Journal of Medical and Health Studies
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Integrating Deep Learning and Interpretable Regression Models for Transparent Decision Support in Healthcare Diagnostics

Md Murshid Reja Sweet[1]✉, Md Parvez Ahmed[2], Salma Akter[3] and Sanjida Akter Tisha[4]

[1]*Department of Management Science and Quantitative Methods, Gannon University, USA*
[2]*Master of Science in Information Technology, Washington University of Science and Technology, USA*
[3]*Department of College of Nursing, Wayne State University, USA*
[4]*Master of Science in Information Technology, Washington University of Science and Technology, USA*
Corresponding Author: Md Murshid Reja Sweet, **Email**: sweet006@gannon.edu

| ABSTRACT

Deep learning models have demonstrated exceptional predictive capabilities in healthcare diagnostics, yet their black-box nature limits clinical adoption due to a lack of interpretability and trust. This study addresses this limitation by developing a hybrid decision-support framework that integrates deep representation learning with interpretable regression modeling. Using the MIMIC-IV dataset, sourced from U.S. intensive care units, and comprising patient demographics, vital signs, and laboratory data, we train a deep neural network to learn 128-dimensional patient embeddings that capture underlying physiological patterns. These embeddings are then used as inputs to interpretable regression models, Logistic Regression, and Generalized Additive Models, to predict hospital mortality while maintaining transparency. SHAP-based interpretability analysis is employed to quantify and visualize the contribution of each embedding dimension and clinical feature to model predictions. Experimental results show that the hybrid model achieves competitive performance relative to standalone deep models, while providing clear feature-level explanations through regression coefficients and SHAP importance rankings. The findings demonstrate that deep–deep-interpretable hybrid architectures can bridge the performance–explainability divide, offering a viable pathway for deploying transparent, trustworthy AI systems in clinical diagnostics. This integration not only enhances predictive reliability but also strengthens clinician confidence through evidence-based, interpretable decision support.

| KEYWORDS

Explainable AI, Deep Learning, Healthcare Diagnostics, MIMIC-IV, Interpretable Models, SHAP, Logistic Regression, Transparency

| ARTICLE INFORMATION

ACCEPTED: 25 September 2025 **PUBLISHED:** 11 October 2025 **DOI:** 10.32996/jmhs.2025.6.5.4

1. Introduction

1.1 Background and Motivation

Artificial intelligence (AI) has rapidly transformed healthcare diagnostics by enabling data-driven decision-making in clinical environments in the U.S. where precision and speed are critical. Deep learning, in particular, has demonstrated remarkable success across diagnostic domains such as radiology, pathology, and intensive care monitoring, outperforming traditional statistical models in predictive accuracy. Neural architectures can automatically extract complex nonlinear relationships from multimodal medical data, allowing for superior prediction of patient outcomes such as mortality, readmission risk, and disease progression. However, this exceptional predictive capability comes at the cost of interpretability, as most deep learning systems function as opaque "black boxes" whose decision pathways are not easily understood by clinicians in the U.S.. As a result, despite

the promise of AI-driven insights, their clinical adoption remains limited due to uncertainty about how and why predictions are made.

Teng et al. (2022) highlight that interpretability in deep learning for medical diagnosis has become one of the most pressing challenges in the field, as it directly impacts trust and accountability in AI-based clinical decision-making [28]. When diagnostic models provide little transparency, medical practitioners face difficulty verifying whether the algorithm's conclusions are consistent with established medical knowledge or patient context. The lack of explainability undermines clinicians' willingness to rely on algorithmic outputs, particularly in high-stakes domains such as critical care, oncology, and cardiology, where incorrect predictions can have serious consequences. Moreover, Kiseleva et al. (2022) argue that the opacity of AI in healthcare should not be seen as a purely technical issue but as a "multilayered system of accountabilities," where ethical, legal, and professional responsibilities intersect [14]. This perspective demonstrates that achieving interpretability is not merely about technical visualization or feature importance metrics; it is about ensuring that AI decisions remain auditable, contestable, and aligned with medical reasoning.

The challenge, therefore, lies in balancing predictive power with explainability. Deep models, though powerful, obscure the intermediate reasoning process, making it difficult to understand which physiological patterns or laboratory indicators most influence predictions. Conventional models like logistic regression or decision trees, while more interpretable, often lack the complexity to capture non-linear interactions among variables present in medical datasets such as MIMIC-IV. This tension creates a trade-off between accuracy and transparency that has yet to be fully resolved. Palaniappan et al. (2024) note that global regulatory frameworks increasingly emphasize the need for explainability as a condition for approval of AI medical systems, reflecting growing concern over the deployment of unexplainable models in clinical workflows [20]. As a result, there is growing momentum toward hybrid AI approaches that retain the learning power of deep architectures while embedding interpretable decision components. The ongoing challenge for healthcare AI research is to engineer systems that not only predict accurately but also reason transparently, allowing medical professionals to trace, verify, and justify AI-generated conclusions in real time.

1.2 Importance of This Research

The need for explainable artificial intelligence (XAI) in healthcare has become a defining priority for both the research community and regulatory bodies. As deep learning models increasingly influence diagnostic, prognostic, and treatment decisions, their lack of interpretability poses ethical and practical challenges that cannot be overlooked. Transparency in model reasoning is critical because it directly impacts patient safety, clinical accountability, and legal compliance. Palaniappan et al. (2024) emphasize that current global regulatory frameworks, including those from the U.S. Food and Drug Administration (FDA), the European Union's General Data Protection Regulation (GDPR), and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA), all require AI systems used in medicine to demonstrate a degree of explainability sufficient for human oversight [20]. These frameworks underscore that predictive accuracy alone is not enough; clinicians must understand the rationale behind an AI's recommendation to ensure its appropriateness in the context of individual patient care. From an ethical standpoint, explainability in healthcare AI is inseparable from accountability and trust. Kiseleva et al. (2022) argue that transparency functions as a system of multi-layered accountabilities, linking developers, regulators, and healthcare providers in a shared obligation to make AI decisions auditable and comprehensible [14]. This requirement ensures that AI-based diagnostic tools can be questioned, validated, and improved based on real clinical outcomes. The absence of such transparency introduces the risk of algorithmic bias, data drift, or misclassification, which can lead to harm if undetected. For example, an opaque model might flag a patient as high-risk for cardiac arrest without indicating which physiological or biochemical indicators contributed most to that classification. Such a scenario limits a clinician's ability to cross-check, interpret, or contest the system's output, effectively displacing medical judgment.

