Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prospective Multicenter Validation of Machine Learning Models for Mortality Prediction in Adult Critically Ill Patients using Transfer Learning

Ioannis Papapanagiotou
IB Hellas
Charikleia S. Vrettou
National & Kapodistrian University of Athens, Evangelismos Hospital
Maria Theodorakopoulou
General Hospital Attiki KAT
Zafiria Mastora
General Hospital of Thoracic Diseases "Sotiria"
Vassiliki Giannopoulou
National & Kapodistrian University of Athens, Evangelismos Hospital
Olga Kampouropoulou
National & Kapodistrian University of Athens, Evangelismos Hospital
Apostolos Karalis
IB Hellas
Maria Pratikaki
Evangelismos Hospital
Spyretta Golemati
National & Kapodistrian University of Athens, Evangelismos Hospital
Georgios Poupouzas
National & Kapodistrian University of Athens, Evangelismos Hospital
Vasileios Issaris
National & Kapodistrian University of Athens, Evangelismos Hospital
Kyriakos Karkoulias
General Hospital Attiki KAT
Sofia Pouriki
General Hospital of Thoracic Diseases "Sotiria"
Chrysi Keskinidou
National & Kapodistrian University of Athens, Evangelismos Hospital
Nikolaos S. Lotsios
National & Kapodistrian University of Athens, Evangelismos Hospital
Anastasia Kotanidou

National & Kapodistrian University of Athens, Evangelismos Hospital

Alice G. Vassiliou

alvass75@gmail.com

National & Kapodistrian University of Athens, Evangelismos Hospital

Stavros Papapanagiotou

IB Hellas

Ioanna Dimopoulou

National & Kapodistrian University of Athens, Evangelismos Hospital

---

Article

Keywords: machine learning, mortality prediction, intensive care unit, adults, domain adaptation, inductive transfer learning, prospective study

Posted Date: February 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8872055/v1

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

Additional Declarations: No competing interests reported.

Abstract

Mortality prediction in critically ill patients remains challenging due to poor cross-institutional performance and limited generalizability of machine learning models. This study addresses this, by systematically benchmarking and prospectively validating transfer learning frameworks. We trained our models on MIMIC-IV and validated them on a multicenter prospective cohort of 539 patients from three hospitals. We compared tree-based methods and modern deep learning architectures for tabular data. Results demonstrated that both Domain Adaptation (DA) and Inductive Transfer Learning (ITL) significantly enhanced model performance under realistic conditions where target-domain data are limited. DA consistently improved discrimination across all evaluated models, with LightGBM showing the most significant gains in Area Under the Receiver Operating Characteristic Curve (AUC) ($p = 0.0010$), and XGBoost yielding the largest improvements in Area Under the Precision-Recall Curve (AUPRC) ($p = 0.0419$). Among all evaluated models, Random Forest (RF) achieved the highest discriminative performance, achieving 90.7% AUC with DA and 81.3% AUPRC with ITL. Notably, the domain-adapted models significantly outperformed APACHE II ($p = 0.0044$) and SOFA ($p = 0.0077$). These findings suggest that transfer learning provides a robust and data-efficient pathway for improving model generalizability across heterogeneous populations, offering a pragmatic solution to the challenge of model degradation in clinical deployment.

1. Introduction

Accurate and timely prediction of mortality in critically ill adults remains a central challenge in modern intensive care medicine. Despite decades of re- search and the widespread adoption of scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE) II [1], the Simplified Acute Physiology Score (SAPS) [2], and the Sequential Organ Failure Assessment (SOFA) [3] score, these tools often exhibit limited discrimination across institutions and patient populations due to their constraints as linear models [4]. Recently, machine learning (ML)-based models have shown promise in capturing complex, nonlinear relationships in tabular Intensive Care Unit (ICU) data often outperforming traditional scoring systems [5, 6, 7].

A wide range of ML approaches have been applied to the Intensive Care Unit (ICU) mortality prediction from admission data, with tree-based ensemble methods frequently outperforming linear models and deep learning architectures in statistical metrics. In their real-time mortality prediction study on the Medical Information Mart for Intensive Care (MIMIC)-III [8] data, Johnson et al. showed that gradient boosting outperformed traditional severity scores [9]. Similarly, Pang et al. reported that Extreme Gradient Boosting (XGBoost) outperformed Logistic Regression (LR) and Support Vector Machines (SVMs) trained on MIMIC-IV [10] data [11]. Large institutional datasets beyond MIMIC have also been leveraged; Min Hyuk Choi et al. analyzed ICU admissions from two major South Korean hospitals, finding that Light Gradient-Boosting Machine (LGBM) and XGBoost yielded the best predictive performance [12]. Random Forests (RF) have also shown promising results in a number of studies. Hu et al. evaluated nine (9) ML models on MIMIC-IV and identified RF as the best performer [5]. Kim et al. applied ML methods to

nursing records in MIMIC-IV and again found RF superior to LR, SVMs, and Naïve Bayes [13], while Alghatani et al. likewise demonstrated that RF outperformed XGBoost and linear models when trained on MIMIC-III [14]. Variability in the best-performing models across MIMIC-III and MIMIC-IV studies likely reflects differences in cohort selection, feature selection, and evaluation setups, rather than fundamental inconsistencies in model capabilities, where tree-based methods generally outperform other methods.

In addition to tree-based classical machine learning, deep learning architectures have also been increasingly used in recent years. However, standard neural networks often struggle with heterogeneous tabular data due to mixed feature types, missing values, and the lack of inherent spatial or sequential structure. To combat this, architectures such as Self-Attention and Intersample Attention Transformer (SAINT) [15] and Neural Oblivious Decision Ensembles (NODE) [16] have been proposed. Studies on SAINT and NODE have shown competitive performance against traditional boosting models, often outperforming them, indicating they could be promising architectures for tasks such as ICU mortality prediction, where capturing subtle relationships among vital signs, laboratory data, and comorbidities is crucial [17, 18].

