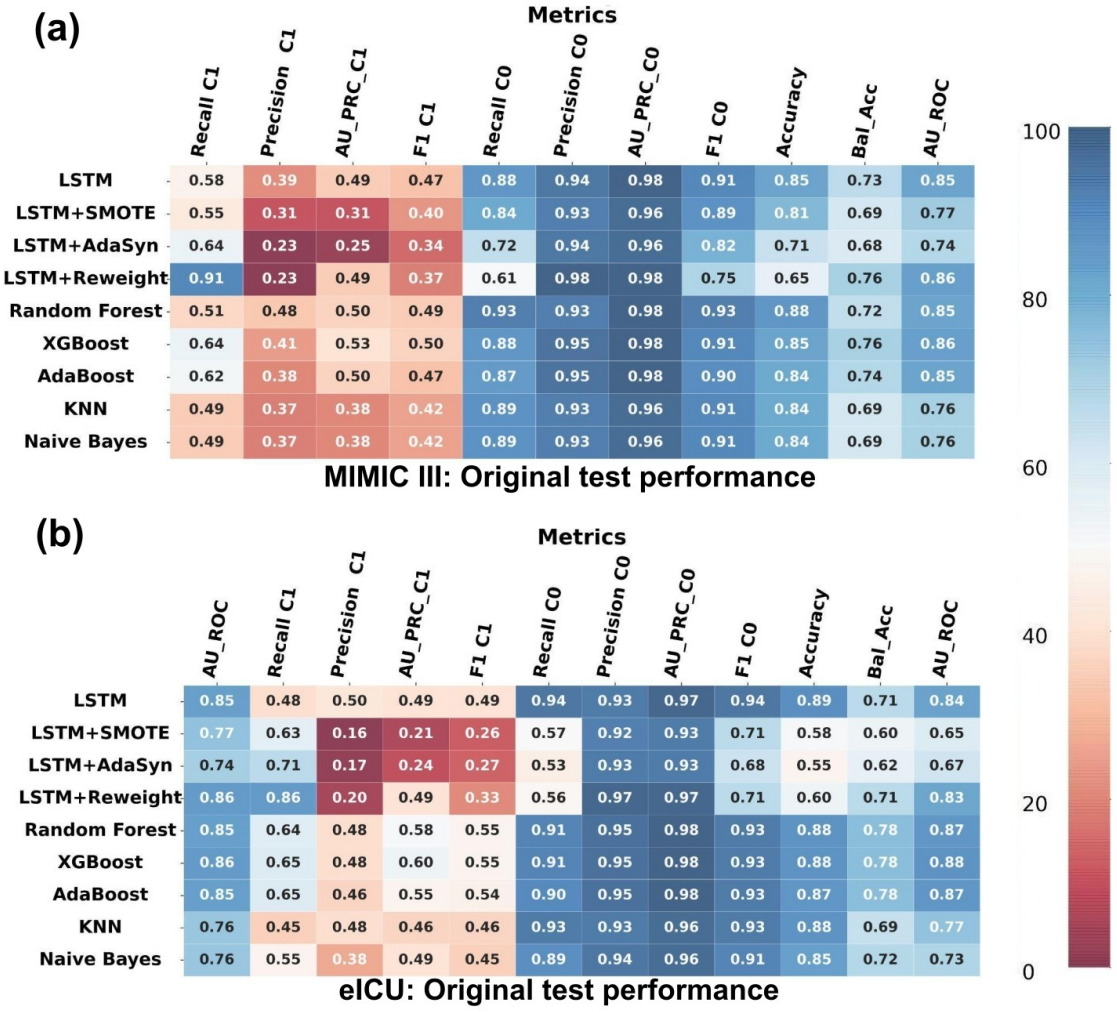


Figure 3.5: Machine learning model performance on the original test set. The death class is the minority class in MIMIC III, eICU. Death class is represented by 1 in in-hospital mortality risk prediction datasets.

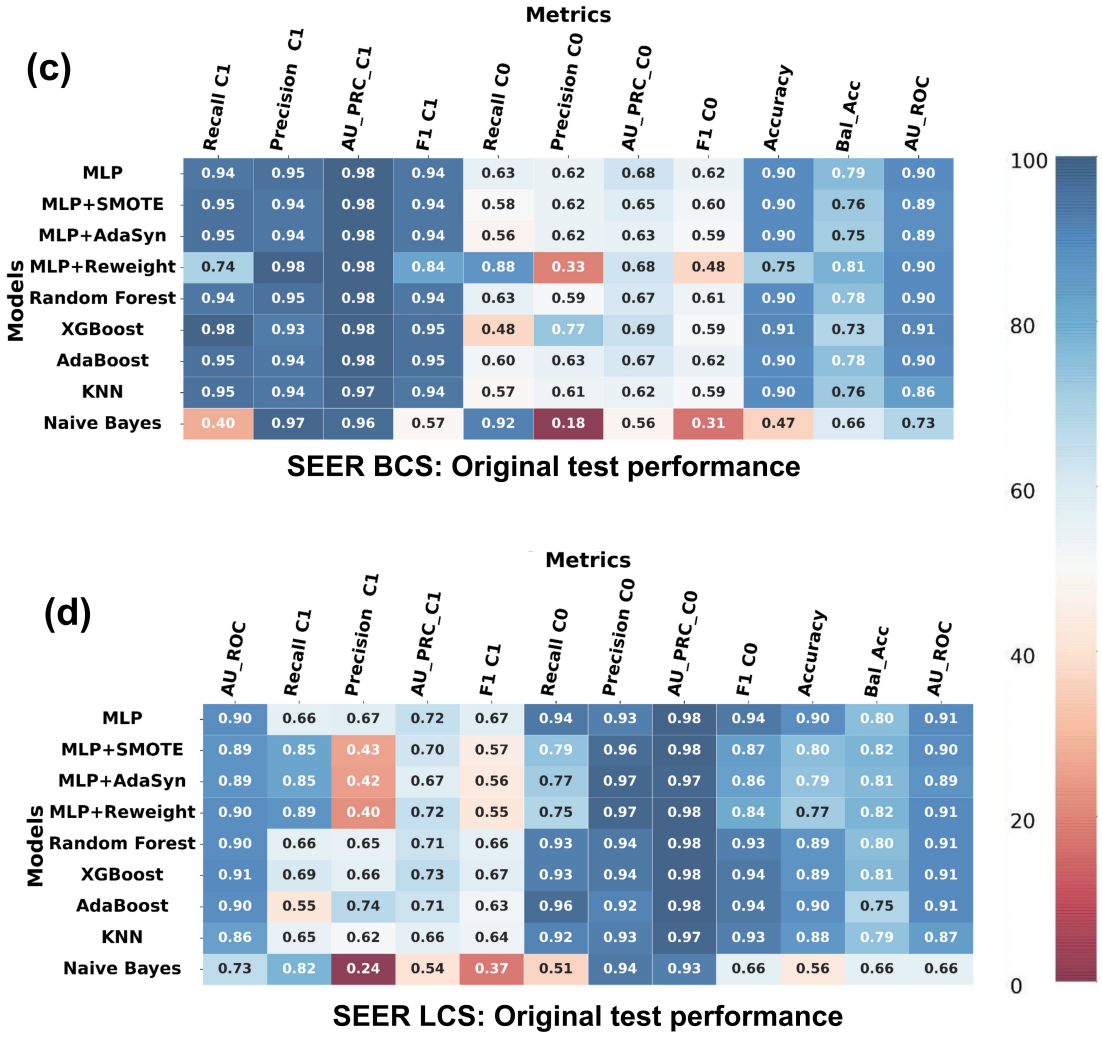


Figure 3.6: Machine learning model performance on the original test set. The death class is the minority class in SEER BCS datasets, whereas the majority in the SEER LCS dataset. Death class is represented by 0 in cancer survivability prediction datasets.

Based on the SEER 5-year breast cancer survivability dataset, we generated 205,414 test cases to represent different patient conditions. Among them, 120,077 cases are generated by changing single attributes (including 7,686 cases representing different N stages, 12,891 cases for the T stage, and 99,500 cases for grades), 60,462 cases by modifying double attributes, and 24,875 cases by changing triple attributes from the seed cases. Based on the LCS

dataset, we generated three sets of single-attribute test cases totaling 31,136 cases, which include 8,367 cases representing different T stages, 24,264 cases representing N stages, and 1,835 cases representing ELNs (Table **??**). We manually assigned labels to synthesized test cases guided by the literature [55, 56, 57, 58, 59, 65, 66].

3.3.1 ML performance under Glasgow Coma Scale (GCS) testing

For in-hospital mortality prediction, we assess MIMIC III-based LSTM, CW-LSTM, and LR models with test cases containing varying GCS scores (Figure 3.7 and 3.8), including severe injury cases with GCS scores 3 to 8, moderate injury with 9 to 12, and mild or no injury with 13 to 15. A low GCS score indicates a poor health condition [67] (medical meanings of each category are shown in Table A.11). The CW-LSTM model gives near zero mortality risk values for 15 severe injury cases, for example, E4M1V3 in Figure 3.7 a, i.e., a case with an eye response score of 4 out of 4, a motor response of 1 out of 6, and a verbal response score of 3 out of 5. For a moderate injury case E4M1V5, CW-LSTM also gives an unexpectedly low mortality risk (0.01) prediction, i.e., predicting the healthy outcome of the patient. The model's prediction is inconsistent, as another moderate injury case E4M3V5 receives a high mortality risk of 0.58.

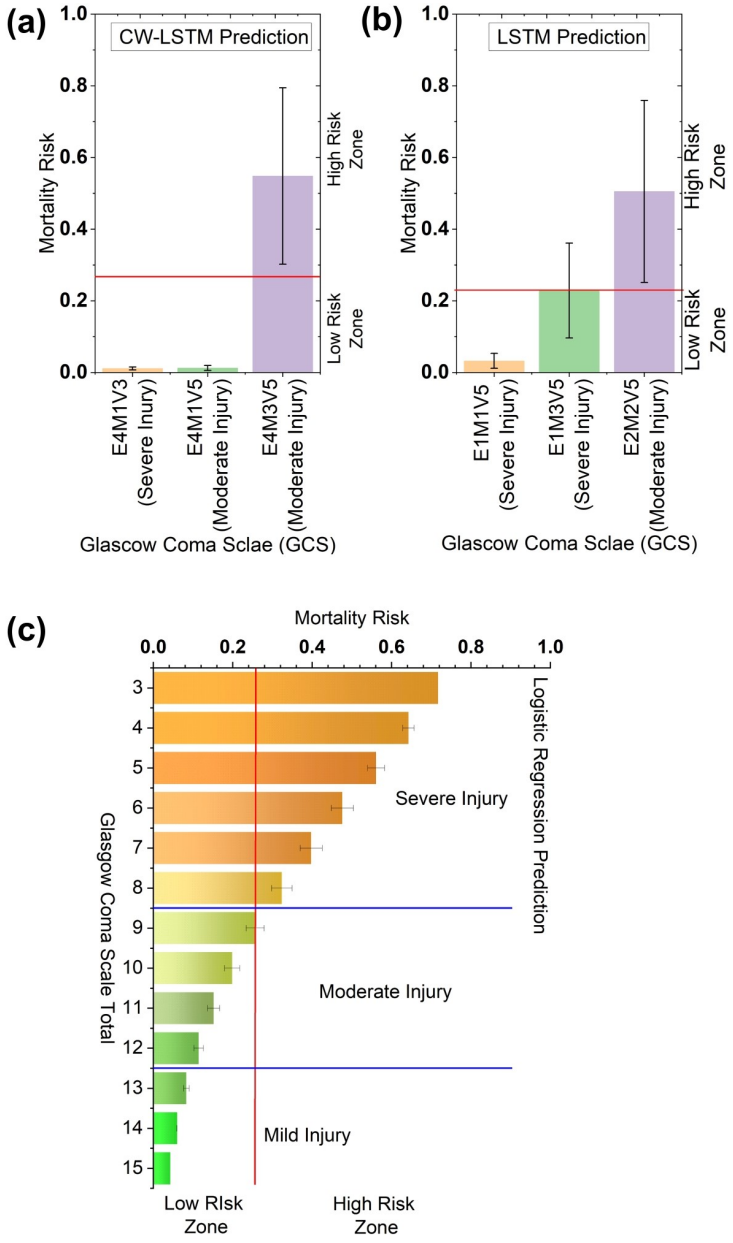


Figure 3.7: Mortality risk (MR) prediction for Glasgow Coma Scale for different combinations using three machine learning models. MR predicted by (a) channel-wise LSTM model for three injury cases, (b) LSTM model for three injury cases, and (c) Logistic regression for injury cases indicated by all combinations of GCS scores.

Similar inaccuracies and inconsistencies are also observed for the LSTM (MIMIC III) model tested. For instance, the LSTM model mistakenly considers a severe injury case E1M1V5

to be much more likely to survive than a moderate injury case E2M2V5 (Figure 3.7 b). In contrast, the LR model consistently predicts at least 0.3 mortality risk for severe injury cases and responds well (Figure 3.7 c). For mild injury cases, the LR model consistently predicts a low mortality risk. The 3D surfaces of the LR model appear smooth and the model reacts to decreased eye and motor signals (Figure 3.8 a). In contrast, LSTM's 3D plots are less monotonic, exhibiting bumps (Figure 3.8). For the most severe cases (subscores being 1 or 2), LSTM's risk predictions incorrectly drop.

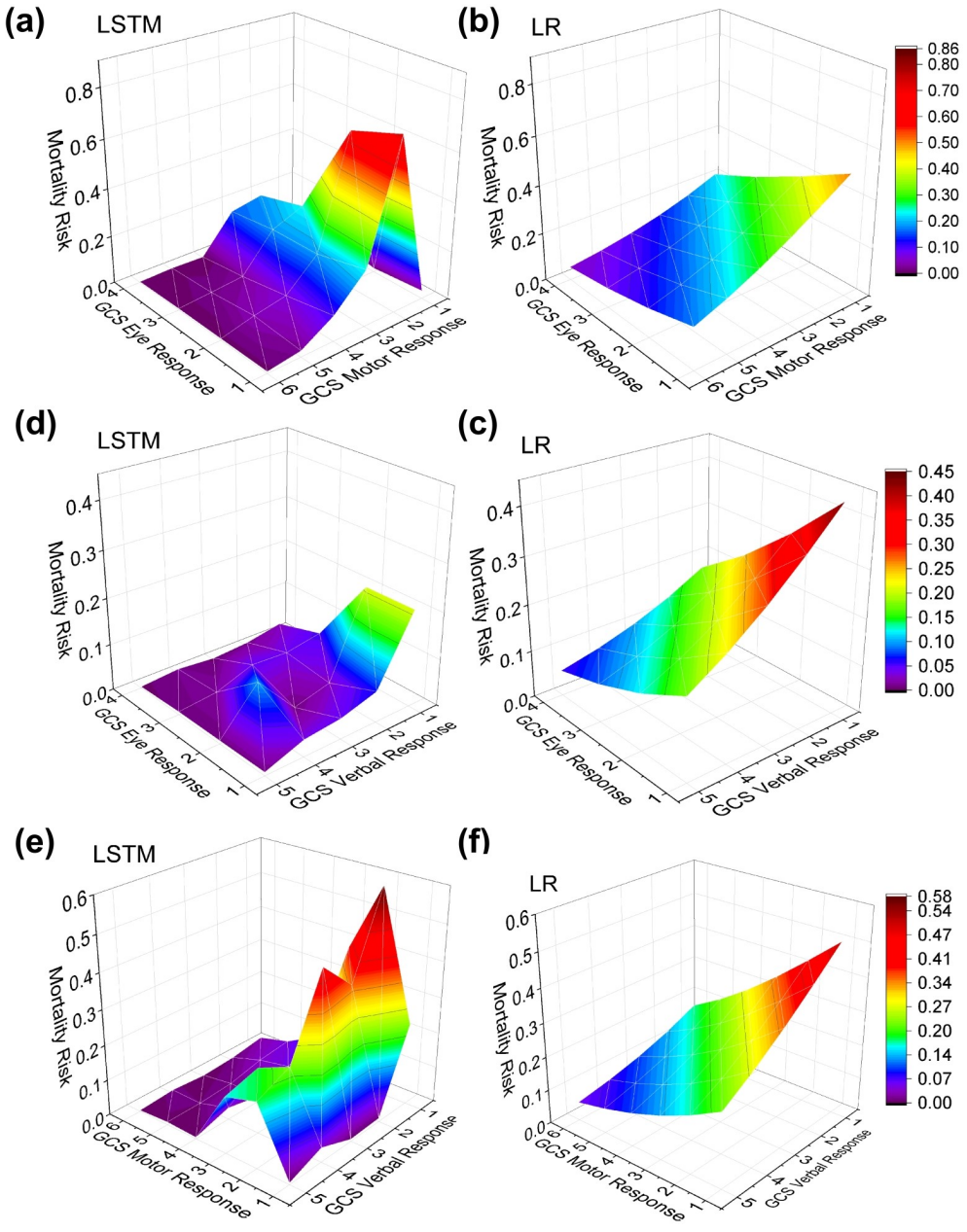


Figure 3.8: Mortality risk (MR) prediction for Glasgow Coma Scale for different combinations using three machine learning models. MR prediction of injury cases defined by different combinations of GCS eye and motor response scores by (a) LSTM and (b) logistic regression model. MR predicted by (c) LSTM and (d) logistic regression using injury cases defined by different combinations of GCS eye and verbal response scores. MR prediction of injury cases defined by different combinations of GCS motor and motor response scores by (e) LSTM and (f) logistic regression.

3.3.2 ML performance under critical zone tests

Single-attribute Critical Zone Test Results

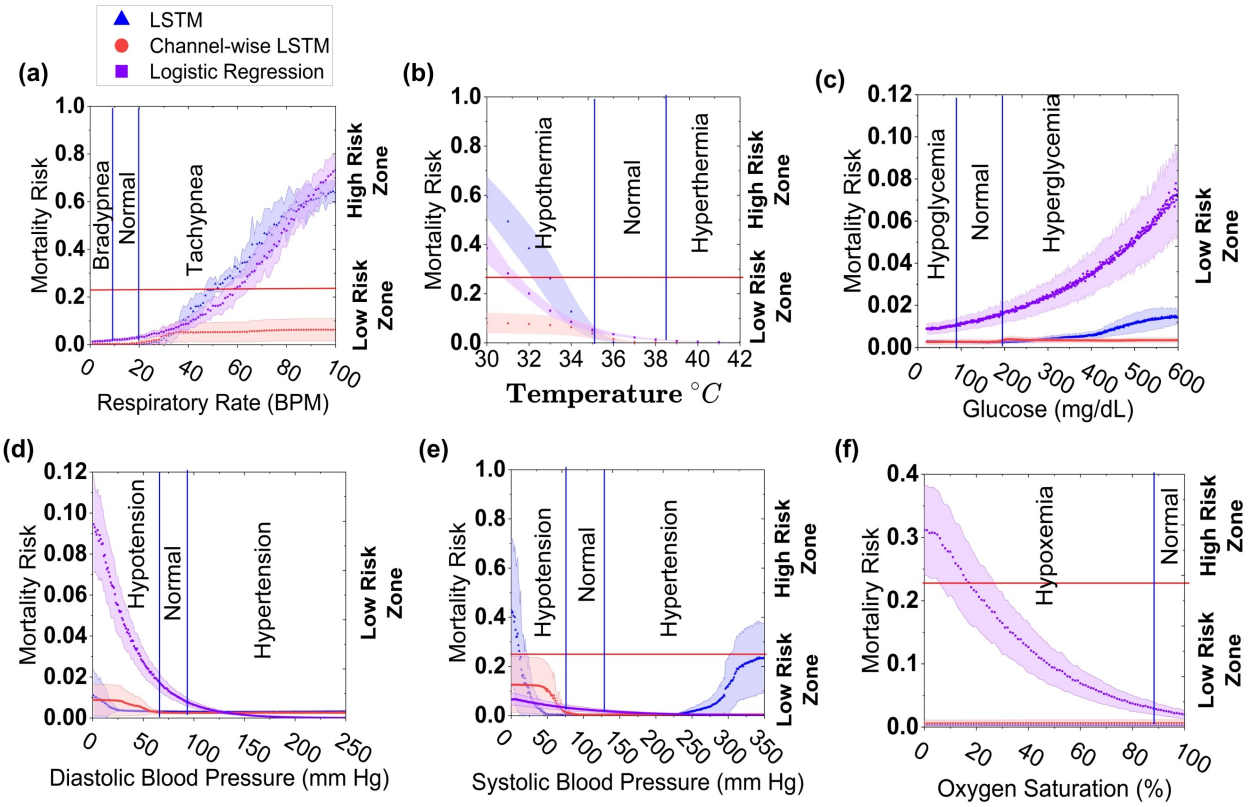


Figure 3.9: Mortality risk prediction for single vital-sign tests using three machine learning models (LSTM, Channel-wise LSTM, and Logistic Regression). LSTM, Channel-wise LSTM, and Logistic Regression (LR), predict the mortality risk (MR) of (a) respiratory rate, (b) body temperature, (c) glucose, (d) diastolic blood pressure, (e) systolic blood pressure, and (f) oxygen saturation test sets (synthesized). The mortality risk (MR) is represented by X-axis and MR above and below a red horizontal line (threshold = 0.22) indicates a high or low mortality risk zone respectively. The entire range of each vital sign (except oxygen saturation) value is divided into three segments, low, normal, and high, by the blue vertical lines. The low and high values within these ranges indicate critical health conditions.

We evaluate the MIMIC III-based LSTM, CW-LSTM, and LR models' ability to respond to a single deteriorating attribute while keeping other attributes stable as in the seed (Figure 3.9). The CW-LSTM model fails to recognize bradypnea, i.e., an abnormally slow breathing

rate and gives only slightly elevated mortality risk prediction (mean mortality risk 0.05 and standard deviation 0.04) for tachypnea, i.e., rapid breathing (Figure 3.9 a), insufficient to trigger an alert. Similarly, CW-LSTM is unable to recognize most of the abnormal vitals. Its mortality risk prediction gives a negligible change to an abnormal patient's glucose level (Figure 3.9 c) and oxygen saturation rate (Figure 3.9 f). For the other 3 attributes tested, CW-LSTM gives small partial responses to either a high critical zone or a low critical zone, but not both. For example, it is unable to recognize high body temperature anomalies and only slightly raises the mortality risk to 0.01 to 0.08 (standard deviation 0.033) for severe hypothermia below 34 degree C (Figure 3.9 b), which is still much below the classification threshold (0.22). CW-LSTM's response to abnormal diastolic and systolic blood pressure is also inadequate (Figures 3.9 d and e).

The LSTM model gives much more elevated risk prediction than CW-LSTM for tachypnea (Figure 3.9a) and hypothermia conditions (Figure 3.9b). Out of the 3 models tested, LSTM is the only machine-learning model that responds to both systolic hypotension and hypertension conditions, producing a U-shaped curve (Figure 3.9e). However, LSTM consistently gives an ultra-low risk prediction for abnormal diastolic blood pressure (Figure 3.9d). Similar to CW-LSTM, LSTM does not recognize hypoxemia, i.e., low blood oxygen level (Figure 3.9f), bradypnea, hyperthermia, and abnormal glucose level (Figure 3.9c), exhibiting either monotonic or near-flat risk prediction curves insensitive to abnormal vitals. Compared to the other 2 models, logistic regression gives a substantially higher risk prediction for hypoxemia. It also computes elevated risk scores in response to increasing hyperglycemia and diastolic hypotension conditions. For all attributes, logistic regression is only able to recognize one end of the critical zones, but not both. Overall, logistic regression, LSTM, and CW-LSTM correctly predict 37.7%, 37.8%, and 22.4% of the single-attribute critical zone test cases on average (Table A.17).

Neuron activation analysis

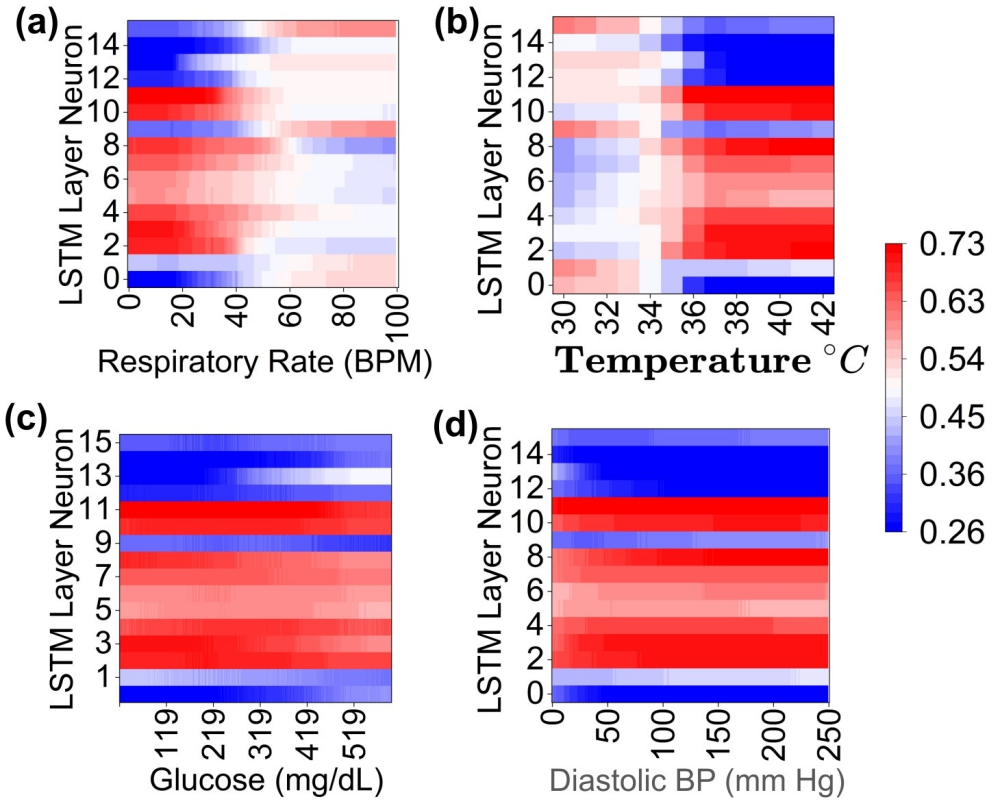


Figure 3.10: Visualizing the neural activation map of the LSTM layer (LSTM model) consisting of 16 neurons. Figures (g)-(j) represent the neural activation map. These are the neural activation values, calculated after applying the sigmoid function, when the model is fed with test cases varying a single vital, such as (g) glucose, (h) diastolic blood pressure, (i) temperature, and (j) respiratory rate.

We visualized neuron output from intermediate layers of the MIMIC III-based LSTM model. Neurons whose activations change with changing variable values are the responsible neurons for recognizing that variable. We found most of the LSTM neurons have low or no responses to varying glucose and diastolic blood pressure values (Figures 3.10i and j). In contrast, neurons are more responsive to temperature and respiratory rate changes, e.g., sharp changes in all neuron activation between 34°C to 36°C (Figure 3.10h) and around or above 40 bpm (Figure 3.10g). However, neurons exhibit minimal or no changes in activation for higher

temperatures or for critically low respiration rates.

To quantify changes in neuron activation, we computed Neural Zone Activation (NZA) and average zone difference NAZ, new metrics defined by us (Equations 3.1 and 3.2). NZA averages neurons' activations within a zone, where a zone is a critically low, critically high, or normal range (Equation 3.1). NAZ computes the averaged NZA difference between zones (Equation 3.2), indicating how much neurons react to zone changes. The LSTM model shows low (0.01 to 0.04) ΔNZA in most cases (Table A.18). In a few cases, e.g., temperature ΔNZA(low, normal) and respiratory rate ΔNZA(high, normal), the values are relatively high (0.14 to 0.16).

Multi-attribute Critical Zone Test Results

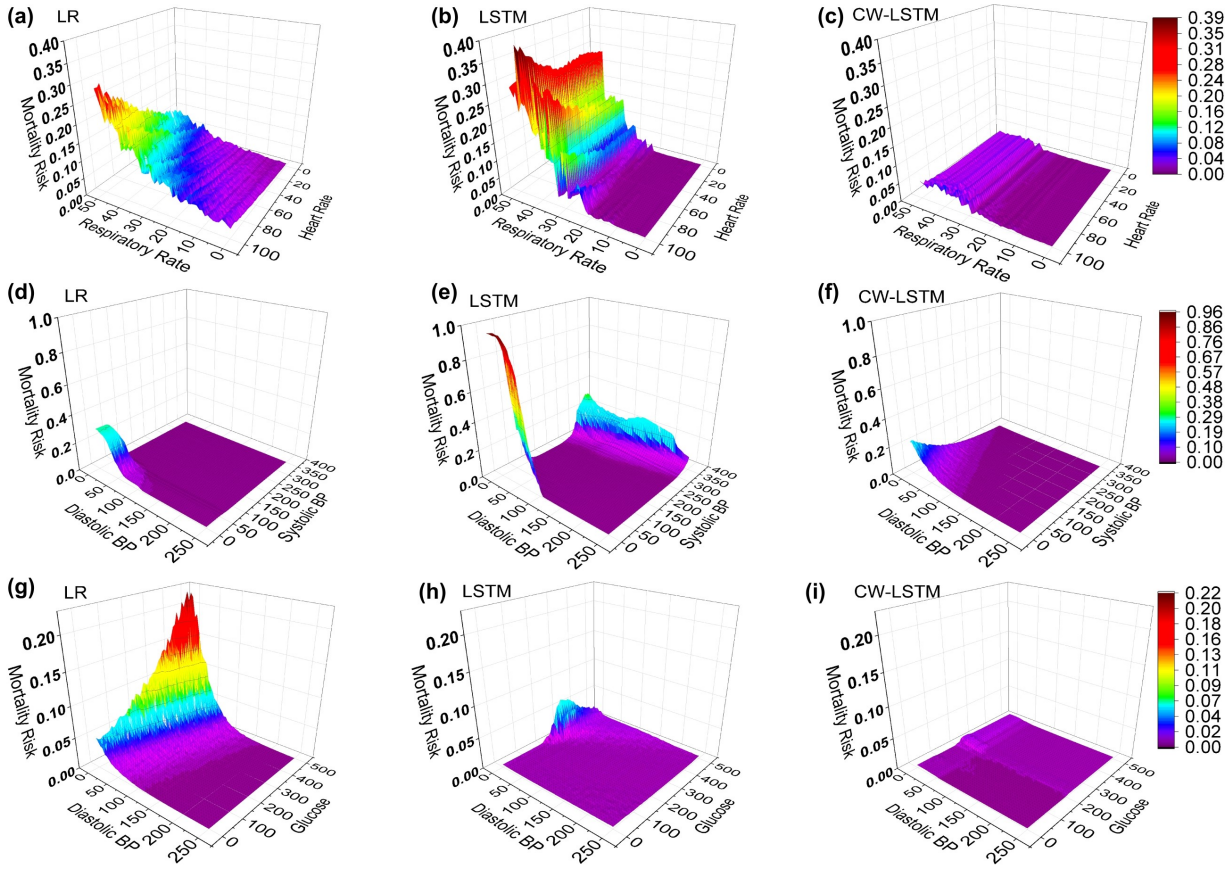


Figure 3.11: (a)-(i) show the mortality risk prediction of patients attributed by double vital signs and test pair generated by altering 6 attributes at the same time. Risk prediction under varying respiratory rate and heart rate by (a) logistic regression, (b) LSTM model, and (c) CW-LSTM model. Risk prediction under varying systolic and diastolic blood pressure by (d) logistic regression, (e) LSTM model, and (f) CW-LSTM model. Risk prediction under varying glucose and diastolic blood pressure by (g) logistic regression, (h) LSTM model, and (i) CW-LSTM model.

We evaluated the 3 MIMIC III-based machine learning models under 42,500 double attribute varying test cases (Figure 3.11), including respiratory rate and heart rate pair (first row), systolic and diastolic blood pressure pair (middle row), and glucose and diastolic blood pressure pair (last row). The CW-LSTM model does not generate high mortality risk predictions for most critical zone cases, consistent with its single-attribute test performance in Figure 3.9.

The logistic regression model gives better performance than CW-LSTM, predicting higher risks for some critical zone combinations (Figures 3.11a, d, and g). However, its prediction is monotonic, thus, unable to recognize both high and low critical zones of an attribute pair. For example, logistic regression fails to alert when patients have low respiratory rate and low heart rate. LSTM model exhibits prediction behaviors (Figures 3.11b, e, and h) consistent with its single attribute performances in Figure 3.9.

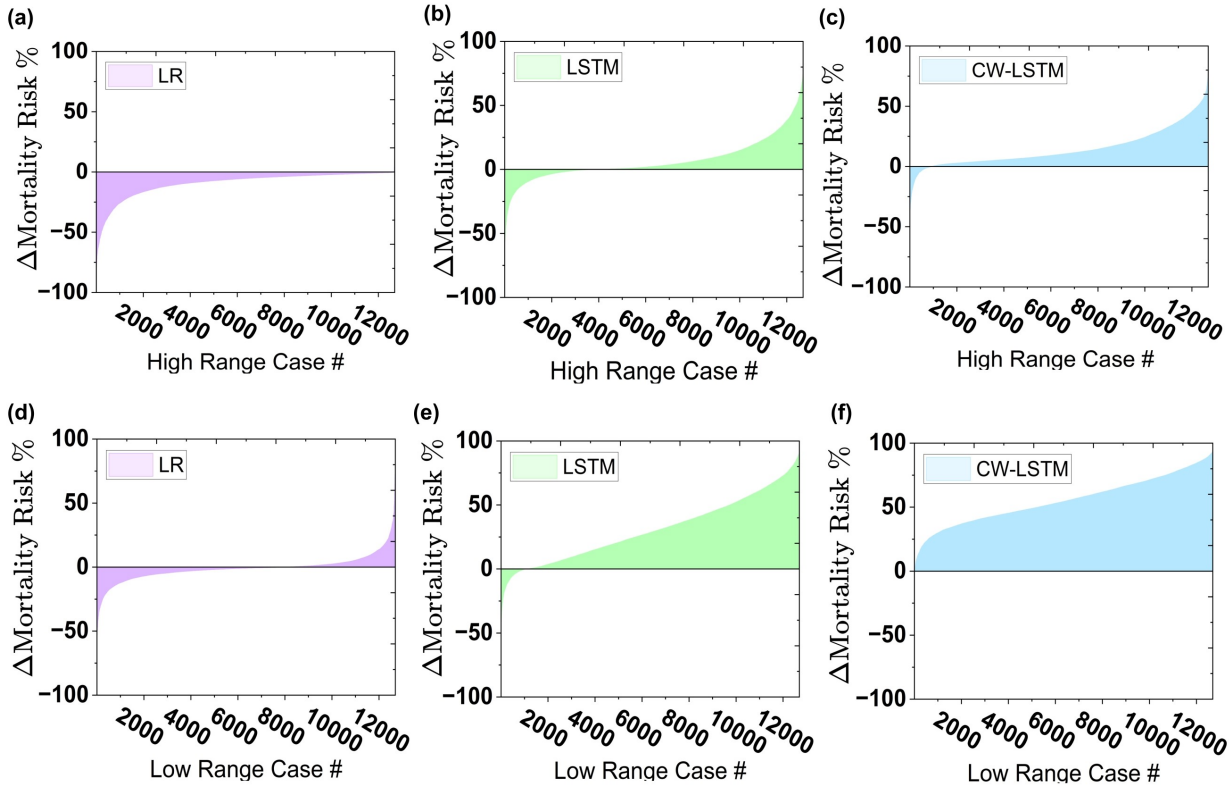


Figure 3.12: This shows the prediction difference between original case and it's corresponding critical version by multi-attribute variation. (j), (k), and (l) represent ΔMR for high critical range cases and (m), (n), and (o) represent ΔMR for low critical range cases. The test set is generated by simultaneously varying systolic blood pressure, diastolic blood pressure, blood glucose level, respiratory rate, heart rate, and body temperature and values are randomly selected from the critical zone. The graph shows mortality risk difference (ΔMR) calculated by subtracting predicted mortality risk of the seed data from the predicted mortality risk of its corresponding critical case. The X-axis represents the case numbers and the Y-axis represents ΔMR. It is expected to get a positive MR difference and the negative ΔMR cases represent failed test cases.

In a 6-attribute varying test setting, we evaluated the responsiveness of MIMIC III-based machine learning models under 6 changing vitals, where test cases have abnormal systolic and diastolic blood pressures, blood glucose level, respiratory rate, heart rate, and body temperature values in their respective critical zones. We recorded how much mortality risk scores changed and showed the distributions in Figures 3.12. Figures 3.12a, b, and c show the mortality risk difference (ΔMR) between each high critical zone test case and its corresponding seed. Medically speaking, the risk should increase under worse health conditions. The CW-LSTM model consistently predicts high mortality risk for most cases, resulting in a positive ΔMR for over 90% of the cases (Figures 3.12c and 4f). The logistic regression model (Figure 3.12a) consistently produces negative ΔMR for all cases, which is incorrect. The LSTM model generates positive ΔMR for two-thirds of the 12,694 test cases.

The 3 models under low critical zone tests performed similarly (Figures 4m, 4n, and 4o), where CW-LSTM responds to multi-attribute critical conditions the most effectively and logistic regression the least. Overall, logistic regression, LSTM, and CW-LSTM correctly predict 6.2%, 45.7%, and 69.3% of the multi-attribute critical zone test cases on average (Table A.17), respectively.

3.3.3 Results on test cases with deteriorating conditions

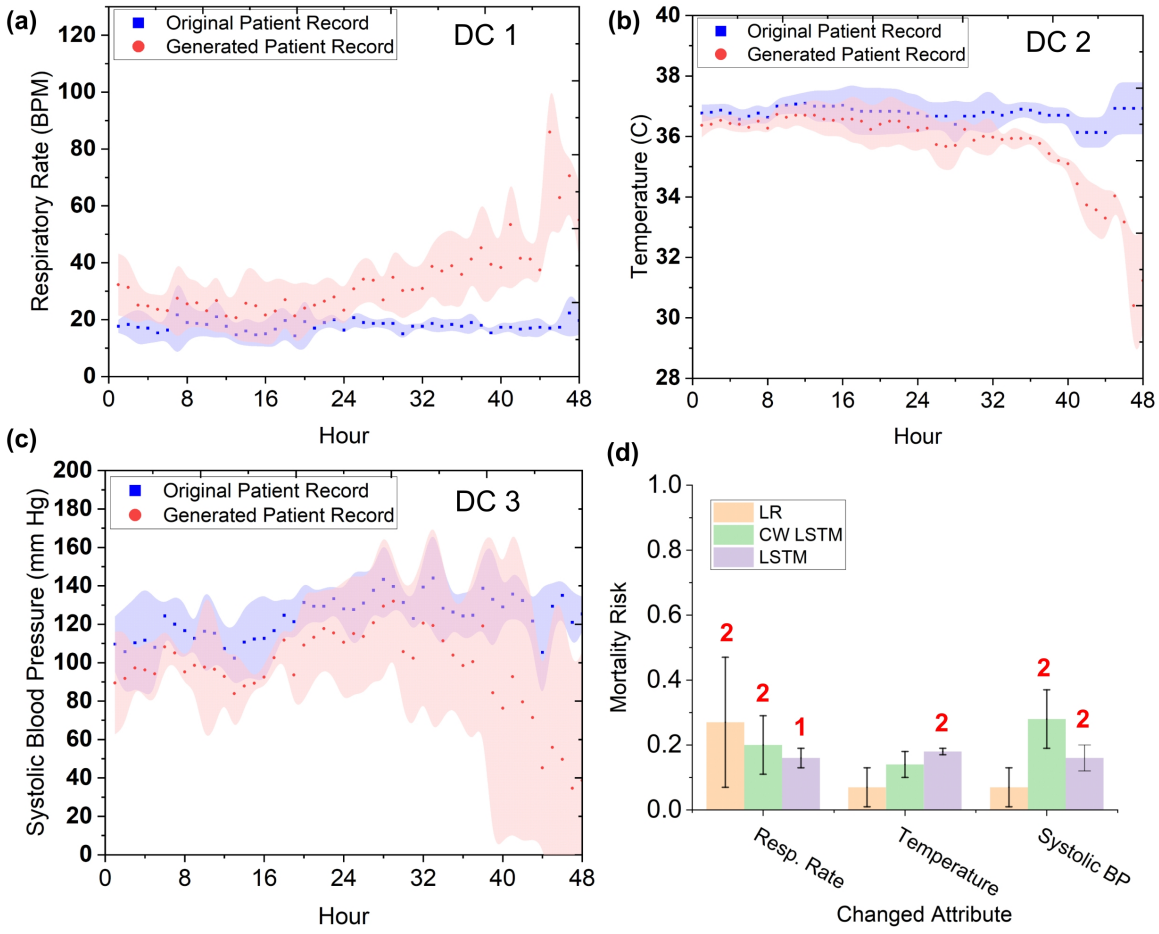


Figure 3.13: Gradient-generated deteriorating test cases and machine learning models' mortality risk predictions by LR, CW-LSTM, and LSTM models. (a)-(c) show the average time series of the generated abnormal test cases (in red area curves) and the normal seed cases used (in blue area curves) for each of the 3 attributes. (d) Models' predicted average mortality risks for each deteriorating attribute. The standard deviation is indicated by the error bar. The numbers (red) at each of the bars represent the number of detected alerts out of input 3 cases.

For in-hospital mortality prediction (using the MIMIC III dataset), we used a gradient ascent method to generate 12 time-series test cases with deteriorating health conditions. 9 of the 12 test cases contain one vital that worsens during the 48 hours and is in the critical zone

during the last 24 to 48 hours, including 3 cases of decreasing systolic blood pressure (Figure 3.13a), 3 cases of increasing respiratory rate cases (Figure 3.13b), and 3 cases of decreasing body temperature (Figure 3.13c). The other 3 test cases have multiple (3) worsening vital signs, with vitals being in critical zones during the last hours (ranging from 30 to 48 hours). All 12 test cases should receive a high mortality risk prediction, i.e., Class 1. We confirmed these labels with two medical doctors who manually reviewed the time series data. Average mortality risks predicted by machine learning models on the 9 single-attribute deteriorating test cases are in Figure 3.13d. Out of the 9 single-attribute deteriorating test cases, logistic regression only detects 2 (22%) respiratory rate cases (average risk 0.38) and fails to detect the other 7. CW-LSTM (MIMIC III) detects 4 (44%) out of the 9 deteriorating test cases, including 2 systolic BP cases (average risk 0.32) and 2 respiratory rate cases (average risk 0.39). However, it is unable to detect deteriorating temperature cases. LSTM reports 5 (56%) out of the 9 deteriorating test cases, including 2 systolic BP cases (average risk 0.22), 2 temperature cases (average risk 0.23), and 1 respiratory rate case (risk 0.22). Collectively, the models detect 41% out of single-attribute deteriorating test cases. Multi-attribute test cases have 3 deteriorating vitals, including oxygen saturation, temperature, and diastolic blood pressure. The LSTM (MIMIC III) model detects all 3 cases (average risk 0.23), whereas CW-LSTM fails to generate any alerts. Logistic regression (MIMIC III) issues alerts for 2 out of 3 cases (average risk 0.56). Collectively, the 3 models detect 56% out of the multi-attribute deteriorating test cases. Overall, the models' average accuracy under all deteriorating test cases is 44%.