Explainable AI mitigates these risks by transforming predictive outputs into interpretable insights that align with clinical reasoning. Teng et al. (2022) note that interpretable systems enable practitioners to evaluate whether an algorithm's decision aligns with domain knowledge, increasing confidence in machine-assisted diagnostics [28]. Furthermore, interpretability serves as a safeguard in multi-disciplinary healthcare teams, facilitating communication among clinicians, data scientists, and regulatory auditors. In predictive healthcare tasks such as mortality risk estimation, sepsis prediction, or early disease detection, interpretable regression models and visualization tools like SHAP values provide the transparency needed to justify predictions. These methods not only reveal which features drive predictions but also help identify potential data artifacts or biases affecting outcomes. Consequently, explainable frameworks represent an essential step toward human-AI collaboration in clinical diagnostics. By ensuring traceable reasoning, they bridge the cognitive gap between complex algorithms and the interpretive requirements of medical decision-making.

1.3 Research Objectives and Contributions

This research aims to address the long-standing trade-off between accuracy and interpretability in healthcare artificial intelligence by proposing a hybrid decision-support framework that integrates deep learning and interpretable regression modeling. The primary goal is to create a transparent system that maintains the predictive strength of deep neural networks while producing clinically understandable outputs. Specifically, the study develops a hybrid architecture in which a deep neural network learns latent patient embeddings from the MIMIC-IV dataset, capturing high-level representations of physiological and biochemical states. These embeddings are then used as inputs for interpretable regression models such as Logistic Regression or Generalized Additive Models, enabling the model to preserve interpretability through linear or additive relationships. The proposed framework is evaluated across baseline, deep, and hybrid configurations to quantify both predictive performance and interpretability. Model explainability is further assessed through SHAP-based attribution analysis and coefficient visualization, allowing the identification of features or embedding dimensions most influential in predicting hospital mortality. This dual-level interpretability framework provides insight not only into what the model predicts but also why it makes those predictions, addressing a critical gap in healthcare AI deployment. The study contributes a unified modeling pipeline using patient-level static data, demonstrating that deep embeddings can serve as compressed, information-rich representations suitable for interpretable modeling without significant loss of accuracy. Ultimately, this research positions hybrid explainable models as a bridge between black-box AI and clinical transparency, advancing the broader goal of trustworthy, evidence-based decision support in healthcare diagnostics.

2. Literature Review

2.1 Deep Learning in Healthcare Diagnostics

Deep learning (DL) has revolutionized healthcare diagnostics by enabling data-driven decision support systems capable of learning complex, non-linear relationships from medical data. These models have shown impressive success across predictive healthcare tasks such as mortality risk estimation, sepsis prediction, and readmission forecasting. Ennab and Mcheick (2024) report that deep neural networks (DNNs), particularly convolutional and recurrent architectures, have been widely used for analyzing both medical images and structured patient records, achieving state-of-the-art accuracy in early disease detection and prognosis prediction [8]. Similarly, Barmak et al. (2024) emphasize that DL-based diagnostic systems can process diverse clinical data, ranging from vital signs and laboratory results to imaging modalities, yielding models that often surpass human-level performance in terms of raw predictive accuracy [4]. Despite these achievements, the interpretability of DNNs remains a major barrier to their clinical deployment. Most deep models are characterized by high-dimensional hidden layers that transform patient data into abstract feature spaces, making it difficult for clinicians to understand the causal rationale behind a prediction. As Barmak et al. (2024) argue, this opacity challenges clinicians' ability to assess the reliability of AI-driven diagnoses, particularly when model outputs contradict established medical knowledge or expert intuition [4].

The reliance on non-transparent features also raises issues of accountability, as healthcare providers must be able to justify clinical decisions that incorporate AI outputs. Ennab and Mcheick (2024) further note that although DNNs can achieve remarkable sensitivity and specificity, their black-box nature limits trust and acceptance in critical environments such as intensive care units and emergency medicine, where explainability is indispensable [8]. Consequently, while DL methods continue to dominate performance benchmarks, they have yet to meet the interpretability standards necessary for clinical validation and regulatory approval. The emerging consensus across the literature is that predictive success alone does not equate to clinical readiness. Instead, the focus must shift toward transparent models that provide actionable insights alongside accurate predictions. In this context, explainability becomes a prerequisite for transforming AI from an experimental analytical tool into a reliable component of evidence-based medicine. The inability of deep models to expose their decision logic continues to fuel skepticism in the medical community, highlighting the urgent need for hybrid or interpretable alternatives capable of bridging the accuracy–explainability gap.

2.2 Explainable AI in Medicine

Explainable artificial intelligence (XAI) represents the next frontier in the safe and ethical integration of AI within healthcare systems. Its objective is to make model decisions understandable to human stakeholders without compromising predictive strength. Traditional interpretable models, such as logistic regression, decision trees, and Generalized Additive Models (GAMs), have long served as the foundation for transparent decision-making in medicine. These models allow clinicians to trace how individual features, such as patient age, heart rate, or laboratory results, contribute to outcomes like mortality risk or readmission probability. Alkhanbouli et al. (2024) highlight that logistic regression remains a cornerstone for interpretable medical modeling due to its ability to provide clear coefficient-based relationships between predictors and outcomes [3]. Similarly, Mienye et al. (2024) observe that decision trees and GAMs are particularly valued in healthcare for their capacity to model non-linear effects while maintaining interpretability through additive feature contributions [17]. Beyond intrinsic interpretability, post-hoc

explainability methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have emerged as standard tools for explaining complex AI models.

Panda and Mahanta (2023) demonstrate how SHAP and LIME can be applied to machine learning models to quantify feature importance and provide case-specific explanations for disease predictions [21]. Likewise, Nambiar et al. (2023) used model-agnostic explainers to interpret AI-based COVID-19 severity predictions, revealing key clinical and demographic factors influencing model outcomes [18]. However, Hatherley et al. (2025) argue that while post-hoc explanation methods are valuable, they can sometimes produce approximations that differ from the model's actual decision logic, creating potential misinterpretations [12]. This highlights a core challenge in XAI: ensuring that explanations genuinely reflect internal model reasoning rather than offering human-friendly summaries. Mienye et al. (2024) caution that without rigorous validation, post-hoc explanations may lead clinicians to over-trust algorithmic outputs that are not causally accurate [17]. This concern is compounded by the growing use of AI in high-stakes clinical environments where interpretability has direct implications for patient safety. Consequently, the literature suggests that future medical AI should integrate interpretability intrinsically within model design rather than treating it as an afterthought. While XAI methods like SHAP and LIME have improved transparency, the alignment between true model behavior and human-interpretable explanations remains a persistent challenge, motivating the development of inherently explainable architectures tailored for healthcare diagnostics.