Although ML approaches have shown promising results, a major barrier to their clinical deployment is their susceptibility to performance degradation when applied prospectively or across heterogeneous healthcare settings. To overcome these limitations, methodological frameworks such as Domain Adaptation (DA) and Inductive Transfer Learning (ITL) aim to improve cross-site robustness by leveraging shared structure across diverse clinical environments. Mutnuri et al. applied DA and ITL by fine-tuning the last hidden layer of their neural networks [19]. This allowed the model to adapt to population-specific patterns while maintaining the structure learned from the larger internal dataset. The resulting model achieved markedly better external discrimination and calibration compared to the baseline model trained solely on internal data. DA has also shown very promising results in dynamic temporal data using convolutional and recurrent layers [20] and also in cross-institutional ICU outcome prediction, as demonstrated by Zhu et al., who pretrained a gradient boosting algorithm on adult MIMIC-II data and fine-tuned on a pediatric cohort, significantly improving ICU-mortality prediction on the target pediatric population [21]. Despite these advances, several important gaps remain unaddressed in the literature. To the best of our knowledge, no ICU mortality prediction model using DA or ITL has yet undergone prospective, real-time validation in adult ICUs, when restricted to the first 24 hours of admission data. The first 24 hours of ICU admission constitute a critical period for early risk assessment, as decisions made during this window can substantially influence patient management and outcomes.

Many existing mortality prediction models leverage longitudinal data collected throughout the ICU stay which reduces their utility [19, 22]. Although DA and ITL have shown promise, existing ICU outcome studies have focused predominantly on neural-network–based architectures [19, 20, 22]. Apart from Zhu et al., who combined DA with gradient boosting [21] using temporal data, we found no systematic benchmarking of DA/ITL frameworks using modern tree-based boosting models such as XGBoost or LightGBM, despite their strong performance in ICU mortality prediction. SAINT and NODE-based

architectures have also recently been explored in similar domains [23, 24, 25, 26, 27], but lack evaluation within DA or ITL cross-domain frameworks or prospective studies.

As such, the contributions of this work are as follows:

- We provide a prospectively validated study of ICU mortality prediction models using DA and ITL in adult ICU populations, focusing specifically on the clinically critical first 24 hours of admission data.
- We benchmark DA and ITL using tree-based models as well as SAINT and NODE for adult ICU mortality prediction, which have not been investigated in the context of 24-hour ICU admission data.

2. Methods

2.1. Data sources

The publicly available MIMIC-IV v3.1 [10], which contains clinical and physiological measurements from a heterogeneous cohort of ICU patients, was used to train our ML algorithms. MIMIC-IV, which contains a total of 94,458 unique adult ICU stays, was used for model training and validation.

The prospective data were collected from the ICUs of “Evangelismos” General Hospital, General Hospital of Thoracic Diseases “Sotiria”, and “KAT” Attica General Hospital. The Ethics Committees of the three Hospitals approved the study (study approval numbers 14–17/01/2025; 27 – 5/02/2025 & 8260-26/03/2025). Informed consent was obtained from all patients’ next-of-kin before inclusion, and all procedures carried out complied with the Helsinki Declaration.

The prospective data collected from the 3 ICUs yielded a combined cohort of 539 adult patients admitted to the ICU. Of these, 141 patients died during hospitalization. Harmonized data collection protocols ensured consistent variable definitions across centers. Demographic characteristics of the MIMIC-IV and the prospective data are shown in Table 1.

Table 1: Comparison of demographic characteristics between MIMIC-IV and prospective data.

Characteristic	MIMIC-IV	Prospective
ICU admissions	94458	539
ICU mortality (%)	7329 (7.76%)	141 (26.16%)
Median age	66	63
Male (%)	56.0%	63.5%

2.2.Data preprocessing and feature alignment

All clinical and laboratory measurements were obtained within the first 24 hours of ICU admission. To ensure compatibility with our feature space, we selected a subset of 44 features from both datasets. Collected variables included demographics (age, sex), vital signs and physiologic measures

(temperature, heart rate, respiratory rate, oxygen saturation, Glasgow coma scale, lactate), and an extensive panel of laboratory values including hematologic indices (white blood cell count, hematocrit, platelet count), renal function tests (blood urea nitrogen, creatinine), metabolic and electrolyte markers (glucose, sodium, potassium), and liver function markers (total bilirubin, albumin). The complete variable lists for the MIMIC-IV and our prospective data are provided in the supplementary material Tables S2 and S3.

2.3. Model development and validation

We trained supervised learning tree-based models (RF, XGBoost, LightGBM, CatBoost) as well as deep learning models (NODE, SAINT). We selected these architectures since the literature had previously identified them among the most promising approaches for ICU mortality prediction. The MIMIC- IV (source) dataset was first split into 80/20, with 80% used for model training and the remaining 20% held out as a test set. We performed a five-fold cross-validation within the training portion to avoid overfitting and ensure model robustness. Model evaluation was conducted primarily using AUC and AUPRC, given the class imbalance in ICU mortality. We have also reported the Brier score, sensitivity, and specificity. Hyperparameters used in all ML experiments were optimized with Optuna, a Bayesian optimizer [28].

In addition to the transfer learning experiments, we trained models exclusively on the prospective (target) dataset. These locally trained models served as comparators to assess the added value of transfer learning beyond what could be achieved using institution-specific data alone. To ensure a fair comparison, identical feature sets, preprocessing steps, performance metrics, and hyperparameter optimization procedures were applied.

2.4. ITL implementation

In the ITL setting, models were trained on the internal MIMIC-IV cohort (source domain) and were further fine-tuned using the outcome labels from the prospective data (target domain). For boosting models, ITL was performed by continuing gradient boosting from the internally trained model using labeled external data. Specifically, the boosting model trained on MIMIC-IV was first fitted on the internal training data to obtain a base booster. This booster was then updated by training additional trees on the external transfer learning subset. To control the extent of adaptation and mitigate overfitting to the small external sample, only a limited number of new trees were added, and the learning rate for these trees was reduced relative to the original model. For each external split, the ITL procedure introduced three tunable hyperparameters: *(i)* the learning-rate scaling factor κ, *(ii)* the number of additional boosting rounds, and *(iii)* the class imbalance correction term (*scale_pos_weight*) applied during fine-tuning. These parameters were optimized using Optuna. Additionally, Platt scaling was applied as a post-hoc calibration step.

Random forests were adapted using a model-based ensemble expansion strategy enabled by warm-start training. The source domain was reinitialized with warm-start enabled, and additional decision trees were appended to the existing ensemble. These newly added trees were trained solely on labeled target-domain data, while the original source-domain trees remained fixed. The tunable hyperparameters were *(i)* number of trees, *(ii)* maximum tree depth, and *(iii)* minimum leaf size.