3.3.4 5-year cancer survivability results

We found similar deficiencies in the machine learning model, in terms of the model's ability to respond to test cases representing serious cancer conditions.

Single-attribute test results

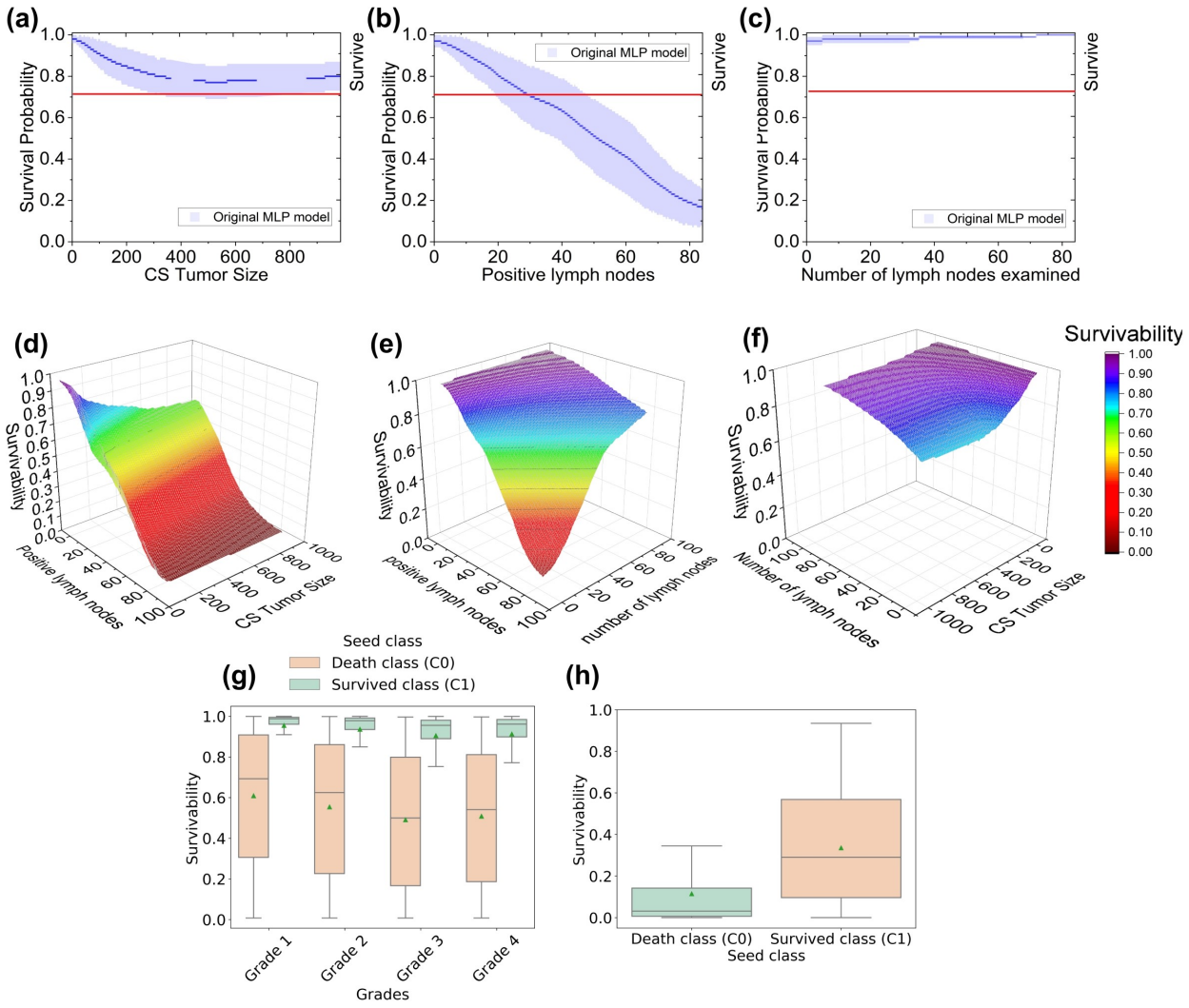


Figure 3.14: Predicted 5-year breast cancer survivability results of a multi-layer perceptron (MLP) model on test cases. Four major breast cancer screen attributes are involved, including CS tumor size, number of positive lymph nodes, number of lymph nodes examined, and grade. (a)-(c) and (g) Predicted survivability results on single-attribute varying test cases. The blue area of (a)-(c) represents the standard deviation. (d)-(f) Predicted survivability results on double-attribute varying test cases. (h) Predicted survivability results on triple-attribute varying test cases involving CS tumor size, number of positive lymph nodes, and grade. In the boxplots (g) and (h) the horizontal line within the box represents the median value, while the box itself encompasses the interquartile range (IQR), containing the middle 50% of the data. The whiskers extend to the values within 1.5 times the IQR from the box (upper and lower quartiles). The green triangle point on the box represents the mean of the distribution.

We evaluated the responsiveness of MLP models trained on the BCS dataset to a single deteriorating attribute (Figure 3.14 and Table A.19) while keeping other attributes the same as the seed. The BCS-MLP model shows some responsiveness with varying tumor size (Figure 3.14a), however, remains above the survivability threshold (0.71) in all cases. As a result, the model fails to trigger an alert for tumor sizes representing critical stage T1 (tumor size less than 20 mm) to T3 (tumor size larger than 50 mm). On the other hand, the model triggers alerts for 74.4% of the 6,780 N3 stage (Figure 3.14b). It fails to generate any alert for 270 N1 and 546 N2 stage cases. The BCS-MLP model accurately generates alerts for 66.4% of critical cases (N1-N3). It decreases the survivability for lower numbers of examined lymph nodes (ELNs), however, still fails to trigger any alerts (Figure 3.14c). For the grade test sets (1-4) generated from 21,723 surviving patient seeds, the BCS-MLP model generates alerts for 989 cases (4.6%) for grade 2 test, 1,616 cases (7.4%) for grade 3 test, and 1,446 cases (6.7%) for grade 4 test (Figure 3.14g). On the other hand, for the grade test generated from 3,152 death events, the model generates alerts for 1,826 cases (57.9%) for grade 2 test, 2,073 cases (65.8%) for grade 3 test, and 2,013 (63.9%) cases for grade 4 test. The model did not generate any alerts for 97% and 48.7% of grade 1 cases generated from seeds of survived and death events, respectively.

The LCS-MLP model shows higher responsiveness with variations in tumor size (Figure 3.18 d), generating alerts in 80.1% of the test cases (Figure 3.20). It also responds well to varying positive lymph node numbers (Figure 3.18 e) with alerts generated in 92.9% of the cases. However, the LCS-BCS model does not react to the increasing number of lymph nodes examined (Figure 3.18 f).

Double-attribute test results

The BCS-MLP model was tested under 60,462 double attribute varying test cases, including i) tumor size (T) and positive lymph node (N) combination test (Figure 6d), ii) number of examined lymph node (ENL) and positive lymph node (N) combination test (Figure 6e), and iii) tumor size (T) and number of examined lymph nodes (ENL) combination test (Figure 6f). The predicted survivability decreases with an increasing number of positive lymph nodes. However, collaborative staging (CS) tumor size or number of examined lymph nodes does not significantly decrease the predicted survivability. The MLP model accurately predicts 93% of T-N cases, 19.6% of N-ENL cases, and 0% of T-ENL cases (Table A.19). In a 3-attribute varying test, the BCS-MLP model is evaluated using cases with T4 tumor size, N3 number of positive lymph nodes, and grade 4 condition at the same time (Figure 6h). The BCS-MLP accurately predicts 90% of cases generated from surviving seeds (Class 1) and 98.9% of cases generated from death event seeds (Class 0).

Tree-based ensemble

AdaBoost, Random Forest, and XGBoost, produced somewhat similar results across all four datasets (Figure 3.15). The random forest model demonstrates good responsiveness for most attributes in both MIMIC III and eICU datasets. However, it does not respond to critically high respiratory rates. XGBoost and AdaBoost have little reaction to attribute changes. In cancer survivability tasks, none of the models responds to worsening patient conditions, except AdaBoost for LCS-positive lymph nodes.

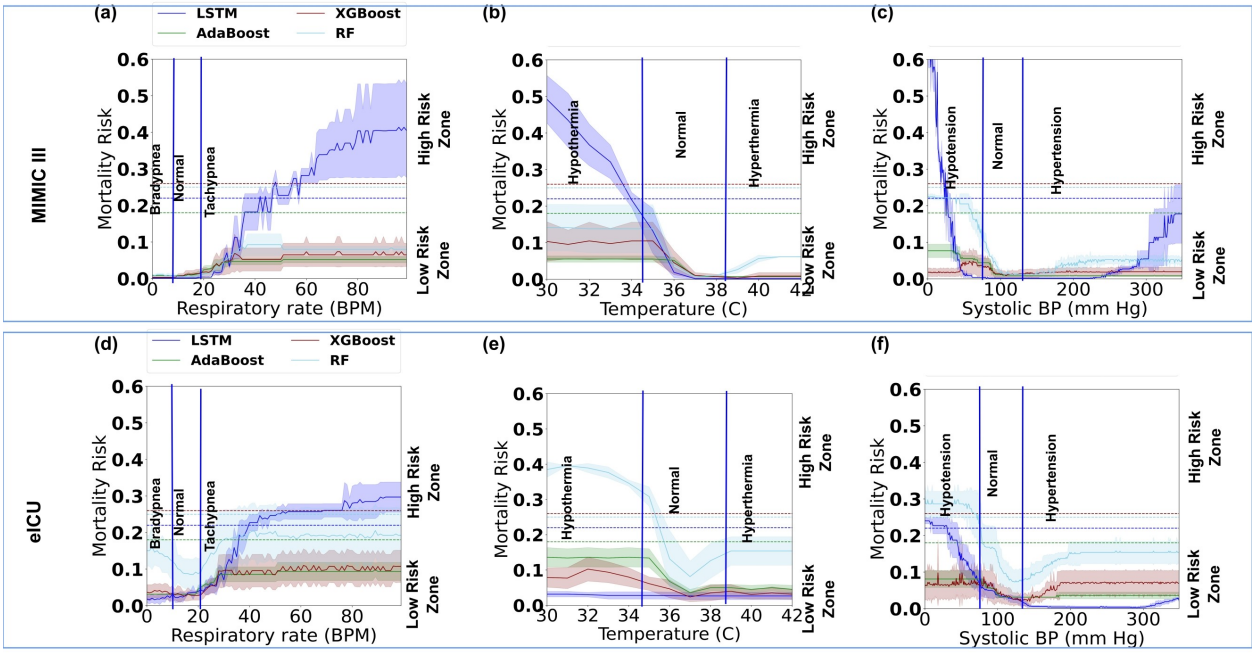


Figure 3.15: Performance comparison between tree-based ensemble methods, AdaBoost, XGBoost, and Random Forest (RF), with LSTM and MLP models under single-attribute varying tests. (a)-(c) and (d)-(f) Mortality risk prediction results by the models under MIMIC-III and eICU test cases for respiratory rate, temperature, and systolic blood pressure, respectively. Horizontal dashed lines represent model-specific thresholds.

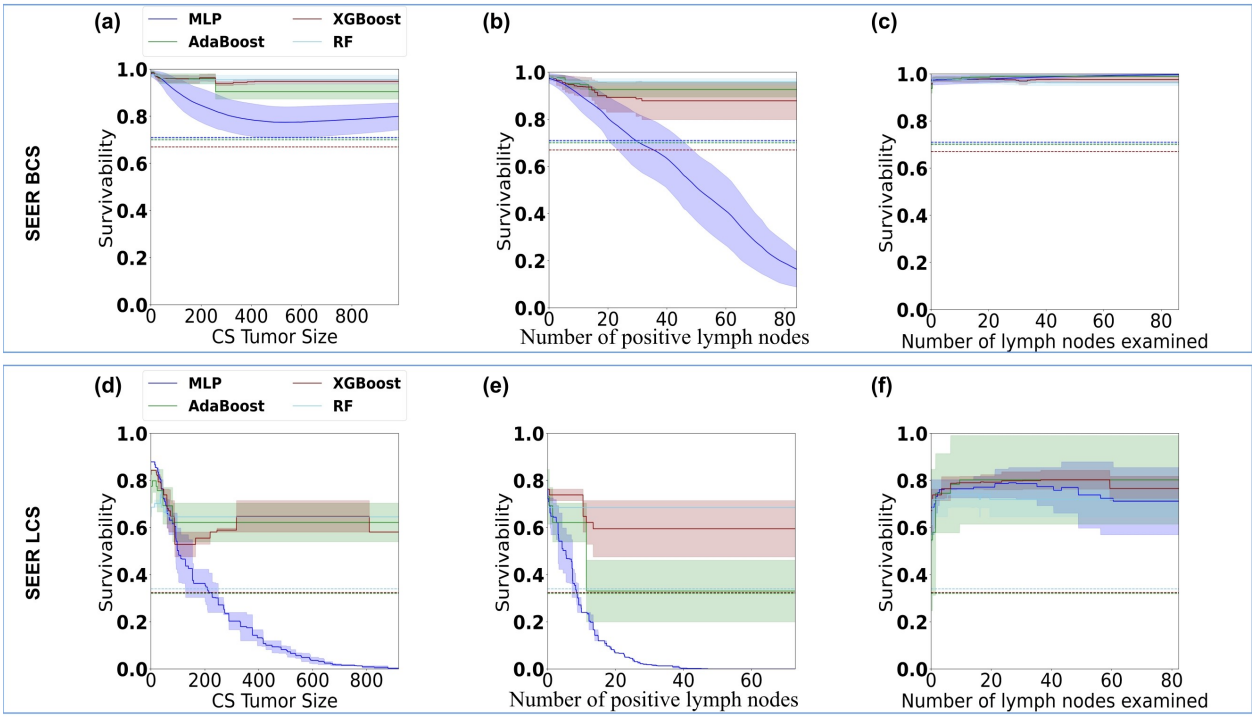


Figure 3.16: Performance comparison between tree-based ensemble methods, AdaBoost, XGBoost, and Random Forest (RF), with LSTM and MLP models under single-attribute varying tests. (a)-(c) and (d)-(f) 5-year cancer survivability prediction results by the models under SEER BCS and LCS test cases for CS tumor size, the number of positive lymph nodes, and the number of lymph nodes examined, respectively. Horizontal dashed lines represent model-specific thresholds.

3.3.5 Comparison of Wasserstein distances

We computed the Wasserstein distance between the original dataset and the generated test cases. Wasserstein distance captures the probability distribution shift given a metric space [53, 54]. For in-hospital mortality prediction, the Wasserstein distance between the original MIMIC-III training set and the synthesized multi-attribute tests is 33.4. This value is much larger than the Wasserstein distance (12.4) between the training data and test data split within the original MIMIC-III. In comparison, for breast cancer survivability prediction, the distribution shift of the generated triple-attribute-based test cases from the original SEER

dataset is smaller, with the Wasserstein distance being 9.8. WD for the original SEER training and test set is 2.1 (Table A.20).

3.3.6 Impacts of resampling and reweighting methods

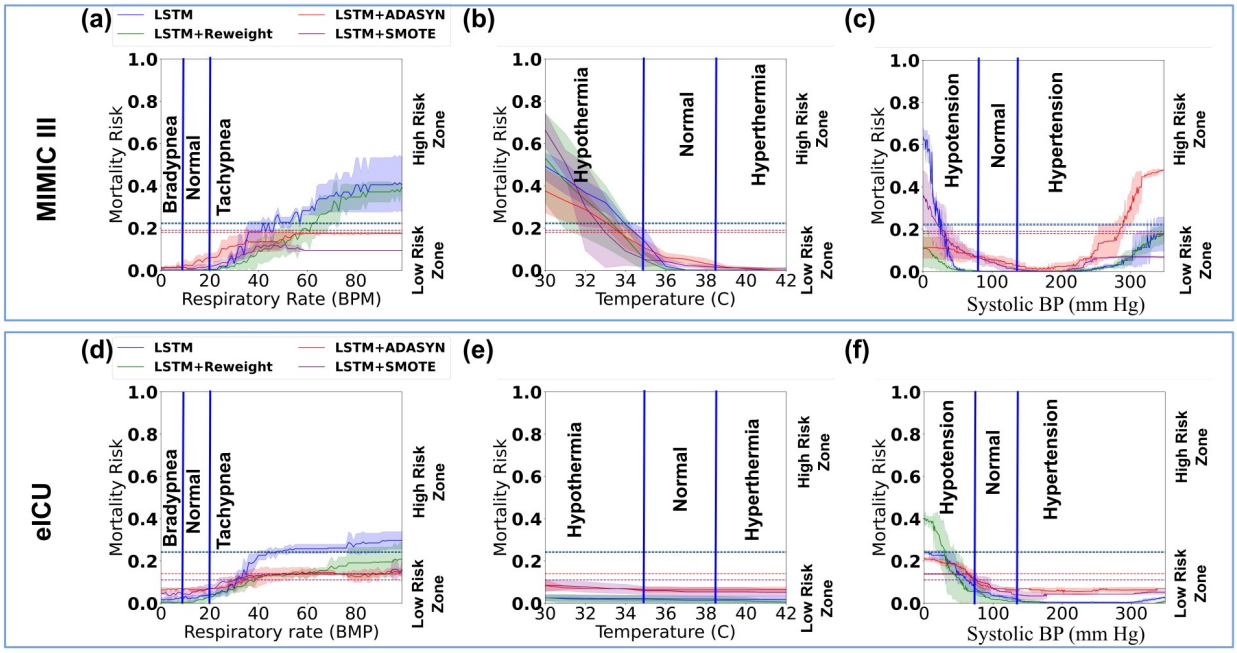


Figure 3.17: Performance comparison between the original machine learning models and the resampled (SMOTE or AdaSyn) or reweighted models under single-attribute varying tests. (a)-(c) and (d)-(f) Mortality risk prediction results by the original LSTM model and the resampled or reweighted LSTM models under MIMIC-III and eICU test cases for respiratory rate, temperature, and systolic blood pressure, respectively. Horizontal dashed lines represent model-specific thresholds.

We trained and tested new ML models to assess the impact of resampling and reweighting methods on models' responsiveness. SMOTE and AdaSyn oversampling methods are used to enrich the minority prediction class. For mortality prediction, resampled LSTM models are tested with our single-attribute critical zone test cases. Overall, the new models remain to have low responsiveness to high-risk patient conditions (Figure 3.17 a-c). Similar to the

original models, models with resampling are still unable to recognize critical patient conditions. For example, LSTM with SMOTE consistently assigns low mortality risk scores to patients with critically high vitals (e.g., respiratory rate, temperature, systolic blood pressure). LSTM with AdaSyn is better at responding to elevated systolic blood pressures than the original model, however, it performs poorly in other tests. Tests with the eICU dataset give a similar or worse performance (Figures 3.17 7d-7f). For BCS and LCS prediction, new MLP models trained with SMOTE or AdaSyn oversampling methods exhibit similar trends as the original MLP model (Figures 3.17). The new models fail to recognize many critical cancerous conditions. In addition, for LCS prediction, SMOTE and AdaSyn methods make the LCS-MLP model less sensitive to increasing CS tumor size (Figure 3.18 d).

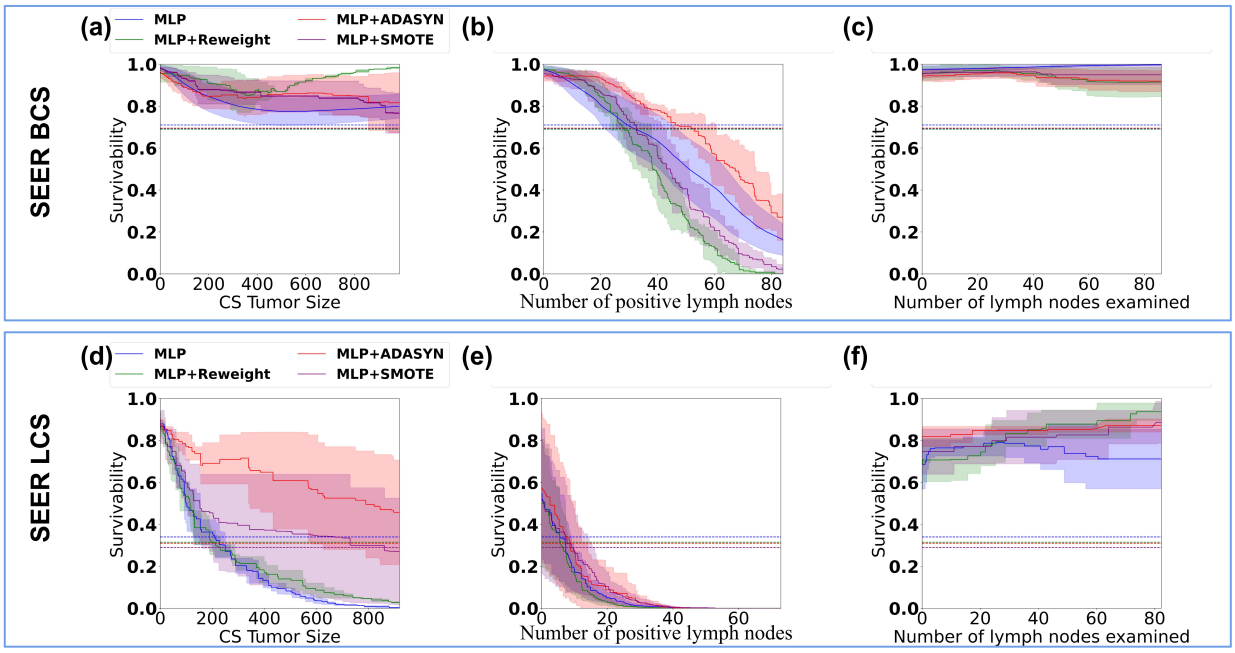


Figure 3.18: Performance comparison between the original machine learning models and the resampled (SMOTE or AdaSyn) or reweighted models under single-attribute varying tests. (a)-(i) and (j)-(l) 5-year cancer survivability prediction results by the original MLP model and the resampled or reweighted MLP models under SEER BCS and LCS test cases for CS tumor size, the number of positive lymph nodes, and the number of lymph nodes examined, respectively. Horizontal dashed lines represent model-specific thresholds.

We also applied the reweighting approach to training. Table A.4 shows the cost parameters used. For mortality prediction, LSTM with reweighting gives comparable performance to the original model (Figures 7a-7g), except in one testing scenario. For eICU critically low systolic BP tests, reweighted LSTM generates elevated risk scores and is slightly better at responding to abnormal patient conditions. However, reweighted LSTM performs worse for similar MIMIC-III test cases (Figure 7c). For cancer survivability prediction, reweighting does not impact MLP's performance in most testing scenarios (Figure 7). For BCS test cases, the reweighted MLP model has slightly better responses to the increasing number of positive lymph nodes than the original MLP, however, it performs worse than the original MLP in terms of recognizing larger tumor sizes.

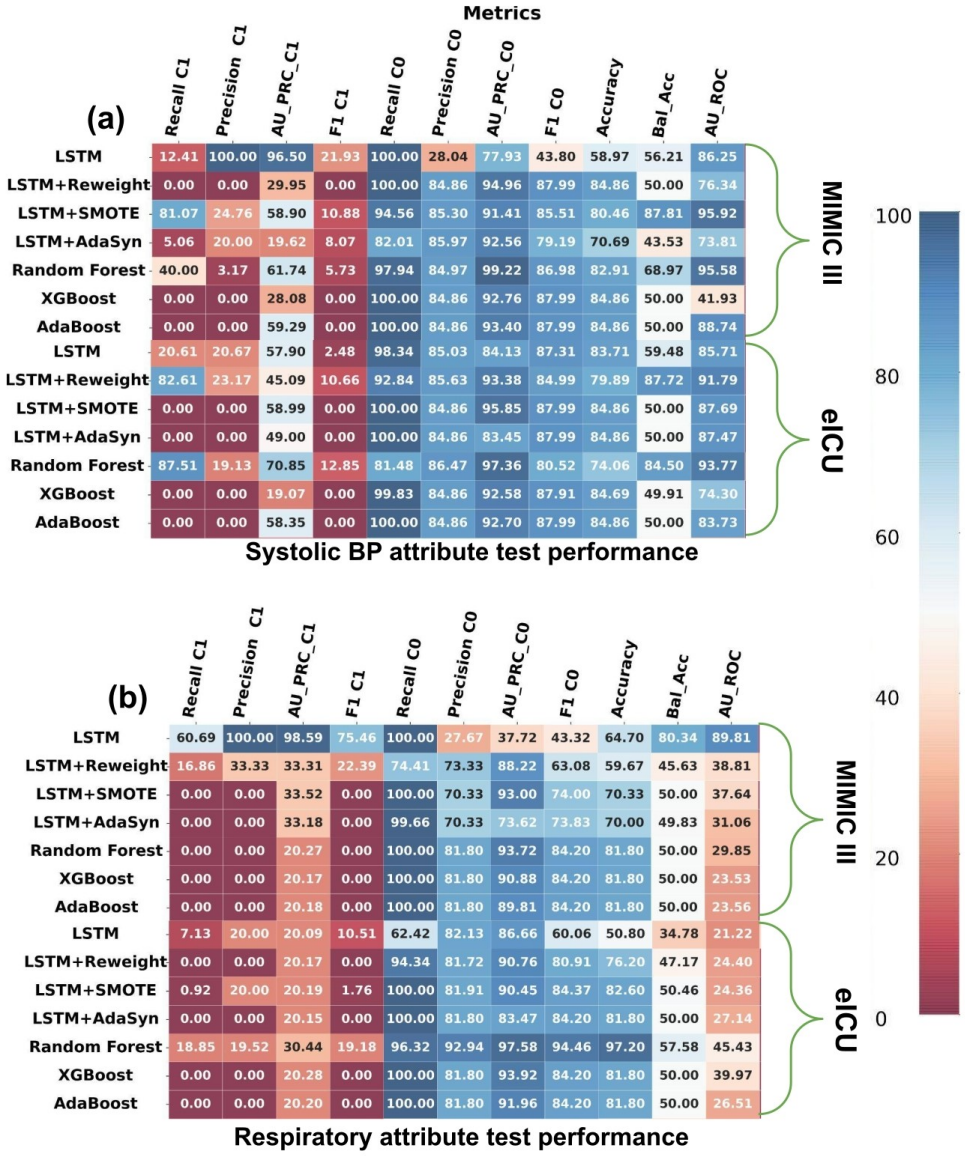


Figure 3.19: The performance of machine learning models was evaluated on synthesized single-attribute test sets, including (a) ICU vital systolic blood pressure and (b) ICU vital respiratory rate. Models were trained on the original datasets, with the top halves of (a) and (b) representing MIMIC III and the bottom halves representing eICU. Model names are aligned along the Y-axis. For ICU mortality prediction (a) and (b), Class 1 corresponds to the death class.

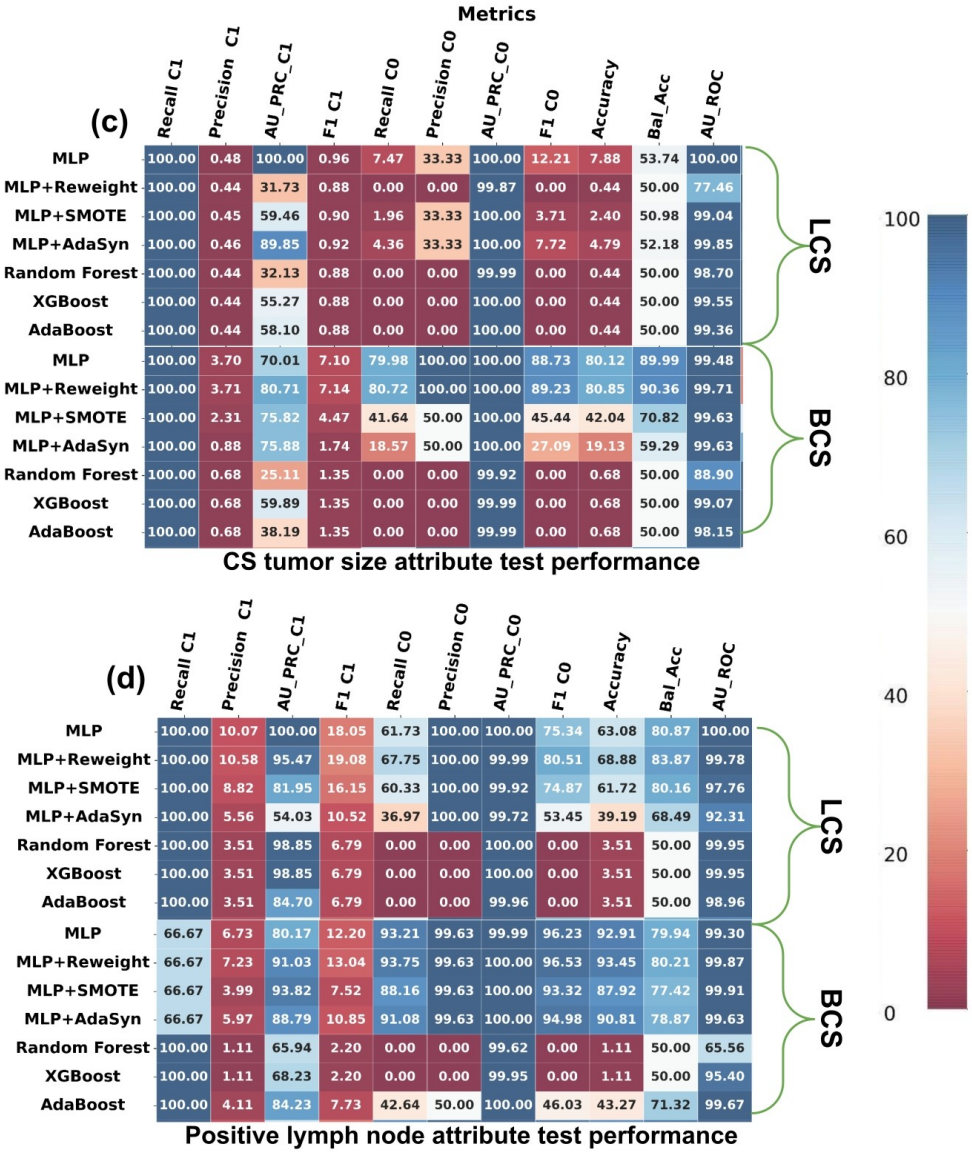


Figure 3.20: The performance of machine learning models was evaluated on synthesized single-attribute test sets, including (a) cancer attribute CS tumor size, and (b) cancer attribute number of positive lymph nodes. Models were trained on the original datasets, with the top halves of (c) and (d) representing SEER BCS and the bottom halves representing SEER LCS, as indicated on the left. Model names are aligned along the Y-axis. For SEER cancer survivability prediction (a) and (b), Class 0 represents the death class.

3.3.7 Responsiveness results of transformer models

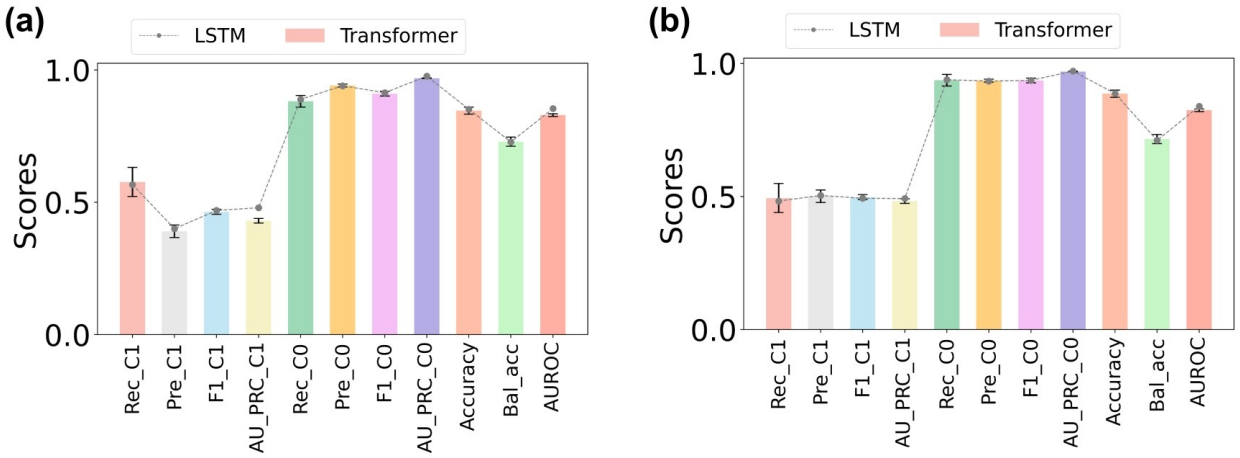


Figure 3.21: The performance of the transformer model compared with the LSTM. Figures (a) and (b) show the various Class 1 (death) and Class 0 (survival) performance metrics of the transformer model trained and tested on the original MIMIC-III and eICU datasets, respectively, with error bars indicating the standard deviation from three experimental trials. The dashed line represents the performance of the LSTM model. `Rec_C1`, `Pre_C1`, `F1_C1`, `AU_PRC_C1`, `Rec_C0`, `Pre_C0`, `F1_C0`, `AU_PRC_C0`, `Accuracy`, `Bal_Acc`, and `AUROC` stand for Recall Class 1, Precision Class 1, F1 score Class 1, Area Under the Precision-Recall Curve Class 1, Recall Class 0, Precision Class 0, F1 score Class 0, Area Under the Precision-Recall Curve Class 0, Accuracy, Balanced Accuracy, and Area Under the Receiver Operating Curve, respectively.

The transformer models exhibit more responsiveness than LSTM in mortality prediction though the performance on original test set are similar (Figure 3.22). They show elevated response in critically high zones for respiratory rate and systolic blood pressure, as well as in the critically low zone for temperature (Figure 3.22). This trend is observed for the single-attribute test cases of both MIMIC-III and eICU datasets. In addition, transformer models recognize both critical zones of systolic blood pressure, yielding a desired U-shaped response curve (Figures 3.22 e and h). However, the transformer models fail to recognize critically low respiratory rates and critically high temperatures and exhibit low responsiveness to critically low systolic blood pressure. It also has delayed response to abnormally high respiratory rates.

The transformer model's risk prediction fluctuates significantly for eICU test cases.

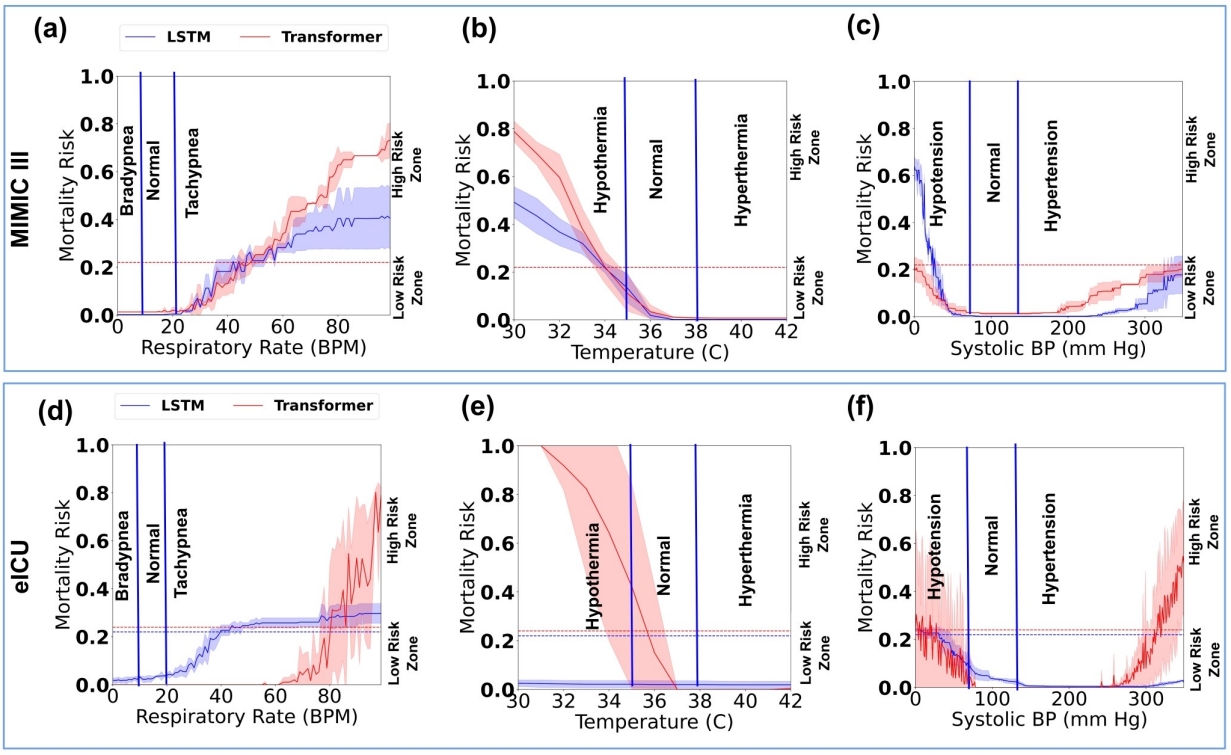


Figure 3.22: Responsiveness of the transformer model compared with LSTM. Figures (a)-(c) show the predicted mortality risk by the transformer model for respiratory rate, temperature, and systolic blood pressure on MIMIC-III single-attribute test cases, while (d)-(f) show the same for eICU test cases. Horizontal dashed lines denote model-specific thresholds for mortality risk prediction.

3.4 Discussion

3.4.1 Need for measuring machine learning (ML) responsiveness and metrics

Our findings highlight the importance of measuring how clinical machine learning (ML) models respond to serious patient conditions. Our results show that most ML models tested

are unable to adequately respond to patients who are seriously ill, even when multiple vital signs are extremely abnormal. For time-sensitive in-hospital mortality prediction, the lack of response to disease conditions is particularly troublesome. ML responsiveness is somewhat related to feature importance in some cases, e.g., the low responsiveness of LSTM to oxygen saturation tests (Figure 3.9f and Table A.17) is consistent with that feature's low (15th) ranking (Figure 3.23). However, for high-ranking features such as glucose and temperature, ML responsiveness to them is still inadequate. This poor responsiveness is also observed in the lack of responses in neural activation values (Figure 3.10 and Table A.18) to important vital changes, such as extremely low respiratory rate or high body temperature.

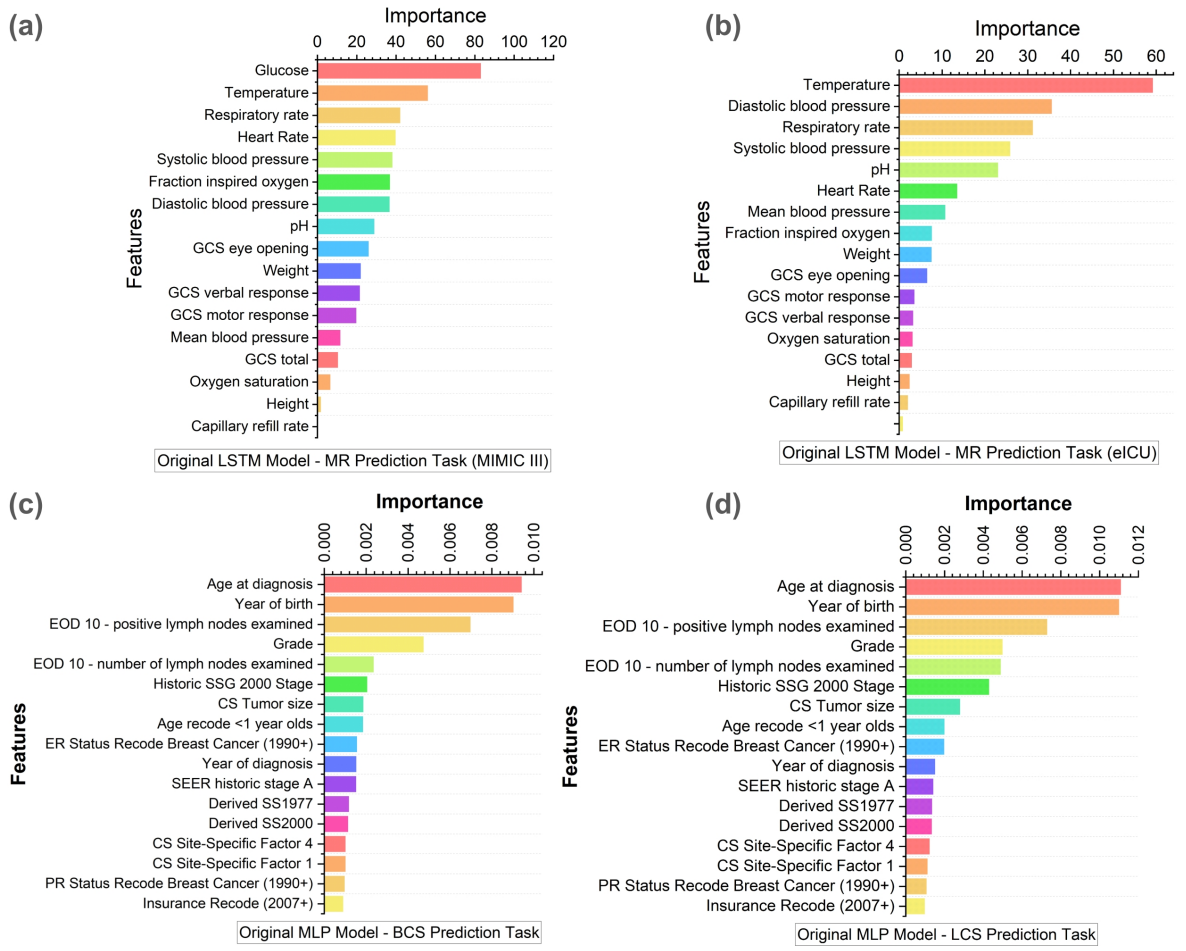


Figure 3.23: SHAP-average (of one-hot encoded features) feature importance of the (a) LSTM trained on MIMIC III dataset, (b) LSTM trained on eICU dataset, (c) MLP trained on BCS dataset. (d) MLP trained on LCS dataset.