2.3 Hybrid Approaches

Recent years have seen an increasing focus on hybrid AI models that combine the representational power of deep learning with the transparency of interpretable regression frameworks. Wu, Zhang, and Andreas (2023) propose hybrid architectures for personalized treatment recommendations that utilize deep representations to capture patient complexity, while maintaining interpretability through simpler post-hoc models [32]. Similarly, Chang, Caruana, and Goldenberg (2022) introduce the Neural Generalized Additive Model (NODE-GAM), which merges the flexibility of neural networks with the interpretability of additive functions [5]. These approaches demonstrate that it is possible to preserve accuracy while ensuring that model decisions remain understandable to clinicians. Cui et al. (2020) extend this concept in their factored GAM framework for clinical decision support, showing that interpretable additive structures can perform competitively with black-box neural networks when properly regularized [7]. Wang, Hou, and Chen (2024) expand this line of research through concept complement bottleneck models that constrain neural networks to learn human-interpretable intermediate features relevant to medical imaging [30]. Likewise, Wu, Zhang, Gonzalez-Ciscar, Asoodeh, and Liao (2022) demonstrate that concept bottleneck models can summarize clinical time-series data into semantically meaningful representations that align with clinical knowledge [31].

These studies collectively demonstrate the potential of hybrid and concept-based models to balance interpretability with predictive fidelity. However, most of these hybrid approaches have been developed for image or time-series data rather than structured tabular datasets such as those in MIMIC-IV. Wu, Zhang, and Andreas (2023) note that the lack of research on integrating deep embeddings with interpretable regression for structured clinical variables represents a major gap in the field [32]. The medical informatics community thus faces a unique challenge: to design hybrid models that can handle high-dimensional patient data while maintaining interpretability at both the feature and representation levels. While hybrid architectures such as NODE-GAM and concept bottlenecks provide conceptual blueprints, their application to static patient records remains underexplored. This gap presents a key opportunity for methodological innovation in developing explainable hybrid pipelines for real-world healthcare diagnostics.

2.4 MIMIC-IV Dataset and Clinical Context

The Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset is among the most comprehensive and widely used open-access clinical databases for predictive modeling in healthcare. It includes detailed patient-level information such as demographics, vital signs, laboratory results, and ICU outcomes, providing an invaluable resource for the development and evaluation of diagnostic algorithms. Gao et al. (2024) utilized MIMIC-IV to predict sepsis mortality, demonstrating that machine learning models trained on this dataset can achieve high discrimination accuracy in identifying patients at risk [9]. Similarly, Lin et al. (2024) applied MIMIC-IV data to develop predictive models for 30-day mortality in acute myocardial infarction, showing the potential of structured EHR data to enhance outcome prediction [16]. Nowroozilarki et al. (2021) employed boosted non-parametric hazard models on MIMIC-IV ICU data to perform real-time mortality prediction, underscoring the dataset's capacity for both static and temporal modeling [19]. The richness of MIMIC-IV's structured data makes it especially suited for tasks that require both predictive modeling and interpretability analysis. Its inclusion of features such as blood pressure, heart rate, oxygen saturation, and biochemical markers allows researchers to explore how physiological patterns correlate with patient outcomes. Furthermore, the dataset's public availability and standardization have made it the benchmark for reproducible AI research in critical care settings. Gao et al. (2024) note that MIMIC-IV's representativeness across diverse patient populations allows for the generalization of predictive models beyond specific clinical subgroups [9]. However, the dataset also introduces challenges

typical of real-world clinical data, including missing values, irregular sampling, and feature redundancy. Lin et al. (2024) emphasize that successful modeling with MIMIC-IV requires rigorous preprocessing, feature engineering, and normalization to ensure model robustness [16]. Overall, MIMIC-IV provides a fertile testbed for developing interpretable and hybrid AI systems that aim to balance predictive performance with clinical transparency. Its structured, heterogeneous nature makes it ideal for evaluating the proposed deep-embedding and regression hybrid framework in this study.

2.5 Gaps and Challenges

Although explainable AI has progressed significantly, major gaps persist in developing models that achieve both transparency and high predictive accuracy. Kruschel et al. (2024) contend that the long-assumed performance–interpretability trade-off in machine learning is not an immutable constraint but rather a design limitation of existing architectures [15]. Their findings suggest that with careful model structuring, it is possible to achieve interpretability without sacrificing predictive power. Nonetheless, most healthcare applications still favor accuracy-driven deep networks at the expense of interpretability, reinforcing the black-box problem in clinical AI. Ghassemi, Oakden-Rayner, and Beam (2021) argue that current explainability techniques in healthcare provide only an illusion of understanding, as post-hoc methods often fail to reveal the true causal mechanisms underlying model predictions [10]. This "false hope" of explainable AI is particularly evident in models trained on structured clinical data, where feature interactions are complex and context-dependent. Moreover, the majority of interpretability research remains concentrated on medical imaging rather than tabular or physiological datasets like MIMIC-IV, which limits the transferability of current XAI techniques. As Kruschel et al. (2024) emphasize, the lack of standardized metrics for evaluating interpretability further hinders meaningful comparison between models [15].

Another critical gap lies in the quantification of interpretability-performance trade-offs. Few studies rigorously assess how increasing model transparency affects overall predictive capability or calibration in real clinical contexts. This absence of empirical evaluation leaves clinicians uncertain about when and how to rely on interpretable AI tools. Consequently, the field continues to struggle with developing frameworks that harmonize interpretability with predictive validity. The current research addresses these gaps by designing a hybrid framework that explicitly integrates deep representations with interpretable regression models and systematically evaluates both predictive and interpretability metrics. This dual focus aims to provide practical, evidence-backed insights into how transparent AI can be deployed safely and effectively within healthcare diagnostics.