In the NODE architecture, ITL was performed by initializing a classifier with the internal base weights learned on the original internal cohort, and then fine-tuning a restricted subset of parameters on the external cohort. A new NODE instance (stacked Oblivious Decision Trees-ODST layers with additive per-layer readouts) was loaded with the stored base checkpoint; all parameters were frozen, and we unfroze only *(i)* the readout layers, *(ii)* the global logit bias term, and *(iii)* the final ODST layer. The final fine-tuned model was then trained using AdamW and weighted binary cross-entropy with logits.

Similarly to tree-based methods, SAINT was adapted via parameter- restricted fine-tuning starting from internal base weights learned on the original internal cohort and then fine-tuning a restricted subset of parameters

on the external cohort. A new SAINT instance (feature-wise tokenization via learnable per-feature affine embeddings followed by multi-head self-attention blocks and mean pooling) was then loaded with the stored base checkpoint; all parameters were frozen, and we unfroze *(i)* the classification head and *(ii)* the final transformer block.

2.5. DA implementation

For DA with all tree-based models as well as NODE, we estimated each internal training instance by fitting a domain classifier that discriminates between internal and external feature distributions. We constructed a pooled dataset with corresponding binary domain labels, then a logistic regression model was trained on $\mathbf{X}_{\text{domain}}$ to estimate the posterior probability

$P(d = 1 \mid \mathbf{x})$, which represents the probability that an observation **x** originates from the external cohort. Using the trained domain classifier, we

computed an importance weight for each internal training sample $\mathbf{x}_i$ as a monotonic function of the odds of originating from the external cohort. To ensure numerical stability, the probabilities p_i were clipped to the interval

$p_i \in [10^{-6}, 1 - 10^{-6}]$, and the resulting importance weights were further

constrained to lie within a bounded range. A DA model was then

trained on the same internal training data, using the computed importance weights. Thus, adaptation was achieved without using external labels for training the outcome model, as external data are only used to learn the domain shift and derive weights.

Unlike the tree-based approaches, SAINT was trained once to obtain a baseline model on the internal training set. Starting from the baseline weights, we fine-tuned an alignment model using labeled internal source batches together with unlabeled external TL batches by minimizing a joint objective:

$$L = L_{\text{task}}(X_s, y_s) + \lambda\, L_{\text{align}}(f(X_s), f(X_t)),$$

where L_{task} is binary cross-entropy with logits on internal labels, $f(\cdot)$ denotes the pooled SAINT representation, and L_{align} is a Radial Basis Function – Maximum Mean Discrepancy loss computed between source and target rep-

resentations. Alignment hyperparameters were fixed from a prior Optuna run: $\lambda = 9.7454 \times 10^{-4}$, $\sigma = 13.8262$, and freeze_epochs = 3, where the feature extractor was frozen for the first 3 epochs, and the classification head remained trainable throughout.

2.6. Performance aggregation strategy

To account for the relatively small size of the external prospective cohort ($n = 539$), both ITL and DA procedures were repeated across 5 random 70/30 splits, each controlled by a sequential independent seed. For every seed, the adapted model was trained, optimized by Optuna, and evaluated on its corresponding split. The resulting performance metrics were aggregated across all repetitions. Final reported results, therefore, reflect the mean and 95% confidence intervals computed over five sequential seeds. APACHE II and SOFA scores used to benchmark our ML models were also computed over the same seeds.

2.7. Statistical analysis

Paired hypothesis testing was conducted to compare model performance across repeated data splits, using performance metrics computed from five sequential seeds. Specifically, paired *t*-tests were used to assess whether the mean difference in performance between models differed significantly from zero, and differences were considered statistically significant when $p < 0.05$.

2.8. Model interpretation

To enhance the interpretability of our findings, we used SHapley Additive exPlanations (SHAP) [29], an interpretability tool that helps to elucidate how our models compute their predictions. We used a SHAP Kernel Explainer to derive feature attributions, followed by a SHAP Beeswarm plot to visualize the distribution of these effects and explore how individual predictors influence the model's outcome predictions.

2.9. Software

Our machine learning models were developed and trained using Python 3.12. Experiments were conducted on an Intel Core i7-14700k CPU and RTX 5600 GPU.

3. Results

3.1. Assessment of DA and ITL over baseline models

Across all evaluated models, DA consistently improved discrimination relative to baseline in 6 out of 6 models, while ITL improved discrimination in 4 out of 6. The largest AUC improvements were observed

for the LightGBM model (Fig. 1a), while the largest AUPRC gains were observed for the XGBoost model (Fig. 1b). Both models showed improvements in discrimination relative to the baseline that were statistically significant for both adaptation strategies. The domain-adapted LGBM model achieved a statistically significant performance gain in AUC (Fig. 1a) (paired t-test; $p = 0.0010$). Similarly, XGBoost achieved a statistically significant performance in AUPRC using DA (Fig. 1b) (paired t-test; $p = 0.0419$). All evaluated models are summarized in Table 2, while additional performance measures, including the Brier score, sensitivity, and specificity, are reported in the supplementary material Table S1.

3.2. Comparison with scoring systems and locally trained models

Figure 2 compares the performance of the domain-adapted LGBM (Fig. 2a) and XGBoost (Fig. 2b) models with the APACHE II and SOFA scores, as well as locally trained baselines, using AUC and AUPRC. We choose to plot LGBM and XGBoost as they demonstrated the largest gains in AUC and AUPRC, respectively. The domain-adapted models significantly outperformed APACHE II, showing higher AUC ($p = 0.0044$) and AUPRC ($p = 0.00026$). They also achieved superior discrimination relative to SOFA, with statistically significant improvements in both AUC ($p = 0.0077$) and AUPRC ($p = 0.0013$). Moreover, demonstrated significant gains over the locally trained models in AUC ($p = 0.033$) and AUPRC ($p = 0.022$).

Table 2: Performance comparison of all ML models used in this study.