New ML responsiveness metrics, especially for the healthcare domain, are urgently needed. ML responsiveness is a new problem. It differs from the well-studied ML robustness [68]. ML robustness aims to ensure model stability and the ability to resist sample perturbations so that small (maliciously injected) noises to samples cannot change the prediction results. Lipschitzness, a common ML robustness metric, measures the model's resilience to noisy data and perturbations [69]. However, for healthcare applications, optimizing Lipschitzness may lead to models being even more insensitive to changes in patient conditions, as adherence

to Lipschitz continuity may hinder the model's ability to capture crucial input variations. In image and natural language domains, a common testing approach is adversarial attacks [70, 71, 72, 73]. That testing approach involves intentionally manipulating input data to deceive the model's predictions and does not apply to our medical settings.

3.4.2 Engineered testing data

Our results identified serious deficiencies in conventionally trained binary classification models in recognizing seriously abnormal medical conditions. For example, in-hospital mortality prediction models fail to generate alerts for bradypnea (low respiratory rates) or hypoglycemia conditions (Figure 3.9). Similarly, the models also consistently underestimate some of the mortality risks when given multiple abnormal vital time series in conjunction (Figure 3.11 and 3.12). When given test cases representing various injury levels, neural network models (namely, LSTM and CW-LSTM) gave inconsistent risk predictions – assigning higher mortality risk (> 0.5) to cases of moderate injury (e.g., GCS score 12), while assigning disproportionately lower risk (< 0.05) to severe injuries (e.g., GCS score 7). The two neural network models exhibit insensitivity to changes in eye response (Figures 2d and 2f). For most attributes, we found the training data's distribution is highly centered, not sufficiently representing high or low critical zones (Figure 3.2 and 3.3). Death and non-death cases exhibit somewhat similar value distributions, means, and standard deviations (Table A.15) for individual attributes, despite the drastically different outcome. Machine learning methods produced by supervised training approaches are unable to recognize the meanings of vitals in dangerous zones. This semantic deficiency of ML models was also reported in image recognition studies, e.g., melanoma classification overinterpreting surgical skin markings [36]. We found similar kinds of semantic deficiencies in an ML model predicting 5-year breast cancer survivability. These findings indicate the importance of using crafted test cases to assess

clinical ML models.

The conventional test set is limited in its distribution shift from the training data. For example, for in-hospital mortality prediction, our generated multi-attribute test cases present a high Wasserstein distance (33.4) from the original MIMIC-III training data, much larger than the split test set's Wasserstein distance (12.4). A similar distribution shift pattern is observed for the breast cancer survivability model. For triple-attribute test cases, the breast cancer survivability prediction model performs better (89-98% triple-attribute test accuracy) than in-hospital mortality prediction models (6-69% multi-attribute test accuracy). This difference in accuracy may be partly due to the different distribution shift in generated test data – there is a much smaller distribution shift in triple-attribute breast cancer test cases (Wasserstein distance 9.8) than in multi-vital test cases (Wasserstein distance 33.4), with respect to their original training data. The LSTM model is slightly better at recognizing multi-attribute test cases (45.7% accuracy) than single-attribute ones (37.8%) and CW-LSTM exhibits a similar pattern (69.3% vs. 22.4%, Table A.17). Multiple abnormal vitals likely provide more clues for the ML models to classify, whereas single isolated attribute changes appear more difficult.

The poor performance of the ML models is somewhat expected because of the distribution shift between training data and our synthetic test data. Yet, these deficiencies are unacceptable from a clinical deployment perspective, as the test cases represent potential real-life medical conditions. Our work points out a fundamental limitation of pure data-driven machine-learning models – models purely trained by patient data do not perform well for tasks that require implicit medical knowledge (e.g., normal vital ranges). This limitation has not been reported in the literature.

3.4.3 Accuracy comparison across models

For in-hospital mortality prediction, all 3 models have multiple deficiencies under single-attribute critical zone test cases and are unable to generate high enough risk predictions for serious patient conditions (Figure 3.9 and Table A.17). Out of all the dual critical zone attributes such as body temperature, only one model recognizes one such attribute – LSTM exhibits a U-shape risk curve for systolic blood pressure (Figure 3.9e). The other risk curves are all monotonic or flat, failing to raise alerts for both ends of the abnormal conditions (Figure 3.9). For multiple attribute testing, logistic regression performs the worst (Figure 3.11 and 3.12 and Table A.17), mispredicting 93.7% of test cases. CW-LSTM's accuracy is the lowest (22.4%) in single-attribute testing, however, it gives the highest average accuracy (69.3%) for multiple-attribute testing. Brain injury-related Glasgow Coma Scale test cases (Figure 3.7 and 3.8) involve simple categorical data (as opposed to numerical data). Logistic regression gives the best performance, generating appropriate and consistent risk estimates, and substantially outperforms the two neural network models (Table A.17). These results suggest that for categorical attributes such as GCS, a simpler model like logistic regression may be more suitable than complex deep learning models, indicating the importance of evaluating a wide variety of machine learning models before clinical use. Deficiencies in ML responsiveness were also observed in the 5-year breast cancer prediction task – the MLP model gave an average of 48.9% prediction accuracy under our test case (Table A.19). This accuracy is much lower than the widely reported death class accuracy of 90% (standard deviation 0.45), computed based on the original test data from SEER [9, 46].

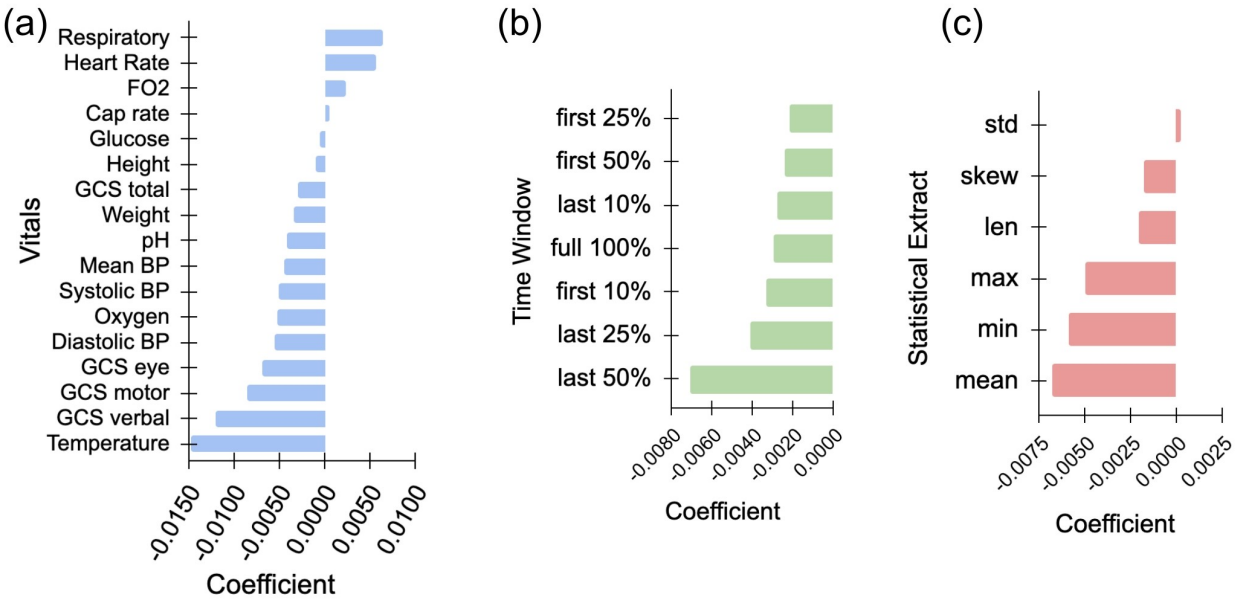


Figure 3.24: Logistic regression coefficients (averaged) for (a) vitals, (b) time-period, and (c) statistical-extract features.

The linear logistic regression model tested is unsuitable for analyzing vitals due to multiple reasons. It reduces time series to statistical summaries, e.g., mean, min, max, etc. (Figure 3.24 c), and is unable to capture data's dynamic details. Linear logistic regression is unable to model non-linear (e.g. U-shaped curve) features, as it responds monotonically to features. In multi-attribute tests, the model gives poor performance (6.2% accuracy, Table A.17), partly because of its many negative coefficients associated with attributes. 13 out of the 17 attributes have negative coefficients, e.g., the temperature is strongly inversely correlated with the predicted probability (Figure 3.24a), resulting in underestimated risk prediction. Gaussian Naive Bayes and KNN show weaker performance on the original test sets than the others (Figure 3.5 and 3.6) and, thus, are excluded from subsequent attribute tests. For completeness, all models' performance on the original test set was given in Figure 3.5 and 3.6.

3.4.4 Deteriorating trends vs. steady values in critical zones

For in-hospital mortality prediction, the LSTM models are slightly better at recognizing deteriorating trends (average 44%, Figure 3.13) than cases with steadily low or steadily high vitals in critical zones (average 36.5%, Figures 3 and 4). When test cases contain 3 simultaneously deteriorating attributes, the models detected 56% of them on average, which is better than their performance of 41% on a single deteriorating attribute. When using LSTM gradient ascent to automatically generate multi-attribute deteriorating test cases, we found the resulting test cases all have significantly decreased oxygen saturation and body temperature values in the last 24 hours. Because the gradient ascent process follows the shortest path within the loss function space of the model, these findings indicate that i) oxygen saturation and body temperature are top LSTM features and ii) the last 24 hours (out of the entire 48-hour timespan) are important in the model's decision-making process, which is also consistent with logistic regression feature ranking (Figure 3.24 b).

3.4.5 ML responsiveness in cancer survivability prediction

The BCS multilayer perceptron model (MLP) model exhibited responsiveness to critical attributes, such as tumor size and lymph node involvement (Figure 6). For example, for N3 stage (extensive lymph node involvement) test cases, the model was able to raise alerts for 74.4% of them (Table A.19). The model also performed well (nearly 100% alerts) when all three critical features (T4 tumor size, N3 lymph node stage, and grade 4) were high. These observations suggest the MLP model's prediction capability in extreme cases is good. However, the model does not respond to severe tumor sizes (T3 stage) – generating no alerts. The consistency in predicted survivability scores is also low, as the model generated slightly more alerts for grade 3 cancer (65.8%) than grade 4 terminal cancer (63.9%). This

inconsistency may be the outcome of the imbalanced dataset (Figure 3.4), as the SEER dataset contains a total of 81,749 (death 15,628 and survived 66,121) grade 3 cases, while only 3,002 (death 640 and survived 2,362) grade 4 cases. The number of examined lymph nodes (ELNs) does not directly indicate a cancerous condition, thus, the model's lack of response to ELN is somewhat expected.

Despite overall good performance on the original test set (Figure 3.5 and 3.6), tree-based ensemble methods such as XGBoost, AdaBoost, and Random Forest exhibit low responsiveness to critical zone tests (Figure 3.15 and 3.16). Ensemble methods perform much worse than MLP for SEER BCS and LCS settings, which is likely due to the sparsity in the one-hot encoded input space. The SEER dataset has much larger feature dimensions (56 for BCS and 47 for LCS) than the MIMIC III and eICU time-series data (17 features). Using the one-hot encoding to encode categorical features leads to an expansive number of sparse encoded representations (1,423 for encoded BCS and 1,315 for encoded LCS), posing challenges to tree-based models.

3.4.6 Countermeasures to reduce blind spots in ML models

One clinical mitigation is to deploy a filter-then-predict 2-step workflow where domain-specific rules are first applied to identify cases with obvious disease conditions. Thus, corner-case scenarios will never reach machine learning models. However, designing such rule-based classifiers, especially under time-series data, is challenging and may require substantial manual efforts. A more efficient approach is out-of-distribution detection, which identifies cases that present a significantly large distribution shift from the model's training data. Existing solutions for detecting out-of-distribution images, e.g., [74], cannot be directly applied to clinical settings. For medical applications, out-of-distribution detection

is challenging. For example, out-of-distribution patient cases still need to be examined and classified. Thus, an overly strict detection may produce too many such out-of-distribution cases for the downstream examination. Finding the right balance will facilitate clinical translation.

A promising direction is medical foundation models based on clinical large language models (LLMs) [75, 76, 77, 78]. Our findings suggest that statistical machine-learning models solely trained from patient data are grossly inadequate. They are unable to capture basic clinical knowledge, e.g., patients with extremely low Glasgow Coma Scale values have a high mortality risk. LLMs are likely able to recognize common sense health conditions and serve as a filter mechanism before ML classifiers. However, it is crucial to quantitatively characterize the trustworthiness of medical LLMs before clinical adoption. Our work suggests the urgent need for new clinical decision-making workflows, as existing models solely trained from patient samples are extremely limited. At clinical time, a human-friendly interface is also important for interpreting ML results. Conventional interpretability techniques were designed for ML experts, not for clinicians, e.g., SHapley Additive exPlanations (SHAP) [79], Local Interpretable Model-Agnostic Explanations (LIME) [80], or TRUSTEE [81]. Therefore, these tools cannot be directly used in clinical settings. An innovative clinical workflow needs to place large language models (LLMs) as the final component to generate narrative explanations based on ML predictions and interpretability results. An interesting research direction is how to fine-tune LLMs for these specific tasks.

The boost provided by conventional resampling and reweighting methods is very limited (Figure 7). Under some scenarios (e.g., tachypnea and increasing CS tumor size), they may perform even worse than the original models. This poor performance is expected, as these methods rely on existing minority class samples in the training set, which are limited in their ranges and variations. The root problem is that the space of all possible minority class

samples is vast. Attempting to cover all or most of them through training data engineering (such as oversampling) is not feasible. Thus, data engineering does not appear to be a feasible direction for the ML responsiveness problem. A more promising approach is to directly encode vital semantics into the clinical decision workflow as discussed above.

Our work provides the first look into ML responsiveness. Comprehensive measurement studies in other medical settings are needed. Our gradient ascent testing methodology can be extended to other health conditions (e.g., rare diseases or comorbidities). Scalability is the key to testing in medicine, because of the complex high-dimensional space. Innovative methods that prioritize testing are needed to reveal the most critical blind spots in a model.

3.4.7 Data and Code Availability

The MIMIC-III, eICU, and SEER data used in this study are not publicly downloadable, but can be requested at their original sites after completing proper training. Parties interested in data access should visit the MIMIC-III website (https://mimic.physionet.org/gettingstarted/access/), eICU website (https://eicu-crd.mit.edu/)

The SEER website (https://seer.cancer.gov/data/access.html) to submit access requests. Because our test cases are generated from these access-controlled datasets, they cannot be publicly released. However, we have released the code for reproducing all our test cases.

We have released all our code used on GitHub, which can be used to generate the test cases and reproduce our experiments. https://github.com/PiasTanmoy/TRUSTWORTHY-ML

Chapter 4

Improving Responsiveness of Machine Learning Model by Integrating Medical Domain Knowledge

4.1 Introduction

This study proposes a novel methodology for targeted mitigation techniques aimed at enhancing model responsiveness and robustness. Our solutions incorporate adaptive training adjustments and data-specific augmentation strategies tailored to reinforce model performance in high-risk areas. These techniques are inspired by recent advances in robust ML training and data augmentation, yet are adapted specifically to address the complexities of clinical datasets [5, 6].

Finally, we conduct a rigorous evaluation of our methodology across multiple prediction tasks, examining its effectiveness and generalizability in diverse AI contexts. By systematically assessing model performance in these high-risk areas, this work provides insights into the trustworthiness of ML models across applications and underscores the importance of robust model evaluation for safe and ethical AI deployment in healthcare [7, 8].

- We propose two methodologies to integrate domain knowledge: i) by building custom

loss function for training machine learning models; ii) crafting a custom rule-based decision tree model along with a data-driven machine learning model.

- We integrate domain knowledge in different types of machine learning models: i) custom loss function for LSTM and Transformer models; ii) custom decision tree model with data-driven XGBoost model.

- We evaluated our approach on two different domains: i) Mortality risk prediction using the MIMIC III dataset, ii) Early prediction of Sepsis from the PhysioNet in cardiology challenge 2019 dataset.

4.2 Method

4.2.1 Prediction tasks, datasets, and model selection

Our work aims to test medical ML models for their binary classification accuracy under serious disease conditions. We focus on two binary prediction tasks, namely 48-h In-hospital mortality risk prediction and early sepsis prediction using time-series vitals and lab tests data.

The datasets in our study include a 2019 benchmark based on the MIMIC III dataset. The first two datasets contain patients' 48-h time series data in critical care units (ICU). Our study excludes clinical free text notes. As with many medical datasets, the MIMIC-III dataset for IHM, containing 21,139 samples, is imbalanced, with 13.2% death cases (Class 1), and 86.8% non-death cases (Class 0). S1 shows the distributions of key attributes of both MIMIC III.

For sepsis prediction task, we selected PhysioNet/Computing in Cardiology Challenge 2019

dataset.

We select ML models that are commonly used in the medical literature for these prediction tasks. Specifically, we select long short term memory (LSTM) as it is widely used for predicting mortality risk in a 48-h ICU time series dataset—in recent literature. For sepsis prediction, we selected the XGBoost model as it showed the highes performance.

4.2.2 Data preprocessing

We train ML models with benchmark datasets of MIMIC-III, following the conventional pre-training processing (e.g., encoding, standardization). As MIMIC-III benchmark datasets contain missing values, we imputed the values that are missing using the most recent observation (within 48h) if it exists, otherwise, a value from the normal range of corresponding vitals is mentioned in ref. 11. Masking was used to indicate whether the vital value was original or imputed. The categorical variables, including binary ones, were encoded using a one-hot vector. The numerical features, such as diastolic blood pressure and glucose level, were converted to their standardized form. After preprocessing, each time-series data point became a 76-by-48 matrix (76 computed features and 48h). The processed dataset was used for training and testing neural network-based models such as LSTM.

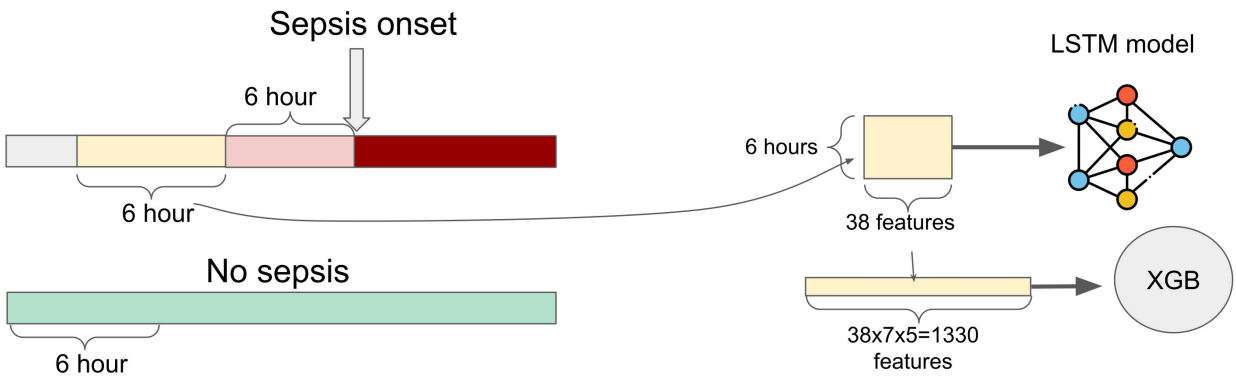


Figure 4.1: Sepsis dataset preprocessing

Table 4.1: Model architecture

	Input layer	BiLSTM layer1	LSTM layer 2	Batch norm	Drop out	Dense layer	Epochs
LSTM	76	8 units	16 units	Yes	0.3	16 units	100
	Input layer	Transformer encoder block	Num of heads	Batch norm	Drop out	Dense layer	Epochs
Transformer	76	3	4	Yes	0.3	64 units	100

For early sepsis prediction, we selected 6 hours of window 6 hours before sepsis onset. For non-neural network models such as XGBoost that cannot directly process time series, we extracted 6 statistical features (mean, min, max, standard deviation, skew, and number of measurements) from various sub-periods (first/last 10%, 25%, 50%, and full 100%). We did not encode the categorical variables, as they contain values with a meaningful scale. The missing values were replaced with mean values computed on the training set and numerical variables were standardized. The continuous variables were standardized before training.

4.2.3 Configuration of machine learning models

For IHM risk prediction, we utilized the LSTM model. We utilized the optimal settings of neural network models (i.e., layers, activation, hyperparameters) for each of the tasks from corresponding benchmarks11,16. The LSTM model consisted of an input layer (76 dimensions), a masking layer (76 dimensions), a bidirectional LSTM layer (16 dimensions), an LSTM layer (16 dimensions), a dropout layer, and finally a dense layer (1 dimension). In total, the LSTM had 7569 trainable parameters.

4.2.4 Model training and threshold tuning

For IHM prediction, LSTM models and transformer models were trained for 100 epochs using the MIMIC-III datasets. The models were trained using binary cross-entropy loss. An epoch was selected based on the threshold-agnostic validation area under the precision-recall curve (AUPRC) and validation loss to avoid overfitting. Specifically, we first selected the top 3 epochs with the highest validation AUPRC and then selected the epoch with the minimum validation loss (Supplementary Tables 2 and 3). We monitored the validation loss and training loss difference to prevent overfitting. In all experiments, the chosen ML model demonstrated a small loss difference (Supplementary Tables 2 and 3).

The training, validation, and test set breakdown for MIMIC-III datasets is 70%, 15%, and 15% . After model calibration, a threshold-tuning process is conducted on the validation set, and an optimal threshold is selected based on balanced accuracy and F1 score for the minority class. Specifically, after training, we first conducted model calibration by applying Isotonic Regression using the validation set. Model calibration mapped the predicted probabilities to actual probabilities. Then, we performed threshold tuning to determine the optimal threshold. The minority F1 score and balanced accuracy were computed on the validation set for each threshold ranging from 0.0 to 1.0 with a step size of 0.01. Subsequently, the top three thresholds yielding the highest minority F1 scores were identified, and the optimal threshold maximizing balanced accuracy across all validation samples was selected. This process was repeated for 3 independently trained models of each type, and the average threshold was calculated from these independent trials. Thresholds are shown in Supplementary Table 6. The tasks were executed on a machine with Ubuntu 18.04 operating system, x86-64 core-i9 architecture, 8 physical cores (16 virtual cores), and 32 GB RAM. The experimental code and models were written using Python 3.7, TensorFlow 1.15, and Keras 2.1.2. The cancer survivability prediction MLP model was trained on a machine with x86_64 Intel(R) Xeon(R)

CPU 2.40 GHz (40 cores) and 125 GB RAM. The experimental code and model were written using Python 3.6, TensorFlow 2.9.0, and Keras 2.9.0.

4.2.5 Custom test set generation

We created new cases by increasing or decreasing one or multiple vital health parameters in the seeding records. To reduce computing complexity, we prioritized by focusing on the most influential features. Relevant medical terminologies are explained in the Supplementary Notes.

In the single-attribute variation, we generated new test cases by varying a single attribute at a time while keeping other attributes unchanged. We then evaluated how the model reacts to these changes and its ability to recognize associated risks (e.g., hypoglycemia). Specifically, given an attribute A, single-attribute variation for time series involved the following operations. First, we identified A's minimum and maximum values in the MIMIC-III, which defined the observed range. Then, the mean and the variance of attribute A were computed from the entire dataset. Using the variance and the observed range, we generated a series of random values for every value from that range, one value for each of the 48h. Then, the new test case was formed by having these generated values for attribute A and other attribute values directly inherited from the seed. We repeat this process for every possible attribute value from the observed range with step 1.

Multi-attribute variation generated new test cases by modifying two or more attributes, aiming to represent medical conditions that were characterized by variations in multiple related attributes. We further differentiated two scenarios: (a) a single set of medically correlated attributes driven by one underlying disease condition, e.g., high diastolic and systolic blood pressure due to hypertension, and (b) medically correlated attributes due to

multiple underlying conditions, e.g., hypertension and diabetes. These test cases were used to assess the ML model's ability to respond to the risks of multiple disease conditions in patients. One of the test sets was created by changing multiple vitals such as systolic blood pressure, diastolic blood pressure, blood glucose level, respiratory rate, heart rate, and body temperature at the same time. A test case was assigned a ground truth label using existing literature or under the guidance of medical doctors. 6 multi-attribute test cases and 12 deteriorating test cases were directly labeled by the medical doctor (Supplementary Tables 7 and 8).

4.2.6 Knowledge Infused Custom Loss Function

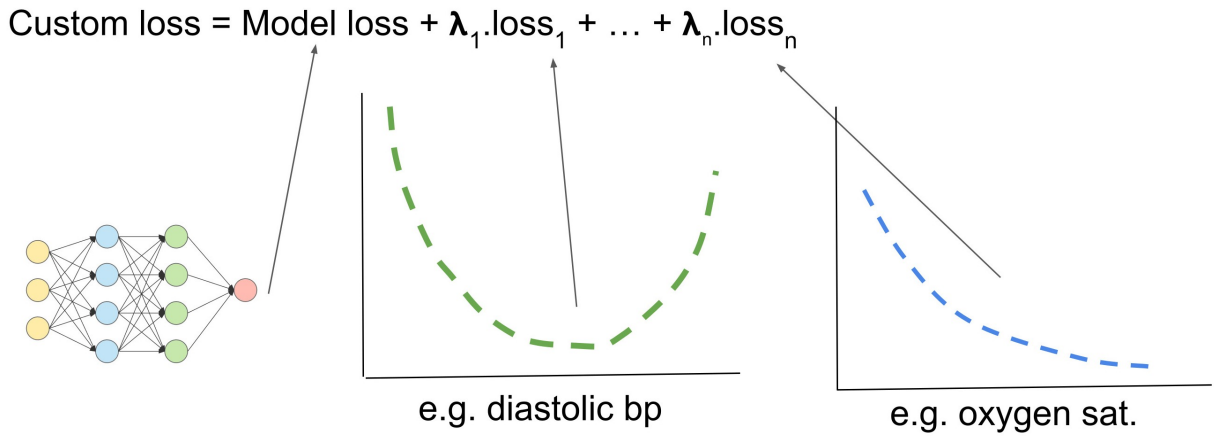


Figure 4.2: Knowledge guided machine learning model with custom loss function

A critical advancement in domain knowledge integration is modifying the model's loss function to incorporate domain-specific constraints, a method that helps guide the model towards clinically relevant outcomes [82]. For instance, by adjusting the loss function to penalize mispredictions in high-stakes clinical scenarios, we can make the model more responsive to specific medical priorities, such as accurately predicting deterioration in critical patients. Along with the traditional loss function, which penalizes the model based on deviation of

the prediction from the ground truth, custom loss units are added for each input feature which penalize the model when the prediction deviates from the domain knowledge. Each loss units are added using a scaler variable to control the impact of each domain feature. In other words, these loss units can be called as regularizer units as well. The loss functions are custom made based on the characteristics of each domain features. By embedding clinical priorities and constraints directly into the training objective, this technique effectively aligns the model's optimization process with healthcare objectives. In practice, this modified loss function can help the model prioritize detecting severe cases, reducing false negatives in critical zones, and ensuring that high-risk patients are less likely to be overlooked.

Mapping a Quadratic Function to Loss Function

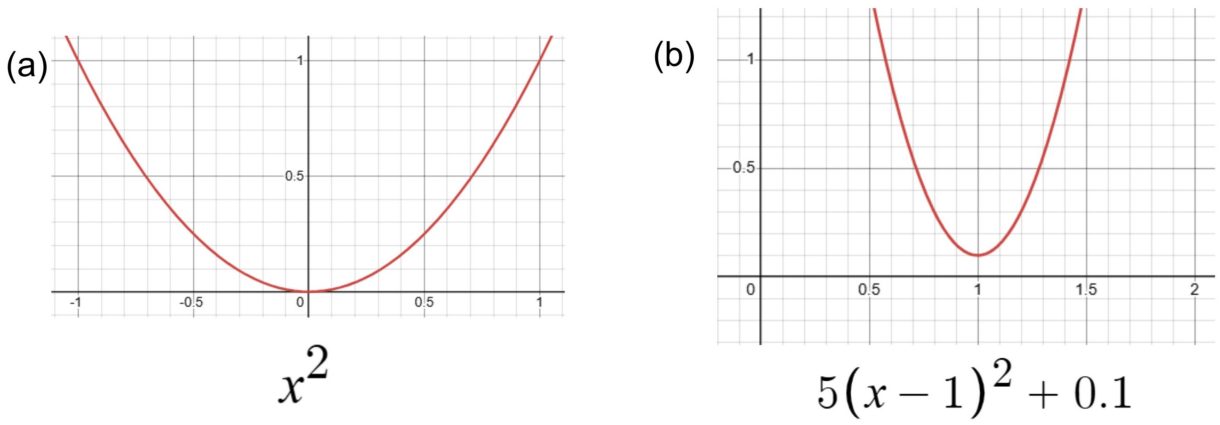


Figure 4.3: Generalized quadratic function

We apply either a *monotonic* or *U-shaped* transformation to each feature before computing the loss. The generic form is

$$F_{\text{transformed}} = a\left(f - f_{\text{optimal}}\right)^n + b, \tag{4.1}$$

where

- f — raw feature value,
- f_{optimal} — domain-defined optimal value of the feature,
- a — scaling hyperparameter that controls the strength of the penalty,
- b — baseline offset,
- n — shape exponent ($n = 1$ yields a linear, monotonic penalty; $n = 2$ produces a symmetric U-shape; higher even values sharpen the U).

Equation (4.1) is applied element-wise to the feature vector and the resulting $F_{\text{transformed}}$ is passed to the loss function, ensuring that deviations from f_{optimal} incur a smoothly scaled penalty.

4.2.7 Custom Loss Function Types

We introduce three custom loss functions that penalize predictions based on their deviation from medically preferred ranges. These loss functions are incorporated as additional penalty terms in the overall loss formulation, encouraging the model to align with clinical reasoning.

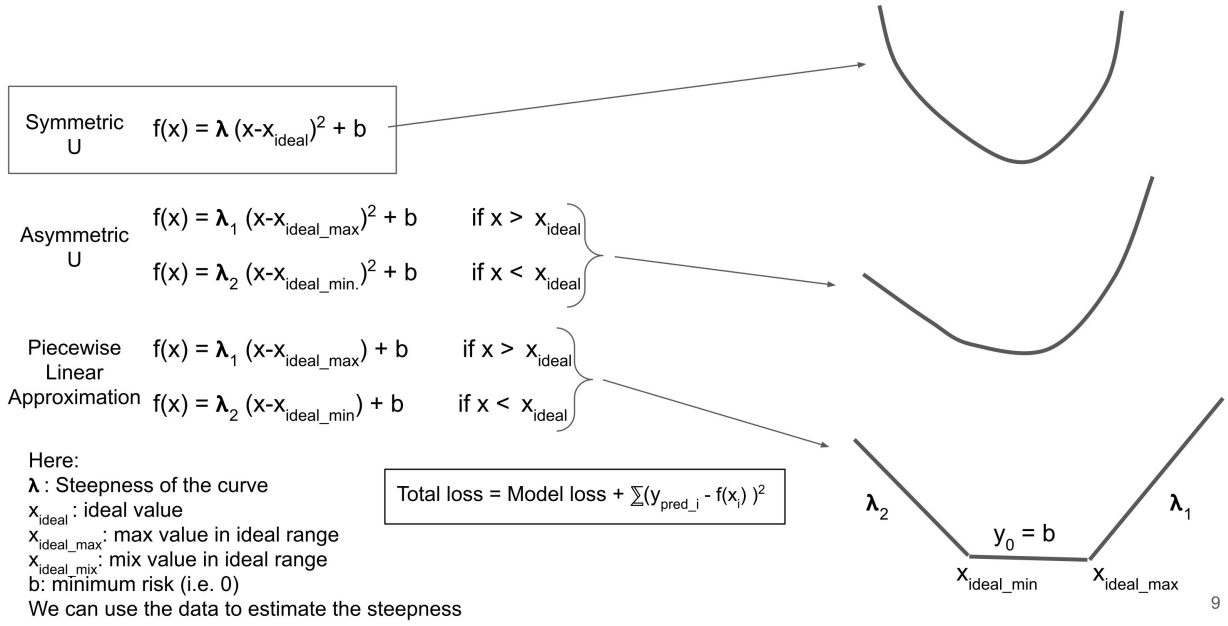


Figure 4.4: Custom loss functions

Let x represent the predicted (or observed) feature value, and x_{ideal} denote the clinically ideal target value. The penalty function $f(x)$ is constructed based on the magnitude and direction of deviation from x_{ideal}, using the following three loss formulations:

Symmetric U-shaped Loss

This loss penalizes deviations equally on both sides of the ideal value and is defined as:

$$f(x) = \lambda(x - x_{\text{ideal}})^2 + b \tag{4.2}$$

where λ controls the steepness of the penalty curve, and b is the minimum baseline risk, typically set to zero. This formulation is appropriate for vitals such as heart rate and temperature, where both increases and decreases from the ideal range are equally undesirable.

Asymmetric U-shaped Loss

To account for scenarios where one direction of deviation is more harmful than the other (e.g., hyperglycemia being more dangerous than mild hypoglycemia), we define an asymmetric penalty function:

$$f(x) = \begin{cases} \lambda_1(x - x_{\text{ideal\_max}})^2 + b, & x > x_{\text{ideal}} \\ \lambda_2(x - x_{\text{ideal\_min}})^2 + b, & x < x_{\text{ideal}} \end{cases} \tag{4.3}$$

Here, λ_1 and λ_2 allow different penalty rates for high and low deviations, and $x_{\text{ideal\_max}}$, $x_{\text{ideal\_min}}$ define the upper and lower bounds of the ideal range, respectively.

Piecewise Linear Approximation

To provide a simpler yet interpretable approximation of risk, we also define a piecewise linear loss function:

$$f(x) = \begin{cases} \lambda_1(x - x_{\text{ideal\_max}}) + b, & x > x_{\text{ideal}} \\ \lambda_2(x - x_{\text{ideal\_min}}) + b, & x < x_{\text{ideal}} \end{cases} \tag{4.4}$$

This is useful when risk increases approximately linearly with deviation from the ideal range, or when clinical guidelines are based on thresholds.

Total Loss Function

$$\text{Total Loss} = \text{Model Loss} + \text{Knowledge guided (KG) loss} \tag{4.5}$$

Model Loss

$$\mathcal{L}_{\text{BCE}} = \frac{1}{N} \sum_{j=1}^{N} [-y_{\text{true},j} \log(y_{\text{pred},j}) - (1 - y_{\text{true},j}) \log(1 - y_{\text{pred},j})] \tag{4.6}$$

Knowledge guided (KG) loss

Deviation from ideal value:

$$u_{\text{expected},i,j} = a_i \cdot \left(\text{vital\_values}_{j,i} - x_{\text{opt},i}\right)^2 + b_i \tag{4.7}$$

The penalty for the i-th column is:

$$\mathcal{L}_{\text{penalty},i} = \frac{1}{N} \sum_{j=1}^{N} \lambda_{\text{mini},i} \cdot \left(y_{\text{pred},j} - u_{\text{expected},i,j}\right)^2 \tag{4.8}$$

The total penalty combines penalties across all vital columns:

$$\mathcal{L}_U = \sum_{i=1}^{C} \mathcal{L}_{\text{penalty},i} \tag{4.9}$$

The total loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{penalty}} \cdot \mathcal{L}_U \tag{4.10}$$

$$\begin{aligned} \mathcal{L} \; = & \frac{1}{N} \sum_{j=1}^{N} \left[-y_{\text{true},j} \log(y_{\text{pred},j}) - (1 - y_{\text{true},j}) \log\left(1 - y_{\text{pred},j}\right) \right] \\ & + \; \lambda_{\text{penalty}} \sum_{i=1}^{C} \left\{ \frac{1}{N} \sum_{j=1}^{N} \lambda_{\text{mini},i} \Big[y_{\text{pred},j} - \left(a_i \left(\text{vital\_values}_{j,i} - x_{\text{opt},i} \right)^2 + b_i \right) \Big]^2 \right\} \end{aligned} \tag{5.12}$$

Let:

y_{true}: Ground truth labels (binary: 0 or 1).

y_{pred}: Predicted probabilities from the model, $y_{\text{pred}} \in [0, 1]$.

$\text{BCE}(y_{\text{true}}, y_{\text{pred}})$: Binary Cross-Entropy Loss.

C: Number of vital columns (length of `vital_column_index`).

$x_{\text{opt},i}$: Optimal value for the i-th vital column.

a_i: Steepness parameter for the U-shaped curve for the i-th vital column.

b_i: Baseline risk for the i-th vital column.

$\lambda_{\text{mini},i}$: Penalty weight for the i-th vital column.

λ_{penalty}: Overall weight for the U-shaped penalty.

Illustrative Example:

Consider a patient with two abnormal vital signs:

- Systolic Blood Pressure (SBP) = 200 (mean = 120, std = 20)
- Glucose = 300 (mean = 110, std = 40)

We standardize the input:

$$z_{\text{SBP}} = \frac{200 - 120}{20} = 4.0$$
$$z_{\text{Glucose}} = \frac{300 - 110}{40} = 4.75$$

We compute the expected risk using a U-shaped function:

$$u_{\text{expected},i} = a_i \cdot (z_i)^2 + b_i \tag{4.11}$$

where $a_i = 0.01$, $b_i = 0.1$ for both features.

Expected risk values:

$$u_{\text{expected,SBP}} = 0.01 \cdot (4.0)^2 + 0.1 = 0.26$$

$$u_{\text{expected,Glucose}} = 0.01 \cdot (4.75)^2 + 0.1 \approx 0.3256$$

Assume the model predicts $y_{\text{pred}} = 0.7$ and $y_{\text{true}} = 1$, the binary cross-entropy is:

$$\mathcal{L}_{\text{BCE}} = -\log(0.7) \approx 0.357 \tag{4.12}$$

Penalty terms:

$$\mathcal{L}_{\text{penalty,SBP}} = (0.7 - 0.26)^2 = 0.1936$$

$$\mathcal{L}_{\text{penalty,Glucose}} = (0.7 - 0.3256)^2 \approx 0.1402$$

Total loss (with $\lambda_{\text{penalty}} = 0.5$):

$$\mathcal{L}_U = 0.1936 + 0.1402 = 0.3338$$

$$\mathcal{L}_{\text{total}} = 0.357 + 0.5 \cdot 0.3338 = \mathbf{0.5239}$$

This method allows the model to stay grounded in learned patterns while gently encouraging outputs that align with clinically meaningful risk profiles.

4.2.8 Rule-based Decision Tree Integration

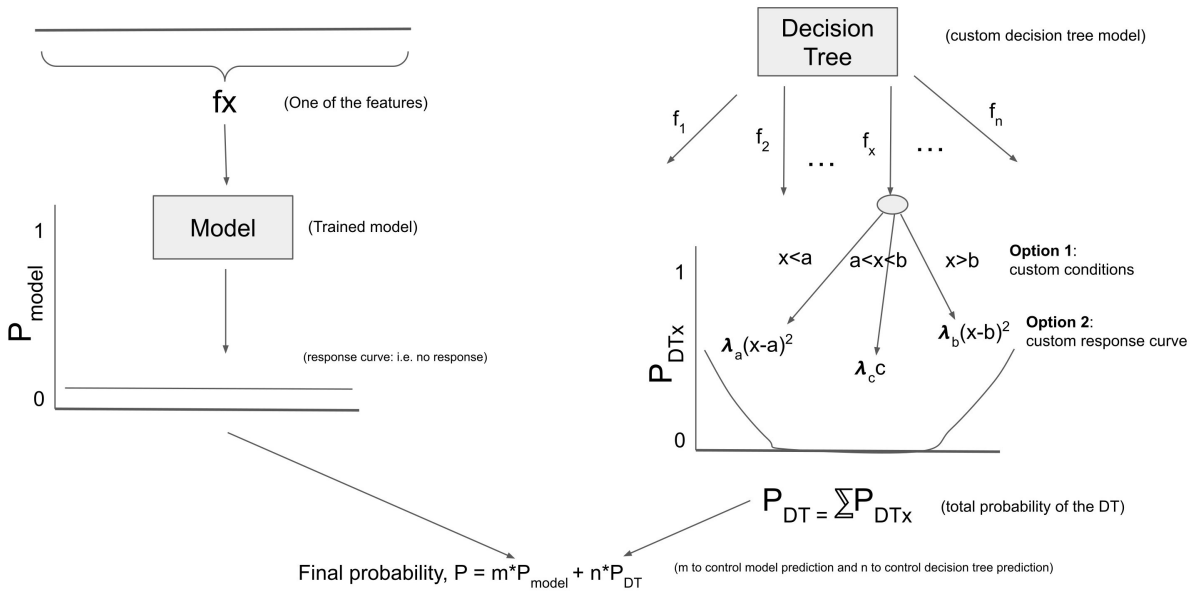


Figure 4.5: Outline of merging rule-based custom decision tree with data driven trained machine learning model

Established medical rules are converted to a decision tree with custom response settings and integrated with a data driven machine learning model. The response of the decision tree is mapped using a mathematical function representative of the realistic output. Here, two different types of monotonic mapping functions are used: i) linear step function, ii) quadratic non-linear function. The output of the custom decision tree is merged with a data-driven (i.e., trained) machine learning model using a weighted average.

To integrate expert clinical knowledge with data-driven predictions, we designed a custom rule-based decision tree module and merged its output with that of a trained XGBoost model. This approach enables interpretability from medical rules while leveraging the predictive power of machine learning.

Established medical rules, such as critical value thresholds for lab measurements (e.g., biliru-

bin, creatinine, platelet count), were encoded into a decision tree structure. Each node of the tree represents a feature condition (e.g., $x < a$, $a \leq x < b$, $x > b$), derived directly from clinical guidelines or expert-defined cutoffs.

Each path through the tree results in a leaf node where a probability score P_{DTx} is assigned. This score quantifies the likelihood of the target condition (e.g., sepsis) based on rule-based logic. The structure is designed such that the sum of all P_{DTx} values for a given instance yields the total decision tree prediction:

$$P_{DT} = \sum P_{DTx} \tag{4.13}$$

To convert binary rules into a smooth probabilistic response, each branch or node applies a monotonic mathematical mapping function. Two types of mapping functions were used:

- **Step Function**: A linear step assigns fixed probabilities to discrete rule zones. For example, $P = 0$ if $x < a$, and $P = \lambda$ if $x \geq a$.
- **Quadratic Function**: A soft boundary is applied by penalizing the distance from the optimal value using a quadratic function, such as $P = \lambda(x - a)^2$, where λ is a tunable scaling factor.