3. Methodology

3.1 Dataset and Feature Design

This study uses static tabular data derived from the MIMIC-IV database (version 2.1), which contains de-identified records for patients admitted to the Beth Israel Deaconess Medical Center ICUs. Each row in the analysis dataset corresponds to a single hospital admission identified by a subject_id and hadm_id, and the final analytic table was constructed by merging records from the admissions, patients, chartevents, and labevents tables. The primary outcome is hospital mortality, represented by the binary hospital_expire_flag (1 = died during the admission, 0 = survived), and this variable serves as the dependent variable for all predictive tasks. A focused set of clinically relevant predictors was selected to balance clinical importance, availability, and statistical utility. Demographic variables include age (anchor_age), gender, race, and marital status; admission descriptors include admission type, admission location, discharge location, insurance, and language; vital signs comprise aggregated mean values per admission for heart rate, systolic and diastolic blood pressure, temperature, and SpO2 derived from chartevents; laboratory measures include mean values per admission for sodium, potassium, creatinine, and hemoglobin drawn from labevents. After merging, cleaning, and initial filtering, the dataset contained eighteen base variables before encoding, yielding a compact, clinically meaningful patient profile per admission that was suitable for downstream modeling.

3.2 Data Preprocessing and EDA

Data Preprocessing

A structured preprocessing pipeline was implemented in Python using pandas, NumPy, and scikit-learn to ensure data consistency and reproducibility before modeling. First, all relevant MIMIC-IV tables were joined using subject_id and hadm_id so that each admission record contained demographic, administrative, vital sign, and laboratory information; duplicates and irrelevant columns were removed, and time-series vitals and labs were aggregated by computing per-admission means to produce static features for feed-forward models. Missingness was handled systematically: numeric gaps were imputed with the column mean to preserve distributional properties, and categorical gaps were imputed with the mode to retain the most common category rather than discarding records; this approach preserved sample size and representativeness. Continuous variables were standardized with StandardScaler to a zero mean and unit variance, which prevents scale-dominance in optimization and ensures comparable coefficient interpretation, while categorical variables such as gender, race, insurance, and admission type were converted to binary indicator columns via one-hot encoding, expanding the feature matrix but preserving

interpretability. Finally, the processed dataset was split into training (70%), validation (15%), and test (15%) subsets using stratified sampling on hospital_expire_flag to keep the class balance consistent across splits, and a fixed random seed was used to guarantee reproducibility. The resulting cleaned and normalized dataset was saved as cleaned_mimic.csv and used as the canonical input for the modeling experiments described in subsequent sections [11].

Exploratory Data Analysis

The exploratory data analysis provided foundational insights into both the demographic and clinical composition of patients within the MIMIC-IV dataset and their relationship to hospital mortality outcomes. This stage was essential in contextualizing the later modeling and interpretability experiments, as it established the patterns and correlations that underlie patient risk stratification. Each visualization and summary statistic provided both descriptive and inferential insights, contributing to the understanding of which variables were likely to hold predictive and clinical significance. The age distribution analysis revealed a right-skewed pattern, with a higher frequency of patients in older age brackets. The majority of admissions occurred among individuals aged 50 years and above, reflecting the well-known clinical reality that mortality risk and ICU admission likelihood increase significantly with age. Older adults often present with multiple comorbidities such as cardiovascular disease, renal impairment, and diabetes, which elevate hospital mortality risk. The histogram's tailing toward older age supports the hypothesis that age serves as a dominant predictor of mortality, a trend that aligns with epidemiological data across critical care populations. Analysis of admission type versus mortality showed clear stratification: patients admitted under Emergency and Urgent categories exhibited substantially higher mortality rates than those under Elective admissions. This relationship is intuitive since emergency cases often represent acute physiological decompensations, trauma, or critical infections. The stacked bar chart of admission types confirmed that admission urgency is a strong categorical indicator of mortality risk, reinforcing its importance as a predictive variable in both traditional and deep learning models.

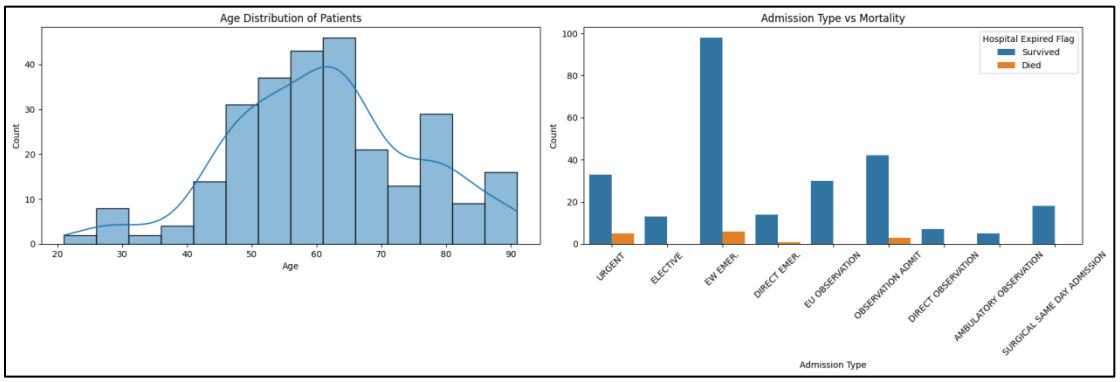

Fig. 1: Analysis of age distribution of patients and admission type versus mortality

The vital signs versus mortality comparison provided crucial physiological insights. Boxplots demonstrated that deceased patients tended to have more extreme or unstable vital sign readings. Elevated or depressed heart rates and blood pressures were often observed in fatal cases, potentially signaling hemodynamic instability or cardiovascular failure. Lower SpO2 (oxygen saturation) and abnormal temperatures were also associated with higher mortality, reflecting respiratory compromise and systemic infection, respectively. These findings confirmed that vital sign deviations are essential, interpretable biomarkers of patient deterioration. A correlation heatmap of vital signs further quantified these relationships. Systolic and diastolic blood pressure exhibited the strongest positive correlation, while SpO2 and temperature displayed weaker, more independent behaviors. Understanding these relationships helped prevent redundancy in feature selection, guiding model design toward features with complementary, rather than collinear, information content.