RF	**0.920**	**0.602**	**0.897 [0.885, 0.910]**	**0.791 [0.763, 0.818]**
ITL			0.903 [0.893, 0.914]	**0.813 [0.788, 0.838]**
DA			**0.907 [0.895, 0.918]**	0.802 [0.773, 0.832]
XGBoost	**0.922**	**0.608**	0.892 [0.878, 0.905]	0.771 [0.747, 0.795]
ITL			0.891 [0.874, 0.907]	0.771 [0.741, 0.800]
DA			0.905 [0.895, 0.914]	0.803 [0.779, 0.827]
LightGBM	0.920	0.607	0.877 [0.863, 0.891]	0.761 [0.734, 0.789]
ITL			0.881 [0.865, 0.897]	0.768 [0.737, 0.799]
DA			0.892 [0.882, 0.902]	0.782 [0.755, 0.809]
CatBoost	0.920	0.600	0.893 [0.884, 0.903]	0.772 [0.738, 0.806]
ITL			0.896 [0.883, 0.908]	0.782 [0.757, 0.807]
DA			0.896 [0.885, 0.906]	0.776 [0.746, 0.807]
NODE	0.913	0.577	0.876 [0.854, 0.898]	0.734 [0.686, 0.783]
ITL			0.882 [0.859, 0.905]	0.750 [0.692, 0.807]
DA			0.878 [0.860, 0.896]	0.724 [0.700, 0.749]
SAINT	0.911	0.581	0.896 [0.883, 0.909]	0.793 [0.758, 0.828]
ITL			0.889 [0.877, 0.900]	0.785 [0.755, 0.814]
DA			0.897 [0.883, 0.910]	0.794 [0.760, 0.829]

Best AUROC and AUPRC values are shown in bold. Means and CIs were calculated for the prospective cohort due to its lower sample size (n = 539). *Abbreviations: DA, Domain Adaptation; ITL, Inductive Transfer Learning; RF, Random Forest.*

3.3. Model interpretability

Feature-level interpretability analyses for the domain-adapted RF model are illustrated in Figs. 3 and 4, including Mean Decrease in Impurity (MDI)- based feature importance rankings and SHAP Beeswarm plots, respectively. We chose to plot RF, as it achieved the highest AUC after domain adaptation.

4. Discussion

This study demonstrates that both DA and ITL can significantly improve the discriminatory performance of ML models for predicting mortality in ICU patients compared to baseline approaches. Across all evaluated architectures, DA consistently improved discrimination, while ITL yielded improvements in most models. These improvements were systematic and, in several cases, statistically significant.

Among the evaluated models, LightGBM exhibited the largest improvements in AUC, and XGBoost showed the most pronounced gains in AUPRC, suggesting that these models may be particularly well suited to benefit from DA or ITL in this setting. LightGBM's leaf-wise tree growth and strong handling of heterogeneous feature interactions may enable more effective refinement of decision boundaries under covariate shift, leading to larger gains in overall ranking performance.

Several studies have investigated DA in temporal and dynamic clinical data using deep learning approaches [20, 21, 22, 19]. However, these works differ substantially from our study in both scope and methodology. Prior approaches primarily rely on time-series data, focus on pediatric populations, or restrict DA to deep neural architectures, without considering tree-based or transformer-based models, and without benchmarking DA alongside ITL. Most closely related to our work, Mutnuri et al. evaluated both DA and ITL within fully connected neural networks [19]. In contrast to our setting, their source and target domains were of comparable size, their data contained temporal features, and adaptation was performed exclusively within fully contained neural networks. NODE and SAINT have been mainly assessed on synthetic and real-world benchmark datasets [23, 24], focusing on representation learning and feature interaction modeling, rather than on clinical DA or transfer learning across heterogeneous healthcare cohorts. To our knowledge, neither NODE nor SAINT has previously been systematically evaluated for DA or ITL in ICU mortality prediction, nor compared directly against classical tree-based methods under similar experimental conditions. Our study, therefore, extends prior work [25, 26, 27] by examining how modern deep tabular architectures respond to both DA and ITL in a clinically realistic, cross-cohort setting, and furthermore by benchmarking their behavior alongside established methods.

Notably, beyond comparisons with baseline machine learning models, the

domain-adapted LGBM and XGBoost models substantially outperformed

established clinical scoring systems, including APACHE II and SOFA, in both AUC and AUPRC. These findings underscore the added value of data-driven, adapted ML approaches over traditional rule-based scores. In addition, the domain-adapted models demonstrated significant improvements over their respective baselines trained on our prospective data, indicating that performance gains were not solely attributable to local retraining but rather to the explicit incorporation of information from external domains.

Interpretability analyses of the RF model showed a coherent and clinically plausible set of dominant predictors, with strong agreement between global Gini importance and local SHAP attributions. Across both representations, APACHE II, lactate, and SOFA score emerged as top predictors in most models. SHAP values further demonstrated directionally consistent effects, in which higher lactate, higher severity scores, and lower saturation could enhance the predicted mortality (Fig. 4). Importantly, DA primarily reweighted these same core clinical variables rather than introducing new drivers, suggesting that performance gains were achieved by refining the relative importance of clinically meaningful signals to better match the external cohort.

In our study, there is an imbalance between source and target sample sizes. With 94,458 samples available in the source domain and 539 in the target domain, instance-level weighting or selection could become inherently unstable, as similarity estimates and importance weights are dominated by high-dimensional noise and may be overly sensitive to a small number of target observations. In contrast, DA operates at a distributional level, enabling the model to leverage the large source dataset while explicitly adjusting for systematic differences between domains, rather than relying on sparse target instances to guide transfer. Under these conditions, DA could provide a more robust and data-efficient mechanism for aligning source and target domains, which likely explains its more consistent performance compared with ITL in our study.

The limitations of our study should be addressed. First, the target domain was substantially smaller than the source dataset, reflecting a common real-world constraint but potentially limiting model stability and the generalizability of conclusions regarding adaptation effectiveness. Second, performance estimation relied on repeating the ITL and DA procedures across five random data splits. Although this approach improves robustness compared with a single split, the limited number of repetitions, which were driven by computational constraints, may reduce the precision of estimated confidence intervals and the power of statistical comparisons. Lastly, the use of a logistic regression domain classifier may inadequately capture complex, nonlinear domain shifts typical of ICU data, potentially resulting in suboptimal importance weighting.