These mappings allow the decision tree to output continuous probabilities, maintaining clinical interpretability while approximating real-world trends.

Model Fusion via Weighted Averaging

The outputs of the XGBoost model (P_{model}) and the custom decision tree (P_{DT}) were fused using a weighted average. A merge coefficient m controls the relative contribution of the

data-driven model:

$$P_{\text{final}} = m \cdot P_{\text{model}} + (1 - m) \cdot P_{DT} \tag{4.14}$$

By varying m from 0 to 1, we explore a continuum of models, ranging from fully rule-based ($m = 0$) to fully data-driven ($m = 1$). This flexible design enables detailed analysis of trade-offs between performance and interpretability.

Listing 4.1: Function to estimate SOFA sub-score

```
def estimate_sofa_score(Creatinine, Bilirubin_total, Platelets):

    score = 0

    if Platelets < 20: score += 4
    elif Platelets < 50: score += 3
    elif Platelets < 100: score += 2
    elif Platelets < 150: score += 1

    if Bilirubin_total > 12.0: score += 4
    elif Bilirubin_total > 6.0: score += 3
    elif Bilirubin_total > 2.0: score += 2
    elif Bilirubin_total > 1.2: score += 1

    if Creatinine > 5.0: score += 4
    elif Creatinine > 3.5: score += 3
    elif Creatinine > 2.0: score += 2
    elif Creatinine > 1.2: score += 1
```

```
19
20      return score
```

In this formulation we replace the discrete SOFA thresholds with a *monotonic quadratic* score that grows smoothly as the physiological value x departs from its normal range.

$$\mathrm{SOFA}_{\mathrm{curve}}(x) = \begin{cases} \alpha\,(x-\beta)^2, & x \notin [x_{\min}^{\mathrm{norm}},\, x_{\max}^{\mathrm{norm}}], \\ 0, & x \in [x_{\min}^{\mathrm{norm}},\, x_{\max}^{\mathrm{norm}}], \end{cases} \tag{4.15}$$

- x – observed laboratory or vital-sign measurement.
- β – nearest boundary of the normal range (either $x_{\min}^{\mathrm{norm}}$ or $x_{\max}^{\mathrm{norm}}$).
- α – scaling factor chosen so that $\mathrm{SOFA}_{\mathrm{curve}}$ attains the maximum sub-score (conventionally 4) at the clinically worst value x_{worst}.

Selecting the scaling factor α. To ensure the curve rises smoothly from 0 and reaches the canonical maximum of 4 at the worst physiological state, set

$$\alpha = \frac{4}{\left(x_{\mathrm{worst}} - \beta\right)^2}. \tag{4.16}$$

Equation (4.16) is applied separately for each SOFA attribute (e.g. platelets, bilirubin, creatinine), yielding attribute-specific scaling constants that preserve both monotonicity and the original 0–4 scoring range.

Applying Eq. (4.16) to each SOFA attribute yields the following scaling constants:

$$\alpha_{\text{Platelets}} = \frac{4}{\big(150 - 20\big)^2} = 0.0002, \tag{4.17}$$

$$\alpha_{\text{Bilirubin}} = \frac{4}{\big(12 - 1.2\big)^2} \approx 0.03, \tag{4.18}$$

$$\alpha_{\text{Creatinine}} = \frac{4}{\big(5.0 - 1.2\big)^2} \approx 0.30. \tag{4.19}$$

These values ensure that each continuous SOFA curve rises smoothly from 0 at the edge of the normal range to the maximum sub-score of 4 at the corresponding worst clinical condition.

To estimate final risk, we combine the probabilistic output of an XGBoost model with the normalized SOFA score (derived from the three continuous components: platelets, bilirubin, and creatinine).

Since each component contributes a maximum of 4 points, the total SOFA score has an upper bound of:

$$\text{SOFA}_{\text{max}} = 4_{\text{platelets}} + 4_{\text{bilirubin}} + 4_{\text{creatinine}} = 12 \tag{4.20}$$

We define the final prediction probability as a weighted combination of the XGBoost model output and the normalized SOFA score:

$$\text{Final Probability} = m \cdot \text{XGB}_{\text{prob}} + n \cdot \left(\frac{\text{SOFA}}{12}\right) \tag{4.21}$$

where:

- XGB_{prob} is the probability output by the XGBoost model,

- SOFA is the combined continuous SOFA score from the three selected attributes,

- m and n are scalar weights such that $m + n = 1$.

This fusion leverages both the machine-learned model's predictive power and the interpretable clinical scoring system to enhance robustness and interpretability. Here we show some sample case studies.

Case 1: No Dysfunction

Inputs: Platelets = 140, Bilirubin = 1.5, Creatinine = 1.5, XGBoost Probability = 0.01

$$\text{Platelets SOFA} = 0.0002 \times (150 - 140)^2 = 0.2$$

$$\text{Bilirubin SOFA} = 0.03 \times (1.5 - 1.2)^2 = 0.027$$

$$\text{Creatinine SOFA} = 0.3 \times (1.5 - 1.2)^2 = 0.027$$

$$\textbf{Total SOFA} = 0.254$$

$$\textbf{Final Probability} = 0.5 \times 0.01 + 0.5 \times \left(\frac{0.254}{12}\right) \approx 0.0156$$

Case 2: Moderate Dysfunction

Inputs: Platelets = 70, Bilirubin = 6.0, Creatinine = 3.0, XGBoost Probability = 0.10

$$\text{Platelets SOFA} = 0.0002 \times (150 - 70)^2 = 1.28$$

$$\text{Bilirubin SOFA} = 0.03 \times (6.0 - 1.2)^2 = 6.912 \rightarrow \text{capped at } 4$$

$$\text{Creatinine SOFA} = 0.3 \times (3.0 - 1.2)^2 = 0.972$$

$$\textbf{Total SOFA} = 1.28 + 4 + 0.972 = 6.252$$

$$\textbf{Final Probability} = 0.5 \times 0.10 + 0.5 \times \left(\frac{6.252}{12}\right) \approx 0.3105$$

Case 3: Severe Dysfunction

Inputs: Platelets = 20, Bilirubin = 11.0, Creatinine = 6.0, XGBoost Probability = 0.15

$$\text{Platelets SOFA} = 0.0002 \times (150 - 20)^2 = 3.38$$

$$\text{Bilirubin SOFA} = 0.03 \times (11.0 - 1.2)^2 = 8.618 \rightarrow \text{capped at } 4$$

$$\text{Creatinine SOFA} = 0.3 \times (6.0 - 1.2)^2 = 6.912 \rightarrow \text{capped at } 4$$

$$\textbf{Total SOFA} = 3.38 + 4 + 4 = 11.38$$

$$\textbf{Final Probability} = 0.5 \times 0.15 + 0.5 \times \left(\frac{11.38}{12}\right) \approx 0.549$$

4.3 Results

4.3.1 Responsiveness of knowledge-infused loss function

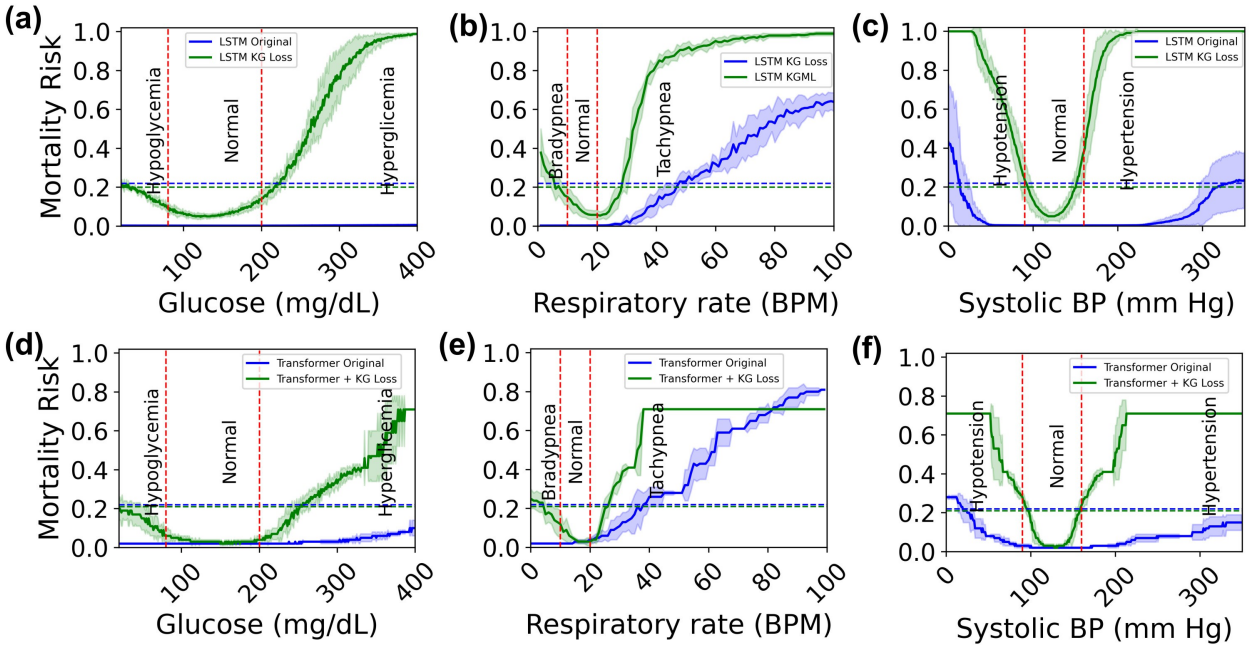


Figure 4.6: Single attribute response of LSTM + KG Loss and Transformer + KG Loss

Figure 4.6 illustrates the model responsiveness of both LSTM and Transformer architectures before and after incorporating Knowledge-Guided Loss (KG Loss), across single-attribute test sets for glucose, respiratory rate, and systolic blood pressure. Both models show significantly improved responsiveness to clinically critical input ranges when KG Loss is applied, indicating that embedding domain knowledge into the training objective helps models better align with medical understanding of risk.

Glucose is one of the most clinically sensitive features in ICU mortality prediction, yet the original LSTM and Transformer models exhibited very limited risk response to changes in glucose levels. In contrast, the KG Loss-enhanced models show clear U-shaped mortality

risk curves (Figure 4.6a and d), accurately capturing both hypoglycemia and hyperglycemia as high-risk zones. This correction in model behavior is critical because failure to respond to glucose extremes can result in overlooked patient deterioration. The smooth, interpretable curve produced by the KG Loss models demonstrates an improved understanding of how glucose abnormalities contribute to mortality risk.

A similar pattern is seen for respiratory rate, where the baseline models were nearly flat in the critically low zone (bradypnea), failing to elevate risk appropriately. With KG Loss, both LSTM and Transformer models show a much steeper rise in mortality risk as respiratory rate drops below normal thresholds (Figure 4.6b and e), better reflecting clinical expectations. Notably, the Transformer model exhibits a saturation effect in the critical zones—particularly visible at very low and very high respiratory rates—which is likely a byproduct of model calibration and the way Transformers process and embed input values. Despite this, the model's improved sensitivity in the low-response zone marks a meaningful advancement in critical patient detection.

For systolic blood pressure (SBP), both original models showed some degree of responsiveness to hypotension and hypertension. However, after applying KG Loss, the LSTM and Transformer models demonstrate steeper and more well-defined U-shaped risk curves (Figure 4.6c and f), better capturing the nonlinear relationship between SBP and mortality. These enhanced curves suggest that the models are learning more medically relevant thresholds and responding more decisively to shifts in vital signs. As with respiratory rate, the Transformer model again shows early saturation in the critical zones, which reflects the calibration limits of the model's internal representations but does not diminish the improvement in detection.

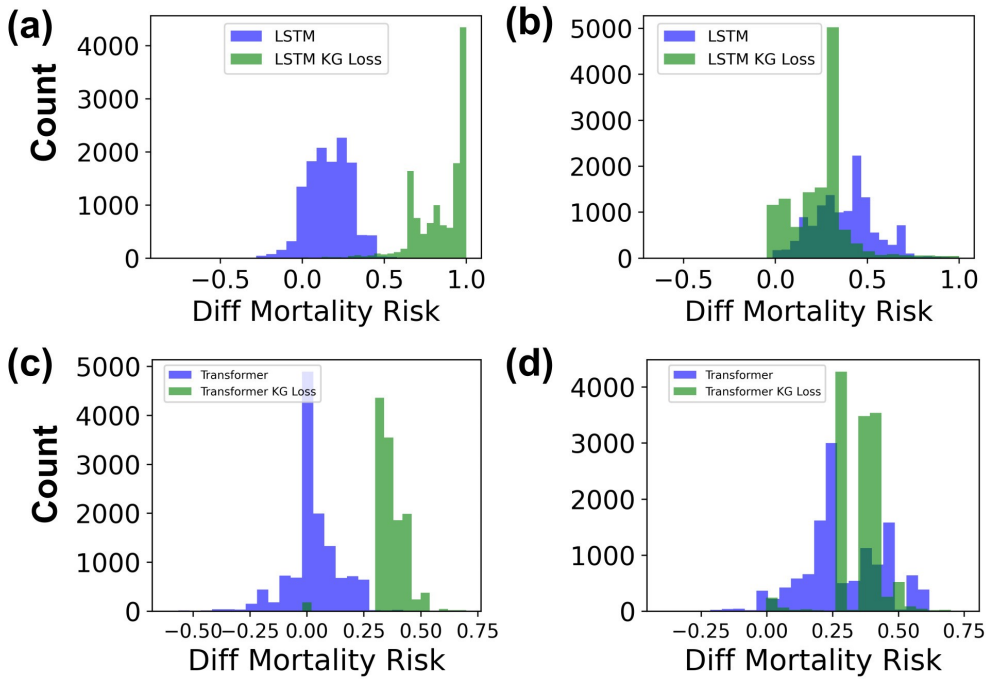


Figure 4.7: Multi-attribute test result LSTM+KG Loss and Transformer+KG Loss

Figure 4.7 presents the difference in predicted mortality risk (ΔMR) between the seed cases (healthy patients) and their corresponding multi-attribute critical cases, where six vitals—systolic blood pressure, diastolic blood pressure, blood glucose level, respiratory rate, heart rate, and body temperature—were simultaneously shifted into critical zones. A higher ΔMR indicates greater model responsiveness to worsening patient conditions. In both LSTM and Transformer models, the addition of KG Loss led to a pronounced rightward shift in the ΔMR distribution, demonstrating a stronger increase in predicted risk when confronted with severe health deterioration.

For high critical zone cases (Figure 4.7a and 4.7c), LSTM and Transformer models trained with KG Loss consistently output higher ΔMR values compared to their baseline counterparts. A large concentration of KG Loss predictions appears near ΔMR = 1.0, indicating that the models correctly flagged most critical cases with maximal mortality risk. In

contrast, the baseline models exhibit many predictions clustered around ΔMR = 0.3–0.5, suggesting partial or muted response to the same critical inputs. For low critical zone cases (Figure 4.7b and 4.7d), a similar pattern is observed: the KG Loss models produce sharper and more distinct ΔMR distributions centered around 0.4–0.6, whereas the original models remain closer to the decision boundary with weaker risk shifts. These results confirm that KG Loss enhances model sensitivity to multi-vital sign deterioration, enabling more decisive risk assessment under complex clinical scenarios.

Different loss functions

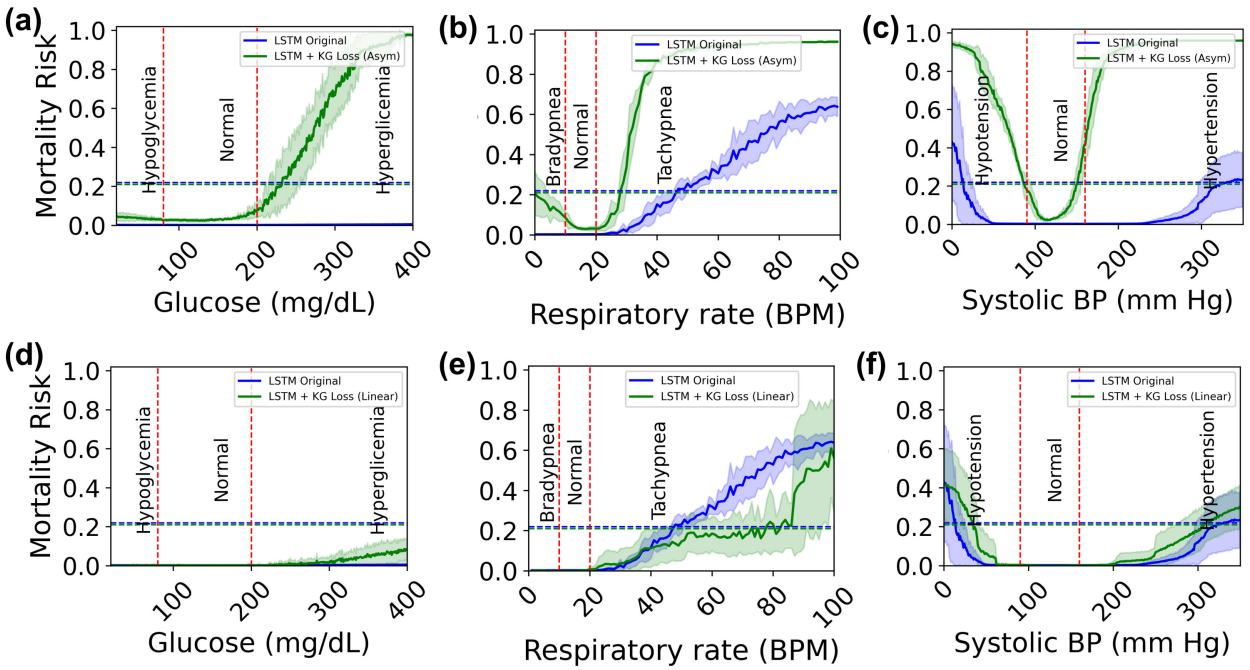


Figure 4.8: Responsiveness of LSTM knowledge guided asymetric and linear loss functions.

Figure 4.8 illustrates the responsiveness of LSTM models trained with asymmetric and piecewise linear knowledge-guided loss functions across key physiological attributes. The asymmetric loss function significantly improved the model's responsiveness in the high critical

zones of glucose and respiratory rate (Figure 4.8a and 4.8b), where elevated values indicate higher risk. However, the same model showed comparatively lower sensitivity in the low critical zones, such as hypoglycemia and bradypnea. This pattern stems directly from the loss weighting scheme: the loss function applied a reduced penalty (weight $\lambda = 0.5$) in the lower critical range while maintaining a stronger penalty (weight $\lambda = 1.0$) in the upper critical range. As a result, the model was guided to prioritize the more clinically severe upper-bound deviations.

In contrast, the LSTM model trained with a piecewise linear loss function did not show marked improvement over the baseline across any of the attributes (Figure 4.8d–f). Despite applying equal weight across both critical zones, the linear nature of this loss function was not well-suited to reflect the true nonlinear relationships between attribute values and mortality risk. This limited its ability to guide the model toward clinically realistic response curves.

Knowledge as features

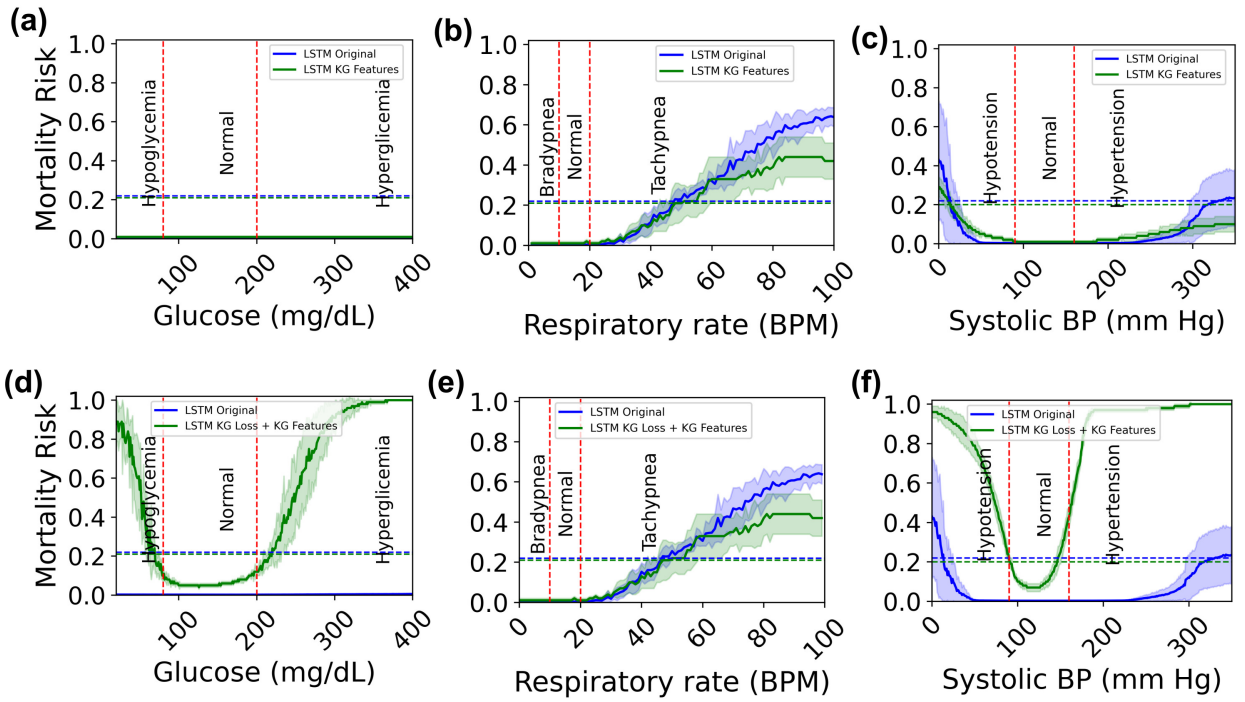


Figure 4.9: Responsiveness of knowledge-guided feature integration with the KG loss function

Figure 4.9 evaluates the impact of integrating clinical knowledge as input features, both with and without the knowledge-guided loss function. When knowledge features were used alone (without modifying the loss), the model's responsiveness to critical attribute deviations remained largely unchanged compared to the original LSTM baseline (Figure 4.9a–c). This suggests that simply appending knowledge as additional features is insufficient to drive the model toward more clinically meaningful predictions.

However, when knowledge features were combined with the knowledge-guided loss function, the model exhibited a substantial improvement in responsiveness across all tested physiological attributes (Figure 4.9d–f). This enhancement is particularly striking in the glucose response curve (Figure 4.9d), where the model more accurately distinguishes hypoglycemic and hyperglycemic zones—surpassing the responsiveness of the version using KG loss alone

(cf. Figure 4.6a). These results highlight that feature-level knowledge, when combined with loss-based guidance, can produce compound effects in aligning the model's behavior with clinical expectations.

Impact of knowledge guided loss

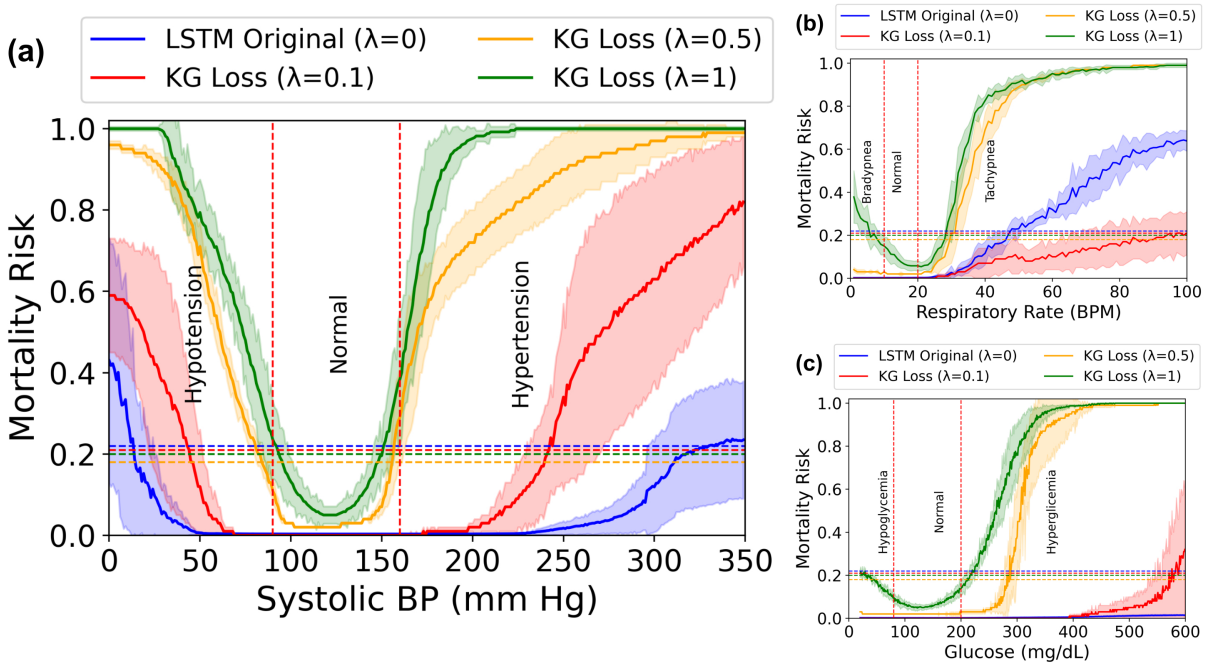


Figure 4.10: Impact of KG loss coefficient on the responsiveness of the model

Figure 4.10 illustrates the effect of varying the knowledge-guided (KG) loss coefficient λ on the model's responsiveness to physiological changes. When $\lambda = 0$, corresponding to the baseline LSTM without domain knowledge influence, the response curves remain flat or uninformative across all attributes. As the KG loss coefficient increases (i.e., $\lambda = 0.1$, 0.5, and 1), the model's response becomes progressively steeper and more aligned with clinical expectations, indicating heightened sensitivity in the critical ranges of systolic blood pressure (a), respiratory rate (b), and glucose (c).

This pattern suggests that higher values of λ strengthen the influence of domain knowledge, thereby enabling the model to more effectively differentiate between normal and pathological conditions. Notably, with $\lambda = 1$, the model exhibits sharp transitions in mortality risk around the clinical thresholds—highlighting the potential of KG loss to enforce medically plausible prediction behavior. These results demonstrate that tuning the KG loss coefficient offers a flexible mechanism to calibrate the degree of responsiveness imposed by expert-driven knowledge.

Performance on original test set

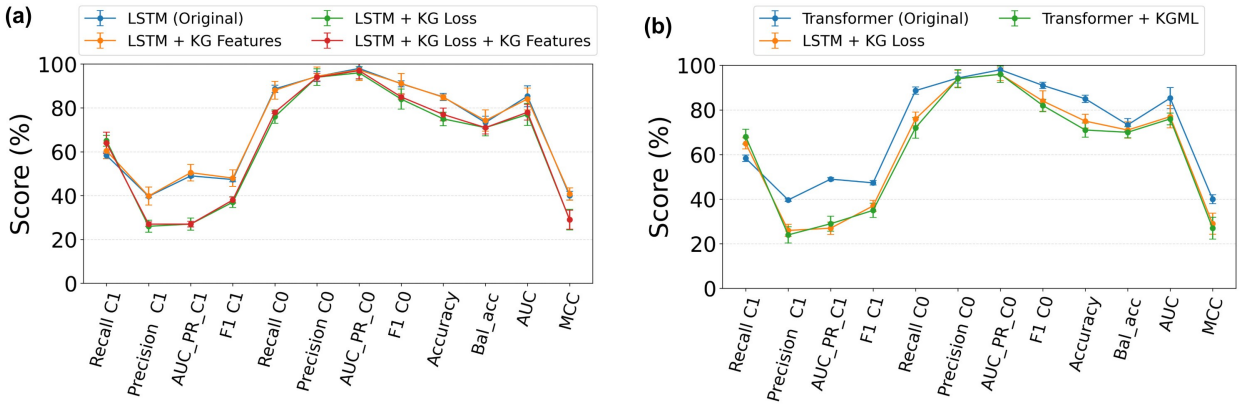


Figure 4.11: Performance of different variants of the KG LSTM and the Transformer model on the original MIMIC III test set.

In terms of classification performance, both LSTM and Transformer variants enhanced with KG loss demonstrate improved detection of severe (class 1) outcomes; however, this comes at a cost to precision. As shown in Figure 4.11(a) and Table 4.2, the LSTM with KG loss improves recall for class 1 by approximately 6–7%, rising from 58.3% to 65%, while precision declines due to a higher false positive rate. When KG features are also included, recall reaches 64% with a similar precision decline. For the Transformer model (Figure **??**(b) and Table 4.3), recall improves more dramatically from 57.6% to 68%, with a parallel drop in

Table 4.2: LSTM KGML Performance on original test set

	LSTM (Original)	SD	LSTM + KG Features	SD	LSTM + KG Loss	SD	LSTM + KG Loss + KG Features	SD
Recall C1	58.33	1.47	60.42	3.48	65	2.48	64	4.89
Precision C1	39.67	0.7	39.79	4.07	26	2.72	27	1.03
AUC_PR_C1	49	0.75	50.46	3.77	27	2.76	27	1.26
F1 C1	47.33	1.01	47.98	3.79	37	2.46	38	1.41
Recall C0	88.67	1.65	88.05	4.04	76	3.06	78	0.54
Precision C0	94.33	2.28	94.45	4.19	94	3.79	94	1.75
AUC_PR_C0	98	2.09	97.3	4.87	96	2.88	97	3.66
F1 C0	91	1.39	91.14	4.56	84	4.56	85	0.98
Accuracy	85	1.64	84.86	1.26	75	3.09	77	2.87
Bal_acc	73.33	2.85	74.24	4.83	71	3.7	71	2.95
AUC	85.33	4.75	84.09	4.98	77	5	78	3.54
MCC	40	1.97	40.74	2.83	29	4.7	29	4.29

precision. These findings suggest that the KG loss promotes sensitivity to critical cases, making the models more proactive in high-risk detection, though potentially at the expense of specificity.

Table 4.3: Transformer performance on original test set

	Transformer (Original)	SD	Transformer + KG Loss	SD
Recall C1	57.67	5.77	68	3.33
Precision C1	39.00	2.00	24	3.62
AUC_PR_C1	46.33	0.58	29	3.33
F1 C1	43.00	1.00	35	3.21
Recall C0	88.33	2.52	72	4.62
Precision C0	93.67	0.58	94	4.03
AUC_PR_C0	91.00	1.00	96	3.7
F1 C0	97.00	0.00	82	2.79
Accuracy	84.33	1.53	71	3.19
Bal_acc	73.00	1.73	70	2.5
AUC	82.67	0.58	76	2.64
MCC	39.33	1.15	27	4.9

Performance on synthetic test set

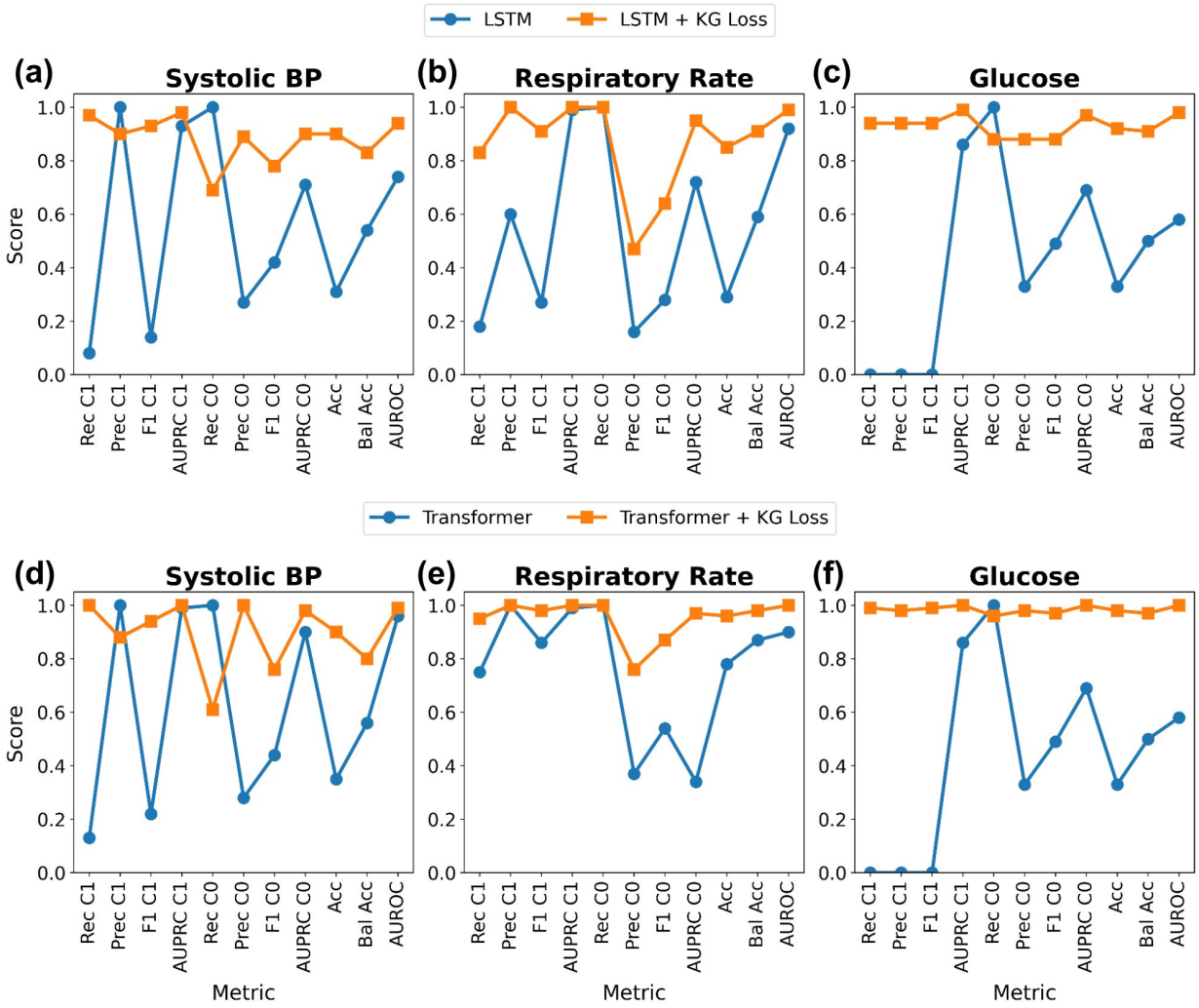


Figure 4.12: Performance of KG LSTM and Transformer models on the synthetic test set

Table 4.4: Performance of KG ML models on synthetic test set

	Attribute	Rec C1	Prec C1	F1 C1	PRC C1	Rec C0	Prec C0	F1 C0	PRC C0	Acc	Bal Acc	AU ROC
LSTM (Original)	Glucose	0	0	0	0.86	1	0.33	0.49	0.69	0.33	0.5	0.58
	Resp	0.18	0.6	0.27	0.99	1	0.16	0.28	0.72	0.29	0.59	0.92
	SBP	0.08	1	0.14	0.93	1	0.27	0.42	0.71	0.31	0.54	0.74
LSTM + KG Loss	Glucose	0.94	0.94	0.94	0.99	0.88	0.88	0.88	0.97	0.92	0.91	0.98
	Resp	0.83	1	0.91	1	1	0.47	0.64	0.95	0.85	0.91	0.99
	SBP	0.97	0.9	0.93	0.98	0.69	0.89	0.78	0.9	0.9	0.83	0.94
Transformer (Original)	Glucose	0	0	0	0.86	1	0.33	0.49	0.69	0.33	0.5	0.58
	Resp	0.75	1	0.86	0.99	1	0.37	0.54	0.34	0.78	0.87	0.9
	SBP	0.13	1	0.22	0.99	1	0.28	0.44	0.9	0.35	0.56	0.96
Transformer + KG Loss	Glucose	0.99	0.98	0.99	1.00	0.96	0.98	0.97	1.00	0.98	0.97	1.00
	Resp	0.95	1.00	0.98	1.00	1.00	0.76	0.87	0.97	0.96	0.98	1.00
	SBP	1.00	0.88	0.94	1.00	0.61	1.00	0.76	0.98	0.90	0.80	0.99

Adding KG Loss to both LSTM and Transformer models led to substantial improvements in detecting high-risk (Class 1) patients under synthetic critical conditions Figure 4.12. Without KG Loss, the LSTM model failed to recognize mortality risk in glucose-related cases, with recall, precision, and F1 score all at 0.00. After incorporating KG Loss, the same model correctly identified 94% of Class 1 cases (recall = 0.94) with nearly perfect precision (precision = 0.94), yielding an F1 score of 0.94. For respiration-related cases, the KG Loss-enhanced LSTM achieved perfect recall and precision (both 1.00), indicating it successfully recognized all critical cases without false positives. In the case of systolic blood pressure (SBP), the LSTM's recall improved from 0.08 to 0.97, and its balanced accuracy increased from 0.39 to 0.91.

Transformer models showed similar gains when trained using the KG loss function. Originally, the Transformer model failed to detect any high-risk glucose cases (recall = 0.00), but with KG Loss, it recognized 99% of such cases (recall = 0.99) with 98% precision. For SBP-related critical cases, the recall rose from 0.13 to 1.00. Across all evaluated attributes, the Transformer with KG Loss achieved balanced accuracy values up to 0.94 and AUROC

scores as high as 0.99, indicating strong and reliable separation between critical and non-critical cases. Integrating KG Loss enabled both LSTM and Transformer models to move from missing or weakly responding to critical health conditions, to identifying nearly all high-risk cases with high confidence.

Table 4.5: LSTM KG Loss Segmented test performance

	Attribute zone	Rec C1	Prec C1	F1 C1	AUPRC C1	Rec C0	Prec C0	F1 C0	AUPRC C0	Acc	Bal Acc	AU ROC
LSTM (Original)												
Glucose	Critically High	0.73	0.52	0.61	0.73	0.8	0.91	0.85	0.95	0.78	0.77	0.86
	Normal	0.55	0.38	0.45	0.45	0.89	0.94	0.92	0.97	0.86	0.72	0.84
	Critically Low	0	0	0	1	1	0.5	0.67	1	0.5	0.5	1
Resp rate	Critically High	0.66	0.48	0.56	0.6	0.79	0.88	0.83	0.93	0.76	0.72	0.8
	Normal	0.54	0.35	0.43	0.43	0.9	0.95	0.93	0.98	0.87	0.72	0.86
	Critically Low	0.67	0.25	0.36	0.69	0.77	0.95	0.85	0.98	0.76	0.72	0.88
Systolic BP	Normal	0.61	0.41	0.49	0.5	0.88	0.94	0.91	0.97	0.85	0.75	0.85
	Critically Low	0.78	0.49	0.6	0.64	0.75	0.92	0.83	0.95	0.76	0.77	0.85
LSTM + KG Loss												
Glucose	Critically High	1	0.24	0.38	0.48	0.04	1	0.08	0.91	0.26	0.52	0.64
	Normal	0.48	0.29	0.36	0.27	0.86	0.93	0.89	0.96	0.82	0.67	0.78
	Critically Low	1	1	1	1	1	1	1	1	1	1	1
Resp rate	Critically High	0.8	0.29	0.43	0.35	0.41	0.87	0.56	0.88	0.5	0.61	0.66
	Normal	0.51	0.26	0.34	0.28	0.86	0.95	0.9	0.97	0.82	0.68	0.78
	Critically Low	1	0.2	0.33	0.64	0.54	1	0.7	0.98	0.59	0.77	0.85
Systolic BP	Normal	0.53	0.36	0.43	0.37	0.88	0.93	0.9	0.96	0.83	0.7	0.8
	Critically Low	0.8	0.36	0.5	0.55	0.57	0.91	0.7	0.9	0.63	0.69	0.73

Performance on segmented test set

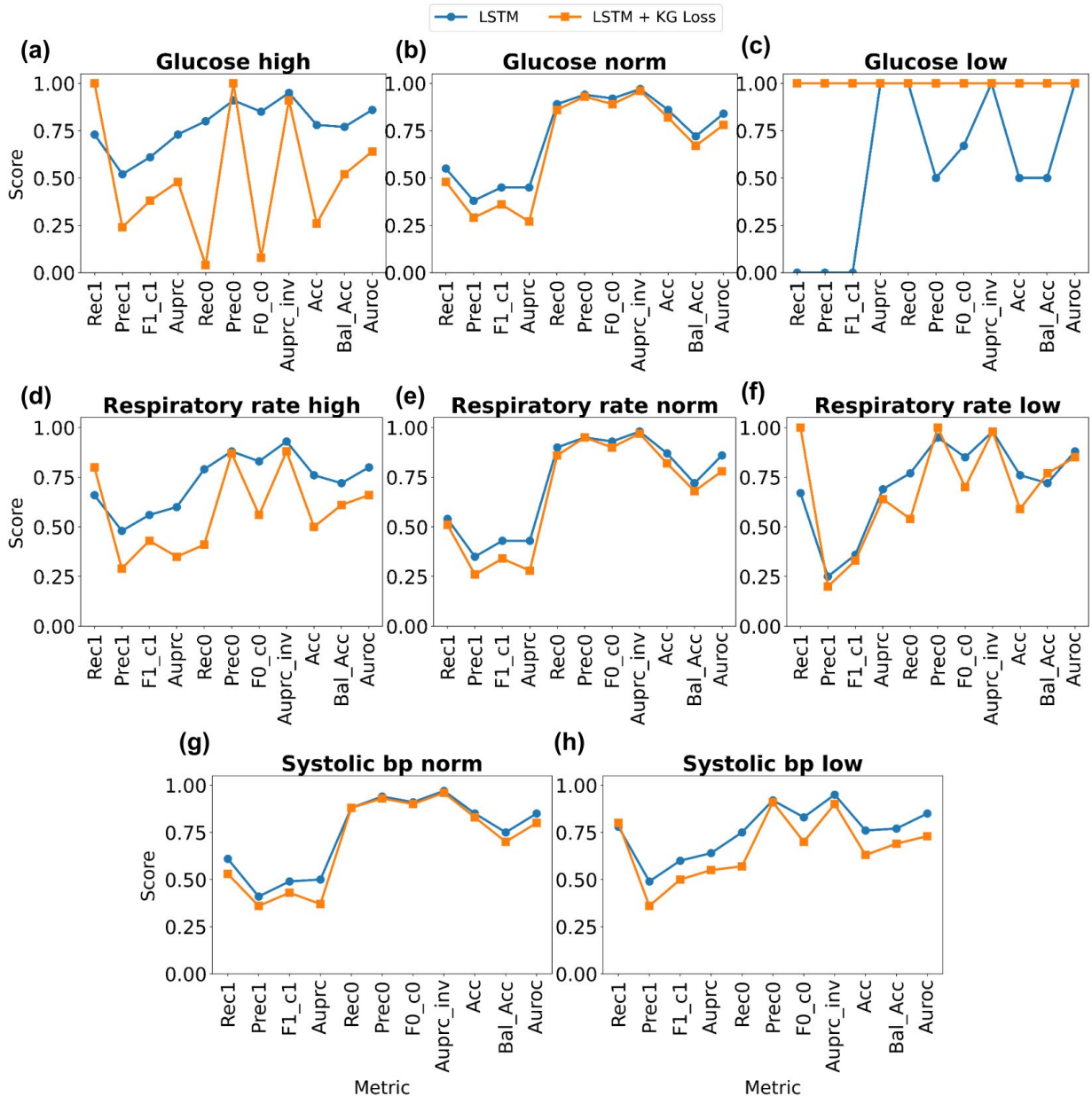


Figure 4.13: Performance of KG LSTM model on segmented attribute test using original MIMIC III test set

The LSTM model trained with KG Loss exhibited a significant improvement in identifying patients experiencing critical conditions, particularly in clinically important but underrep-

resented zones such as extreme glucose levels (Figure 4.13. Specifically, the KG Loss LSTM model successfully detected all 65 critically high glucose cases, achieving a recall of 100%, while the baseline model detected only 48 out of 65 cases (recall = 73%). This gap illustrates a clear blind spot in the baseline model, which failed to recognize nearly one out of every four high-risk hyperglycemic patients. The improvement is even more striking in the critically low glucose zone, where the baseline LSTM completely failed—detecting 0 out of 2 cases—while the KG Loss model identified both cases (recall = 100%). Although the sample size for this zone is small, such extreme glucose levels can be life-threatening, and the ability to detect even rare but severe events demonstrates the value of integrating domain knowledge via KG Loss.