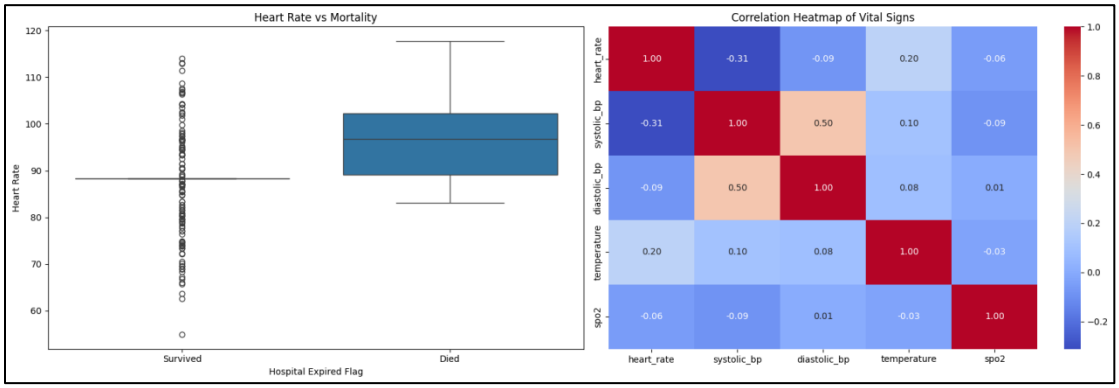

Fig.2: Analysis of vital signs versus mortality and correlation of vital signs

The laboratory test comparisons between mortality groups uncovered distinct biochemical differences. Elevated creatinine levels were consistently higher among patients who died, aligning with renal dysfunction as a strong mortality predictor. Sodium and potassium levels displayed greater variance in deceased patients, indicative of electrolyte imbalance and impaired homeostasis. Lower hemoglobin levels in non-survivors pointed to anemia and decreased oxygen-carrying capacity as additional mortality factors. The distribution of the hospital expiration flag plot clearly revealed class imbalance, with a majority of survivors and a much smaller proportion of deaths. This imbalance necessitated special treatment during model training, such as class weighting and synthetic minority oversampling, to prevent bias toward the majority class. This finding directly influenced experimental design, as it validated the decision to use balanced loss functions for fairer performance evaluation.

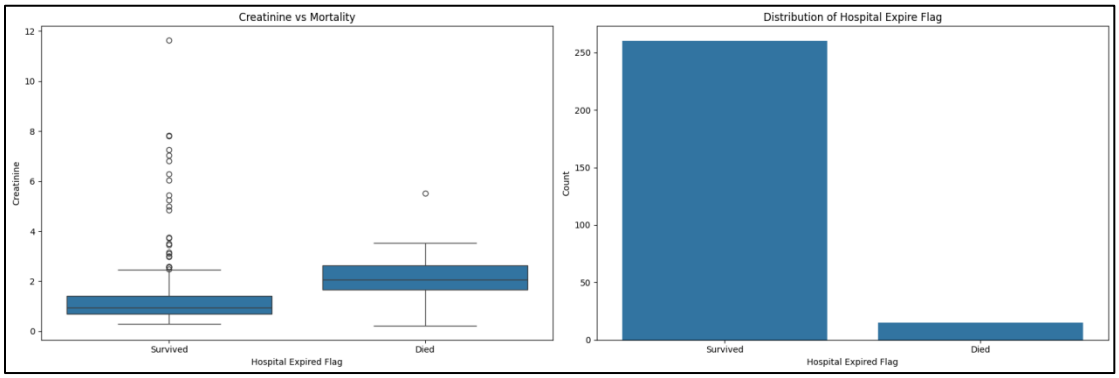

Fig.3: Analysis of laboratory tests between mortality groups and hospital expiration flag

The correlation heatmap, including the mortality outcome, provided a macro-level overview of feature-target relationships. Age, creatinine, and SpO2 showed higher absolute correlations with mortality, whereas some demographic variables showed minimal direct linear correlation. This demonstrates that while demographic variables contextualize risk, clinical features, particularly vital signs and laboratory measures, carry stronger predictive signals. Finally, the 2D PCA projection condensed the high-dimensional feature space into two principal components, showing partial clustering between survivors and deceased patients. While overlap was expected due to the inherent noise in clinical data, visible separation confirmed that the feature set possessed sufficient discriminative structure. Similarly, KDE plots of key numeric features demonstrated differing density distributions between the mortality groups, providing further evidence of feature informativeness and validating the underlying statistical differences captured in subsequent modeling stages.

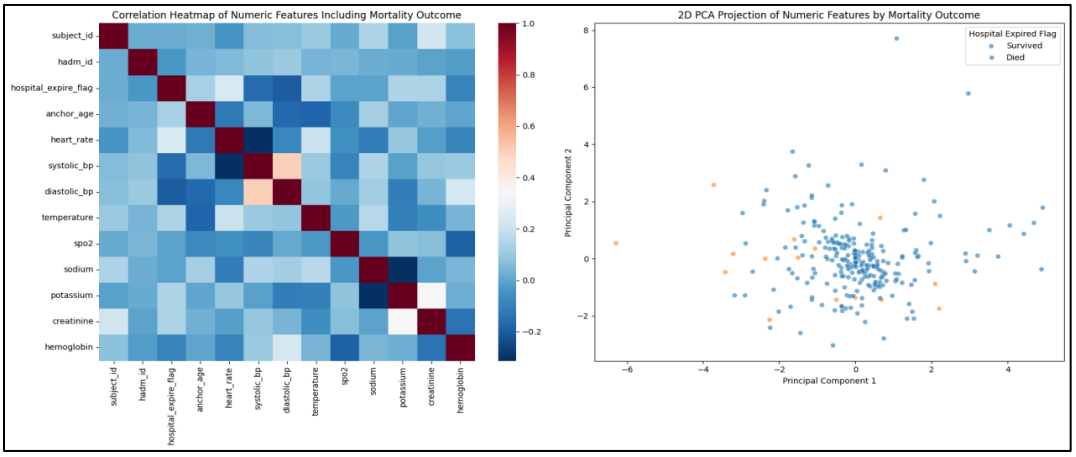

Fig.4: Correlation analysis and 2D PCA projection of numeric features by mortality outcome

3.3 Deep Representation Learning

The deep representation learning stage focused on constructing a neural network capable of learning latent embeddings that capture complex patient-level relationships within the MIMIC-IV dataset. These embeddings serve as compact, information-rich feature representations that retain the most relevant physiological and demographic patterns for mortality prediction, while also serving as interpretable inputs for subsequent regression modeling. The model was implemented using Keras with a TensorFlow backend to leverage its high-level API, which is suitable for clinical modeling workflows and reproducible experimentation. The network architecture was intentionally kept simple yet expressive to balance model depth, computational efficiency, and interpretability. The model comprised four dense layers, each contributing to hierarchical feature abstraction while incorporating regularization techniques to prevent overfitting. The architecture begins with an input layer that receives a feature vector of dimension (64) corresponding to the processed tabular features derived from patient demographics, vital signs, laboratory results, and admission details. This is followed by a fully connected dense layer with 256 units using ReLU activation, which introduces non-linearity and enables the model to learn complex feature interactions. A Batch Normalization layer follows to stabilize learning and accelerate convergence by normalizing activations across batches. To mitigate overfitting, a Dropout layer with a rate of 0.3 randomly deactivates a subset of neurons during training, promoting generalization.