The strengths of our study should also be acknowledged. By leveraging the large-scale, widely recognized MIMIC-IV database for source training and validating our findings on an independent, multicenter prospective ICU cohort, we ensured a rigorous test of model generalizability in a clinically realistic setting. A key strength lies in our systematic benchmarking of modern deep tabular architectures, such as NODE and SAINT, alongside established tree-based methods. Our results highlight the utility of tree-based methods, which demonstrated superior capacity to refine decision boundaries under covariate shift, achieving the most pronounced gains in AUC and AUPRC, respectively. Furthermore, by demonstrating that domain-adapted models substantially outperform traditional clinical scoring systems like APACHE II and SOFA, this study provides strong evidence for the superiority of adaptive, data-driven approaches over static rule-based metrics.

5. Conclusions

This study demonstrates that transfer learning strategies can meaningfully improve ICU mortality prediction under realistic cross-domain conditions, when target-domain data are limited. Across multiple modeling paradigms, domain adaptation consistently improved discrimination, achieving statistically significant improvements over baseline models or traditional scoring systems such as APACHE II or SOFA scores. From a clinical perspective, these findings support DA and ITL as pragmatic pathways to improve model generalizability across heterogeneous ICU populations without requiring large volumes of local retraining data. Such approaches may help mitigate well-documented performance degradation when models are transferred between hospitals, healthcare systems, or patient cohorts. The prospective

validation performed in this study further strengthens the translational relevance of these results. Future work should extend this framework to incorporate multimodal representations to assess whether DA benefits persist in richer data settings. Broader multicenter prospective evaluations are also needed to examine clinical robustness across heterogeneous ICU environments.

Declarations

Data and code availability

The processed MIMIC-IV dataset generated and analyzed during the current study is publicly available in PhysioNet: https://physionet.org/content/mimiciv/3.1. The code to reproduce our experiments will also be available in the Github repository: https://github. com/giannis3p/TL-mortality-pred. The datasets analyzed during the current study are not publicly available due to participant confidentiality and informed consent restrictions, but are available from the corresponding author on reasonable request.

Author Contributions

AKo, AGV, SP, and ID conceived the work; IP, AKa, SG, and SP designed the work; MT, ZM, VG, OK, MP, GP, VI, KK, SP, CK, and NSL acquired the data; IP, CSV, MT, ZM, VG, OK, AKa, MP, SG, GP, VI, KK, SP, CK, NSL, AKo, AGV, SP, ID analyzed and interpreted the data; IP, MT, ZM, VG, OK, AKa, MP, GP, VI, KK, SP, CK, NSL drafted the work; CSV, SG, AKo, AGV, SP, and ID substantively revised it. All authors have approved the submitted version and have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Competing Interests

The authors declare no competing interests.

Funding

The work was implemented under the Clusters of Research Excellence (CREs), funded by the European Union, Recovery and Resilience Facility (RRF), Greece 2.0, Next Generation EU. Grant number: ΥΠ3ΤΑ-0559412.

References

1. Knaus, W. A., Draper, E. A., Wagner, D. P. & Zimmerman, J. E. APACHE II: a severity of disease classification system. *Crit. Care Med.* **13** (10), 818–829 (1985).
2. Le Gall, J. R. et al. A simplified acute physiology score for ICU patients, Crit. *Care Med.* **12** (11), 975–977 (1984).

3. Vincent, J. L. et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Med.* **22** (7), 707–710 (1996).
4. Raith, E. P. et al. D. V. Pilcher, for the Australian, N. Z. I. C. S. A. C. for Outcomes, R. E. (CORE), Prognostic accuracy of the sofa score, sirs criteria, and qsofa score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit, JAMA 317 (3) 290–300. (2017). 10.1001/jama.2016.20328
5. Hu, C., Gao, C., Li, T., Liu, C. & Peng, Z. Explainable artificial intelligence model for mortality risk prediction in the intensive care unit: a derivation and validation study. *Postgrad. Med. J.* **100** (1182), 219–227. 10.1093/postmj/qgad144 (2024).
6. Olang, O. et al. Artificial intelligence-based models for prediction of mortality in icu patients: A scoping review, Journal of Intensive Care Medicine 40 (12) 1240–1246, pMID: 39150821. (2025). 10.1177/ 08850666241277134.
7. Keuning, B. E. et al. H. consortium, Mortality prediction models in the adult critically ill: A scoping review, Acta Anaesthesiologica Scandinavica 64 (4) 424–442. (2020). https: //doi.org/10.1111/aas.13527
8. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data*. **3** (1), 160035 (2016).
9. Johnson, A. & Mark, R. G. Real-time mortality prediction in the intensive care unit. *AMIA Annu. Symp. Proc.* **2017**, 994–1003 (2018).
10. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data*. **10** (1), 1 (2023).
11. Pang, K., Li, L., Ouyang, W., Liu, X. & Tang, Y. Establishment of ICU mortality risk prediction models with machine learning algorithm using MIMIC-IV database. *Diagnostics (Basel)*. **12** (5), 1068 (2022).
12. Choi, M. H. et al. Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. *Sci. Rep.* **12** (1), 7180 (2022).
13. Kim, Y., Kim, Y. & Choi, M. Machine learning-based prediction models of mortality for intensive care unit patients using nursing records, in: Studies in Health Technology and Informatics, Studies in health technology and informatics, IOS, (2024).
14. Alghatani, K., Ammar, N., Rezgui, A. & Shaban-Nejad, A. Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation. *JMIR Med. Inf.* **9** (5), e21347. 10.2196/21347 (2021).
15. Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B. & Goldstein, T. Saint: Improved neural networks for tabular data via row atten- tion and contrastive pre-training arXiv:2106.01342. URL (2021). https://arxiv.org/abs/2106.01342
16. Popov, S., Morozov, S. & Babenko, A. Neural oblivious decision ensembles for deep learning on tabular data (2019). arXiv:1909.06312. URL https://arxiv.org/abs/1909.06312
17. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? arXiv:2207.08815. URL (2022). https://arxiv.org/abs/2207.08815