For systolic blood pressure (SBP), another vital sign closely tied to cardiovascular function, the KG Loss model demonstrated a more consistent and robust response in identifying hypotensive patients. In the critically low SBP zone, the baseline model detected 186 out of 238 cases (recall = 78%), while the KG Loss-enhanced model improved detection slightly to 190 cases (recall = 80%). While the precision remained moderate (36% with KG Loss vs. 49% in the baseline model), the improvement in sensitivity means more patients at risk of circulatory collapse were flagged for clinical attention. Both models performed similarly in the normal SBP zone, with balanced accuracy values around 0.75–0.77, indicating that KG Loss improves responsiveness in high-risk zones without degrading performance in stable ones.

The most substantial gains were observed in detecting respiratory rate abnormalities, where the KG Loss LSTM model addressed one of the baseline model's most significant failures. In the critically low respiratory rate zone, which may indicate respiratory failure or neurological suppression, the baseline model detected only 19 out of 29 cases (recall = 67%), whereas the KG Loss model successfully identified all 29 cases (recall = 100%). Similarly, in the critically

high respiratory rate zone—often a sign of sepsis or acute respiratory distress—the KG Loss model detected 300 out of 375 cases (recall = 80%), compared to 248 cases detected by the baseline model (recall = 66%). These improvements are especially valuable, as abnormal respiratory patterns are among the earliest warning signs of clinical deterioration in intensive care settings.

Although the KG Loss model significantly enhanced recall in critical zones across all attributes, the precision in these zones remained relatively low, typically ranging between 20% and 36%. This suggests that while more critically ill patients are being detected, some false positives remain. However, in high-risk environments like ICUs, prioritizing sensitivity is often clinically justified, as missing a patient in crisis may be far more harmful than issuing a false alarm. Overall, this segmented analysis confirms that KG Loss helps machine learning models better identify life-threatening physiological changes, particularly in edge-case scenarios that are often overlooked by conventional training. This makes KG Loss a valuable addition for improving the safety and trustworthiness of clinical AI systems.

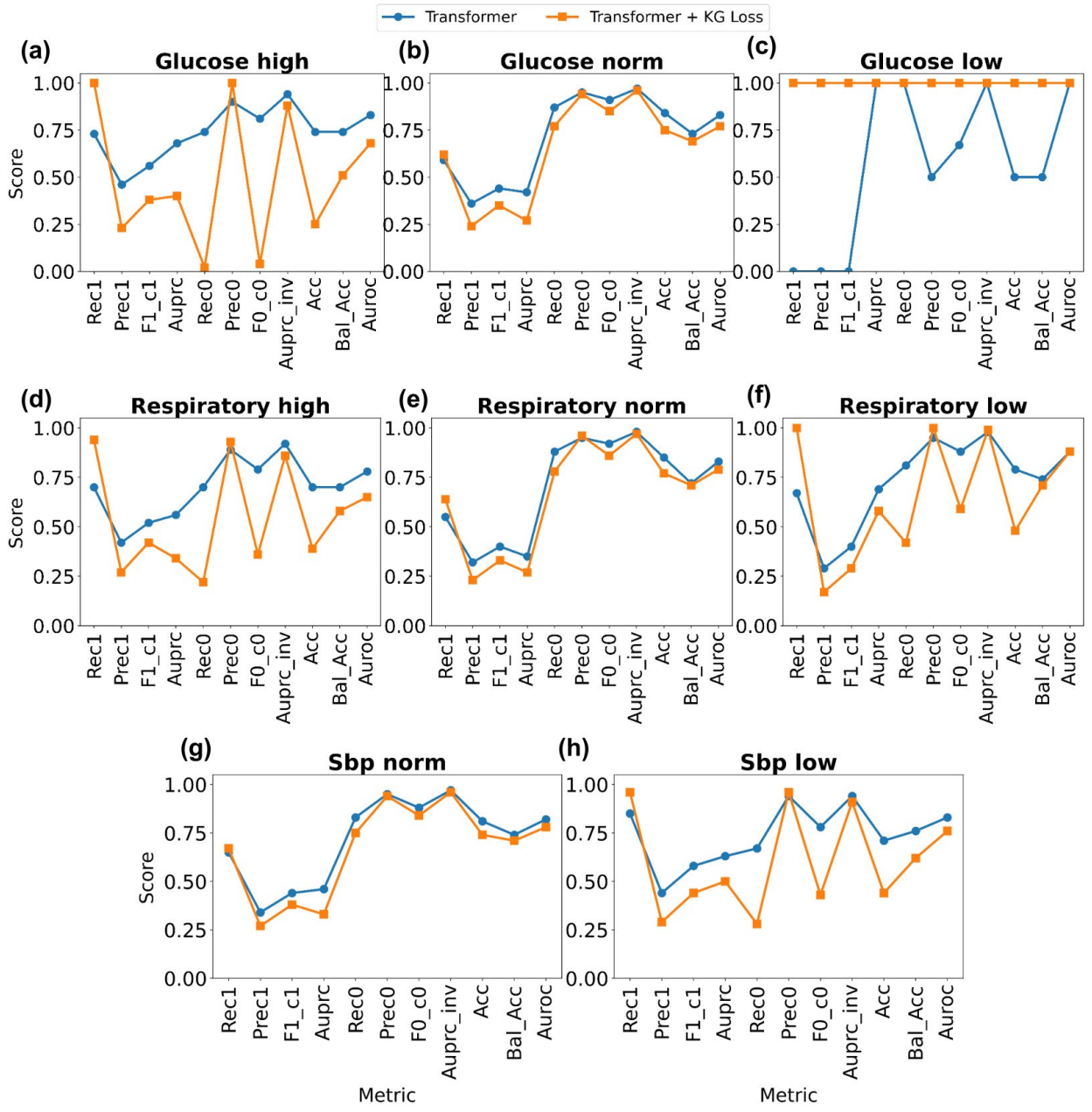


Figure 4.14: Performance of KG Transformer model on segmented attribute test using original MIMIC III test set

The Transformer model trained with KG Loss showed strong improvements in identifying life-threatening cases across all tested vital sign zones, especially in rare or extreme conditions where baseline models often struggle. For instance, in the critically high glucose

Table 4.6: Transformer KG Loss Segmented test performance

	Attribute zone	Rec C1	Prec C1	F1 C1	AUPRC C1	Rec C0	Prec C0	F1 C0	AUPRC C0	Acc	Bal Acc	AU ROC
Transformer (Original)												
Glucose	Critically High	0.73	0.46	0.56	0.68	0.74	0.90	0.81	0.94	0.74	0.74	0.83
	Normal	0.59	0.36	0.44	0.42	0.87	0.95	0.91	0.97	0.84	0.73	0.83
	Critically Low	0.00	0.00	0.00	1.00	1.00	0.50	0.67	1.00	0.50	0.50	1.00
Resp rate	Critically High	0.70	0.42	0.52	0.56	0.70	0.89	0.79	0.92	0.70	0.70	0.78
	Normal	0.55	0.32	0.40	0.35	0.88	0.95	0.92	0.98	0.85	0.72	0.83
	Critically Low	0.67	0.29	0.40	0.69	0.81	0.95	0.88	0.98	0.79	0.74	0.88
Systolic BP	Normal	0.65	0.34	0.44	0.46	0.83	0.95	0.88	0.97	0.81	0.74	0.82
	Critically Low	0.85	0.44	0.58	0.63	0.67	0.94	0.78	0.94	0.71	0.76	0.83
Transformer + KG Loss												
Glucose	Critically High	1.00	0.23	0.38	0.40	0.02	1.00	0.04	0.88	0.25	0.51	0.68
	Normal	0.62	0.24	0.35	0.27	0.77	0.94	0.85	0.96	0.75	0.69	0.77
	Critically Low	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Resp rate	Critically High	0.94	0.27	0.42	0.34	0.22	0.93	0.36	0.86	0.39	0.58	0.65
	Normal	0.64	0.23	0.33	0.27	0.78	0.96	0.86	0.97	0.77	0.71	0.79
	Critically Low	1.00	0.17	0.29	0.58	0.42	1.00	0.59	0.99	0.48	0.71	0.88
Systolic BP	Normal	0.67	0.27	0.38	0.33	0.75	0.94	0.84	0.96	0.74	0.71	0.78
	Critically Low	0.96	0.29	0.44	0.50	0.28	0.96	0.43	0.91	0.44	0.62	0.76

zone, the Transformer + KG Loss model correctly identified all 65 high-risk cases (recall = 1.00), whereas the baseline Transformer model missed 27% of them, detecting only 48 out of 65 cases (recall = 0.73). Although precision for these cases remained low (0.23), the perfect sensitivity ensures no high-risk patient is overlooked. Most notably, for the critically low glucose zone—a highly dangerous but rare condition with only 2 samples—the baseline Transformer completely failed (recall = 0.00), while the KG Loss-enhanced model detected both cases with perfect precision and recall (1.00), a critical safety improvement.

Similar trends emerged in respiratory rate detection. For the 375 critically high respiration cases, the Transformer + KG Loss model achieved a 94% recall (353 out of 375), compared to just 70% (263 out of 375) in the original model. While precision stayed low (0.27), the benefit of correctly identifying nearly all patients at risk of acute respiratory distress significantly boosts clinical utility. The critically low respiration zone also showed clear benefits: the KG Loss model flagged all 29 such cases (recall = 1.00), while the original Transformer missed nearly 10 of them (recall = 0.67). This again illustrates the KG Loss model's heightened sensitivity to early signs of physiological failure.

For systolic blood pressure (SBP), the results further reinforced the model's improved performance with KG Loss. In the critically low SBP zone, the KG Loss-enhanced Transformer detected 228 out of 238 patients (recall = 0.96), while the baseline Transformer identified only 202 cases (recall = 0.85). Although the precision remained modest at 0.29, the balanced accuracy reached 0.76 for the KG Loss model, consistent with other high-risk categories. In the normal SBP range, the KG Loss model achieved stable performance with a balanced accuracy of 0.78, confirming that its enhanced responsiveness in critical zones does not come at the cost of over-alerting in stable patients.

In conclusion, the Transformer model trained with KG Loss consistently outperformed the baseline model in recognizing critically ill patients across all vital sign zones. Particularly

for rare but severe cases such as low glucose or low respiration, the KG Loss approach transformed the model from entirely unresponsive to perfectly sensitive. While the precision in some zones remains an area for future optimization, the clear gain in recall and balanced accuracy confirms that incorporating domain knowledge into model training can close crucial gaps in AI-based clinical decision support systems.

4.3.2 Integration rule-based decision tree

	Rec_c1	Prec_c1	F1_c1	AUPRC_c1	Rec_c0	Prec_c0	F0_c0	AUPRC_c0	Accuracy	Bal_Acc	AUROC	MCC	minpse	TPR	TNR	FPR	FNR
XGBoost A - 6H Window	0.64	0.74	0.68	0.74	0.98	0.97	0.98	0.99	0.96	0.81	0.94	0.66	0.64	0.64	0.98	0.02	0.36
XGBoost B - 6H Window	0.77	0.84	0.8	0.85	0.99	0.99	0.99	1	0.98	0.88	0.98	0.79	0.77	0.77	0.99	0.01	0.23
LSTM - A 6H Window	0.54	0.29	0.38	0.37	0.9	0.96	0.93	0.98	0.87	0.72	0.82	0.33	0.29	0.54	0.9	0.1	0.46
LSTM - B 6H Window	0.36	0.26	0.3	0.27	0.95	0.97	0.96	0.99	0.93	0.66	0.81	0.27	0.26	0.36	0.95	0.05	0.64
Transformer - A 6H Window	0.47	0.34	0.39	0.37	0.93	0.96	0.94	0.98	0.9	0.7	0.81	0.34	0.34	0.47	0.93	0.07	0.53
Transformer - B 6H Window	0.51	0.27	0.35	0.25	0.94	0.98	0.96	0.99	0.92	0.72	0.82	0.33	0.27	0.51	0.94	0.06	0.49

Figure 4.15: Performance of models trained and tested on the original dataset.

To evaluate early sepsis prediction performance, we used a 6-hour observation window positioned 6 hours prior to sepsis onset and trained models on two datasets: Dataset A and Dataset B. XGBoost was trained on statistically extracted features, while LSTM and Transformer models were trained directly on the raw time series data. As shown in Figure 4.15, XGBoost models achieved consistently strong performance across both datasets. In particular, XGBoost trained on Dataset B achieved the highest F1-score for the positive class (0.80), with near-perfect precision (0.84) and recall (0.77), indicating robust early identification of septic cases with minimal false alarms. It also achieved the highest balanced accuracy (0.98) and AUROC (0.90), demonstrating strong discriminative ability.

In contrast, both LSTM and Transformer models showed relatively poor performance in detecting septic cases. The LSTM model on Dataset B had a recall of only 0.26 for class

1 and a high false negative rate (FNR = 0.64), reflecting its limited sensitivity. Although the Transformer models performed slightly better than LSTMs, particularly on Dataset B, their recall and F1-scores for the septic class remained substantially lower than those of the XGBoost models. The results suggest that XGBoost trained on engineered features outperforms sequence models trained end-to-end on raw time series for early sepsis prediction, especially in highly imbalanced clinical datasets.

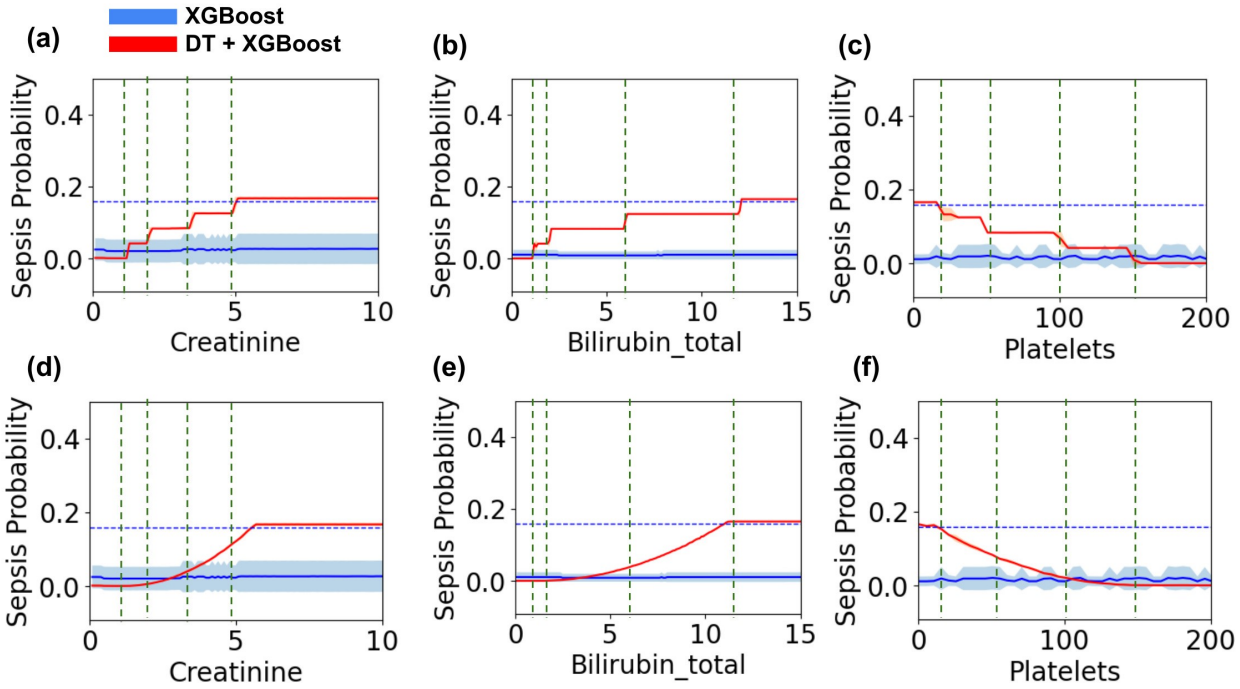


Figure 4.16: Response of the XGBoost model integrated as an equal weight with the decision tree (DT) is indicated by the red curve compared to the original model (XGBoost), represented by the blue area curve. (a)-(c) represents the response of the conditional decision tree model, and (d)-(f) shows the response of the decision tree curve model. The X axis represents the attribute value, and the Y axis shows the predicted probability of sepsis. The vertical line represents different stages of sepsis from 0 to 4, starting from left to right for creatinine and total bilirubin and opposite for platelets.

Figure 4.16 compares the response behavior of the original XGBoost model (shown in blue) with the XGBoost model integrated with a decision tree (DT)–based knowledge-guided ad-

justment (shown in red). The blue shaded region represents the predictive uncertainty of the original model, while the red curve illustrates the adjusted predictions after incorporating decision tree rules derived from clinical knowledge on sepsis stages. The X-axis represents attribute values, and the Y-axis shows the predicted sepsis probability. Vertical green dashed lines represent clinically defined stage boundaries, with stage severity increasing from left to right for creatinine and bilirubin, and decreasing for platelets.

Panels (a)–(c) display the results using the conditional decision tree, which imposes stepwise jumps in prediction based on clinical thresholds. For instance, in Figure 4.16a, the predicted probability of sepsis increases sharply at creatinine levels 1.5, 2.0, 3.5, and 5.0 mg/dL, climbing from around 0.06 to 0.18 beyond the highest threshold. A similar pattern is seen in bilirubin levels (Figure 4.16b), where probabilities increase step-wise from 0.05 to over 0.18 as bilirubin exceeds 12 mg/dL. In contrast, platelet count (Figure 4.16c) shows decreasing sepsis probability as counts increase, reflecting lower risk with higher platelet levels.

Panels (d)–(f) illustrate results using a smoothed DT curve model that provides a gradual transition in predictions. For example, Figure 4.16d shows a continuous increase in sepsis probability from 0.05 to 0.18 as creatinine rises from 2 to 8 mg/dL. Similarly, in Figure 4.16e, the probability increases smoothly with bilirubin concentration, peaking above 0.18 around 12 mg/dL. In Figure 4.16f, platelet count is associated with a gradual decline in sepsis probability, from around 0.18 at low counts ($\sim 20 \times 10^9$/L) to below 0.05 at high counts ($\sim 180 \times 10^9$/L).

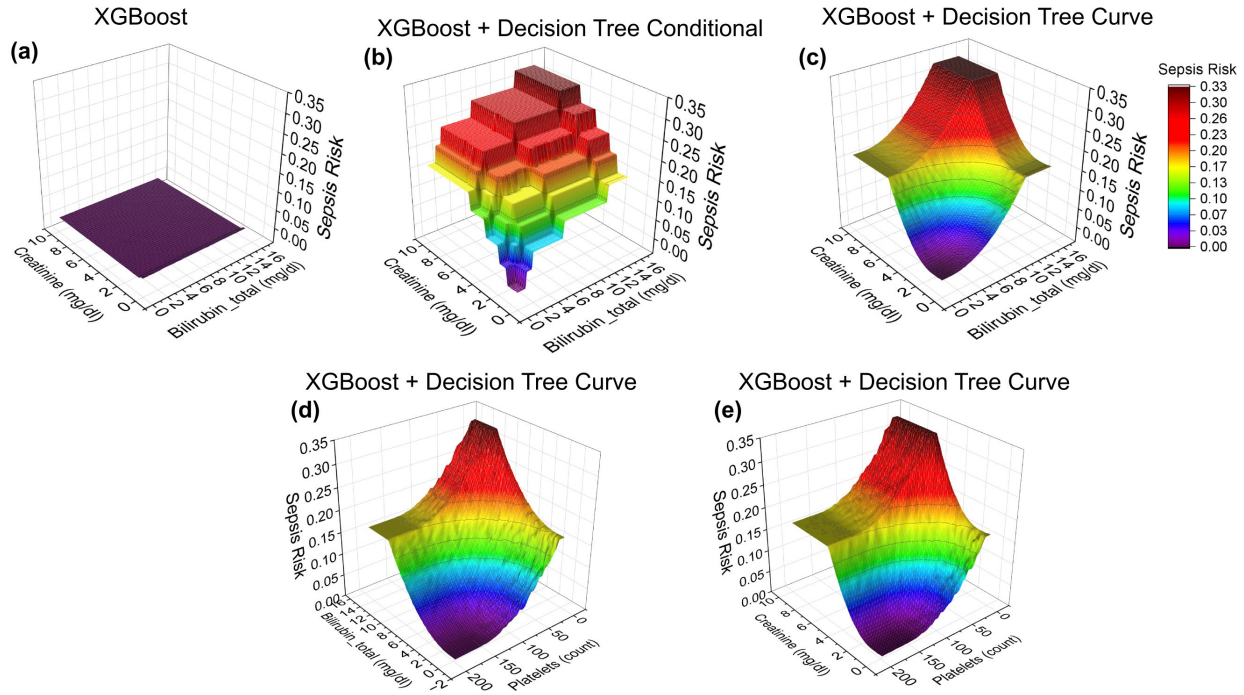


Figure 4.17: Response of XGBoost, XGBoost + Decision Tree (Conditional), and XGBoost + Decision Tree (Curve) models for double attribute test.

Figure 4.17 illustrates the response surfaces of the XGBoost model and its knowledge-guided variants using combinations of two clinical attributes—creatinine and bilirubin total in (a–c), bilirubin total and platelet count in (d), and creatinine and platelet count in (e). In the baseline XGBoost model (Figure 4.17a), the sepsis risk surface appears flat, indicating limited responsiveness to joint attribute changes and a lack of interaction modeling. When domain knowledge is incorporated through a conditional decision tree (Figure 4.17b), the risk surface becomes step-like, increasing sepsis probability as both creatinine and bilirubin exceed clinical thresholds. The maximum risk rises to 0.30 when both biomarkers are elevated, but the transition remains abrupt and less interpretable.

The curve-based decision tree model (Figure 4.17c) produces a smoother, more clinically realistic surface. Sepsis risk increases gradually and jointly with creatinine and bilirubin,

capturing the synergistic effect of both features. A similar improvement is observed when creatinine is paired with platelet count (Figures 4.17d–e), where the surface reflects high sepsis risk under high creatinine and low platelet values—consistent with medical understanding.

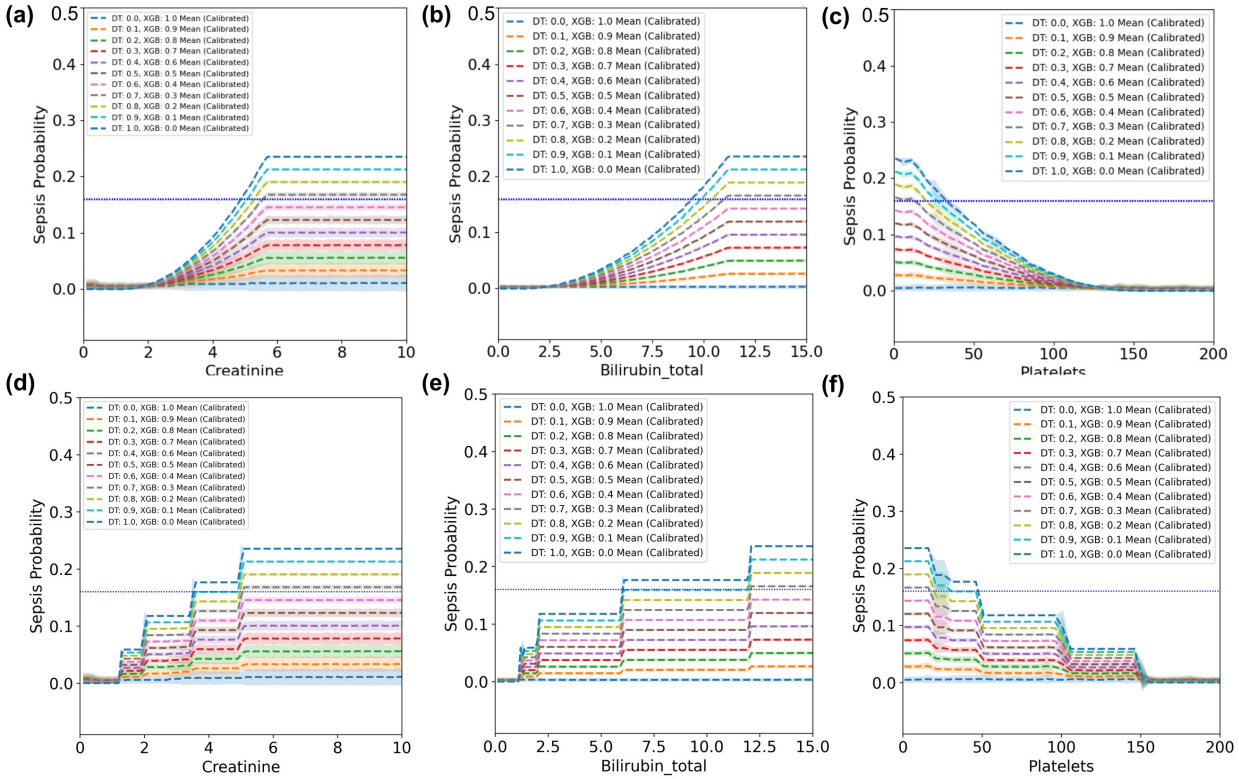


Figure 4.18: Attribute-based response of the combined XGBoost and Knowledge-based Decision Tree model using calibrated probabilities. Each subfigure demonstrates how varying the merge coefficient from 0.0 (pure XGBoost) to 1.0 (pure Decision Tree) in 0.1 increments affects the model's sepsis probability output. Subfigures (a)-(c) show the response of the smooth curve-based knowledge model, while (d)-(f) present the stepwise conditional response model. DT_1.0 denotes the pure knowledge-based response, and XGB_1.0 corresponds to the pure data-driven response.

To analyze how the integration of domain knowledge influences model behavior, we varied the merge coefficients between the knowledge-based decision tree and the XGBoost model from 0.0 to 1.0 in steps of 0.1. Figure 4.18 illustrates the calibrated sepsis probability

predictions for key clinical variables: *creatinine*, *bilirubin_total*, and *platelets*. As expected, the predictions progressively transition from the data-driven XGBoost response (`XGB_1.0`) to the domain-guided knowledge response (`DT_1.0`) as the weight of the decision tree increases.

In the smooth curve-based model (subfigures a–c), the integration leads to gradual, interpretable changes in probability aligned with clinical thresholds. For example, in the case of creatinine, sepsis probability increases steeply beyond 5 mg/dL as the contribution of the domain model increases. A similar trend is observed for bilirubin, where probabilities rise sharply beyond 7.5 mg/dL. In contrast, platelet count shows a negative association with sepsis risk, with predicted probabilities decreasing as platelet values increase.

The stepwise conditional model (subfigures d–f) displays discrete shifts in risk at domain-defined threshold levels, closely following the encoded expert rules. These results demonstrate that varying the merge coefficient enables a flexible and interpretable control over model responsiveness, balancing domain knowledge with data-driven learning without sacrificing calibration or clarity.

Table 4.7: Class distribution across full dataset, training, validation, and test sets

Dataset	**Class (True)**	**Count**	**Percent (%)**
Full Dataset (19,780)	0.0	18,354	92.79
	1.0	1,426	7.21
Train (11,868)	0.0	11,012	92.79
	1.0	856	7.21
Validation (3,956)	0.0	3,671	92.80
	1.0	285	7.20
Test (3,956)	0.0	3,671	92.80
	1.0	285	7.20

Figure 4.19: Performance of XGBoost and Knowledge-based Decision Tree combined model. Each subfigure demonstrates the performance of different combinations of merge coefficients starting from 0 to 1 with step of 0.1. XGB_1.0 represents the original model performance, whereas DT_1.0 represents the Knowledge-based Decision Tree performance only.

To assess the performance of the proposed knowledge-guided XGBoost framework, we con-

Table 4.8: Performance of DT_m_XGB_n_simulation

Model (DT_m_XGM_n)	Rec C1	Prec C1	F1 C1	PRC C1	Rec C0	Prec C0	F1 C0	PRC C0	Acc	Bal Acc	AU ROC
DT_0.0_XGB_1.0	0.71	0.71	0.71	0.75	0.98	0.98	0.98	0.99	0.96	0.84	0.94
DT_0.0_XGB_1.0_curve	0.71	0.71	0.71	0.75	0.98	0.98	0.98	0.99	0.96	0.84	0.94
DT_0.1_XGB_0.9	0.71	0.71	0.71	0.74	0.98	0.98	0.98	0.99	0.96	0.84	0.92
DT_0.1_XGB_0.9_curve	0.71	0.72	0.71	0.74	0.98	0.98	0.98	0.99	0.96	0.84	0.94
DT_0.2_XGB_0.8	0.71	0.68	0.70	0.72	0.97	0.98	0.98	0.99	0.96	0.84	0.90
DT_0.2_XGB_0.8_curve	0.70	0.72	0.71	0.73	0.98	0.98	0.98	0.99	0.96	0.84	0.93
DT_0.3_XGB_0.7	0.72	0.58	0.64	0.70	0.96	0.98	0.97	0.98	0.94	0.84	0.89
DT_0.3_XGB_0.7_curve	0.67	0.71	0.69	0.72	0.98	0.97	0.98	0.99	0.96	0.83	0.93
DT_0.4_XGB_0.6	0.74	0.39	0.51	0.68	0.91	0.98	0.94	0.98	0.90	0.82	0.88
DT_0.4_XGB_0.6_curve	0.66	0.67	0.67	0.70	0.97	0.97	0.97	0.99	0.95	0.82	0.92
DT_0.5_XGB_0.5	0.75	0.26	0.39	0.64	0.83	0.98	0.90	0.98	0.83	0.79	0.86
DT_0.5_XGB_0.5_curve	0.65	0.50	0.56	0.68	0.95	0.97	0.96	0.99	0.93	0.80	0.92
DT_0.6_XGB_0.4	0.76	0.25	0.38	0.56	0.83	0.98	0.90	0.98	0.82	0.79	0.84
DT_0.6_XGB_0.4_curve	0.62	0.48	0.54	0.62	0.95	0.97	0.96	0.99	0.92	0.79	0.91
DT_0.7_XGB_0.3	0.76	0.14	0.23	0.40	0.63	0.97	0.76	0.97	0.64	0.70	0.81
DT_0.7_XGB_0.3_curve	0.56	0.44	0.50	0.48	0.94	0.97	0.95	0.99	0.92	0.75	0.89
DT_0.8_XGB_0.2	0.72	0.13	0.22	0.21	0.63	0.97	0.76	0.96	0.64	0.68	0.72
DT_0.8_XGB_0.2_curve	0.48	0.39	0.43	0.30	0.94	0.96	0.95	0.98	0.91	0.71	0.86
DT_0.9_XGB_0.1	0.60	0.11	0.19	0.14	0.63	0.95	0.76	0.95	0.63	0.61	0.62
DT_0.9_XGB_0.1_curve	0.15	0.17	0.16	0.21	0.94	0.93	0.94	0.98	0.88	0.55	0.82
DT_1.0_XGB_0.0	0.68	0.06	0.11	0.08	0.17	0.87	0.28	0.91	0.21	0.42	0.47
DT_1.0_XGB_0.0_curve	0.11	0.12	0.12	0.09	0.94	0.93	0.93	0.92	0.88	0.53	0.51

ducted a comprehensive simulation where the merge coefficient was varied from 0 to 1 in increments of 0.1. Table 5.8 and Figure 5.19 present the detailed performance metrics across all configurations, evaluated on a test set of 3,956 samples (7.2% positive class, as shown in Table 5.7). The original model (DT_0.0_XGB_1.0) achieved strong performance, with F1 score for class 1 (F1_C1) of 0.71, AUROC of 0.94, and balanced accuracy of 0.84. However, as the merge coefficient increased—shifting more weight from the data-driven XGBoost model to the knowledge-based Decision Tree—the performance on minority class (class 1) metrics degraded substantially. For example, with DT_0.5_XGB_0.5, the F1_C1 dropped to 0.56 and AUROC to 0.86, indicating reduced sensitivity to the positive class.

At the extreme end (DT_1.0_XGB_0.0), the model exhibited a sharp decline across all metrics, with F1_C1 = 0.12, AUROC = 0.53, and balanced accuracy = 0.51—barely better than random guessing. These results, visualized in Figure 5.19, show a consistent performance degradation as the model transitions from purely data-driven to purely rule-based (domain knowledge). Notably, the conditional integration method (step-wise merging based on thresholds) consistently outperformed the curve-based merging (smooth interpolation) in terms of F1_C1 and AUROC for most values of the merge coefficient. This suggests that the conditional method better preserves the predictive strength of the original XGBoost model while introducing domain priors in a controlled manner.

4.4 Discussion

The incorporation of domain knowledge into machine learning models remains a key strategy for improving clinical relevance, robustness, and responsiveness. In this study, we proposed a knowledge-guided modeling framework that integrates vital sign heuristics into the loss function via a U-shaped penalty and encodes rules in a decision tree. This design encour-

ages model predictions to be not only data-driven but also physiologically meaningful. The penalty term effectively embeds clinical expectations by penalizing deviations from optimal values of vital signs—defined as the standardized mean—using a quadratic formulation. This mechanism reflects the non-linear, risk-sensitive nature of vital deviations in clinical decision-making.

The proposed custom loss function is flexible and can be adapted to capture non-linear relationships between vital signs and risk. However, designing this loss function or setting its parameters requires domain knowledge. In our work, we rely on two primary sources: published clinical literature and expert input from medical professionals. One way to conceptualize the relationship between vitals and risk is as monotonic. As a vital sign deviates further from its ideal reference value, the associated risk should consistently increase. For example, a patient with a fever of 104°F should be assigned a higher risk score than one with a fever of 100°F, assuming all other vitals are within normal ranges. We consider models that demonstrate this kind of regular, monotonic behavior to be clinically reasonable. However, a key limitation is that we did not quantitatively estimate this monotonic relationship, as no such established mapping currently exists in the literature.

Our results demonstrate that the inclusion of this penalty improves model calibration and reduces overconfidence in predictions made under abnormal physiological conditions. For example, when systolic blood pressure or glucose levels were excessively high, the baseline model tended to output confident survival predictions, likely due to class imbalance or insufficient coverage in those regions of feature space. In contrast, the knowledge-guided model exhibited more cautious and aligned behavior, adjusting its predictions to reflect increased clinical risk. This shift is critical in high-stakes applications such as mortality prediction, where underestimating risk can lead to harmful outcomes.

Furthermore, the approach retains flexibility by allowing adjustable weights (λ_{penalty}, $\lambda_{\text{mini},i}$)

that tune the strength of the domain-guided regularization. This tunability enables a continuum between purely empirical models and strictly rule-based systems, offering a hybrid formulation that balances generalization with interpretability. The use of standardized inputs and setting the optimal value to zero simplifies the implementation while preserving the core clinical meaning.

One of the key advantages of this approach is its generalizability. The U-shaped penalty can be extended to any tabular feature with a known safe range, making it applicable beyond vital signs to laboratory test results, risk scores, or even derived features. Moreover, this methodology can complement existing fairness-aware or uncertainty-aware frameworks, further enhancing trustworthiness. We also outlined a more generalized version of the U-shaped loss function which is asymmetric. This asymmetric loss can incorporate custom weights across different critical zones, enabling greater flexibility in penalizing deviations. When the weights are equal on both sides, the function reduces to a symmetric form, allowing it to be tailored to various clinical contexts. In other words, the asymmetric function can be considered as more general compared to the symmetric function.

We selected the hyperparameters of the custom loss function manually through simulation, guided by qualitative similarities with known associations reported in the literature. However, a quantitative evaluation of this similarity is currently lacking. This is primarily due to the absence of an established quantitative relationship between vital signs and hazard ratios. Further research is needed to identify a meaningful relationship that could serve as an objective function for automatically determining the custom loss parameters.

Our methods rely on the assumption that the optimal values and true characteristics of each vital sign are well-known and constant across populations, which may not always hold. Future work may explore adaptive or learned optimal characteristics based on subgroups (e.g., age, comorbidities) to make the penalty more context-aware. Additionally, while the cur-

rent formulation uses a quadratic loss, other biologically inspired shapes (e.g., asymmetric or piecewise) could be investigated to better reflect varying clinical thresholds. The knowledge-guided model introduces several additional hyperparameters, such as the steepness of the penalty function, the optimal clinical values for each vital sign, and the weights assigned to different types of penalties. These hyperparameters play a critical role in shaping how strongly the model responds to deviations from medically accepted norms. While our current implementation sets these values based on clinical intuition and manual tuning, future work could adopt more systematic optimization strategies. For instance, grid search, Bayesian optimization, or reinforcement learning-based hyperparameter tuning could be used to identify the most effective configuration. Additionally, AI-driven approaches such as neural architecture search or meta-learning could further refine these parameters in a data-driven yet clinically meaningful way. This would not only improve the robustness and performance of the model but also make the integration of domain knowledge more adaptive and scalable across diverse clinical tasks.

The integration of a Knowledge-based Decision Tree (KDT) with machine learning models offers a transparent and clinically grounded mechanism for improving decision quality, especially in high-risk healthcare settings. In our framework, we employed a human-crafted decision tree that encodes domain rules derived from clinical heuristics and merged its output with a data-driven model such as XGBoost using a tunable coefficient. This hybrid approach enables the system to balance learned patterns with medically validated logic, mitigating the risk of spurious correlations and data-driven overfitting.

The decision tree contributes responsiveness and robustness, particularly in regions of the feature space that are poorly represented in the training data or inherently uncertain. For example, in cases where vital signs lie in extreme or clinically abnormal ranges, the KDT provides a strong corrective influence that reflects medical knowledge, steering predictions

toward outcomes that align with clinical risk. This is especially important when the data-driven model may underrepresent rare but high-risk cases due to class imbalance.

We also observed that the model's behavior varies meaningfully as we adjust the merge coefficient α, which controls the relative weight of the KDT and XGBoost predictions. When $\alpha = 1$, the model purely follows the decision tree, acting as a rule-based expert system; when $\alpha = 0$, it relies entirely on XGBoost.

Another important future direction for this research is the inclusion of appropriate baseline models from the literature to strengthen comparative analysis and contextualize the performance of our proposed approach. For instance, causal inference models, such as structural causal models, could be used to explicitly model the causal relationships between vital signs and clinical outcomes [83]. These models can help disentangle correlation from causation and may offer more robust predictions, especially when dealing with confounding variables or evaluating interventions. In addition, counterfactual models can be leveraged to explore "what-if" scenarios—for example, estimating how a patient's risk profile might change if a particular vital sign were within a normal range [84]. This perspective aligns well with clinical reasoning and could offer meaningful insights into patient management strategies or treatment prioritization. Another promising avenue is the inclusion of Kernelized Attention Networks (KAN), a more generalized form of neural networks that employs learnable activation functions [85]. KANs are particularly well-suited for capturing complex, non-linear, and domain-informed relationships between inputs (such as vital signs) and outcomes, which could potentially improve model interpretability and fit, especially in critical care settings where physiological signals often follow irregular patterns. Furthermore, integrating out-of-distribution (OOD) testing into the evaluation pipeline would make the testing framework more robust and realistic. This would involve testing the model on data that differ in distribution from the training set—such as data from different hospitals, patient subgroups,

or time periods—to assess generalizability and identify failure modes. OOD evaluation is particularly critical in healthcare, where models must perform reliably across diverse patient populations and clinical environments. Also, making the models more interpretable can help to understand the causal relationships [86, 87].

Chapter 5

Enhancing Fairness and Accuracy in Diagnosing Type 2 Diabetes in Young Population

5.1 Introduction

Diagnosing chronic diseases like diabetes is crucial, given the substantial global burden of diabetes-related complications and deaths, particularly in low- and middle-income countries [88, 89]. At the present rate of expansion, the International Diabetes Federation predicts that by 2045, a staggering 693 million individuals globally will be affected by diabetes [90]. The prevalence of diabetes, specifically type 2 diabetes, has been steadily increasing over the past few decades [91, 92]. While diabetes has traditionally been associated with the elderly population, recent studies indicate a rising prevalence of diabetes among the younger population as well [93]. According to CDC, by 2060, type 2 diabetes cases might increase by about 70-700% in the young population [94]. Young adults with diabetes face a significantly elevated risk of early health complications and even premature death compared to their counterparts without diabetes [95]. Younger people diagnosed with diabetes face an increased risk of developing early and severe complications, encompassing microvascular (retinopathy, neuropathy, ulceration, nephropathy) and macrovascular (cardiovascular, cerebrovascular,

peripheral vascular) diseases [96, 97, 98]. Early detection and awareness can help the young population at risk take steps to prevent or delay type 2 diabetes, and early intervention can even reverse prediabetes [99, 100].

Machine learning (ML) has been increasingly integrated into the healthcare systems [101, 102, 103] because of its potential to assist clinicians and medical doctors in taking better care of patients. Many machine learning models have been applied to diagnose diabetes [92]. However, health data can be imbalanced, which could potentially lead machine learning models to learn patterns with existing bias from the provided data [104, 105, 106, 107, 108, 109]. Data bias, if not addressed, can exacerbate and perpetuate inequalities in the performance of algorithms in different subgroups [110, 111, 112], particularly in historically underserved populations like female patients [113], black patients, or those with low socioeconomic status [114]. AI models can be susceptible to digital ageism, potentially leading to biased diagnoses that could harm patients [115, 116]. Moreover, AI models existing in the healthcare domain can show faithfulness issues [117]. It is of high importance that we evaluate the machine learning model before deployment ensuring social fairness [118, 119].