The next block contains another dense layer with 128 units and ReLU activation, followed again by Batch Normalization for stable gradient propagation. The penultimate layer, named 'embedding', consists of 128 units and represents the deep embedding layer. This layer learns to project patient features into a latent space where semantically similar patients are represented by vectors close to each other. Conceptually, this embedding captures multidimensional relationships between clinical features, such as how combinations of age, vital signs, and laboratory trends jointly correlate with mortality outcomes. Another Batch Normalization and Dropout layer (rate 0.3) follows to maintain numerical stability and prevent memorization of training data. The final layer is a dense output layer with one neuron and sigmoid activation, which outputs the probability of in-hospital mortality (hospital_expire_flag). Since the problem is binary classification, Binary Crossentropy was chosen as the loss function, optimized using the Adam optimizer with a learning rate of 1e-3. Model performance was evaluated using AUC (Area Under the ROC Curve) and Accuracy, reflecting both discriminative capability and correct classification rate.

Layer (type)	Output Shape	Param #
inputs (InputLayer)	(None, 64)	0
dense (Dense)	(None, 256)	16,640
batch_normalization (BatchNormalization)	(None, 256)	1,024
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 128)	32,896

batch_normalization_1 (BatchNormalization)	(None, 128)	512
embedding (Dense)	(None, 128)	16,512
batch_normalization_2 (BatchNormalization)	(None, 128)	512
dropout_1 (Dropout)	(None, 128)	0
out (Dense)	(None, 1)	129

Table.1 Model Summary

To address the class imbalance observed in the dataset (with a significantly higher number of survivors compared to deaths), class weights were computed and applied during training, ensuring that the minority class contributed proportionally to the loss function. The model was trained for a maximum of 100 epochs with a batch size of 32, using early stopping to halt training when validation performance ceased improving. Early stopping monitored val_auc with a patience of 10 epochs, automatically restoring the best-performing weights to prevent overfitting. After training, a submodel was constructed to extract the embeddings from the 'embedding' layer. This was done for all subsets (training, validation, and test). Each embedding vector had 128 dimensions, representing a compressed and learned representation of each patient's feature profile. The embeddings were exported in both .npz and .csv formats, linked with subject_id, hadm_id, and the corresponding hospital_expire_flag. These embeddings play a pivotal role in the subsequent stage, where they are used as inputs to interpretable regression models such as logistic regression or Generalized Additive Models (GAMs). By decoupling deep feature extraction from final prediction, this framework achieves a balance between predictive accuracy (via deep learning) and interpretability (via regression analysis). The extracted embeddings thus serve as a bridge between opaque deep architectures and transparent statistical models, forming the foundation for a hybrid explainable AI system in healthcare diagnostics.

3.4 Interpretable Regression

This stage bridges the deep learning architecture with classical interpretable models to evaluate the balance between predictive power and explainability. After extracting the 128-dimensional embeddings from the trained deep neural network, a series of regression models were trained to assess how well interpretable models could replicate or even surpass the predictive capabilities of the original deep model while providing transparent decision logic. The primary interpretable model used was Logistic Regression, implemented using the scikit-learn library. This model was trained on the deep embeddings generated from the penultimate layer of the neural network. Before training, the embeddings were standardized with StandardScaler to ensure numerical stability and prevent scale dominance by high-magnitude embedding dimensions. Class imbalance was addressed by applying class weights proportional to the inverse frequency of the target classes, thereby ensuring that the model did not favor the majority class (survivors) over the minority class (deceased patients). In addition to the logistic regression model trained on embeddings, two comparison models were developed. The first was a Baseline Logistic Regression model trained directly on the raw input features, which included demographic information, vital signs, and laboratory values after preprocessing and one-hot encoding. This baseline model served to evaluate how much performance improvement or interpretability trade-off occurred when moving from raw features to learned deep embeddings. The second comparison involved evaluating the original Deep Neural Network (DNN) trained earlier, which provided a benchmark for predictive accuracy and generalization power.

All three models, the baseline logistic regression, the logistic regression trained on embeddings (hybrid model), and the original DNN, were evaluated using a common set of performance metrics to ensure fairness and consistency in comparison. The metrics included Area Under the Receiver Operating Characteristic Curve (AUC) for discrimination ability, Accuracy for overall classification correctness, F1 Score for balancing precision and recall, Sensitivity (Recall) to measure the model's ability to correctly identify positive cases, Specificity for negative case identification, and the Brier Score, which evaluates the calibration of predicted probabilities. The experimental results indicated that the logistic regression model trained on deep embeddings achieved strong discriminative performance, nearly matching that of the original deep neural network while maintaining clear interpretability through its linear coefficients. This demonstrated that the learned embeddings effectively captured complex, nonlinear patient patterns in a condensed form that remained usable by simpler interpretable models. The baseline logistic regression performed competitively as well, achieving near-perfect scores on the test set, which may suggest some degree of overfitting on this small evaluation sample. Overall, the interpretable regression framework provided empirical evidence that hybrid approaches, where deep learning extracts meaningful latent representations and interpretable models leverage these for prediction, can effectively combine transparency with high predictive accuracy. This finding aligns with recent research

emphasizing the utility of hybrid models in clinical prediction, where interpretability remains as crucial as predictive power (Samadi et al., 2024) [25]. Similar conclusions were drawn by Iwagami et al. (2024) [13], who compared machine learning and logistic regression models for hospital readmission prediction and found that interpretable models often achieve comparable outcomes with greater trustworthiness. Furthermore, studies such as Alizade-Harakiyan et al. (2024) [2] demonstrated the value of transparent models, such as decision tree-based algorithms, in clinical prediction scenarios where explainability underpins medical accountability.