18. Chen, K. Y., Chiang, P. H., Chou, H. R., Chen, T. W. & Chang, T. H. Trompt: Towards a better deep neural network for tabular data arXiv:2305.18446. URL (2023). https://arxiv.org/abs/2305.18446
19. Mutnuri, M. K., Stelfox, H. T., Forkert, N. D. & Lee, J. Using domain adaptation and inductive transfer learning to improve patient outcome prediction in the intensive care unit: Retrospective observational study, J Med Internet Res 26 e52730. 10.2196/52730. URL (2024). https://www.jmir.org/2024/1/e52730
20. Alves, T., Laender, A., Veloso, A. & Ziviani, N. Dynamic prediction of icu mortality risk using domain adaptation, in: IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 1328–1336. (2018). 10.1109/bigdata.2018.8621927. URL http://dx.doi.org/10.1109/BigData.2018.8621927.
21. Zhu, Y. et al. Domain adaptation using convolutional autoencoder and gradient boosting for adverse events prediction in the intensive care unit, Frontiers in Artificial Intelligence Volume 5–2022 (2022). 10.3389/frai.2022.640926
22. Shickel, B. et al. Deep multi-modal transfer learning for augmented patient acuity assessment in the intelligent icu. *Front. Digit. Health Volume*. 3–2021. 10.3389/fdgth.2021.640685 (2021).
23. Hwang, Y. & Song, J. Recent deep learning methods for tabular data. *Commun. Stat. Appl. Methods*. **30**, 215–226. 10.29220/CSAM.2023.30.2.215 (2023).
24. Gardner, J., Popovic, Z. & Schmidt, L. Benchmarking distribution shift in tabular data with tableshift arXiv:2312.07577. URL (2024). https://arxiv.org/abs/2312.07577
25. Gutheil, J. & Donsa, K. SAINTENS: Self-Attention and Intersample Attention Transformer for Digital Biomarker Development Using Tabular Healthcare Real World Data, Vol. 293, (2022). 10.3233/SHTI220371
26. Heikal, M. et al. Using machine learning and electronic health records to identify neuropsychiatric risk scores for delirium in ICU and general hospital settings. *Neuropsychiatr Dis. Treat.* **20**, 1861–1876 (2024).
27. Lin, Y. T., Deng, Y. X., Tsai, C. L., Huang, C. H. & Fu, L. C. Interpretable deep learning system for identifying critical patients through the pre- diction of triage level, hospitalization, and length of stay: Prospective study, JMIR Med Inform 12 e48862. (2024). 10.2196/48862. URL https://medinform.jmir.org/2024/1/e48862.
28. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next- generation hyperparameter optimization framework (2019). arXiv: 1907.10902. URL https://arxiv.org/abs/1907.10902
29. Lundberg, S. & Lee, S. I. A unified approach to interpreting model predic- tions (2017). arXiv:1705.07874.URL. https://arxiv.org/abs/1705.07874

Figures

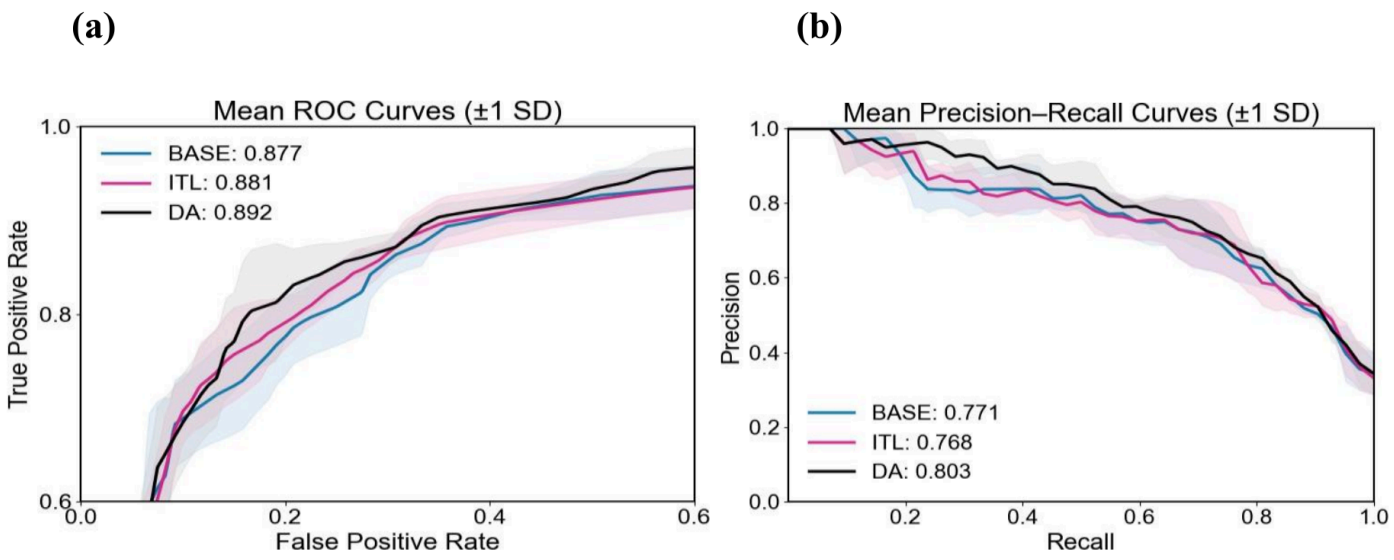


Figure 1

Performance comparison between Baseline (BASE) (blue), ITL (magenta), and DA (black). (a) Mean ROC curves for BASE, ITL, and DA LGBM models. Solid lines represent the mean across 5 seeds, and shaded regions show ±1 standard deviation. (b) Mean Precision–Recall curves for BASE, ITL, and DA XGBoost models. Solid lines represent the mean across 5 seeds, and shaded regions show ±1 standard deviation. Abbreviations: BASE, Baseline; DA, Domain Adaptation; ITL, Inductive Transfer Learning, LGBM, Light Gradient-Boosting Machine; XGBoost, Extreme Gradient Boosting.