This paper identifies that the traditional machine learning models such as Logistic regression (LR), Multi-Layer Perceptron (MLP), Naive Bayes (NB), AdaBoost (AB), Random Forest (RF), and K Nearest Neighbor (KNN), trained on imbalance BRFSS (with only 15% representing the diabetes population) dataset [20], tend to misdiagnose diabetes more frequently in the younger population (30-44 years) compared to other subgroups. The recall of the positive class (Rec_C1) in the 30-34 age group is only 30%, whereas the positive class (Rec_C1) is 68-72% in the gender group and 66-84% in the ethnic group. Moreover, we bring attention to the fact that solely relying on AUROC can be misleading – while the overall group's AUROC is 82%, the age group 30-34 exhibits a seemingly high 84% AUROC that masks the poor diabetes detection rate (recall C1) within that specific age group. So, we focus mostly

on recall of class 1, balanced accuracy (average recall of class 1 and class 0), and AUROC for a fair comparison.

We propose an effective solution which successfully mitigates bias from young groups and increases type 2 diabetes (T2D) diagnosing sensitivity. None of the existing papers developed any effective machine-learning-based approach for effectively diagnosing T2D in the young population (30-44 years). To the best of our knowledge, our work is the first precision T2D diagnosis paper using machine learning and improving the diagnostic performance of machine learning models to diagnose T2D in young populations.

Our major contributions are:

1. We address a critical gap in the literature by developing a machine learning approach specifically designed for precision type 2 diabetes (T2D) diagnosis in a young adult population (30-44 years old). This is a new study to target this specific age group for T2D diagnosis using machine learning.

2. We show that multiple machine learning models and a number of sampling techniques (SMOTE, random sampling, etc.) fail to achieve fair performance in terms of detecting T2D in young adults.

3. We propose a bias correction technique specifically for improving T2D diagnosis in young adults. This demonstrates the effectiveness of subgroup-focused bias correction, promoting fairer and more accurate machine learning models in healthcare settings.

In this study, we train diverse machine learning models, incorporating a transparent model like logistic regression, to uncover the roots of bias and missed detections. This is achieved through a thorough analysis of association coefficients responsible for shaping model decisions.

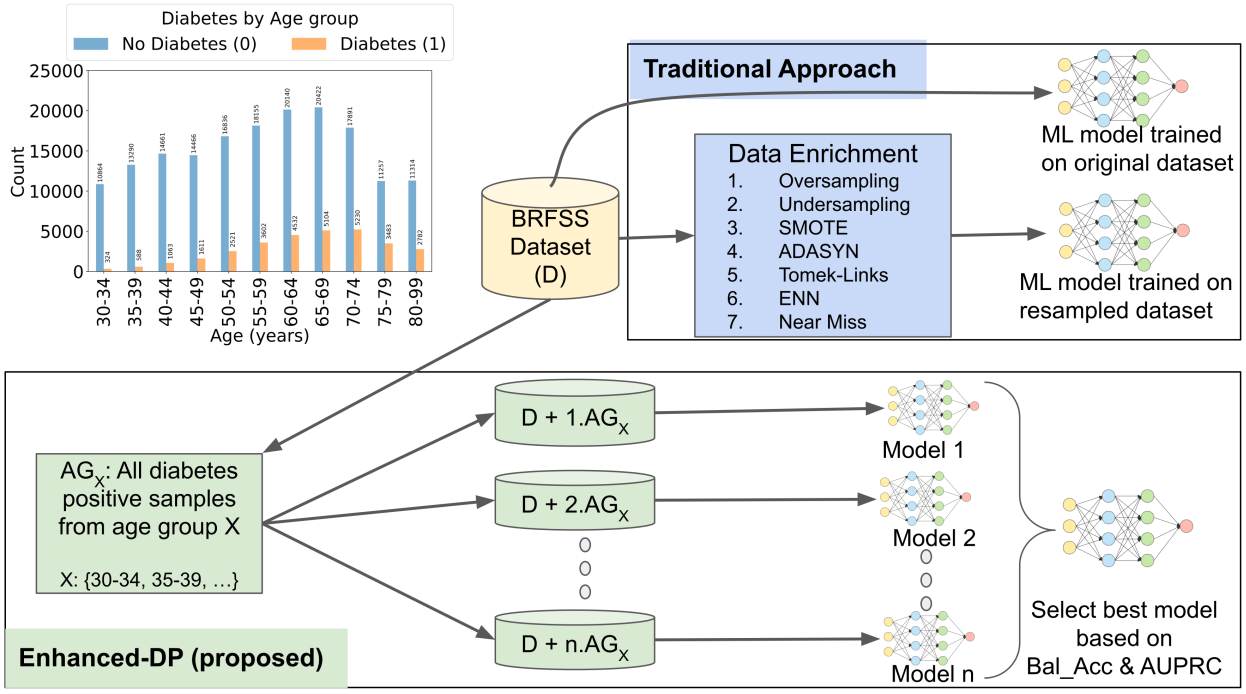


Figure 5.1: The Enhanced-DP approach in contrast to traditional approaches, enriches the minority age groups and creates new training sets by replicating diabetic samples (1 to n times) from a minority age group. n machine learning models are trained on each of the n versions of the training sets. The best model is selected based on performance metric balance accuracy (Bal_Acc) and area under precision and recall curve (AUPRC). The top left bar chart represents age distribution (histogram) in the original dataset.

5.2 Methods

5.2.1 Dataset

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual health survey conducted in the United States to monitor various health behaviors such as cardiovascular diseases, chronic diseases, diabetes, obesity, and other risk factors that contribute to the leading causes of death and disability [20]. The BRFSS datasets have been collected from all 50 states in the U.S. and the District of Columbia and included responses from over 400,000 participants which makes it one of the largest publicly available datasets related to public

health. Each record contains an individual's BRFSS survey responses on various health behaviors and risk factors such as tobacco use, physical activity, alcohol consumption, existing chronic diseases, and mental health. The survey also gathered demographic information such as age, gender, and race/ethnicity, which are helpful to explore important correlations and even causation.

This survey is conducted every year, and the CDC makes it publicly available for research. In this study, we selected the BRFSS dataset from 2021, which contains more than 400,000 subject information and 330+ attributes from each subject. The BRFSS dataset is a valuable resource for identifying health disparities and evaluating the effectiveness of public health programs and policies. However, the dataset is not free from challenges because a large portion of attribute values are missing (25%) and this dataset can be a highly imbalanced data imbalance (diabetes class 15%).

5.2.2 Data Preprocessing

According to the literature on the risk factors for diabetes, [92, 120], we selected 30 attributes including diabetes labels. Subjects aged over 30 are selected for this study [92, 121, 122]. We create age groups spanning 5 years starting from 30 up to 80+ years. The selected cohort contains 200,136 subjects information where 169,296 subjects (84.6%) are diabetes negative and 30,840 subjects (15.4%) are positive.

As we mentioned before, this dataset is not balanced for the age, gender, and racial subgroups. In the age group 30-34, the number of negative cases is 10,864 but the number of positive cases is only 324. The age groups 35-39 and 40-44 are also highly imbalanced as the positive-negative ratio is only 4.4% and 7.2%.

The selected variables fall into three distinct types: nominal, ordinal, and binary. Nominal

variables lack any inherent order, ordinal variables possess a meaningful order, and binary variables exclusively hold two distinct values. For example, the presence of high blood pressure, cholesterol, or heart disease can be represented by a binary variable. BMI category, education level, and income level are considered ordinal variables. On the other hand, marital status and race are nominal variables. Binary variables are represented using a binary encoding, where 1 signifies a positive outcome and 0 represents a negative outcome. Ordinal variables are encoded using integers, preserving their meaningful order. Nominal variables are transformed into one-hot encoding for appropriate representation. The selected variables are listed in Fig. 5.8(c) .

5.2.3 Bias Mitigation Approach

We utilized a modified and enhanced version of the prioritized (DP) bias correction method (Fig. 5.1) which is inspired by [123] prioritizes a specific subgroup, the young age group in this case, that suffers from data imbalance. We incrementally replicate data points of the minority class (diabetes positive indicated by class 1) and choose an optimal unit of replication based on the model performance. As a result, the enriched training set contains the original samples as well as the replicated samples. However, the vanilla DP technique by [123] has several shortcomings, including the selection of replication units.

Our proposed Enhanced-DP technique replicates all samples of the diabetes class from the young population up to n times. Each time the duplicated sample units are merged with the original dataset, including the original set. In our experiment, we set the maximum DP unit "n" based on the lowest subgroup positive-negative sample ratio. In order to achieve a balanced set, the lowest subgroup ratio is for the age group 30-34. The number of negative cases is 10,864 but the number of positive cases is 324. So, the ratio is 34, which is the limit of

DP unit n. We also employed early stopping to select the best n quickly as the performance curve contains a global maximum point. This makes our Enhanced-DP algorithm faster compared to the original DP algorithm [123].

Each training set containing the original set and the duplicated n 1 to 34) units is used to train a single model. In this way, we train 35 models (34 models trained on the enriched training sets and 1 model trained on the original training set) and select the best-performing model based on balanced accuracy and the area under the curve (AUC) of the minority class (C1) precision-recall curve, denoted by PRC_C1. We identify the top three machine learning models with the highest balanced accuracy values and select the model that gives the highest PR_C1 for that particular age group. This selected model is used only for diagnosing diabetes for that particular group, e.g., the Enhanced-DP model trained with duplicated age-group 30-34 samples is used to diagnose new patients from age-group 30-34 years.

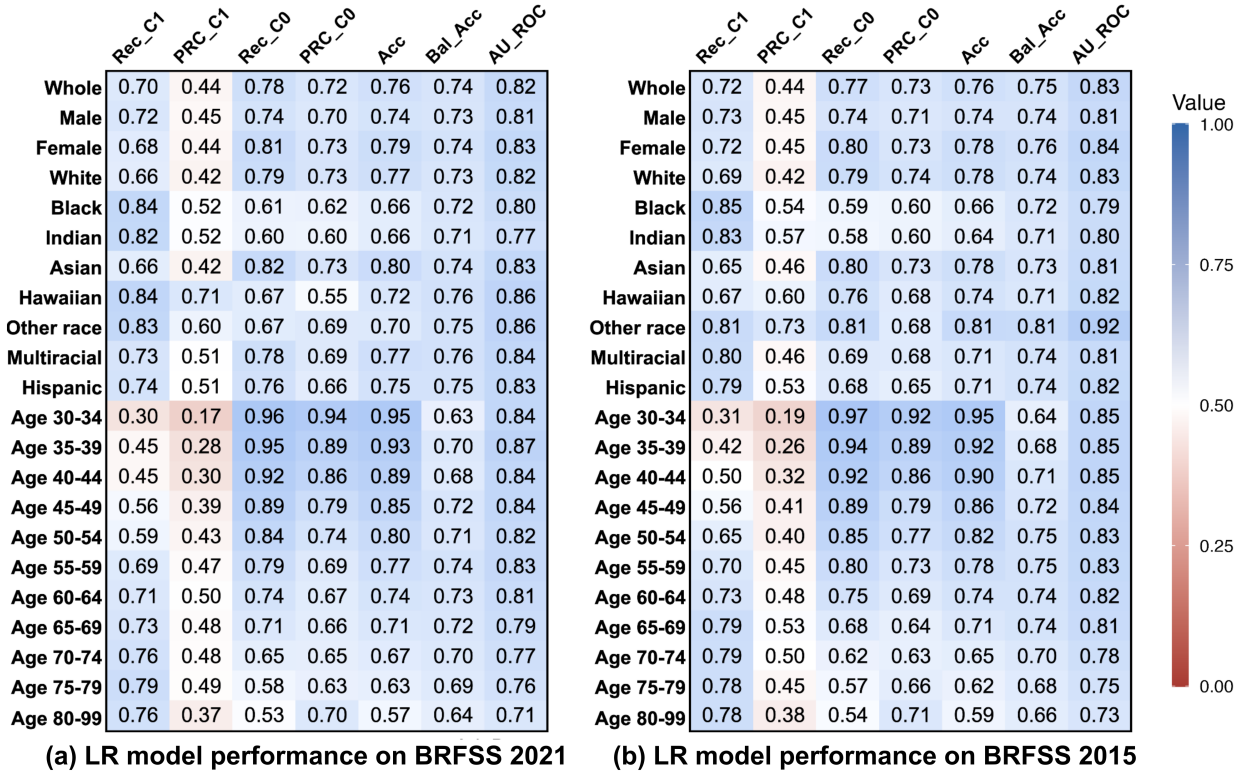


Figure 5.2: Performance of logistic regression model trained and tested on the original (a) BRFSS 2021 and (b) BRFSS 2015. (c) Performance of multiple machine learning models for the young adult age group (30-44 years) along with the whole group. The x-axis represents performance metrics where Rec, PRC, Acc, Bal_Acc, AUROC, represent Recall, Area Under the Precision-Recall Curve, Accuracy, Balanced Accuracy, and area under the ROC curve respectively. C1 and C0 stand for class 1 (diabetes positive) and class 0 (diabetes negative) respectively. The Y-axis represents the different subgroups.

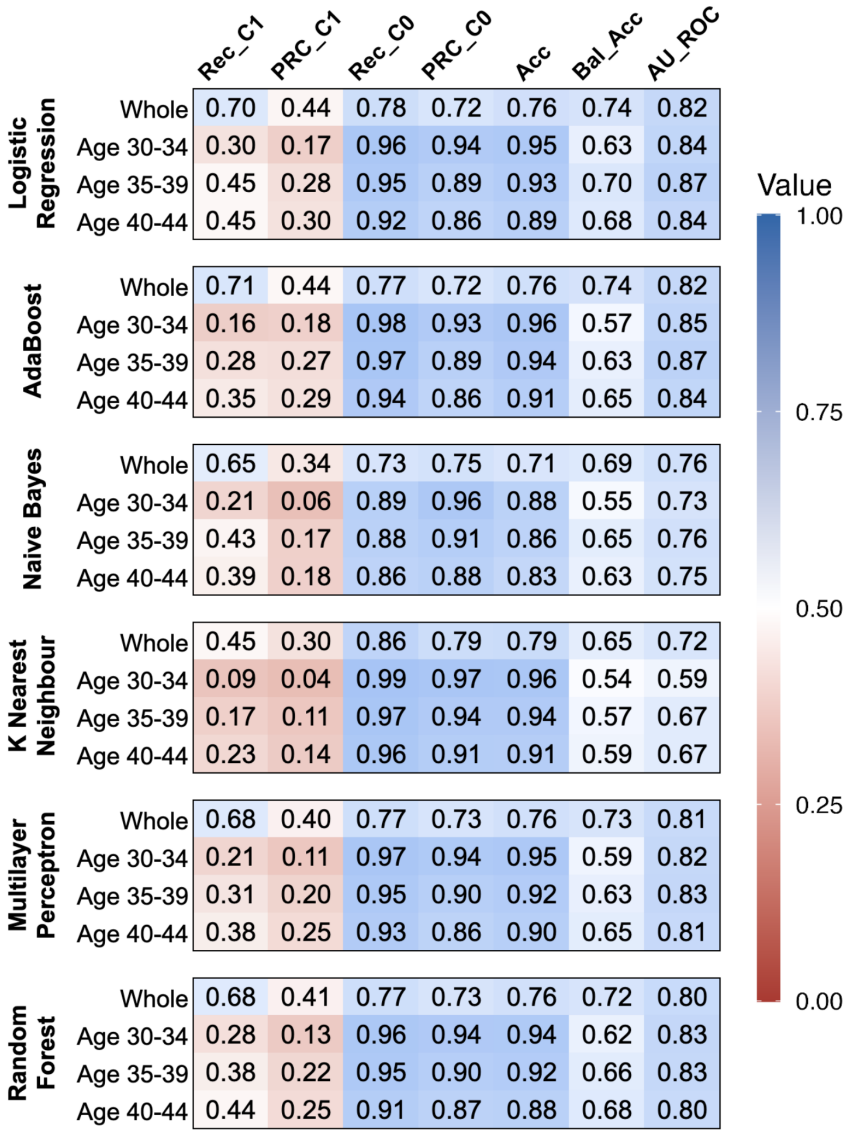


Figure 5.3: Performance of logistic regression model trained and tested on the original (a) BRFSS 2021 and (b) BRFSS 2015. (c) Performance of multiple machine learning models for the young adult age group (30-44 years) along with the whole group. The x-axis represents performance metrics where Rec, PRC, Acc, Bal_Acc, AUROC, represent Recall, Area Under the Precision-Recall Curve, Accuracy, Balanced Accuracy, and area under the ROC curve respectively. C1 and C0 stand for class 1 (diabetes positive) and class 0 (diabetes negative) respectively. The Y-axis represents the different subgroups.

5.2.4 Sampling Algorithms

A frequently employed technique for addressing the challenges of data imbalance is the utilization of sampling methods. To evaluate the effectiveness of the sampling algorithms detecting diabetes in young groups, we first compared multiple sampling algorithms, including (1) random oversampling, randomly duplicating instances from the minority class [124]; (2) random undersampling, randomly removing instances from the majority class [124]; (3) SMOTE (Synthetic Minority Over-sampling Technique), creating synthetic samples for the minority class by interpolating between existing instances [125]; (4) ADASYN (Adaptive Synthetic Sampling), similar to SMOTE, but generating more synthetic samples for difficult-to-learn instances [126]; (5) Tomek Links, removing instances from the majority class that are close to instances in the minority class [127]; (6) ENN (Edited Nearest Neighbors), removing instances from the majority class that are misclassified based on the nearest neighbors from both classes [128]; and (7) NearMiss, selecting instances from the majority class based on their distance to instances in the minority class [129]. These sampling techniques have been proven to be effective in the literature, however, these techniques are not tailored to tackle subgroup biases. As a result, diabetes is misdiagnosed in the young population. Therefore, we utilize a new concept of using one model for a single group which deviates from the traditional one model-fits-all ideology.

5.2.5 Machine Learning Models

We selected six commonly used ML models including Logistic Regression (LR) [130], Random Forest (RF) [131], Adaboost (AB), Multilayer Perceptron (MLP) [132], and Naive Bayes (NB) to evaluate the effectiveness of the sampling strategy. We purposely selected simple models such as logistic regression because of their interpretability. It is very important

for the science community, especially for healthcare to create interpretable models to find out the root cause of a prediction, however, [133] evaluated 511 scientific papers across different ML domains and identified a notable deficiency in reproducibility metrics, including dataset and code accessibility in clinical ML domain papers. Moreover, complex and high-end models which have recently been applied to the healthcare domain might pose difficulties in reproducibility [134]. Using simple models like logistic regression will pave the way both for explaining the results and making it easy for other researchers to reproduce the model and the same results.

In the RF, we selected 100 trees or estimators to achieve optimal performance. In our MLP configuration, we used a single hidden layer with 100 ReLU-activated neurons and optimized training with the Adam gradient descent optimizer for efficiency and effectiveness. For Naive bayes we utilized the Gaussian kernel. For Adaboost we select the default settings from scikit-learn Python library. We repeated the training process at least five times with different test train splits to ensure consistent performance and calculate the standard deviation of the performance.

For LR, we utilized the liblinear equation to establish the decision boundary between positive and negative classes. This analysis can provide valuable insights into feature importance, offering a comprehensive understanding of how each variable contributes to the model's outcome. We investigate the features responsible for biased outcomes from the original model. The logistic regression estimates the probability from the coefficients and the corresponding feature value which makes it a white box and fully interpretable.

To examine the potential unfairness in the datasets, we calculated class-based accuracy, recall, AUROC, and AU-PRC. Nonetheless, our primary focus lies in prioritizing the recall of the positive class, also known as sensitivity, as the detection of diabetes holds paramount importance in our specific case. The dataset undergoes a random division into three disjoint

sets namely training (60% or 120,081 samples), validation (20% or 40,027 samples), and testing set (20% or 40,028 samples) through the utilization of the Python sklearn library. The training dataset serves as the foundation for model training, while the testing dataset remains consistent across all the experiments, ensuring the robustness and comparability of our results. Only the training set is used for DP, ensuring no data leakage. The performance of the models in each experiment is recorded as the average value of 3-5 independent trials.

5.2.6 Model Calibration and Threshold Tuning

Before mapping probabilities into labels, we calibrate the predicted probabilities for each model on the validation set. Calibration involves adjusting the distribution of probabilities to enhance accuracy. We calibrate the output probabilities using the Isotonic Regression [135] technique. We then perform threshold tuning to find the optimal threshold based on balanced accuracy and the F1_C1 score. Threshold tuning is the process of selecting an optimal threshold value in binary classification models. In these models, predictions are typically expressed as probabilities, and a threshold is applied to determine the class label. Adjusting this threshold can impact the trade-off between false positives and false negatives.

At first, we identify the top 3 thresholds that produce the highest F1_C1 scores and then select the optimal threshold that generates the highest balanced accuracy for all samples in the validation dataset. For some subgroups, only a few samples (< 100) are present in the validation set. Selecting the threshold based on subgroup samples may overfit the validation set. So, we select a threshold based on the whole validation performance. We calculate the optimal threshold to be 0.195 from 10 independent trials by averaging the best thresholds from each trial. We use the F1 score and balanced accuracy jointly to select a model with a balanced performance.

5.3 Results

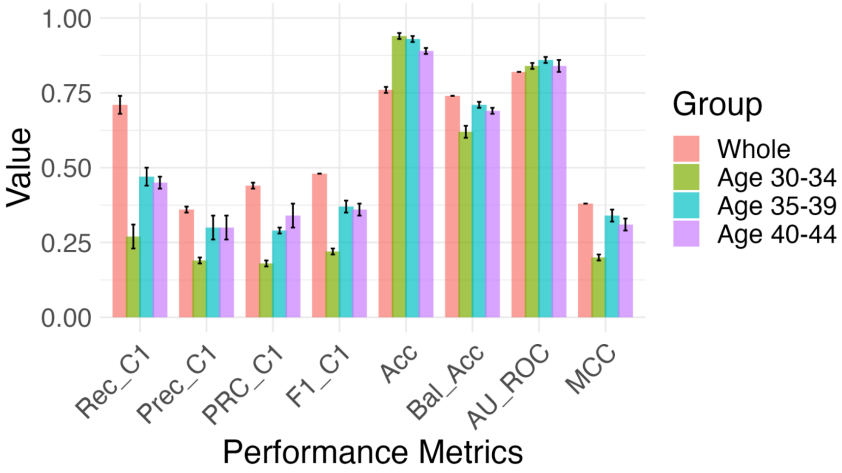
5.3.1 Performance of Original Model

The machine learning models trained on the original training set show different performances for the minority diabetes-positive group (C1) and the majority-negative group (C0). Fig. 5.3 (a) and (b) shows the performance of the original logistic regression model on 2021 and 2015 datasets. The whole group recall C1 is 0.70-0.72 whereas recall C0 is 0.77-0.78. For males, recall_C1 is 0.72 and for Female recall_C1 is 0.68. The ethnic groups such as White, Asian, and multicultural show recall_C1 of 0.66, 0.66, and 0.73 respectively. On the other hand, Black, Indian, and Hawaiian show recall_C1 of 0.84, 0.82, and 0.84 respectively.

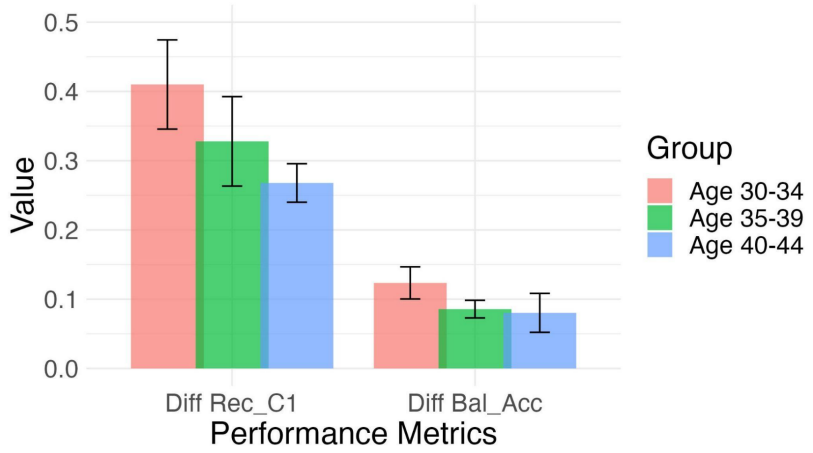
However, the young age group suffers the most from missed detection of diabetes. For age groups [30-34], [35-39], and [40-44] recall C1 is only 0.30-0.31, 0.42-0.45, and 0.45-0.50 respectively. This means out of 100 diabetic patients aged 30-34, 70 patients are misdiagnosed. The Adaboost, Naive bias, random forest, and KNN models also show similar performance, represented in Fig. 5.3 (c). The standard deviation of each metric is less than 0.04 (from multiple experimental trials) unless mentioned otherwise. The model behaves similarly in the BRFSS 2015 dataset (this is not shown due to space constraints).

In imbalanced datasets, commonly employed metrics like AUC_ROC and accuracy can be misleading and do not accurately represent the performance of the minority class. Despite potentially poor performance in the minority class, these metrics might indicate falsely elevated values. The AUROC of the whole population, age groups [30-34], [35-39], and [40-44] are 0.82, 0.84, 0.87, and 0.84. However, these age groups show very poor recall_C1 which is not reflected with AUROC. In contrast to the AUROC and accuracy metrics which can be overly optimistic, we use Recall_C1 to measure the true detection rate reflecting the

performance of the model.



(a) Original model performance



(b) Performance difference between DP and original model

Figure 5.4: Performance of the original logistic regression model (a) when tested on the whole population and minority age group. (b) Performance difference between Enhanced-DP and original model Logistic Regression model. "Diff Rec_C1" means subtracting the recall of class 1 of the original model from the Enhanced-DP model and "Diff Bal_Acc" means subtracting the balanced accuracy of the original model from the Enhanced-DP model. Positive values indicate performance improvement from the original model. The error bars represent the standard deviation of the experiment results.

5.3.2 Enhanced-DP model improves diagnostic accuracy

The Enhanced-DP model for three age groups improves the diabetes detection accuracy significantly (Fig. 5.4 (a)). Fig. 5.4 (b) shows the performance difference between the Enhanced-DP model and the original model in terms of positive recall and balanced accuracy. Diff Rec_C1 means subtracting the recall of class 1 of the original model from the Enhanced-DP model. The Enhanced-DP model for age groups [30-34], [35-39], and [40-44] improves the positive recall by 41% (SD 6.4%), 32% (SD 6.4%), and 24% (SD 2.8%) respectively. This means the Enhanced-DP models are better at detecting diabetes in the young population.

On the other hand, Diff Bal_Acc means subtracting the balanced accuracy of the original model from the Enhanced-DP model. The balanced accuracy (average recall of both C0 and C1 classes) is also improved by 13% (SD 2.3%), 10.5% (SD 1%), and 7.7% (SD 2.8%) for the Enhanced-DP model.

5.3.3 Whole-population sampling

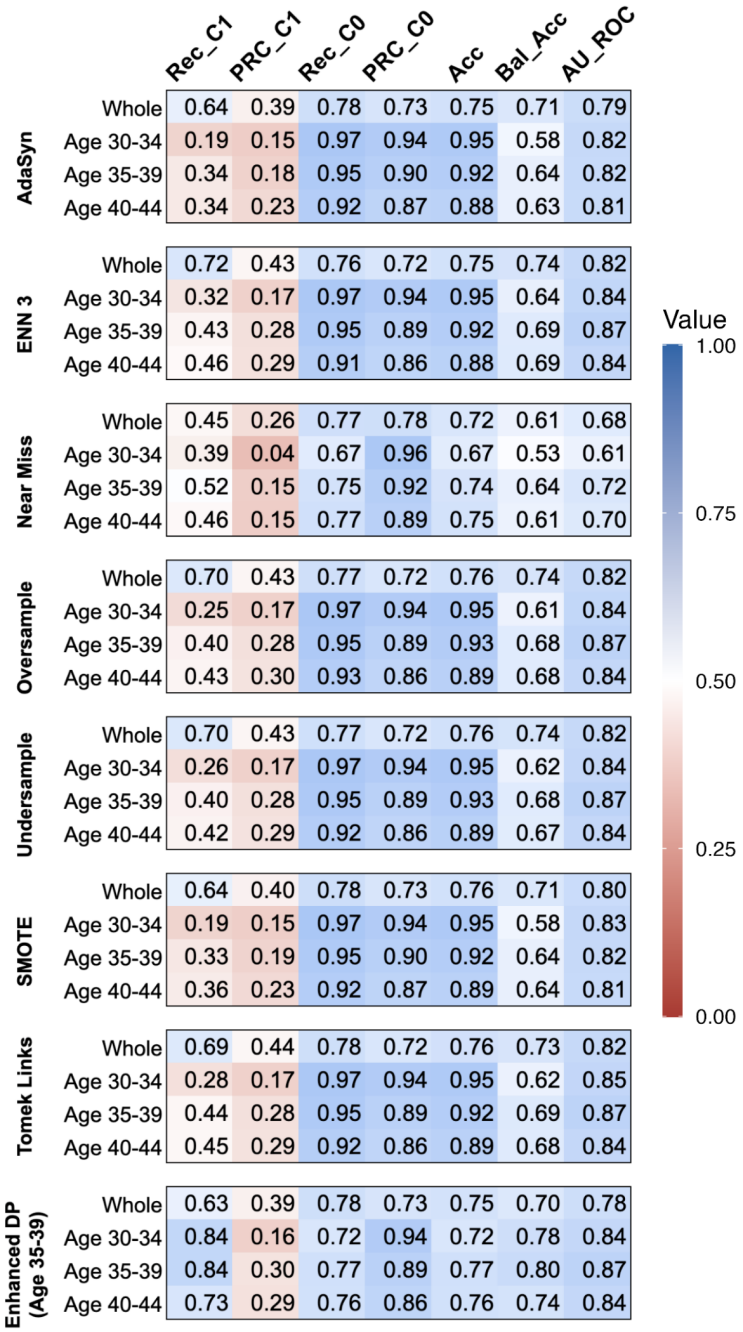


Figure 5.5: Comparing whole group sampling method performance with Enhanced-DP (age group 35-39) for the young adult age groups 30-34, 35-39, and 40-44 along with the whole group.

The whole group-based sampling approach doesn't improve the detection rate in the young group Fig. 5.5. Moreover, AdaSyn (Rec_C1 0.19), SMOTE (Rec_C1 0.19), Tomek-Link (Rec_C1 0.28), Random oversampling (Rec_C1 0.25) and undersampling (Rec_C1 0.26) methods decrease the original detection accuracy (Rec_C1 0.30) for age group [30-34]. A similar performance decrease is observed in the other two age groups [35-39] and [40-44]. Near Miss and ENN merely improve the recall C1 by 9% and 2%. On the other hand, the respective Enhanced-DP models (i.e. trained on age group [30-34] and applied to age group [30-34]) improve the positive recall by at least 24%.

The whole group sampling methods also show poor performance in terms of balanced accuracy. For example, for the age group [35-39] all of the methods decrease the balanced accuracy from the original value. On the other hand, the Enhanced-DP model increases balanced accuracy by 3-9%.

5.3.4 Segmented training

Models	Age Span	Rec_C1			Bal_Acc		
		Age 30-34	Age 35-39	Age 40-44	Age 30-34	Age 35-39	Age 40-44
Baseline	ALL	0.30	0.45	0.45	0.63	0.70	0.68
AdaBoost	5	0.12	0.26	0.35	0.55	0.62	0.64
	15	0.41	0.48	0.51	0.66	0.70	0.69
Logistic Regression	5	0.11	0.20	0.36	0.55	0.59	0.65
	15	0.38	0.52	0.56	0.64	0.70	0.71
Multilayer Perceptron	5	0.02	0.15	0.23	0.51	0.56	0.59
	15	0.36	0.45	0.50	0.63	0.66	0.66
Random Forest	5	0.04	0.22	0.30	0.52	0.60	0.63
	15	0.38	0.48	0.54	0.64	0.69	0.70
K Nearest Neighbour	5	0.00	0.07	0.10	0.50	0.53	0.54
	15	0.48	0.60	0.61	0.62	0.68	0.67
Naive Bayes	5	0.00	0.00	0.15	0.50	0.50	0.55
	15	0.42	0.50	0.51	0.60	0.64	0.63
DP (ours)	ALL	**0.75**	**0.84**	**0.69**	**0.77**	**0.80**	**0.76**

Figure 5.6: Models trained on age-segmented training sets composed of age groups spanning 5 years or 15 years. The models are trained and tested on the same or overlapping age group. We utilized all 7 resampling techniques (Fig. 5.5) and reported the best performance for each model. The baseline model represents logistic regression model trained on the whole training set whereas DP is our proposed model.

We compared the performance of models trained on age-segmented datasets, where the original dataset was divided into groups with 5-year and 15-year spans. Each model was trained and tested on data from the same or overlapping age groups. For example, a model trained on individuals aged 30-34 was evaluated using a test set from the same age range. Similarly, a model trained on the 30-44 age group was tested across the 30-34, 35-39, and 40-44 age groups. Each model was calibrated using the Isotonic Regression and model-specific decision threshold was calculated. We applied all seven resampling techniques outlined in Fig. 5.5 and reported the best performance for each model. As shown in Fig. 5.6, the

DP method consistently performed the best. Interestingly, models trained on 15-year spans outperformed the baseline model trained on the entire dataset. These findings were also consistent with results obtained from the BRFSS'15 data.

5.3.5 Cross-group performance

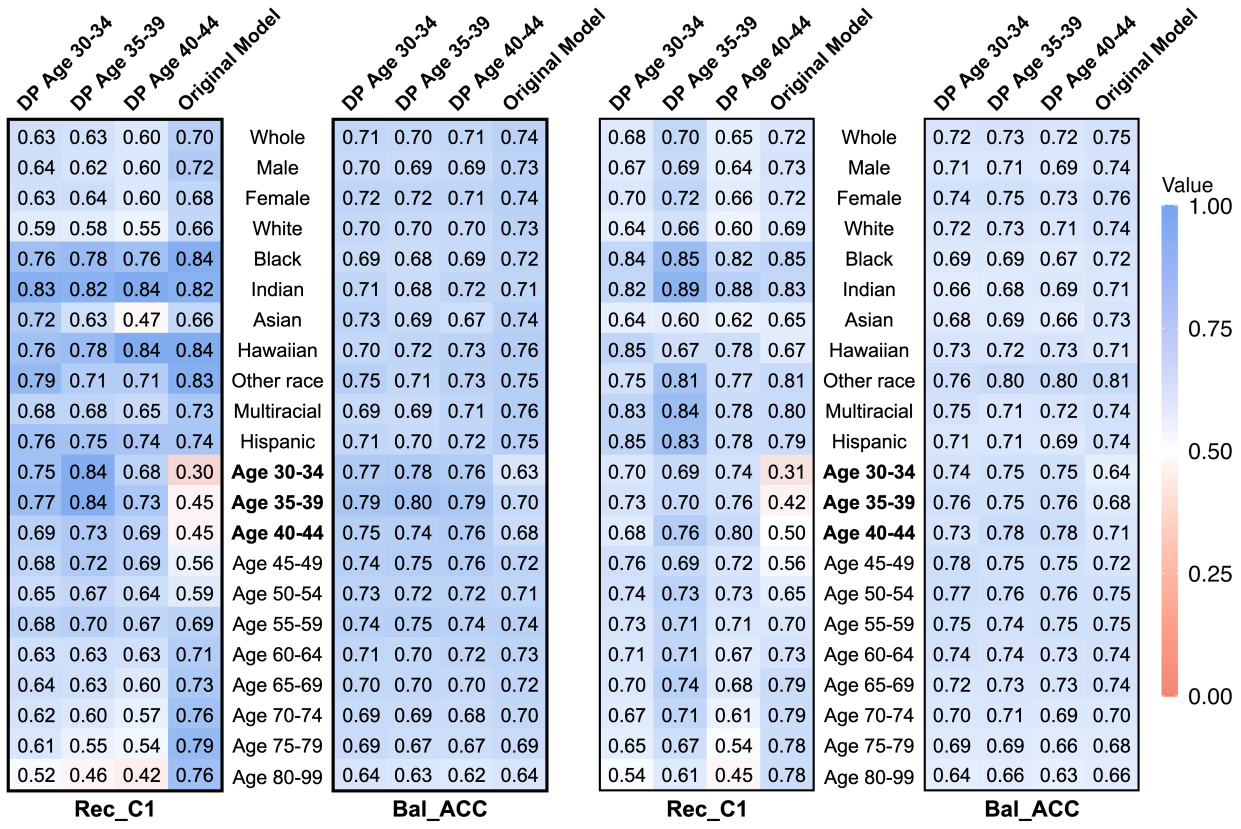


Figure 5.7: Cross-group performance analysis using class 1 recall (Rec_C1) and balanced accuracy (Bal_Acc) on (a) BRFSS 2021 and (b) BRFSS 2015. In each subfigure, each column corresponds to an Enhanced-DP model trained for a specific subgroup. Each row represents a subgroup that a model is evaluated on. (c) Represents logistic regression model coefficient values which are associated with each feature from the original model and Enhanced-DP models. The one hot encoded feature (marital status, employment status, and race) coefficients are averaged

Enhanced-DP model trained for a specific age group is applied to all other age groups Fig. 5.8 (a) and (b). For example, a DP model trained for age group [30-34], is tested on age groups [30-34], [35-39], and [40-44]. We also compare the whole group's performance with the DP models. The original model shows very poor recall C1 for all three age groups. Interestingly, the DP model trained for the age group [30-34] can also be applied to age groups [35-39] and [40-44]. DP model for age group [35-39] shows the highest recall C1 of 73% - 84% for all age groups. The balanced accuracy is also high when one DP model is applied to another age group.

On the other hand, the DP model trained for the young age group shows poor performance when applied to gender or ethnic groups for which the DP model is not trained. The recall of class 1 goes down by 8% and 19% when the DP model for the age group [40-44] is applied to the female and Asian groups respectively. The balanced accuracy also declines if age group DP models are applied to female or Asian groups.

5.3.6 Feature analysis

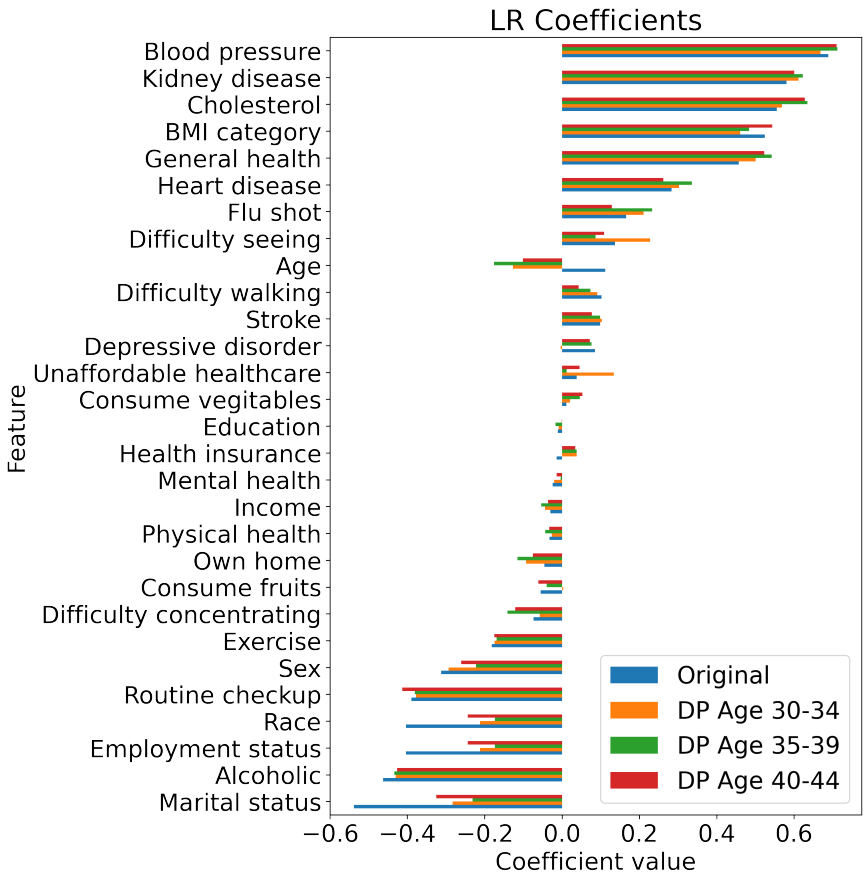


Figure 5.8: Cross-group performance analysis using class 1 recall (Rec_C1) and balanced accuracy (Bal_Acc) on (a) BRFSS 2021 and (b) BRFSS 2015. In each subfigure, each column corresponds to an Enhanced-DP model trained for a specific subgroup. Each row represents a subgroup that a model is evaluated on. (c) Represents logistic regression model coefficient values which are associated with each feature from the original model and Enhanced-DP models. The one hot encoded feature (marital status, employment status, and race) coefficients are averaged

One of the major motivations for selecting a logistic regression model is its interpretability. [136] showed that deep learning models are mostly black-box and cannot always be correctly interpretable.

Fig. 5.8 (c) shows the logistic regression model coefficients of the original and the Enhanced-

DP models. It shows that the original model had a strong correlation with the subject's race (-0.40), employment status (-0.40), and marital status (-0.54). The Enhanced-DP model reduces the strength of these attributes by at least 19%. Moreover, age was positively correlated in the original model while it was negatively correlated in the Enhanced-DP models. Other attribute coefficients show minor changes.

5.4 Discussion

The machine learning trained on the original dataset shows poor diabetes diagnosis performance in the young group. Because the diabetes data is limited in the original dataset the traditional machine learning model tends to pick the general statistics to build a diagnostic model. As a result, the models show poor performance in the minor young group. Whole group-based data enrichment such as sampling methods cannot overcome the problem of poor performance as it doesn't enhance the young group. Some of them such as SMOTE, a popular sampling method which is well known for removing bias from imbalanced datasets, decrease the performance. This is because these sampling methods are not equipped to reduce disparate ratios in the minority group. Moreover, we tested multiple and diverse machine learning models. However, all the models show consistently poor performance in this scenario.