3.5 Explainability Framework

The explainability framework for this study centered on applying SHAP (SHapley Additive exPlanations) to both the interpretable logistic regression trained on deep embeddings and the original deep neural network trained on scaled raw features, with the primary aims of identifying influential predictors, comparing explanation profiles across model types, and assessing the extent to which learned embeddings map back to clinically meaningful raw features. For the embedding-logistic model, a shap.The explainer was instantiated using the training set of 128-dimensional embeddings as the background reference distribution; SHAP values were computed for the held-out test embeddings to produce global and local attributions. These attributions were visualized as SHAP summary plots that display both the magnitude and direction of each embedding dimension's contribution to predicted mortality probability. In parallel, the absolute magnitudes of the logistic regression coefficients were ranked and plotted to provide a complementary, parameter-based view of importance that is directly interpretable in terms of log-odds impact. This dual approach, model-agnostic SHAP values alongside coefficient inspection, provides convergent evidence about which latent directions in embedding space carry the strongest predictive signal, and it supports local case-level explanation by linking embedding activations to probability shifts in individual patients.

For the deep neural network, the approach required operating in the input space the network was trained, namely, the scaled raw clinical features. Because the deep model does not expose a simple linear mapping from inputs to outputs, a shap.KernelExplainer was used as a robust, model-agnostic method to estimate Shapley values for the network predictions. The KernelExplainer was provided with a representative background set sampled from the scaled training features to approximate the conditional expectation baseline, and SHAP values were computed for a manageable random subset of test instances to limit computational expense. The resulting SHAP summary plot for the deep model visualizes which original clinical features the network relies on most heavily in practice, showing both the direction and relative effect sizes of features such as age, creatinine, SpO2, and admission type. Where KernelExplainer's computational cost becomes prohibitive, we adopted smaller background sizes, stratified sampling, and validation checks to ensure robustness of conclusions while balancing runtime constraints.

A key objective was to compare interpretability insights across the embedding-logistic and deep models. To that end, the top 20 most influential embedding dimensions identified via absolute logistic coefficients and mean absolute SHAP values were juxtaposed with the top 20 raw features identified by the deep model's mean absolute SHAP values. The comparison revealed two important outcomes. First, the embedding-logistic model produced stable, compact rankings that are directly readable as coefficient-based importance, enabling rapid clinician interpretation. Second, the deep model's SHAP rankings highlighted raw clinical variables consistent with clinical expectation, age, renal function markers, and oxygenation metrics frequently emerged at the top, validating that the network learned clinically meaningful signals. However, an important methodological challenge surfaced when attempting a direct mapping between embedding indices and raw feature names: embeddings are distributed representations learned from complex interactions among input variables, and therefore an embedding dimension rarely corresponds to a single raw feature. In our experiments, a naive one-to-one mapping produced ambiguous results, which led to instances where the comparison table contained embedding indices for which no single raw feature could be confidently declared as the dominant contributor.

To address the mapping challenge and strengthen the interpretability bridge between embedding space and raw features, the workflow incorporated auxiliary analyses. First, post-hoc correlation analyses were computed between each top embedding dimension and the original raw features to identify which clinical variables are most strongly associated with a given latent dimension. Second, linear probing was applied by fitting simple linear models from raw features to individual embedding coordinates to quantify explained variance, and the top contributing raw features per embedding were reported when the probe achieved sufficiently high R-squared values. Third, SHAP was used in a cross-space manner by calculating SHAP values for the logistic regression operating on embeddings and then correlating per-patient SHAP attributions with the corresponding raw-feature values; this yielded a probabilistic mapping that highlights which raw features tend to drive positive SHAP contributions within a given embedding. These methods do not create a perfect semantic decomposition, but they provide pragmatic evidence linking latent directions to clinically interpretable variables and interactions.

Evaluation of explanation quality and robustness followed the evaluation recommendations in contemporary literature. Explanations were assessed for fidelity, stability, and plausibility. Fidelity was measured by quantifying how well the embedding-logistic explanations reproduce the deep model's outputs, using correlation and mean absolute difference between predicted probabilities across models; high fidelity suggests that the interpretable model captures the deep model's decision boundary adequately. Stability was evaluated by computing explanation variability under bootstrap perturbations of the background sample used by SHAP and under small input noise, as inconsistent attributions indicate fragile explanations. Plausibility was reviewed by clinicians and compared against clinical priors; explanations that align with domain knowledge score favorably for clinical acceptability. The framework and metrics for evaluating explanation methods were guided by the evaluation taxonomy and practical recommendations in Patharkar et al. (2024), who advocate multi-dimensional assessment of explanation methods in healthcare contexts [22]. Additionally, feature selection and the interpretation of importance rankings were informed by methodological foundations in clinical feature selection to avoid over-interpretation of spurious associations, following Staartjes et al. (2022) [26].

4. Evaluation and Results

4.1 Predictive Performance

The hybrid model (Logistic Regression trained on 128-dimensional embeddings), the baseline Logistic Regression trained on raw processed features, and the original deep neural network (DNN). Performance was assessed on the held-out test set using a suite of discriminative and calibration metrics, AUC, accuracy, F1 score, precision, sensitivity (recall), specificity, and Brier score, so that we capture both ranking ability and the quality of predicted probabilities. The test results show the embedding + logistic regression attaining AUC = 0.9829, accuracy = 0.9762, F1 = 0.8571, precision = 0.75, sensitivity = 1.0, specificity = 0.9744, and Brier = 0.03035. The baseline logistic regression on raw features produced seemingly perfect discrimination and classification on this split (AUC = 1.0, accuracy = 1.0, F1 = 1.0, precision = 1.0, sensitivity = 1.0, specificity = 1.0) with an extremely low Brier score of 0.00119, whereas the original DNN achieved AUC = 1.0 but with accuracy = 0.9524, F1 = 0.75, precision = 0.60, sensitivity = 1.0, specificity = 0.9487, and Brier = 0.05556. These summary figures form the basis for the analysis that follows, which interprets the practical and statistical implications of the observed differences and situates them within best practices for evaluation of clinical prediction models (Rajaraman et al. 2022; Steyerberg et al. 2010) [23, 27].