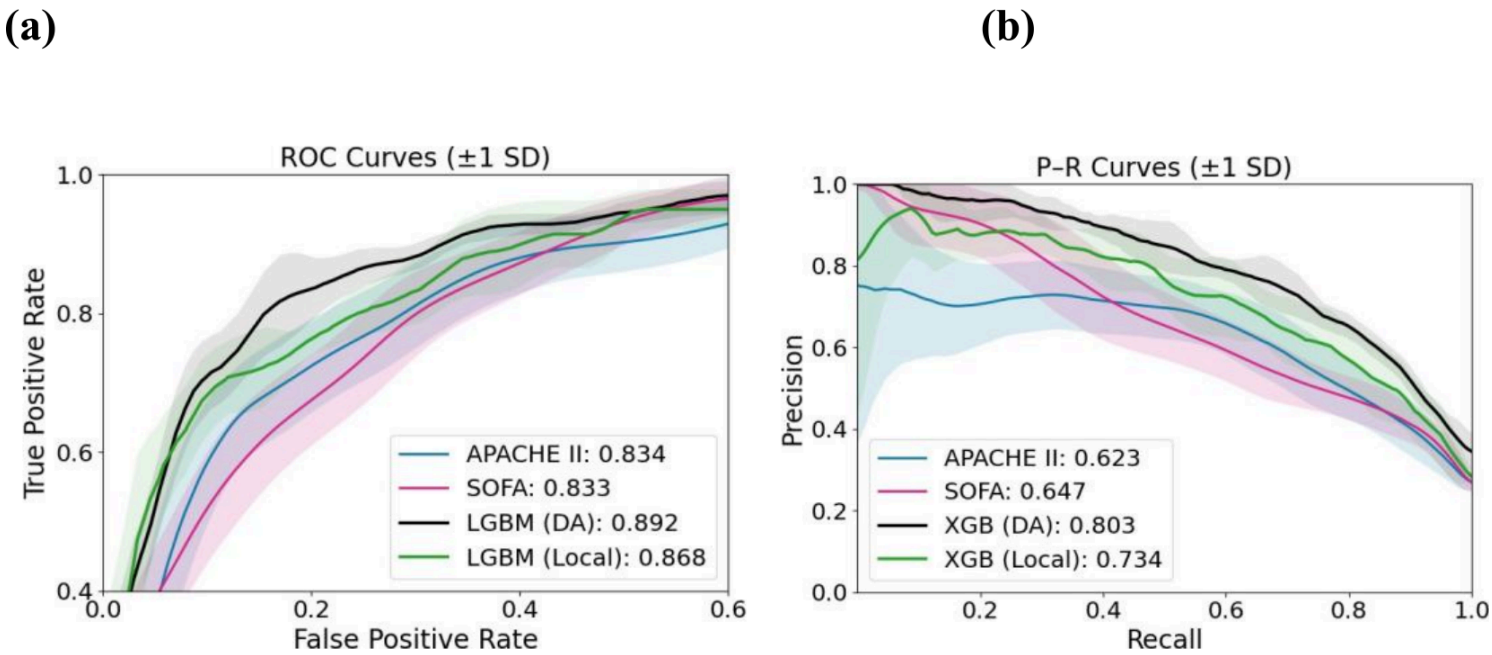


Figure 2

Performance comparison between APACHE II (blue), SOFA (magenta), domain- adapted LGBM (black), and models trained on local data (green). (a) Mean ROC curves for APACHE II, SOFA, domain-adapted

LGBM, and local LGBM. Solid lines represent the mean across 5 seeds, and shadedregions show ±1 standard deviation (SD). (b) Mean Precision–Recall curves for APACHE II, SOFA, domain-adapted XGB, and local XGB. Solid lines represent the mean across 5 seeds, and shaded regions show ±1 standard deviation (SD). Abbreviations: APACHE II, Acute Physiology and Chronic Health Evaluation II; LGBM, Light Gradient-Boosting Machine; SOFA, Sequential Organ Failure Assessment; XGBoost, Extreme Gradient Boosting.