Finally, we proposed an enhanced version of the double prioritized bias correction technique (DP) to make the model effective and useful for the young age group. By replicating the minor group population dynamically, the technique improves the model's performance. However, one DP model can be targeted for one particular group. The results show that the DP model trained for the young age group is not applicable for both gender or ethnic groups. This limitation can be easily overcome by using multiple DP models for each minor group.

However, one of the limitations of this approach is the use of multiple models compared to a single model. This limitation only applies in the training phase but not in the application phase. Training multiple DP models will take more time than a single model approach but the DP model will take the same time during diagnosis as only one model is going to be used for predicting a particular sample. Interestingly, this technique has broader applications beyond the specific dataset evaluated and can be utilized in any domain to overcome bias or data limitation challenges.

We also investigated the root cause of bias in the original model by visualizing the coefficients of the logistic regression model. Being a white box model, we can easily understand the feature importance and the correlation of each feature with the detection probability. The original model shows a strong correlation with non-medical attributes such as marital status, employment status, and race. The DP model decreases the strength of correlation with these factors significantly. Moreover, the original model is positively correlated with age. It means the diabetes-positive probability increases with age. However, this positive correlation was one of the key factors why the original shows poor performance in the young age group.

However, changing label prevalence through oversampling introduces several important challenges. In particular, it can lead the model to overestimate the likelihood of rare outcomes, such as diabetes in the young population. It also increases the risk of model overfitting to the oversampled group and introduces bias, especially when the duplicated positive cases dominate the training data. Additionally, there is a clear precision-recall tradeoff as oversampling tends to improve recall by making the model more sensitive to the minority class, but at the cost of reduced precision. For example, in our case, the model becomes more accurate for the oversampled young subgroup (which is intentional), but this comes at the cost of reduced performance and potential bias against the older subgroup. To overcome these issues, we are creating custom models for each of the underperforming subgroups.

In conclusion, we identified a major deficiency in the traditional machine learning and commonly used sampling approach when it comes to diagnosing diabetes in the young age group which is the minority population in this case. We proposed an enhanced DP method to overcome this issue and improved diagnostic performance significantly. This approach has the potential to mitigate bias issues in the diagnosis of diabetes among young adults, offering a pathway toward more equitable and accurate healthcare practices in this demographic.

Chapter 6

Conclusion

This research addresses critical challenges in the deployment of machine learning (ML) models in healthcare, focusing on improving model responsiveness, reducing bias, and enhancing trustworthiness in high-stakes clinical applications. Through the development of novel methodologies for detecting and mitigating misprediction regions, we propose systematic approaches that allow ML models to adapt and perform reliably within diverse clinical scenarios, particularly in areas associated with high-risk or minority groups often underrepresented in training data. Our work also introduces targeted mitigation techniques, including adaptive re-weighting and data-specific augmentation, to reinforce model accuracy and robustness where it matters most.

Our findings underscore the limitations of traditional model evaluation approaches, which often fail to capture the real-world complexities of healthcare data, especially in time-sensitive and ethically sensitive domains. By employing clustering and density estimation methods, this research successfully identifies misprediction regions in the model's hyperspace, allowing for precise interventions in areas likely to impact patient outcomes.

This study presents a novel knowledge-guided machine learning framework that integrates clinical domain expertise with data-driven modeling to enhance prediction reliability and interpretability in healthcare. By embedding structured knowledge through U-shaped penalty functions and custom decision trees, the model aligns its predictions with physiologically meaningful patterns, particularly in regions where data is sparse, noisy, or clinically ambigu-

ous. The hybrid architecture ensures that the learning process not only minimizes empirical loss but also adheres to established medical reasoning. This approach improves generalization, enhances clinician trust, and identifies critical "problem regions" where predictions deviate from expected behavior. Our findings underscore the importance of incorporating domain knowledge as a first-class component in machine learning pipelines, especially in high-stakes applications like clinical decision support.

The comprehensive evaluations conducted in this work, including cross-dataset testing and ablation studies, provide a clear pathway for enhancing model trustworthiness in digital health. This research contributes not only to the technical development of more robust ML models but also to the ethical imperative of ensuring fairness and reducing bias in clinical predictions. By advancing model responsiveness and accountability, this dissertation lays a foundation for future work in safe and equitable AI in healthcare, paving the way for models that are both scientifically rigorous and socially responsible.

Bibliography

[1] Suman Jana Yuchi Tian, Kexin Pei and Baishakhi Ray. Deeptest: automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering (ICSE '18)*, page 303–314, 2018.

[2] Yinzhi Cao Junfeng Yang Pei, Kexin and Suman Jana. Deepxplore: Automated white-box testing of deep learning systems. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017.

[3] Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1962–1981, 2022.

[4] I. Ktena, O. Wiles, I. Albuquerque, and et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30:1166–1173, 2024.

[5] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[6] Chunjong Park, Anas Awadalla, Tadayoshi Kohno, and Shwetak Patel. Reliable and trustworthy machine learning for health using dataset shift detection. *Advances in Neural Information Processing Systems*, 34:3043–3056, 2021.

[7] Y. Yang, H. Zhang, J.W. Gichoya, D. Katabi, and M. Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30:2838–2848, 2024.

[8] Haoran Zhang, Harvineet Singh, Marzyeh Ghassemi, and Shalmali Joshi. " why did the model fail?": Attributing model performance changes to distribution shifts. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[9] Wenjia Song Charles B. Nemeroff Chang Lu Afrose, Sharmin and Danfeng Yao. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. *Communications Medicine*, 2(1):111, 2022.

[10] E. Pierson, D.M. Cutler, J. Leskovec, and et al. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27:136–140, 2021.

[11] M. Qi, H. Santos, P. Pinheiro, D.L. McGuinness, and K.P. Bennett. Demographic and socioeconomic determinants of access to care: A subgroup disparity analysis using new equity-focused measurements. *PLOS ONE*, 18(11):e0290692, 2023.

[12] L. Seyyed-Kalantari, H. Zhang, M.B.A. McDermott, and et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nature Medicine*, 27:2176–2182, 2021.

[13] Ahmed Ismail, Hong-Linh Truong, and Wolfgang Kastner. Manufacturing process data analysis pipelines: a requirements analysis and survey. *Journal of Big Data*, 6(1):1–26, 2019.

[14] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[15] Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

[16] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[17] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[18] E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, 2021.

[19] W. Yuan, B. K. Beaulieu-Jones, K.-H. Yu, S. L. Lipnick, N. Palmer, J. Loscalzo, T. Cai, and I. S. Kohane. Temporal bias in case-control design: preventing reliable predictions of the future. *Nature Communications*, 12(1):1107, 2021.

[20] CDC. CDC - BRFSS — cdc.gov. https://www.cdc.gov/brfss/. [Accessed 23-Apr-2023].

[21] Christel Sirocchi, Alessandro Bogliolo, and Sara Montagna. Medical-informed machine learning: integrating prior knowledge into medical decision systems. *BMC Medical Informatics and Decision Making*, 24(Suppl 4):186, 2024.

[22] Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10):2318–2331, 2017.

[23] Nikhil Muralidhar, Mohammad Raihanul Islam, Manish Marwah, Anuj Karpatne, and Naren Ramakrishnan. Incorporating prior domain knowledge into deep neural net-

works. In *2018 IEEE international conference on big data (big data)*, pages 36–45. IEEE, 2018.

[24] Arka Daw, Anuj Karpatne, William D Watkins, Jordan S Read, and Vipin Kumar. Physics-guided neural networks (pgnn): An application in lake temperature modeling. In *Knowledge guided machine learning*, pages 353–372. Chapman and Hall/CRC, 2022.

[25] Xiaowei Jia, Jared Willard, Anuj Karpatne, Jordan Read, Jacob Zwart, Michael Steinbach, and Vipin Kumar. Physics guided rnns for modeling dynamical systems: A case study in simulating lake temperature profiles. In *Proceedings of the 2019 SIAM international conference on data mining*, pages 558–566. SIAM, 2019.

[26] Xiaowei Jia, Jared Willard, Anuj Karpatne, Jordan S Read, Jacob A Zwart, Michael Steinbach, and Vipin Kumar. Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *ACM/IMS Transactions on Data Science*, 2(3):1–26, 2021.

[27] Jordan S Read, Xiaowei Jia, Jared Willard, Alison P Appling, Jacob A Zwart, Samantha K Oliver, Anuj Karpatne, Gretchen JA Hansen, Paul C Hanson, William Watkins, et al. Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, 55(11):9173–9190, 2019.

[28] Arka Daw, R Quinn Thomas, Cayelan C Carey, Jordan S Read, Alison P Appling, and Anuj Karpatne. Physics-guided architecture (pga) of neural networks for quantifying uncertainty in lake temperature modeling. In *Proceedings of the 2020 siam international conference on data mining*, pages 532–540. SIAM, 2020.

[29] Kaan Sel, Amirmohammad Mohammadi, Roderic I Pettigrew, and Roozbeh Jafari. Physics-informed neural networks for modeling physiological time series for cuffless blood pressure estimation. *npj Digital Medicine*, 6(1):110, 2023.

[30] Lavin P.T. Birch M. et al. Abràmoff, M.D. Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, 1:39, 2018.

[31] K. Sennaar. How america's 5 top hospitals are using machine learning today.

[32] Sarro D Alderton E Futoma J Gao M Nichols M Revoir M Yashar F Miller C Kester K Sandhu S Corey K Brajer N Tan C Lin A Brown T Engelbosch S Anstrom K Elish MC Heller K Donohoe R Theiling J Poon E Balu S Bedoya A O'Brien C. Sendak MP, Ratliff W. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Medical Informatics*, 8(7), 2020.

[33] W.L.; Vail C.J.; Daigle T.; Kirk A.D.; Allen P.J.; Henao R.; Buckland D.M. Zaribafzadeh, H.; Webster. Development, deployment, and implementation of a machine learning surgical case length prediction model and prospective evaluation. *Annals of Surgery*, 278(6):890–895, 2023.

[34] Donnelly JP et al. Wong A, Otles E. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*, 181(8):1065–1070, 2021.

[35] Yazdany J Schmajuk G. Gianfrancesco MA, Tamang S. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*, 178(11):1544–1547, 2018.

[36] Toberer F Enk A Deinlein T Hofmann-Wellenhof R Thomas L Lallas A Blum A Stolz W Haenssle HA. Winkler JK, Fink C. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10):1135–1141, 2019.

[37] National Transportation Safety Board. Collision between vehicle controlled by developmental automated driving system and pedestrian. Accident Report NTSB/HAR-19/03, Tempe, Arizona, March 18, 2018.

[38] Dara. Kerr. Driverless car startup cruise's no good, terrible year. Accessed June 6, 2024.

[39] Tadesse G.A. Ho D. et al. Liang, W. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4:669–677, 2022.

[40] T. M. Johnson, J. M. Khoshgoftaar. Survey on deep learning with class imbalance. *J. Big Data*, 6:1–54, 2019.

[41] Khachatrian H. Kale D.C. et al. Harutyunyan, H. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:96, 2019.

[42] Pollard Tom Johnson, Alistair and Roger Mark. Mimic-iii clinical database (version 1.4).

[43] Pollard T. J. Shen L. Lehman L. H. Feng M. Ghassemi M. Moody B. Szolovits P. Celi L. A. Mark R. G. Johnson, A. E. W. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[44] Vevake Balaraman Sheikhalishahi, Seyedmostafa and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *PLOS One*, 15(7):e0235424, 2020.

[45] Raffa JD Celi LA Mark RG Pollard TJ, Johnson AEW and Badawi O. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 2018.

[46] Leonard Gruelich Julian Varghese Hegselmann, Stefan and Martin Dugas. Reproducible survival prediction with seer cancer data. In *Machine Learning for Healthcare Conference*, pages 49–66, 2018.

[47] Epidemiology National Cancer Institute, Surveillance and End Results Program. Seer incidence data, 1975 – 2021. (accessed June 6, 2024).

[48] Karan Aggarwal Shafiq Joty Khadanga, Swaraj and Jaideep Srivastava. Using clinical notes with time series data for icu management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 6432–6437, 2019.

[49] Mohit Iyyer Deznabi, Iman and Madalina Fiterau. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4026–4031, 2021.

[50] Harvineet Singh Marzyeh Ghassemi Zhang, Haoran and Shalmali Joshi. "why did the model fail?": attributing model performance changes to distribution shifts. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, volume 202, page 41550–41578, 2023.

[51] Yuwen Chen Zhou, Helen and Zachary Lipton. Evaluating model performance in medical datasets over time. In *Conference on Health, Inference, and Learning*, pages 498–508, 2023.

[52] Ibomoiye Domor Mienye and Yanxia Sun. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, 25:100690, 2021.

[53] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[54] Cédric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Society, 2021.

[55] Hines R.B. Jhala N.C. et al. Shanmugam, C. Evaluation of lymph node numbers for adequate staging of stage ii and iii colon cancer. *J Hematol Oncol*, 4:25, 2011.

[56] Dong Y Yong J, Ding B and Yang M. Impact of examined lymph node number on lymph node status and prognosis in figo stage ib-iia cervical squamous cell carcinoma: A population-based study. *Frontiers in Oncology*, 12:994105, 2022.

[57] Poon JT. Choi HK, Law WL. The optimal number of lymph nodes examined in stage ii colorectal cancer and its impact on outcomes. *BMC Cancer*, 10:267, Jun 2010.

[58] Chen Y Chang J Jiang Y Zhu D Wei Y. Wu Q, Zhang Z. Impact of inadequate number of lymph nodes examined on survival in stage ii colon cancer. *Frontiers in Oncology*, 11:736678, 2021.

[59] American Cancer Society. Stages of breast cancer: Understand breast cancer staging. (accessed June 6, 2024).

[60] Understanding blood pressure readings. Understanding blood pressure readings. (accessed October 17, 2023).

[61] Johns Hopkins Medicine. Vital signs (body temperature, pulse rate, respiration rate, blood pressure). (accessed June 14, 2022).

[62] Cleveland Clinic. Vital signs: How to check my vitals at home.

[63] University of Rochester Medical Center. Vital signs (body temperature, pulse rate, respiration rate, blood pressure). (accessed June 6, 2024).

[64] Bhandari P. Sapra A, Malik A. Vital sign assessment. [Updated 2023 May 1]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-.

[65] Ren H Liu G Sun L. Sun L, Li P. Quantifying the number of lymph nodes for examination in breast cancer. *J Int Med Res*, 48(2):300060519879594, 2020.

[66] Wang H Wang Z. Chi H, Zhang C. The appropriate number of elns for lymph node negative breast cancer patients underwent mrm: a population-based study. *Oncotarget*, 8(39):65668–65676, Aug 2017.

[67] S. Jain and L. M. Iverson. Glasgow coma scale, 2018.

[68] Timon Gehr Mirman, Matthew and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3578–3586, 2018.

[69] Zhuoqun Fu Chuyun Deng Xiaojing Liao Jia Zhang Qin, Yue and Haixin Duan. Stolen risks of models with security properties. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 756–770, 2023.

[70] Anthony D. Joseph Blaine Nelson Benjamin IP Rubinstein Huang, Ling and J. Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, pages 43–58, 2011.

[71] J. D. Tygar. Adversarial machine learning. *IEEE Internet Computing*, 15(5):4–6, 2011.

[72] Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam. Adversarial attacks against medical deep learning systems. *arXiv preprint*, 2018.

[73] Nur Imtiazul Haque Amit Kumar Sikder Mohammad Ashiqur Rahman Newaz, AKM Iqtidar and A. Selcuk Uluagac. Adversarial attacks to machine learning-based smart healthcare systems. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pages 1–6, 2020.

[74] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint*, 2021.

[75] Law JZF Low LL Kwa ALH Giacomini KM Ting DSW. Ong JCL, Seng BJJ. Artificial intelligence, chatgpt, and other large language models for social determinants of health: Current state and future directions. *Cell Reports Medicine*, 5, 2024.

[76] Chen A. PourNejatian N. et al. Yang, X. A large language model for electronic health records. *npj Digital Medicine*, 5:194, 2022.

[77] Yang X. Chen A. et al. Peng, C. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6:210, 2023.

[78] Nature Medicine. How to support the transition to ai-powered healthcare. *Nature Medicine*, 30:609–610, 2024.

[79] S. Lundberg. A unified approach to interpreting model predictions. *arXiv preprint*, 2017.

[80] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[81] A. S. Jacobs and et al. Ai/ml for network security: The emperor has no clothes. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1537–1551, 2022.

[82] Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10):2318–2331, 2017.

[83] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

[84] Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30, 2017.

[85] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

[86] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34:4699–4711, 2021.

[87] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[88] Kushan De Silva, Wai Kit Lee, Andrew Forbes, Ryan T Demmer, Christopher Barton, and Joanne Enticott. Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. *International journal of medical informatics*, 143:104268, 2020.

[89] WHO. Diabetes — who.int. https://www.who.int/health-topics/diabetes. [Accessed 04-23-2023].

[90] Nam H Cho, JE Shaw, Suvi Karuranga, Yafang Huang, JD da Rocha Fernandes, AW Ohlrogge, and BIDF Malanda. Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes research and clinical practice*, 138:271–281, 2018.

[91] Etienne G Krug. Trends in diabetes: Sounding the alarm. *The Lancet*, 387(10027):1485–1486, 2016.

[92] Zidian Xie, Olga Nikolayeva, Jiebo Luo, and Dongmei Li. Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing chronic disease*, 16, 2019.

[93] Nadia Lascar, James Brown, Helen Pattison, Anthony H Barnett, Clifford J Bailey, and Srikanth Bellary. Type 2 diabetes in adolescents and young adults. *The lancet Diabetes & endocrinology*, 6(1):69–80, 2018.

[94] CDC. Diabetes in young people is on the rise — cdc.gov. https://www.cdc.gov/diabetes/data-research/research/young-people-diabetes-on-rise.html. [Accessed 10-09-2024].

[95] CDC. How to make the leap from type 1 teen to adult — cdc.gov. https://www.cdc.gov/diabetes/about/type-1-teen-adult.html. [Accessed 10-09-2024].

[96] Rozalina G McCoy, Renée SM Kidney, Danette Holznagel, Tina Peters, and Vimbai Madzura. Challenges for younger adults with diabetes. *Minnesota medicine*, 102(2):34, 2019.

[97] Diabetes. Type 2 diabetes and young adults — diabetes.org.uk. https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes/type-2/young-adults. [Accessed 12-09-2024].

[98] GOV NIH. Serious complications from youth-onset type 2 diabetes arise by young adulthood. *National Institutes of Health*, Jul 2021.

[99] CDC. Preventing type 2 diabetes — cdc.gov. https://www.cdc.gov/diabetes/prevention-type-2. [Accessed 10-09-2024].

[100] Diabetes. Diabetes prevention | ADA — diabetes.org. https://diabetes.org/about-diabetes/diabetes-prevention. [Accessed 10-09-2024].

[101] Wei-Chun Lin, Jimmy S Chen, Michael F Chiang, and Michelle R Hribar. Applications of artificial intelligence to electronic health record data in ophthalmology. *Translational vision science & technology*, 9(2):13–13, 2020.

[102] Siaw-Teng Liaw, Harshana Liyanage, Craig Kuziemsky, Amanda L Terry, Richard Schreiber, Jitendra Jonnagaddala, and Simon de Lusignan. Ethical use of electronic health record data and artificial intelligence: Recommendations of the primary care informatics working group of the international medical informatics association. *Yearbook of medical informatics*, 29(01):051–057, 2020.

[103] Thomas H Davenport, T Hongsermeier, and Kimberly Alba Mc Cord. Using AI to improve electronic health record. *Harvard Business Review*, 12:1–6, 2018.

[104] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[105] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 200(01):191–200, 2020.

[106] Young Juhn and Hongfang Liu. Artificial intelligence approaches using natural language processing to advance ehr-based clinical research. *J Allergy Clin Immunol*, 145(2):463–469, 2020.

[107] Kuan Zhang, Bardia Khosravi, Sanaz Vahdati, Shahriar Faghani, Fred Nugen, Seyed Moein Rassoulinejad-Mousavi, Mana Moassefi, Jaidip Manikrao M Jagtap, Yashbir Singh, Pouria Rouzrokh, et al. Mitigating bias in radiology machine learning: 2. model development. *Radiology: Artificial Intelligence*, 4(5):e220010, 2022.

[108] Stephanie S Gervasi, Irene Y Chen, Aaron Smith-McLallen, David Sontag, Ziad Obermeyer, Michael Vennera, and Ravi Chawla. The potential for bias in machine learning and opportunities for health insurers to address it: Article examines the potential for bias in machine learning and opportunities for health insurers to address it. *Health Affairs*, 41(2):212–218, 2022.

[109] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

[110] Jie Xu, Yunyu Xiao, Wendy Hui Wang, Yue Ning, Elizabeth A Shenkman, Jiang Bian, and Fei Wang. Algorithmic fairness in computational medicine. *EBioMedicine*, 84, 2022.

[111] Irene Y Chen. *Machine Learning Approaches for Equitable Healthcare*. PhD thesis, Massachusetts Institute of Technology, 2022.

[112] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[113] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can AI help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

[114] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[115] Charlene H Chu, Simon Donato-Woodger, Shehroz S Khan, Rune Nyrup, Kathleen Leslie, Alexandra Lyn, Tianyu Shi, Andria Bianchi, Samira Abbasgholizadeh Rahimi, and Amanda Grenier. Age-related bias and artificial intelligence: A scoping review. *Humanities and Social Sciences Communications*, 10(1):1–17, 2023.

[116] Weimin Tan, Qiaoling Wei, Zhen Xing, Hao Fu, Hongyu Kong, Yi Lu, Bo Yan, and Chen Zhao. Fairer AI in ophthalmology via implicit fairness learning for mitigating sexism and ageism. *Nature Communications*, 15(1):4750, 2024.

[117] Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. Faithful AI in healthcare and medicine. *medRxiv*, pages 2023–04, 2023.

[118] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4:123–144, 2021.

[119] Matthew BA McDermott, Bret Nestor, and Peter Szolovits. Clinical artificial intelligence: Design principles and fallacies. *Clinics in Laboratory Medicine*, 43(1):29–46, 2023.

[120] Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. *BMC medicine*, 9(1):1–14, 2011.

[121] Patrick W Sullivan, Elaine H Morrato, Vahram Ghushchyan, Holly R Wyatt, and James O Hill. Obesity, inactivity, and the prevalence of diabetes and diabetes-related cardiovascular comorbidities in the us, 2000–2002. *Diabetes care*, 28(7):1599–1603, 2005.

[122] Douglas Noble, Rohini Mathur, Tom Dent, Catherine Meads, and Trisha Greenhalgh. Risk models and scores for type 2 diabetes: Systematic review. *Bmj*, 343, 2011.

[123] Sharmin Afrose, Wenjia Song, Charles B Nemeroff, Chang Lu, and Danfeng Yao. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. *Communications medicine*, 2(1):111, 2022.

[124] Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pages 935–942, 2007.

[125] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[126] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint*

conference on neural networks (IEEE world congress on computational intelligence), pages 1322–1328. Ieee, 2008.

[127] Ivan Tomek. Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772, 1976.

[128] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):408–421, 1972.

[129] Inderjeet Mani and I Zhang. kNN approach to unbalanced data distributions: A case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, pages 1–7. ICML, 2003.

[130] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

[131] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[132] Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron) — A review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[133] Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586):eabb1655, 2021.

[134] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4:123–144, 2021.

[135] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

[136] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

Appendices

Appendix A

Appendix: Chapter 3

A.1 Supplementary Notes

ML responsiveness (defined by us): A machine learning model's ability to recognize changing patient conditions, such as severely impaired medical conditions and (rapidly) deteriorating health, and to adjust the model's prediction accordingly.

In-hospital mortality prediction Class 1 : Based on the first 48 hours of ICU information, the patient dies in the ICU.

Breast cancer survivability Class 1: Patient survives more than 5 years after breast cancer diagnosis.

Lung cancer survivability class 1 : Patient survives more than 5 years after the lung cancer diagnosis.

Glasgow Coma Scale (GCS): A scale for assessing a person's level of consciousness by scoring their eye, verbal, and motor responses [Jain 2018] .

High risk zone (defined by us) : The mortality risk range (threshold, 1.0] is defined as the high-risk zone.

Low risk zone (defined by us) : The mortality risk range [0, threshold] is defined as the low-risk zone.

Bradypnea : Abnormally slow breathing (< 12 BPM).

Tachycardia : Abnormally fast heart rate (> 100 BPM).

Hypothermia: Body temperature below normal (< 35°C / 95°F).

Hyperthermia: Elevated body temperature (> 38.5°C / 101.3°F).

Hypoglycemia : Low blood glucose levels (< 70 mg/dL).

Hyperglycemia : High blood glucose levels (> 180 mg/dL).

Hypotension: Low blood pressure (< 80/60 mmHg).

Hypertension: High blood pressure (> 130/90 mmHg).

Hypoxemia : Low blood oxygen saturation (< 90

CS tumor size: Collaborative Staging (CS) tumor size records the largest dimension (length, width, or height) or the diameter of the primary tumor in millimeters.

T: The extent (size) of the tumor.

- T0: absence of such tumor.
- T1: tumor size less than 20 mm.
- T2: tumor size 20-50 mm.
- T3: tumor size larger than 50 mm.
- T4: any size growing into skin.

N: The spread to nearby lymph nodes.

- N0: 0 positive lymph nodes.

- N1: 1-3 positive lymph nodes.
- N2: 4-9 positive lymph nodes.
- N3: more than 10 positive lymph nodes.

ELNs: Examined number of lymph nodes.

A.2 Tables

Table A.1: Training and validation losses of selected in-hospital mortality risk predictor models.

Prediction task	Model	Trial	Selected Epoch	Validation AUPRC	Training Loss	Validation Loss	Difference
MIMIC III	LSTM	1	39	0.568	0.273	0.278	0.005
		2	29	0.564	0.279	0.278	0.001
		3	33	0.566	0.274	0.279	0.005
	CW-LSTM	1	37	0.556	0.270	0.284	0.014
		2	24	0.556	0.280	0.286	0.006
		3	20	0.561	0.283	0.281	0.002
	LSTM+ SMOTE	1	4	0.410	0.438	0.385	0.054
		2	11	0.410	0.342	0.355	0.013
		3	2	0.366	0.473	0.417	0.056
	LSTM+ AdaSyn	1	18	0.337	0.318	0.411	0.093
		2	6	0.413	0.395	0.353	0.042
		3	33	0.408	0.305	0.343	0.038
	LSTM+ Reweight	1	63	0.568	0.436	0.429	0.007
		2	58	0.559	0.428	0.431	0.003
		3	66	0.561	0.429	0.424	0.005
eICU	LSTM	1	81	0.476	0.277	0.280	0.003
		2	87	0.510	0.274	0.283	0.009
		3	86	0.492	0.279	0.276	0.003
	LSTM+ SMOTE	1	49	0.385	0.495	0.410	0.085
		2	55	0.402	0.445	0.391	0.055
		3	42	0.433	0.412	0.471	0.059
	LSTM+ AdaSyn	1	39	0.320	0.410	0.419	0.009
		2	38	0.376	0.454	0.499	0.045
		3	35	0.336	0.439	0.468	0.029
	LSTM+ Reweight	1	99	0.473	0.532	0.526	0.006
		2	90	0.495	0.512	0.539	0.026
		3	89	0.454	0.494	0.502	0.008

Table A.2: Training and validation losses of selected cancer survivability predictor models.

Prediction task	Model	Trial	Selected Epoch	Validation AUPRC	Training Loss	Validation Loss	Difference
SEER BCS	MLP	1	3	0.673	0.231	0.230	0.002
		2	3	0.672	0.232	0.229	0.003
		3	4	0.675	0.230	0.228	0.002
	MLP+SMOTE	1	4	0.654	0.323	0.340	0.017
		2	2	0.650	0.350	0.340	0.010
		3	5	0.647	0.310	0.338	0.028
	MLP+AdaSyn	1	1	0.643	0.393	0.384	0.010
		2	3	0.622	0.391	0.357	0.034
		3	2	0.618	0.378	0.353	0.025
	MLP+Reweight	1	7	0.677	0.389	0.352	0.037
		2	5	0.666	0.391	0.354	0.037
		3	5	0.672	0.389	0.361	0.028
SEER LCS	MLP	1	5	0.724	0.248	0.254	0.006
		2	4	0.725	0.250	0.254	0.004
		3	4	0.727	0.245	0.244	0.002
	MLP+SMOTE	1	3	0.705	0.311	0.339	0.028
		2	5	0.710	0.322	0.345	0.024
		3	4	0.708	0.341	0.362	0.021
	MLP+AdaSyn	1	3	0.682	0.328	0.391	0.063
		2	3	0.701	0.342	0.366	0.024
		3	4	0.682	0.322	0.375	0.054
	MLP+Reweight	1	3	0.726	0.376	0.351	0.026
		2	3	0.715	0.367	0.354	0.012
		3	5	0.725	0.371	0.362	0.009

Table A.3: Number of samples in the training set after resampling

Dataset	Resampling Method	Total Samples	Class 0	Class 1
MIMIC III	SMOTE	25,222	12,694	12,528
	AdaSyn	24,872	12,694	12,178
eICU	SMOTE	38,030	19,015	19,015
	AdaSyn	37,417	19,015	18,402
BCS	SMOTE	347,436	173,718	173,718
	AdaSyn	344,528	170,810	173,718
LCS	SMOTE	276,218	138,109	138,109
	AdaSyn	277,264	138,109	139,155

Table A.4: Cost-sensitive learning balanced class weights

Dataset	Class 0	Class 1
MIMIC III	0.578	3.694
eICU LSTM	0.565	4.363
LCS MLP	0.595	3.122
BCS MLP	3.936	0.573

Table A.5: Thresholds of various models. Thresholds are identified through the validation process.

Models/ Datasets	NN	NN+ SMOTE	NN+ AdaSyn	NN+ Reweight	XG-Boost	Random Forest	Ada-Boost	KNN	Naive Bayes
eICU (NN: LSTM)	0.24	0.11	0.14	0.24	0.22	0.24	0.24	0.21	0.17
MIMIC-III (NN: LSTM)	0.22	0.19	0.18	0.22	0.26	0.25	0.18	0.21	0.06
SEER-BCS (NN: MLP)	0.71	0.694	0.695	0.69	0.67	0.70	0.70	0.71	0.83
SEER-LCS (NN: MLP)	0.34	0.29	0.31	0.31	0.32	0.34	0.32	0.31	0.11

Table A.6: Mortality risk prediction of two medical experts on attribute-varying test cases. Doctors were given below and told the attribute values may fluctuate.

Case	Attribute values	Predicted risk - Synthesized test cases			
		MR MD 1	MR MD 2	MR (avg) MD	MR LSTM
1	Systolic blood pressure (mm Hg): 52 Diastolic blood pressure (mm Hg): 48 Glucose level (mg/dL): 22 Respiratory rate (BPM): 37 Heart rate (BPM): 21 Oxygen saturation (%): 34	0.75	0.95	0.85	0.084
2	Systolic blood pressure (mm Hg): 79 Diastolic blood pressure (mm Hg): 35 Glucose level (mg/dL): 47 Respiratory rate (BPM): 41 Heart rate (BPM): 20 Oxygen saturation (%): 43	0.80	0.90	0.85	0.137
3	Systolic blood pressure (mm Hg): 61 Diastolic blood pressure (mm Hg): 32 Glucose level (mg/dL): 35 Respiratory rate (BPM): 42 Heart rate (BPM): 48 Oxygen saturation (%): 26	0.85	0.90	0.875	0.611
4	Systolic blood pressure (mm Hg): 61 Diastolic blood pressure (mm Hg): 43 Glucose level (mg/dL): 35 Respiratory rate (BPM): 42 Heart rate (BPM): 48 Oxygen saturation (%): 20	0.85	0.90	0.875	0.56
5	Systolic blood pressure (mm Hg): 223 Diastolic blood pressure (mm Hg): 153 Glucose level (mg/dL): 253 Respiratory rate (BPM): 30 Heart rate (BPM): 117 Oxygen saturation (%): 68	0.50	0.50	0.50	0.117
6	Systolic blood pressure (mm Hg): 295 Diastolic blood pressure (mm Hg): 153 Glucose level (mg/dL): 253 Respiratory rate (BPM): 45 Heart rate (BPM): 136 Oxygen saturation (%): 62	0.55	0.60	0.575	0.455

Table A.7: Deteriorating test cases labeling of mortality risk by 2 medical doctors. Doctors are given the time series.

Attribute name	Case	Predicted risk - Deteriorating test cases			
		MR MD 1	MR MD 2	MR (avg) MD	MR LSTM
Respiratory rate	1	0.85	0.60	0.73	0.13
	2	0.90	0.80	0.85	0.19
	3	0.90	0.60	0.75	0.16
Temperature	1	0.90	0.40	0.65	0.18
	2	0.50	0.35	0.43	0.17
	3	0.40	0.30	0.35	0.18
Systolic BP	1	0.95	0.90	0.93	0.13
	2	0.95	0.90	0.93	0.03
	3	0.95	0.90	0.93	0.20
Multi-attribute	1	0.80	0.90	0.85	0.20
	2	0.85	0.90	0.88	0.18
	3	0.80	0.90	0.85	0.19

Table A.8: The table shows the gradient-based test case generation steps. Each step of gradient ascent creates a new test case by changing a single attribute value.

Attribute	Deteriorating test case	Step factor	# of Steps	Seed attribute value	Final attribute value
Diastolic BP (mm Hg)	1	0.01	16	52.94	10.77
	2	0.01	100	55.89	1.19
	3	0.01	100	89.37	4.49
Systolic BP (mm Hg)	1	0.2	54	119.85	95.00
	2	0.2	35	124.70	105.35
	3	0.2	170	125.59	67.99
Resp. Rate (BPM)	1	0.01	82	15.67	27.51
	2	0.01	138	17.49	31.14
	3	0.01	218	20.49	42.40
Oxygen Sat. (%)	1	0.01	9	69.26	0.00
	2	0.01	21	69.83	0.00
	3	0.01	44	80.68	0.00
Temp (C)	1	0.001	27	36.78	35.96
	2	0.001	38	36.62	35.64
	3	0.001	57	36.79	35.41

Table A.9: The table shows the gradient-based test predictions.

Attribute	Deteriorating test case	MR by LSTM 1	MR by LSTM 2	MR by LSTM 3	MR by LR	MR by CW-LSTM
Diastolic BP (mm Hg)	1	0.04	0.22	0.22	0.03	0.05
	2	0.10	0.04	0.08	0.20	0.02
	3	0.07	0.06	0.09	0.14	0.05
Systolic BP (mm Hg)	1	0.14	0.22	0.06	0.02	0.27
	2	0.07	0.11	0.22	0.13	0.19
	3	0.18	0.21	0.22	0.05	0.37
Resp. Rate (BPM)	1	0.22	0.12	0.06	0.05	0.28
	2	0.17	0.22	0.20	0.45	0.22
	3	0.11	0.15	0.22	0.32	0.11
Oxygen Sat. (%)	1	0.03	0.15	0.15	0.18	0.01
	2	0.03	0.01	0.03	0.67	0.01
	3	0.02	0.01	0.01	0.38	0.03
Temp (C)	1	0.15	0.22	0.20	0.02	0.18
	2	0.15	0.14	0.23	0.14	0.10
	3	0.15	0.17	0.23	0.05	0.13

Table A.10: Number of test cases (excluding deteriorating condition) in one test set created from a single seed case in the original MIMIC-III dataset. Five test sets are produced for each attribute or combination of attributes using five different seeds. The deteriorating conditions test is generated by three seeds resulting in 12 test cases.

Test type	Tested attribute	Number of cases (per set)	Number of total cases
GCS test	GCS total	120	600
	GCS eye-motor	24	120
	GCS eye-verbal	20	100
	GCS motor-verbal	30	150
GCS test total case		194	970
Single attribute critical zone test	Diastolic BP	250	1,250
	Glucose	600	3,000
	Oxygen saturation	101	505
	Respiratory rate	101	505
	Systolic BP	350	1,750
	Temperature	13	65
Critical zone test total case		1415	7,075
Double-attribute critical zone test	Diastolic BP & Glucose	2,500	12,500
	Diastolic BP & Systolic BP	3,500	17,500
	Respiratory rate & Heart rate	2,500	12,500
Double attribute critical zone test total case		8,500	42,500
Multi-attribute (6 attributes) critical zone test	High zone	12,695	63,475
	Low zone	12,695	63,475
Multi-attribute critical zone test (6 attributes) total case		25,390	126,950
Deteriorating conditions test (gradient ascent method)	Single attribute	9	9
	Triple-attribute	3	3
Deteriorating conditions test total case		12	12
Grand total		35,511	177,507

Table A.11: Clinical Assessment of Neurological Function: Glasgow Coma Scale (GCS) Scoring System.

GCS component	Response	GCS score
Eye-opening (E)	No response	1
	To pain	2
	To verbal	3
	Spontaneous	4
Verbal response (V)	No response	1
	Incomprehensible sounds	2
	Inappropriate words	3
	Confused, disoriented	4
	Oriented, converses	5
Motor response (M)	No response	1
	Abnormal extension to pain	2
	Abnormal flexion to pain	3
	Withdraws to pain	4
	Localizes pain	5
	Obeys commands	6

Table A.12: Data distribution of created single attribute test set for SEER 5-year breast cancer survivability (BCS) and lung cancer survivability (LCS).

Single attribute test	EOD 10 - positive lymph nodes examined continuously					
	N stage	Group (count)	Lung cancer set		Breast cancer set	
			# case per set	# total case	# case per set	# total case
	N0	0	111	333	30	90
	N1	(0, 4)	333	999	90	270
	N2	[4, 10)	666	1,998	182	546
	N3	>10	6,978	20,934	2,260	6,780
	Total		8,088	24,264	2,562	7,686
	CS Tumor size continuous					
	T stage	Group (mm)	Lung cancer set		Breast cancer set	
			# case per set	# total case	# case per set	# total case
	T0	0	4	12	6	18
	T1	(0, 20]	57	171	81	243
	T2	(20-50]	91	273	130	390
	T3	(50, 988]	2,637	7,911	4,080	12,240
	Total		2,789	8,367	4,297	12,891
	Grades					
	Grade stage	status	Seed Class		Breast cancer set	
					# case per class	# total case
	1	Well differentiated	Survived (C1)		21,723	24,875
			Death (C0)		3,152	
	2	Moderately differentiated	Survived (C1)		21,723	24,875
			Death (C0)		3,152	
	3	Poorly differentiated	Survived (C1)		21,723	24,875
			Death (C0)		3,152	
	4	Undifferentiated	Survived (C1)		21,723	24,875
			Death (C0)		3,152	
	Total					99,500

Table A.13: Data distribution of created multi-attribute test set for SEER 5-year breast cancer survivability (BCS) and lung cancer survivability (LCS).

Double attribute test	Combinations		Breast cancer set	
			# case per set	# total case
	T-N	Tumor size and positive lymph node	6,177	18,531
	T-ENLs	Tumor size and # examined lymph nodes	6,177	18,531
	N-ENLs	# examined lymph nodes and positive lymph node	7,800	23,400
	Total			60,462
Triple attributes test	Three attributes (T4, N3, Grade 4)		Breast cancer set	
			# case per class	# total case
	Seed class survived (C1)		21,723	24,875
	Seed class death (C0)		3,152	
	Total			24,875

Table A.14: Ranges of important vitals.

Attribute	Ideal range	Critical risk zone	
		Low	High
Diastolic BP (mm Hg)	70-80	0-60	120-250
Glucose (mg/dL)	70-180	0-54	250-600
Respiratory Rate (BPM)	12-18	0-10	24-100
Temperature (C)	37	30-35	39-42
Oxygen Saturation (%)	95-100	0-90	-
Systolic BP (mm Hg)	100-120	0-90	180-350

Table A.15: Mean standard deviation of the attribute of the original MIMIC-III dataset.

Attributes	All		Class 0 (Non-death)		Class 1 (Death)	
	Mean	SD (avg)	Mean	SD (avg)	Mean	SD (avg)
Diastolic bp (mm Hg)	61.45	15.61	61.60	14.20	60.46	24.63
GCS eye-opening	3.22	0.59	3.34	0.58	2.47	0.63
GCS motor response	5.33	0.61	5.46	0.60	4.47	0.70
GCS total	11.85	1.65	12.28	1.67	9.24	1.51
GCS verbal response	3.32	0.67	3.48	0.71	2.30	0.46
Glucose (mg/dL)	141.54	48.34	140.72	49.67	146.81	39.78
Heart Rate (BPM)	86.67	10.08	86.14	9.91	90.03	11.13
Oxygen saturation (%)	97.98	11.50	98.23	12.61	96.44	4.39
Respiratory rate (BPM)	20.96	19.92	21.05	22.27	20.44	4.95
Systolic bp (mm Hg)	120.25	15.67	120.78	15.48	116.85	16.92
Temperature (C)	36.94	0.71	36.95	0.69	36.88	0.86

Table A.16: Selected seeds from the MIMIC-III dataset. These five seeds are used to create attribute-based test cases. Mean and SD represent the average and standard deviation values of the particular attribute. Missing values were ignored for calculating the mean and standard deviation.

Seeds	1		2		3		4		5	
Attribute	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Diastolic BP (mm Hg)	55.40	8.27	55.85	7.00	55.77	4.29	66.08	5.39	66.72	10.05
GCS eye opening	4.00	0.00	4.00	0.00	4.00	0.00	4.00	0.00	4.00	0.00
GCS motor response	6.00	0.00	6.00	0.00	6.00	0.00	6.00	0.00	6.00	0.00
GCS total	15.00	0.00	15.00	0.00	15.00	0.00	15.00	0.00	15.00	0.00
GCS verbal response	5.00	0.00	5.00	0.00	5.00	0.00	5.00	0.00	5.00	0.00
Glucose (mg/dL)	118.75	13.84	153.70	22.60	97.00	8.19	126.50	7.33	109.00	21.21
Heart Rate (BPM)	74.27	8.29	64.40	5.72	66.90	6.22	99.34	9.89	67.67	5.84
Mean BP (mm Hg)	73.15	5.69	84.67	7.32	72.64	4.63	84.88	4.87	84.92	10.53
Oxygen sat. (%)	96.71	1.68	96.81	1.13	98.51	2.01	99.06	0.91	95.10	1.79
Respiratory rate (BPM)	15.33	3.87	16.33	3.50	17.07	2.11	15.52	3.38	14.22	2.83
Systolic BP (mm Hg)	108.00	8.91	116.34	12.62	106.38	6.54	122.48	5.96	121.32	14.61
Temperature (C)	37.38	0.37	36.97	0.56	37.41	0.27	37.21	0.27	37.52	0.50

Table A.17: Average accuracy of machine learning models under various testing conditions for in-hospital mortality prediction.