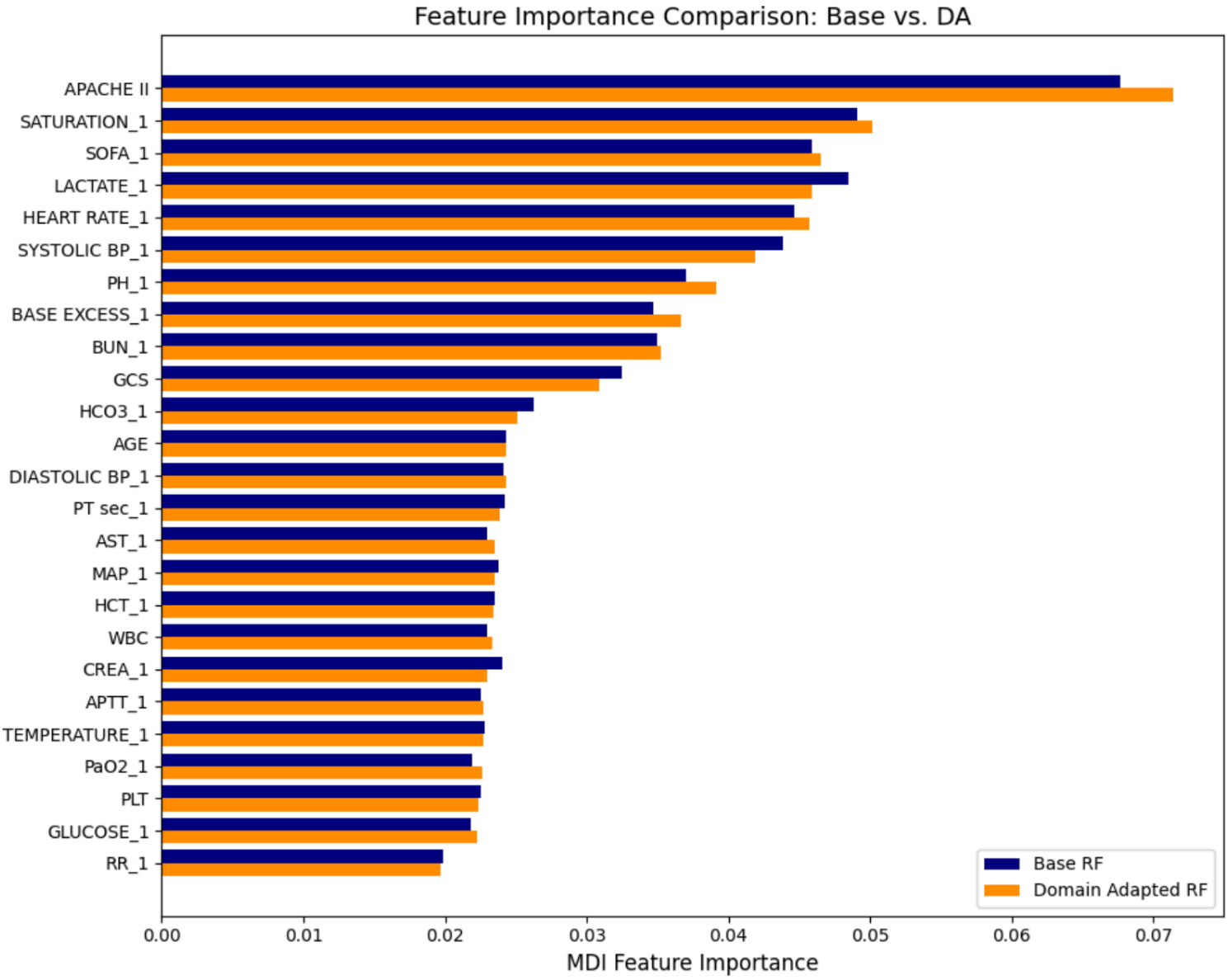


Figure 3

Feature-level MDI importances for baseline and domain-adapted Random Forest models. Abbreviations: APACHE II, Acute Physiology and Chronic Health Evaluation II; AST, Aspartate Aminotransferase; BUN, Blood Urea Nitrogen; GCS, Glasgow Coma Scale; FiO_2, Fraction of Inspired Oxygen; LDH, Lactate Dehydrogenase; HCO_3, Bicarbonate; ALT, Alanine Aminotransferase; NA, Sodium; PCO_2, Partial Pressure of Carbon Dioxide; RR, Respiratory Rate; APTT, Activated Partial Thromboplastin Time; HCT, Hematocrit;

WBC, White blood cell count; CREA, Creatinine; PaO_2, Partial Pressure of Arterial Oxygen; PLT, Platelet count; SOFA, Sequential Organ Failure Assessment.

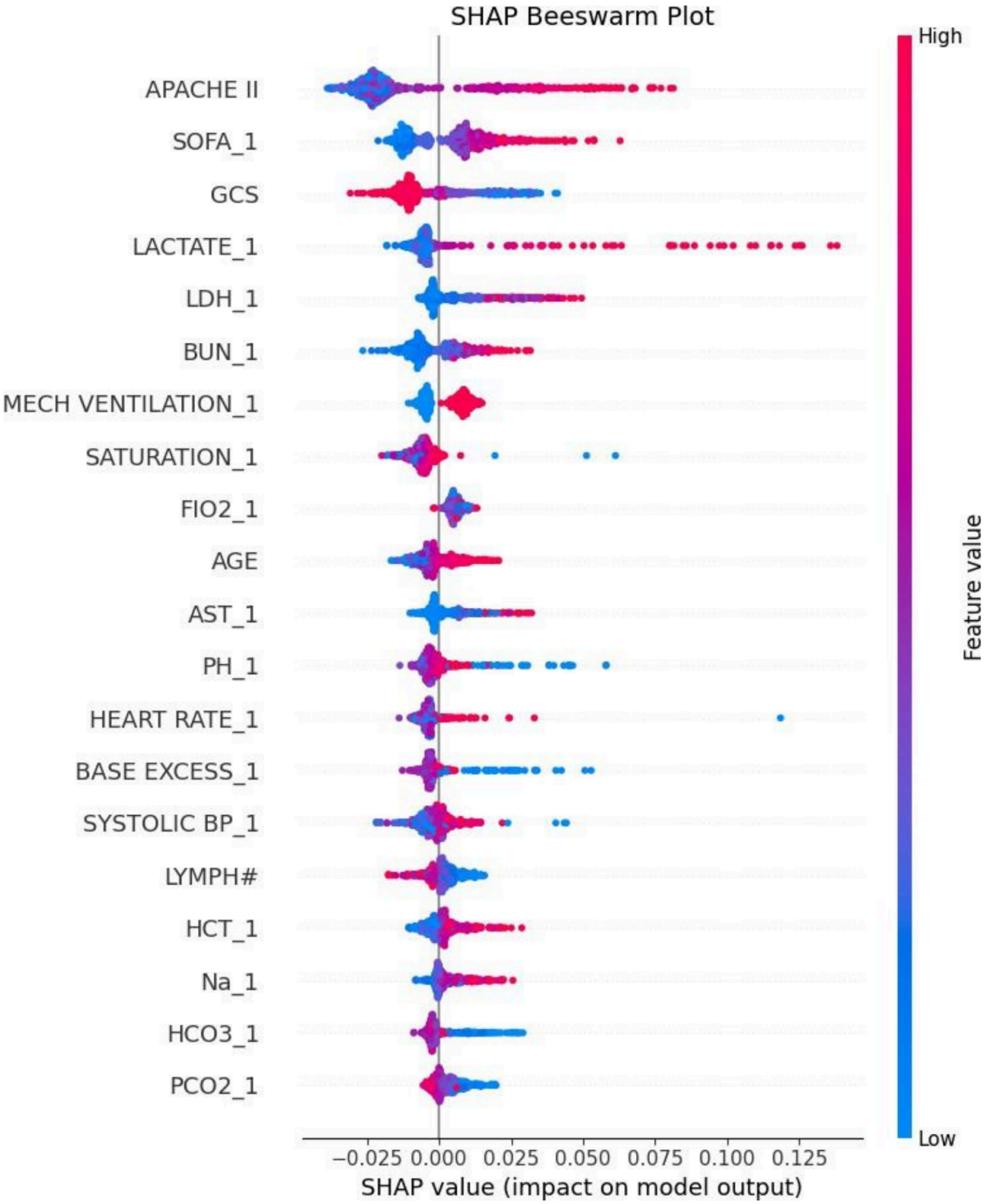


Figure 4

Shap Beeswarm plot for Random Forest models. Abbreviations: APACHE II, Acute Physiology and Chronic Health Evaluation II; AST, Aspartate Aminotransferase; BUN, Blood Urea Nitrogen; GCS, Glasgow

Coma Scale; FiO_2, Fraction of Inspired Oxygen; LDH, Lactate Dehydrogenase; HCO_3, Bicarbonate; ALT, Alanine Aminotransferase; NA, Sodium; PCO_2, Partial Pressure of Carbon Dioxide. Lymph#, Absolute Number of Lymphocytes; HCT, Hematocrit; SOFA, Sequential Organ Failure Assessment; Systolic BP, Systolic Blood Pressure.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryScientificReports.docx