		Accuracy		
	Attributes	LSTM	CW-LSTM	LR
Single attribute critical zone tests	Respiratory rate	64.67%	13.00%	59.00%
	Temperature	58.97%	23.08%	46.15%
	Diastolic BP	23.60%	23.60%	23.60%
	Oxygen saturation	10.89%	10.89%	38.61%
	Systolic BP	35.24%	30.00%	25.43%
	Glucose	33.62%	33.62%	33.62%
Average accuracy of single-attribute tests		37.83%	22.36%	37.74%
GCS		33.33%	33.33%	100%
Multiple attribute critical zone tests	Zones			
	High critical zone	22.50%	40.10%	0.15%
	Low critical zone	68.80%	98.40%	12.30%
Average accuracy of multi-attribute tests		45.65%	69.25%	6.23%
Deteriorating condition tests	Any critical zones	83.33%	66.66%	33.33%

Table A.18: Changes in neural zone activation values of the LSTM model. Values in columns 2 and 3 represent the average difference of a neuron's activation score between two different regions (critically low zone, normal zone, and critically high zone).

Test Attribute	NZA(Low zone, normal zone)	NZA(normal zone, high zone)
Temperature	0.16	0.01
Glucose	0.01	0.04
Diastolic blood pressure	0.02	0.01
Respiratory rate	0.01	0.14

Table A.19: Average accuracy of MLP model under different test scenarios for 5-year breast cancer survivability prediction.

Single attribute test	Positive lymph nodes examined			
	N stage	Group (count)		Accuracy
	N0	0		100%
	N1	(0, 4)		0%
	N2	[4, 10)		0%
	N3	>10		74.4%
	Average			43.6%
	CS Tumor size			
	T stage	Group (mm)		Accuracy
	T0	0		100%
	T1	(0, 20]		0%
	T2	(20-50]		0%
	T3	(50, 988]		0%
	Average			25%
	Grades			
	Grade stage	status	Seed Class	Accuracy
	1	Well differentiated	Survived (C1)	96.97%
			Death (C0)	48.67%
	2	Moderately differentiated	Survived (C1)	4.55%
			Death (C0)	57.93%
	3	Poorly differentiated	Survived (C1)	7.44%
			Death (C0)	65.77%
	4	Undifferentiated	Survived (C1)	6.66%
			Death (C0)	63.86%
	Average			43.98%
Double attribute test	Double Combinations			Accuracy
	T-N: Tumor size and positive lymph node combination			92.97%
	T-ENL: Tumor size and number of examined lymph nodes			0%
	N-ENL: Number of examined lymph nodes and positive lymph node			19.6%
	Average			37.52%
Triple attributes test	Three attributes (T4, N3, Grade 4)			Accuracy
	Seed class survived (C1)			89.89%
	Seed class death (C0)			98.92%
	Average			94.41%

Table A.20: Wasserstein distances between various training and testing datasets.

Training data	Testing data	Wasserstein distances (avg)
Original MIMIC-III train set	Original MIMIC-III test set	12.42
Original MIMIC-III train set	Synthesized MIMIC-III multi-attribute test set	33.38
Original SEER BCS train set	Original SEER BCS test set	2.1
Original SEER BCS train set	Synthesized SEER BCS triple-attribute test set	9.75

Appendix B

Appendix: Chapter 5

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.70	0.36	0.44	0.48	0.78	0.93	0.72	0.85	0.76	0.74	0.82	0.37
Male	0.72	0.36	0.45	0.48	0.74	0.93	0.70	0.82	0.74	0.73	0.81	0.36
Female	0.68	0.37	0.44	0.48	0.81	0.94	0.73	0.87	0.79	0.74	0.83	0.38
White	0.66	0.35	0.42	0.46	0.79	0.93	0.73	0.86	0.77	0.73	0.82	0.36
Black	0.84	0.39	0.52	0.53	0.61	0.92	0.62	0.73	0.66	0.72	0.80	0.37
Indian	0.82	0.43	0.52	0.56	0.60	0.91	0.60	0.72	0.66	0.71	0.77	0.37
Asian	0.66	0.37	0.42	0.48	0.82	0.94	0.73	0.87	0.80	0.74	0.83	0.39
Hawaiian	0.84	0.49	0.71	0.62	0.67	0.92	0.55	0.77	0.72	0.76	0.86	0.46
Other race	0.83	0.33	0.60	0.47	0.67	0.95	0.69	0.79	0.70	0.75	0.86	0.38
Multiracial	0.73	0.40	0.51	0.52	0.78	0.94	0.69	0.85	0.77	0.76	0.84	0.42
Hispanic	0.74	0.41	0.51	0.53	0.76	0.93	0.66	0.83	0.75	0.75	0.83	0.41
Age 30-34	0.30	0.18	0.17	0.23	0.96	0.98	0.94	0.97	0.95	0.63	0.84	0.21
Age 35-39	0.45	0.29	0.28	0.35	0.95	0.97	0.89	0.96	0.93	0.70	0.87	0.32
Age 40-44	0.45	0.27	0.30	0.34	0.92	0.96	0.86	0.94	0.89	0.68	0.84	0.29
Age 45-49	0.56	0.36	0.39	0.44	0.89	0.95	0.79	0.92	0.85	0.72	0.84	0.37
Age 50-54	0.59	0.37	0.43	0.45	0.84	0.93	0.74	0.88	0.80	0.71	0.82	0.35
Age 55-59	0.69	0.40	0.47	0.50	0.79	0.93	0.69	0.85	0.77	0.74	0.83	0.39
Age 60-64	0.71	0.39	0.50	0.50	0.74	0.92	0.67	0.82	0.74	0.73	0.81	0.37
Age 65-69	0.73	0.38	0.48	0.50	0.71	0.91	0.66	0.80	0.71	0.72	0.79	0.36
Age 70-74	0.76	0.38	0.48	0.50	0.65	0.91	0.65	0.76	0.67	0.70	0.77	0.34
Age 75-79	0.79	0.37	0.49	0.51	0.58	0.90	0.63	0.71	0.63	0.69	0.76	0.32
Age 80-99	0.76	0.28	0.37	0.41	0.53	0.90	0.70	0.66	0.57	0.64	0.71	0.23

(a) LR original model performance

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.68	0.35	0.40	0.47	0.77	0.93	0.73	0.84	0.76	0.73	0.81	0.36
Male	0.70	0.35	0.41	0.46	0.74	0.92	0.71	0.82	0.73	0.72	0.79	0.35
Female	0.67	0.36	0.42	0.47	0.80	0.94	0.73	0.86	0.78	0.73	0.82	0.37
White	0.67	0.34	0.39	0.45	0.78	0.93	0.74	0.85	0.77	0.73	0.81	0.35
Black	0.80	0.39	0.46	0.52	0.62	0.91	0.64	0.74	0.66	0.71	0.77	0.35
Indian	0.74	0.44	0.50	0.55	0.66	0.87	0.61	0.75	0.68	0.70	0.75	0.35
Asian	0.64	0.36	0.35	0.46	0.82	0.93	0.75	0.87	0.79	0.73	0.79	0.37
Hawaiian	0.71	0.48	0.60	0.57	0.70	0.86	0.57	0.78	0.71	0.71	0.81	0.38
Other race	0.73	0.34	0.41	0.47	0.73	0.93	0.72	0.82	0.73	0.73	0.80	0.36
Multiracial	0.71	0.39	0.49	0.50	0.78	0.93	0.69	0.85	0.76	0.74	0.82	0.39
Hispanic	0.68	0.40	0.48	0.50	0.76	0.91	0.67	0.83	0.75	0.72	0.81	0.37
Age 30-34	0.21	0.16	0.11	0.18	0.97	0.98	0.94	0.97	0.95	0.59	0.82	0.16
Age 35-39	0.31	0.22	0.20	0.26	0.95	0.97	0.90	0.96	0.92	0.63	0.83	0.22
Age 40-44	0.38	0.27	0.25	0.32	0.93	0.96	0.86	0.94	0.90	0.65	0.81	0.27
Age 45-49	0.51	0.35	0.33	0.42	0.90	0.94	0.80	0.92	0.86	0.70	0.82	0.35
Age 50-54	0.57	0.37	0.40	0.45	0.85	0.93	0.74	0.88	0.81	0.71	0.81	0.35
Age 55-59	0.68	0.38	0.42	0.49	0.78	0.92	0.70	0.84	0.76	0.73	0.80	0.38
Age 60-64	0.72	0.38	0.44	0.49	0.72	0.92	0.68	0.81	0.72	0.72	0.79	0.36
Age 65-69	0.75	0.36	0.46	0.48	0.66	0.91	0.67	0.77	0.68	0.71	0.78	0.33
Age 70-74	0.76	0.37	0.46	0.50	0.63	0.90	0.65	0.75	0.66	0.70	0.76	0.33
Age 75-79	0.77	0.37	0.46	0.50	0.59	0.89	0.63	0.71	0.63	0.68	0.74	0.31
Age 80-99	0.72	0.28	0.34	0.41	0.55	0.89	0.71	0.68	0.58	0.64	0.69	0.22

(a) MLP original model performance

Figure B.1: Performance of the logistic regression and multi-layer perceptron models trained on the original training set (BRFSS 2021).

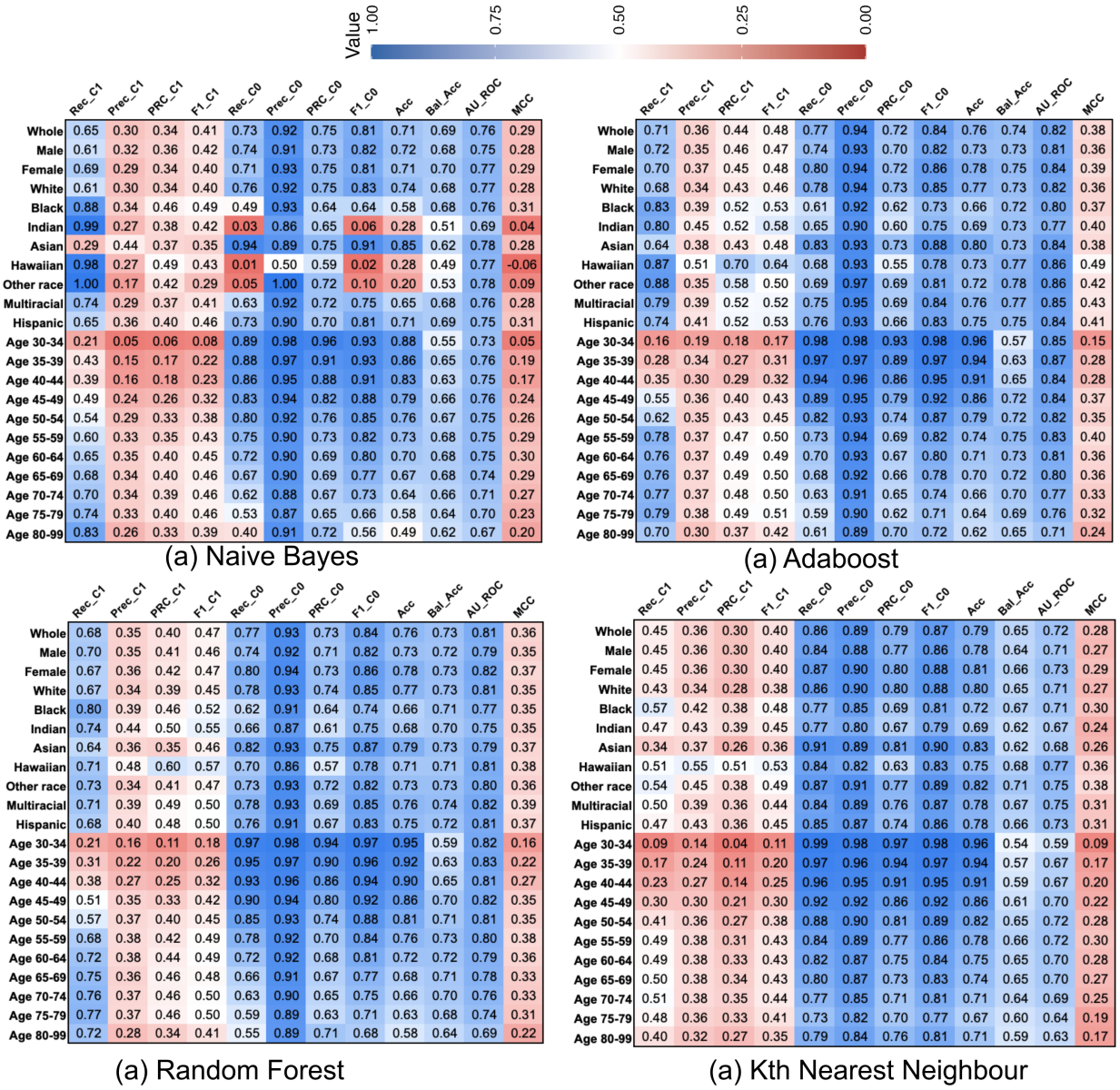
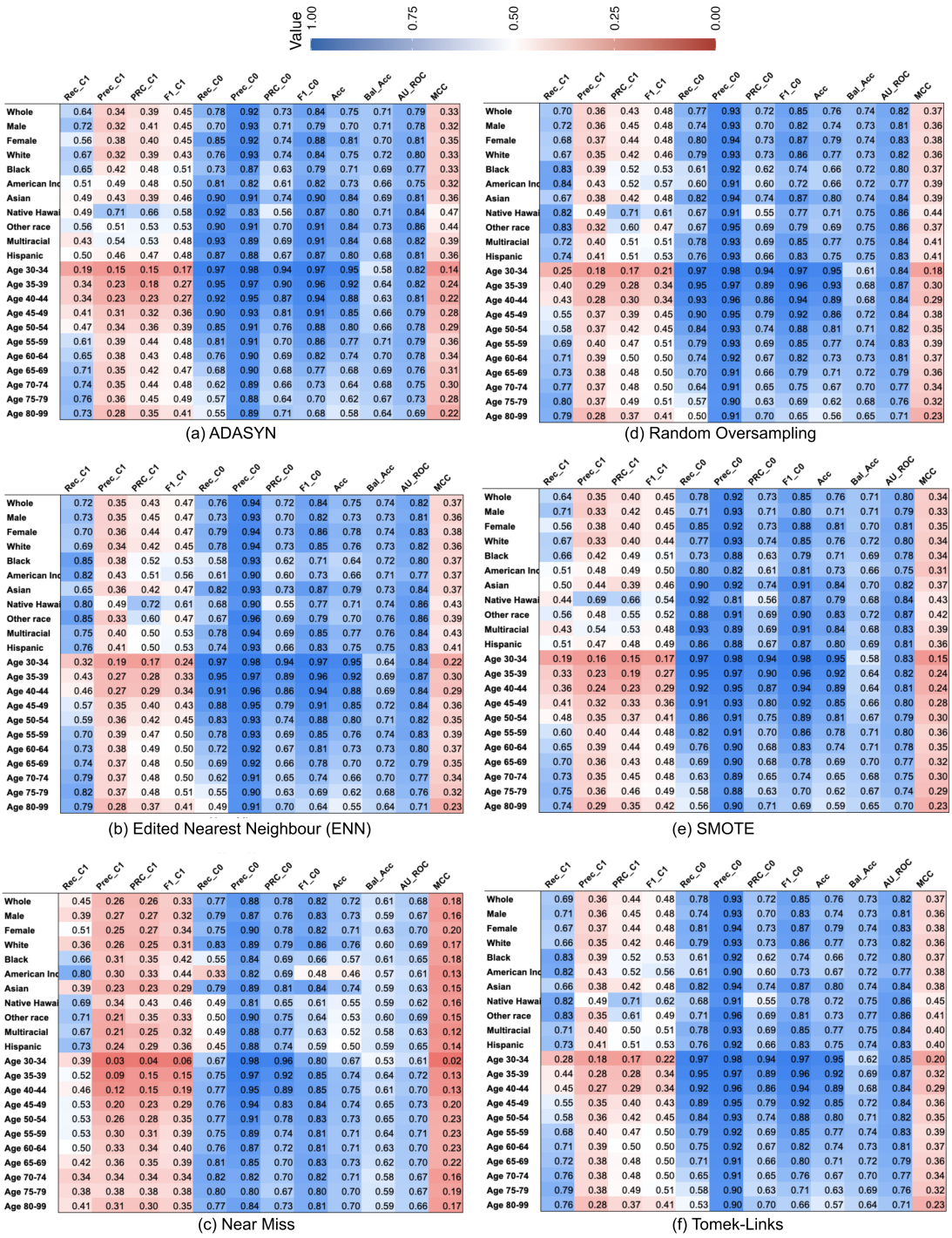


Figure B.2: Performance of the machine learning models trained on the original training set (BRFSS 2021).



(a) ADASYN

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.64	0.34	0.39	0.45	0.78	0.92	0.73	0.84	0.75	0.71	0.79	0.33
Male	0.72	0.32	0.41	0.45	0.70	0.93	0.71	0.79	0.70	0.71	0.78	0.32
Female	0.56	0.38	0.40	0.45	0.85	0.92	0.74	0.88	0.81	0.70	0.81	0.35
White	0.67	0.32	0.39	0.43	0.76	0.93	0.74	0.84	0.75	0.72	0.80	0.33
Black	0.65	0.42	0.48	0.51	0.73	0.87	0.63	0.79	0.71	0.69	0.77	0.33
American Inc	0.51	0.49	0.48	0.50	0.81	0.82	0.61	0.82	0.73	0.66	0.75	0.32
Asian	0.49	0.43	0.39	0.46	0.90	0.91	0.74	0.90	0.84	0.69	0.81	0.36
Native Hawai	0.49	0.71	0.66	0.58	0.92	0.83	0.56	0.87	0.80	0.71	0.84	0.47
Other race	0.56	0.51	0.53	0.53	0.90	0.91	0.70	0.91	0.84	0.73	0.86	0.44
Multiracial	0.43	0.54	0.53	0.48	0.93	0.89	0.69	0.91	0.84	0.68	0.82	0.39
Hispanic	0.50	0.46	0.47	0.48	0.87	0.88	0.67	0.87	0.80	0.68	0.81	0.36
Age 30-34	0.19	0.15	0.15	0.17	0.97	0.98	0.94	0.97	0.95	0.58	0.82	0.14
Age 35-39	0.34	0.23	0.18	0.27	0.95	0.97	0.90	0.96	0.92	0.64	0.82	0.24
Age 40-44	0.34	0.23	0.23	0.27	0.92	0.95	0.87	0.94	0.88	0.63	0.81	0.22
Age 45-49	0.41	0.31	0.32	0.36	0.90	0.93	0.81	0.91	0.85	0.66	0.79	0.28
Age 50-54	0.47	0.34	0.36	0.39	0.85	0.91	0.76	0.88	0.80	0.66	0.78	0.29
Age 55-59	0.61	0.39	0.44	0.48	0.81	0.91	0.70	0.86	0.77	0.71	0.79	0.36
Age 60-64	0.65	0.38	0.43	0.48	0.76	0.90	0.69	0.82	0.74	0.70	0.78	0.34
Age 65-69	0.71	0.35	0.42	0.47	0.68	0.90	0.68	0.77	0.68	0.69	0.76	0.31
Age 70-74	0.74	0.35	0.44	0.48	0.62	0.89	0.66	0.73	0.64	0.68	0.75	0.30
Age 75-79	0.76	0.36	0.45	0.49	0.57	0.88	0.64	0.70	0.62	0.67	0.73	0.28
Age 80-99	0.73	0.28	0.35	0.41	0.55	0.89	0.71	0.68	0.58	0.64	0.69	0.22

(d) Random Oversampling

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.70	0.36	0.43	0.48	0.77	0.93	0.72	0.85	0.76	0.74	0.82	0.37
Male	0.72	0.36	0.45	0.48	0.74	0.93	0.70	0.82	0.74	0.73	0.81	0.36
Female	0.68	0.37	0.44	0.48	0.80	0.94	0.73	0.87	0.79	0.74	0.83	0.38
White	0.67	0.35	0.42	0.46	0.79	0.93	0.73	0.86	0.77	0.73	0.82	0.36
Black	0.83	0.39	0.52	0.53	0.61	0.92	0.62	0.74	0.66	0.72	0.80	0.37
American Inc	0.84	0.43	0.52	0.57	0.60	0.91	0.60	0.72	0.66	0.72	0.77	0.39
Asian	0.67	0.38	0.42	0.48	0.82	0.94	0.74	0.87	0.80	0.74	0.84	0.39
Native Hawai	0.82	0.49	0.71	0.61	0.67	0.91	0.55	0.77	0.71	0.75	0.86	0.44
Other race	0.83	0.32	0.60	0.47	0.67	0.95	0.69	0.79	0.69	0.75	0.86	0.37
Multiracial	0.72	0.40	0.51	0.51	0.78	0.93	0.69	0.85	0.77	0.75	0.84	0.41
Hispanic	0.74	0.41	0.51	0.53	0.76	0.93	0.66	0.83	0.75	0.75	0.83	0.41
Age 30-34	0.25	0.18	0.17	0.21	0.97	0.98	0.94	0.97	0.95	0.61	0.84	0.18
Age 35-39	0.40	0.29	0.28	0.34	0.95	0.97	0.89	0.96	0.93	0.68	0.87	0.30
Age 40-44	0.43	0.28	0.30	0.34	0.93	0.96	0.86	0.94	0.89	0.68	0.84	0.29
Age 45-49	0.55	0.37	0.39	0.45	0.90	0.95	0.79	0.92	0.86	0.72	0.84	0.38
Age 50-54	0.58	0.37	0.42	0.45	0.84	0.93	0.74	0.88	0.81	0.71	0.82	0.35
Age 55-59	0.69	0.40	0.47	0.51	0.79	0.93	0.69	0.85	0.77	0.74	0.83	0.39
Age 60-64	0.71	0.39	0.50	0.50	0.74	0.92	0.67	0.82	0.73	0.73	0.81	0.37
Age 65-69	0.73	0.38	0.48	0.50	0.70	0.91	0.66	0.79	0.71	0.72	0.79	0.36
Age 70-74	0.77	0.37	0.48	0.50	0.64	0.91	0.65	0.75	0.67	0.70	0.77	0.34
Age 75-79	0.80	0.37	0.49	0.51	0.57	0.90	0.63	0.69	0.62	0.68	0.76	0.32
Age 80-99	0.79	0.28	0.37	0.41	0.50	0.91	0.70	0.65	0.56	0.65	0.71	0.23

(b) Edited Nearest Neighbour (ENN)

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.72	0.35	0.43	0.47	0.76	0.94	0.72	0.84	0.75	0.74	0.82	0.37
Male	0.73	0.35	0.45	0.47	0.73	0.93	0.70	0.82	0.73	0.73	0.81	0.36
Female	0.70	0.36	0.44	0.47	0.79	0.94	0.73	0.86	0.78	0.74	0.83	0.38
White	0.69	0.34	0.42	0.45	0.78	0.94	0.73	0.85	0.76	0.73	0.82	0.36
Black	0.85	0.38	0.52	0.53	0.58	0.93	0.62	0.71	0.64	0.72	0.80	0.37
American Inc	0.82	0.43	0.51	0.56	0.61	0.90	0.60	0.73	0.66	0.71	0.77	0.37
Asian	0.65	0.36	0.42	0.47	0.82	0.93	0.73	0.87	0.79	0.73	0.84	0.37
Native Hawai	0.80	0.49	0.72	0.61	0.68	0.90	0.55	0.77	0.71	0.74	0.86	0.43
Other race	0.85	0.33	0.60	0.47	0.67	0.96	0.69	0.79	0.70	0.76	0.86	0.39
Multiracial	0.75	0.40	0.50	0.53	0.78	0.94	0.69	0.85	0.77	0.76	0.84	0.43
Hispanic	0.76	0.41	0.50	0.53	0.74	0.93	0.66	0.83	0.75	0.75	0.83	0.41
Age 30-34	0.32	0.19	0.17	0.24	0.97	0.98	0.94	0.97	0.95	0.64	0.84	0.22
Age 35-39	0.43	0.27	0.28	0.33	0.95	0.97	0.89	0.96	0.92	0.69	0.87	0.30
Age 40-44	0.46	0.27	0.29	0.34	0.91	0.96	0.86	0.94	0.88	0.69	0.84	0.29
Age 45-49	0.57	0.35	0.40	0.43	0.88	0.95	0.79	0.91	0.85	0.72	0.84	0.36
Age 50-54	0.59	0.36	0.42	0.45	0.83	0.93	0.74	0.88	0.80	0.71	0.82	0.35
Age 55-59	0.70	0.39	0.47	0.50	0.78	0.93	0.69	0.85	0.76	0.74	0.83	0.39
Age 60-64	0.73	0.38	0.49	0.50	0.72	0.92	0.67	0.81	0.73	0.73	0.80	0.37
Age 65-69	0.74	0.37	0.48	0.50	0.69	0.92	0.66	0.78	0.70	0.72	0.79	0.35
Age 70-74	0.79	0.37	0.48	0.50	0.62	0.91	0.65	0.74	0.66	0.70	0.77	0.34
Age 75-79	0.82	0.37	0.48	0.51	0.55	0.90	0.63	0.69	0.62	0.68	0.76	0.32
Age 80-99	0.79	0.28	0.37	0.41	0.49	0.91	0.70	0.64	0.55	0.64	0.71	0.23

(e) SMOTE

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.64	0.35	0.40	0.45	0.78	0.92	0.73	0.85	0.76	0.71	0.80	0.34
Male	0.71	0.33	0.42	0.45	0.71	0.93	0.71	0.80	0.71	0.71	0.79	0.33
Female	0.56	0.38	0.40	0.45	0.85	0.92	0.73	0.88	0.81	0.70	0.81	0.35
White	0.67	0.33	0.40	0.44	0.77	0.93	0.74	0.85	0.76	0.72	0.80	0.34
Black	0.66	0.42	0.49	0.51	0.73	0.88	0.63	0.79	0.71	0.69	0.78	0.34
American Inc	0.51	0.48	0.49	0.50	0.80	0.82	0.61	0.81	0.73	0.66	0.75	0.31
Asian	0.50	0.44	0.39	0.46	0.90	0.92	0.74	0.91	0.84	0.70	0.82	0.37
Native Hawai	0.44	0.69	0.66	0.54	0.92	0.81	0.56	0.87	0.79	0.68	0.84	0.43
Other race	0.56	0.48	0.55	0.52	0.88	0.91	0.69	0.90	0.83	0.72	0.87	0.42
Multiracial	0.43	0.54	0.53	0.48	0.93	0.89	0.69	0.91	0.84	0.68	0.83	0.39
Hispanic	0.51	0.47	0.48	0.49	0.86	0.88	0.67	0.87	0.80	0.69	0.81	0.36
Age 30-34	0.19	0.16	0.15	0.17	0.97	0.98	0.94	0.98	0.95	0.58	0.83	0.15
Age 35-39	0.33	0.23	0.19	0.27	0.95	0.97	0.90	0.96	0.92	0.64	0.82	0.24
Age 40-44	0.36	0.24	0.23	0.29	0.92	0.95	0.87	0.94	0.89	0.64	0.81	0.24
Age 45-49	0.41	0.32	0.33	0.36	0.91	0.93	0.80	0.92	0.85	0.66	0.80	0.28
Age 50-54	0.48	0.35	0.37	0.41	0.86	0.91	0.75	0.89	0.81	0.67	0.79	0.30
Age 55-59	0.60	0.40	0.44	0.48	0.82	0.91	0.70	0.86	0.78	0.71	0.80	0.36
Age 60-64	0.65	0.39	0.44	0.49	0.76	0.90	0.68	0.83	0.74	0.71	0.78	0.35
Age 65-69	0.70	0.36	0.43	0.48	0.69	0.90	0.68	0.78	0.69	0.70	0.77	0.32
Age 70-74	0.73	0.35	0.45	0.48	0.63	0.89	0.65	0.74	0.65	0.68	0.75	0.30
Age 75-79	0.75	0.36	0.46	0.49	0.58	0.88	0.63	0.70	0.62	0.67	0.74	0.29
Age 80-99	0.74	0.29	0.35	0.42	0.56	0.90	0.71	0.69	0.59	0.65	0.70	0.23

(c) Near Miss

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.45	0.26	0.26	0.33	0.77	0.88	0.78	0.82	0.72	0.61	0.68	0.18
Male	0.39	0.27	0.27	0.32	0.79	0.87	0.76	0.83	0.73	0.59	0.67	0.16
Female	0.51	0.25	0.27	0.34	0.75	0.90	0.78	0.82	0.71	0.63	0.70	0.20
White	0.36	0.26	0.25	0.31	0.83	0.89	0.79	0.86	0.76	0.60	0.69	0.17
Black	0.66	0.31	0.35	0.42	0.55	0.84	0.69	0.66	0.57	0.61	0.65	0.18
American Inc	0.80	0.30	0.33	0.44	0.33	0.82	0.69	0.48	0.46	0.57	0.61	0.13
Asian	0.39	0.23	0.23	0.29	0.79	0.89	0.81	0.84	0.74	0.59	0.63	0.15
Native Hawai	0.69	0.34	0.43	0.46	0.49	0.81	0.65	0.61	0.55	0.59	0.62	0.16
Other race	0.71	0.21	0.35	0.33	0.50	0.90	0.75	0.64	0.53	0.60	0.69	0.15
Multiracial	0.67	0.21	0.25	0.32	0.49	0.88	0.77	0.63	0.52	0.58	0.63	0.12
Hispanic	0.73	0.24	0.29	0.36	0.45	0.88	0.74	0.59	0.50	0.59	0.65	0.14
Age 30-34	0.39	0.03	0.04	0.06	0.67	0.98	0.96	0.80	0.67	0.53	0.61	0.02
Age 35-39	0.52	0.09	0.15	0.15	0.75	0.97	0.92	0.85	0.74	0.64	0.72	0.13
Age 40-44	0.46	0.12	0.15	0.19	0.77	0.95	0.89	0.85	0.75	0.61	0.70	0.13
Age 45-49	0.53	0.20	0.23	0.29	0.76	0.94	0.83	0.84	0.74	0.65	0.73	0.20
Age 50-54	0.53	0.26	0.28	0.35	0.77	0.91	0.78	0.83	0.73	0.65	0.70	0.23
Age 55-59	0.53	0.30	0.31	0.39	0.75	0.89	0.74	0.81	0.71	0.64	0.71	0.23
Age 60-64	0.50	0.33	0.34	0.40	0.76	0.87	0.72	0.81	0.71	0.63	0.70	0.23
Age 65-69	0.42	0.36	0.35	0.39	0.81	0.85	0.70	0.83	0.73	0.62	0.70	0.22
Age 70-74	0.34	0.34	0.34	0.34	0.82	0.82	0.70	0.82	0.71	0.58	0.67	0.16
Age 75-79	0.38	0.38	0.38	0.38	0.80	0.80	0.67	0.80	0.70	0.59	0.67	0.19
Age 80-99	0.41	0.31	0.30	0.35	0.77	0.84	0.73	0.81	0.70	0.59	0.66	0.17

(f) Tomek-Links

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.69	0.36	0.44	0.48	0.78	0.93	0.72	0.85	0.76	0.73	0.82	0.37
Male	0.71	0.36	0.45	0.48	0.74	0.93	0.70	0.83	0.74	0.73	0.81	0.36
Female	0.67	0.37	0.44	0.48	0.81	0.94	0.73	0.87	0.79	0.74	0.83	0.38
White	0.66	0.35	0.42	0.46	0.79	0.93	0.73	0.86	0.77	0.73	0.82	0.36
Black	0.83	0.39	0.52	0.53	0.61	0.92	0.62	0.74	0.66	0.72	0.80	0.37
American Inc	0.82	0.43	0.52	0.56	0.61	0.90	0.60	0.73	0.67	0.72	0.77	0.38
Asian	0.66	0.38	0.42	0.48	0.82	0.94	0.74	0.87	0.80	0.74	0.84	0.38
Native Hawai	0.82	0.49	0.71	0.62	0.68	0.91	0.55	0.78	0.72	0.75	0.86	0.45
Other race	0.83	0.35	0.61	0.49	0.71	0.96	0.69	0.81	0.73	0.77	0.86	0.41
Multiracial	0.71	0.40	0.50	0.51	0.78	0.93	0.69	0.85	0.77	0.75	0.84	0.40
Hispanic	0.73	0.41	0.51	0.53	0.76	0.92	0.66	0.83	0.75	0.74	0.83	0.40
Age 30-34	0.28	0.18	0.17	0.22	0.97	0.98	0.94	0.97	0.95	0.62	0.85	0.20
Age 35-39	0.44	0.28	0.28	0.34	0.95	0.97	0.89	0.96	0.92	0.69	0.87	0.32
Age 40-44	0.45	0.27	0.29	0.34	0.92	0.96	0.86	0.94	0.89	0.68	0.84	0.29
Age 45-49	0.55	0.35	0.40	0.43	0.89	0.95	0.79	0.92	0.85	0.72	0.84	0.36
Age 50-54	0.58	0.36	0.42	0.45	0.84	0.93	0.74	0.88	0.80	0.71	0.82	0.35
Age 55-59	0.68	0.40	0.47	0.50	0.79	0.92	0.69	0.85	0.77	0.74	0.83	0.39
Age 60-64	0.71	0.39	0.50	0.50	0.75	0.92	0.67	0.82	0.74	0.73	0.81	0.37
Age 65-69	0.72	0.38	0.48	0.50	0.71	0.91	0.66	0.80	0.71	0.72	0.79	0.36
Age 70-74	0.76	0.38	0.48	0.50	0.65	0.91	0.65	0.76	0.67	0.70	0.77	0.34
Age 75-79	0.79	0.38	0.49	0.51	0.58	0.90	0.63	0.71	0.63	0.69	0.76	0.32
Age 80-99	0.76	0.28	0.37	0.41	0.53	0.90	0.70	0.66	0.57	0.64	0.71	0.23

Figure B.3: Performance of the logistic regression model trained on resampled training set.

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.63	0.35	0.41	0.45	0.79	0.92	0.73	0.85	0.76	0.71	0.80	0.34
Male	0.64	0.34	0.41	0.45	0.76	0.91	0.71	0.83	0.74	0.70	0.77	0.32
Female	0.63	0.36	0.43	0.46	0.82	0.93	0.73	0.87	0.79	0.72	0.82	0.36
White	0.59	0.35	0.40	0.44	0.82	0.92	0.74	0.87	0.79	0.70	0.80	0.33
Black	0.76	0.38	0.49	0.50	0.62	0.90	0.63	0.73	0.65	0.69	0.76	0.32
American Indian	0.83	0.42	0.52	0.56	0.59	0.91	0.60	0.72	0.66	0.71	0.76	0.37
Asian	0.72	0.31	0.39	0.44	0.74	0.94	0.75	0.83	0.74	0.73	0.79	0.34
Native Hawaiian	0.76	0.45	0.64	0.56	0.64	0.87	0.57	0.74	0.67	0.70	0.80	0.36
Other race	0.79	0.33	0.53	0.47	0.70	0.95	0.70	0.80	0.71	0.75	0.84	0.37
Multiracial	0.68	0.32	0.42	0.44	0.71	0.92	0.71	0.80	0.70	0.69	0.78	0.30
Hispanic	0.76	0.34	0.45	0.47	0.66	0.92	0.68	0.77	0.68	0.71	0.79	0.33
Age 30-34	0.75	0.08	0.16	0.15	0.78	0.99	0.94	0.88	0.78	0.77	0.84	0.20
Age 35-39	0.77	0.16	0.29	0.27	0.81	0.99	0.89	0.89	0.81	0.79	0.87	0.29
Age 40-44	0.69	0.20	0.30	0.31	0.81	0.97	0.86	0.88	0.80	0.75	0.84	0.29
Age 45-49	0.68	0.29	0.39	0.40	0.81	0.96	0.79	0.88	0.80	0.74	0.84	0.35
Age 50-54	0.65	0.34	0.43	0.45	0.80	0.94	0.74	0.86	0.78	0.73	0.82	0.35
Age 55-59	0.68	0.41	0.47	0.51	0.80	0.93	0.69	0.86	0.78	0.74	0.83	0.40
Age 60-64	0.63	0.40	0.49	0.49	0.79	0.90	0.67	0.84	0.76	0.71	0.80	0.36
Age 65-69	0.64	0.41	0.47	0.50	0.77	0.89	0.67	0.83	0.74	0.70	0.79	0.35
Age 70-74	0.62	0.42	0.48	0.50	0.76	0.88	0.65	0.81	0.73	0.69	0.77	0.34
Age 75-79	0.61	0.45	0.49	0.52	0.77	0.86	0.63	0.81	0.73	0.69	0.76	0.34
Age 80-99	0.52	0.35	0.37	0.42	0.76	0.87	0.70	0.81	0.71	0.64	0.71	0.25

Figure B.4: Performance of the enhanced-DP logistic regression model optimized for age group 30-34 years patient group.

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.63	0.34	0.39	0.44	0.78	0.92	0.73	0.84	0.75	0.70	0.78	0.32
Male	0.62	0.34	0.39	0.44	0.75	0.91	0.72	0.82	0.73	0.69	0.76	0.30
Female	0.64	0.34	0.41	0.45	0.79	0.93	0.73	0.86	0.77	0.72	0.81	0.34
White	0.58	0.34	0.38	0.43	0.81	0.92	0.74	0.86	0.78	0.70	0.79	0.32
Black	0.78	0.35	0.47	0.49	0.57	0.90	0.64	0.70	0.62	0.68	0.75	0.30
American Indian	0.82	0.39	0.51	0.53	0.54	0.90	0.60	0.68	0.62	0.68	0.76	0.33
Asian	0.63	0.30	0.37	0.40	0.75	0.93	0.75	0.83	0.74	0.69	0.76	0.29
Native Hawaiian	0.78	0.46	0.56	0.58	0.65	0.89	0.58	0.75	0.69	0.72	0.79	0.39
Other race	0.71	0.31	0.52	0.43	0.70	0.93	0.70	0.80	0.70	0.71	0.83	0.31
Multiracial	0.68	0.31	0.38	0.42	0.69	0.91	0.72	0.79	0.69	0.69	0.76	0.29
Hispanic	0.75	0.33	0.42	0.46	0.65	0.92	0.69	0.76	0.67	0.70	0.77	0.31
Age 30-34	0.84	0.07	0.16	0.14	0.72	0.99	0.94	0.83	0.72	0.78	0.84	0.19
Age 35-39	0.84	0.14	0.30	0.25	0.77	0.99	0.89	0.86	0.77	0.80	0.87	0.29
Age 40-44	0.73	0.17	0.29	0.28	0.76	0.98	0.86	0.86	0.76	0.74	0.84	0.27
Age 45-49	0.72	0.27	0.40	0.40	0.79	0.96	0.79	0.86	0.78	0.75	0.84	0.35
Age 50-54	0.67	0.33	0.43	0.44	0.78	0.94	0.74	0.85	0.77	0.72	0.82	0.34
Age 55-59	0.70	0.40	0.47	0.51	0.79	0.93	0.69	0.85	0.77	0.75	0.83	0.40
Age 60-64	0.63	0.40	0.49	0.49	0.78	0.90	0.67	0.83	0.75	0.70	0.80	0.35
Age 65-69	0.63	0.41	0.47	0.50	0.77	0.89	0.67	0.83	0.75	0.70	0.78	0.35
Age 70-74	0.60	0.43	0.48	0.50	0.77	0.87	0.65	0.82	0.74	0.69	0.77	0.34
Age 75-79	0.55	0.46	0.48	0.50	0.79	0.85	0.63	0.82	0.74	0.67	0.75	0.33
Age 80-99	0.46	0.36	0.36	0.40	0.80	0.86	0.70	0.83	0.73	0.63	0.71	0.24

Figure B.5: Performance of the enhanced-DP logistic regression model optimized for age group 35-39 years patient group.

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.60	0.37	0.41	0.46	0.81	0.92	0.73	0.86	0.78	0.71	0.80	0.34
Male	0.60	0.36	0.41	0.45	0.79	0.91	0.71	0.85	0.76	0.69	0.78	0.33
Female	0.60	0.38	0.43	0.46	0.83	0.92	0.73	0.88	0.80	0.71	0.82	0.36
White	0.55	0.37	0.40	0.44	0.84	0.92	0.74	0.88	0.80	0.70	0.80	0.33
Black	0.76	0.38	0.49	0.51	0.63	0.90	0.63	0.74	0.66	0.69	0.76	0.33
American Indian	0.84	0.42	0.52	0.56	0.59	0.91	0.60	0.71	0.65	0.72	0.77	0.38
Asian	0.47	0.39	0.39	0.42	0.88	0.91	0.75	0.89	0.82	0.67	0.79	0.32
Native Hawaiian	0.84	0.46	0.61	0.59	0.62	0.91	0.57	0.74	0.68	0.73	0.81	0.41
Other race	0.71	0.36	0.55	0.48	0.76	0.93	0.70	0.84	0.75	0.73	0.84	0.37
Multiracial	0.65	0.36	0.42	0.46	0.76	0.92	0.71	0.83	0.75	0.71	0.79	0.34
Hispanic	0.74	0.36	0.46	0.48	0.70	0.92	0.68	0.79	0.70	0.72	0.79	0.35
Age 30-34	0.68	0.10	0.17	0.17	0.83	0.99	0.94	0.91	0.83	0.76	0.84	0.21
Age 35-39	0.73	0.18	0.28	0.29	0.84	0.99	0.89	0.91	0.84	0.79	0.87	0.31
Age 40-44	0.69	0.22	0.29	0.33	0.83	0.97	0.86	0.90	0.82	0.76	0.84	0.32
Age 45-49	0.69	0.31	0.40	0.42	0.82	0.96	0.79	0.89	0.81	0.76	0.84	0.37
Age 50-54	0.64	0.35	0.43	0.45	0.81	0.93	0.74	0.87	0.78	0.72	0.82	0.35
Age 55-59	0.67	0.41	0.47	0.51	0.80	0.92	0.69	0.86	0.78	0.74	0.83	0.40
Age 60-64	0.63	0.42	0.50	0.51	0.80	0.90	0.67	0.85	0.77	0.72	0.80	0.37
Age 65-69	0.60	0.43	0.48	0.50	0.80	0.89	0.66	0.84	0.76	0.70	0.79	0.36
Age 70-74	0.57	0.44	0.48	0.49	0.80	0.87	0.65	0.83	0.75	0.68	0.77	0.33
Age 75-79	0.54	0.47	0.48	0.50	0.81	0.85	0.63	0.83	0.74	0.67	0.75	0.33
Age 80-99	0.42	0.36	0.37	0.39	0.81	0.85	0.70	0.83	0.74	0.62	0.71	0.22

Figure B.6: Performance of the enhanced-DP logistic regression model optimized for age group 40-44 years patient group.