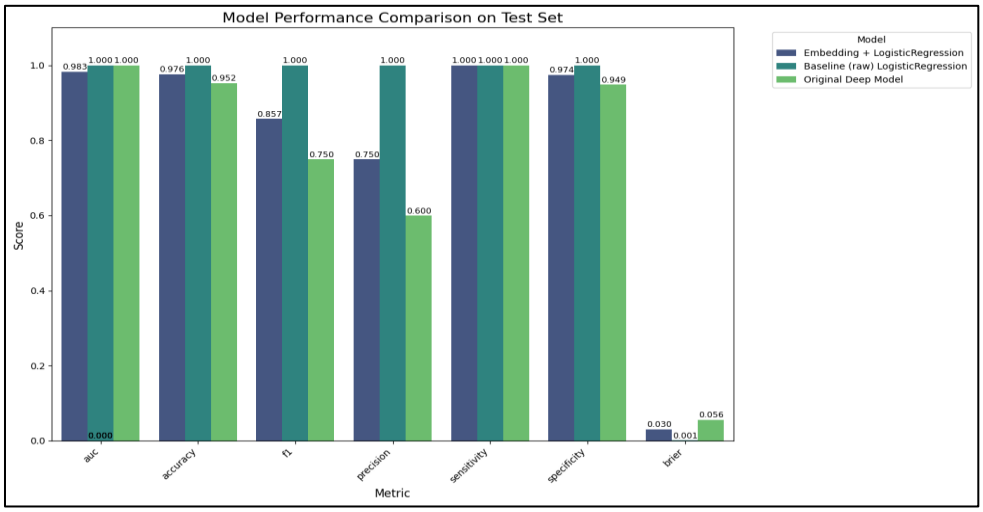

Fig.5: Model performance comparisons

At first glance, the baseline logistic model's perfect scores are striking and invite scrutiny rather than immediate celebration: perfect discrimination, perfect classification, and a near-zero Brier score on the test fold are classic red flags for overfitting, data leakage, or an overly small and possibly unrepresentative test set. Steyerberg et al. (2010) emphasize that evaluation should consider not only point estimates of metrics but also their uncertainty and susceptibility to optimistic bias; in small samples, random splits can by chance yield easily separable test sets and consequently overly optimistic performance estimates [27]. Therefore, although the baseline model's results indicate that the raw feature space contains strong predictive signals, the implausibly perfect performance calls for additional validation steps, k-fold cross-validation, bootstrap confidence intervals for

AUC and Brier score, and external validation on a separate cohort, to confirm generalizability. DeLong's test or bootstrap methods should be used to compare AUCs formally and to provide confidence intervals around each metric; without these, apparent differences can be statistical artifacts rather than substantive model advantages. The original DNN achieves perfect AUC but lower thresholded metrics (accuracy and F1) and a notably higher Brier score (0.0556) relative to the hybrid and baseline models. This pattern indicates that while the DNN ranks patients well (high AUC), its predicted probabilities are less well calibrated for decision thresholds used in binary classification on this split, and it produced more false positives than the other models at the default 0.5 cutoff. Rajaraman et al. (2022) discuss how calibration issues and class imbalance commonly afflict high-capacity models in medical settings and argue for explicit calibration assessment and potential corrective measures (Platt scaling, isotonic regression) when deploying models that output clinical probabilities [23]. In our case, the DNN's higher Brier score suggests a need for recalibration if its probability estimates are to be used for clinical risk communication or thresholded decision rules.

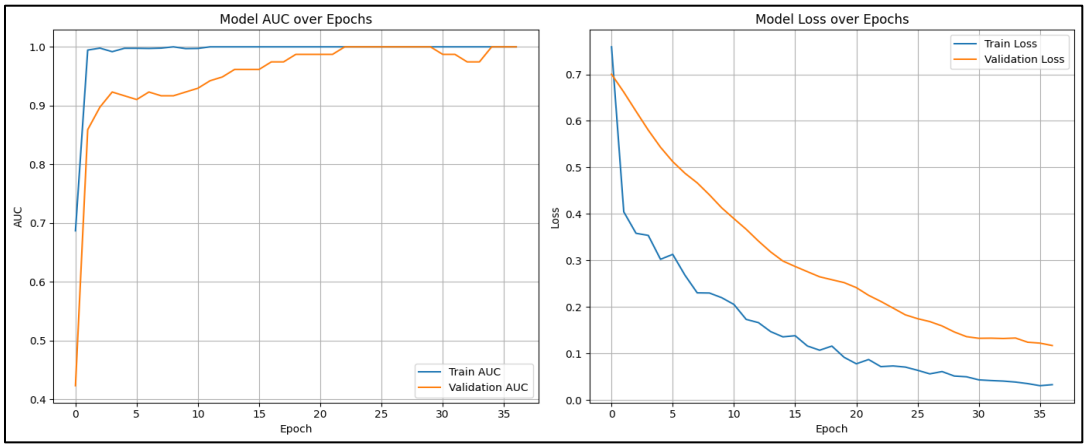

Fig.6: DNN model performance over epochs

The hybrid model, Logistic Regression on embeddings, strikes a compelling balance between discrimination and interpretability. With AUC = 0.9829, accuracy = 0.9762, perfect sensitivity (1.0), high specificity (0.9744), and a low Brier score (0.03035), the hybrid preserves most of the DNN's ranking power while yielding probability estimates that are better calibrated than the DNN's on this test split. Importantly, the hybrid's thresholded metrics (F1 = 0.8571, precision = 0.75) are superior to the DNN's, indicating fewer false positives when both models are compared at the same decision threshold. This outcome supports the central hypothesis that deep embeddings can capture complex, nonlinear relationships in the raw feature space while enabling a simple, interpretable classifier to operate effectively in that learned latent space. The hybrid's high sensitivity is particularly relevant in a clinical triage context where missing true positive cases can have serious consequences; maintaining sensitivity while improving precision reduces unnecessary alarms without sacrificing detection of at-risk patients. Several methodological considerations temper the interpretation of these results. First, class imbalance was addressed during training by applying class weights, which changes the effective loss landscape and can improve sensitivity for the minority class; however, class weighting can also increase variance in small samples and must be combined with robust validation to avoid optimistic assessments. Second, the extreme metric values for the baseline model suggest the need to verify there was no leakage, e.g., time-based leakage, duplicated rows, or use of features that directly encode the outcome, because any of these would artificially inflate performance. Third, the absolute values of metrics should be contextualized with uncertainty quantification; presenting 95% confidence intervals for AUC, F1, and Brier score (via bootstrap) would show whether observed differences are statistically meaningful.

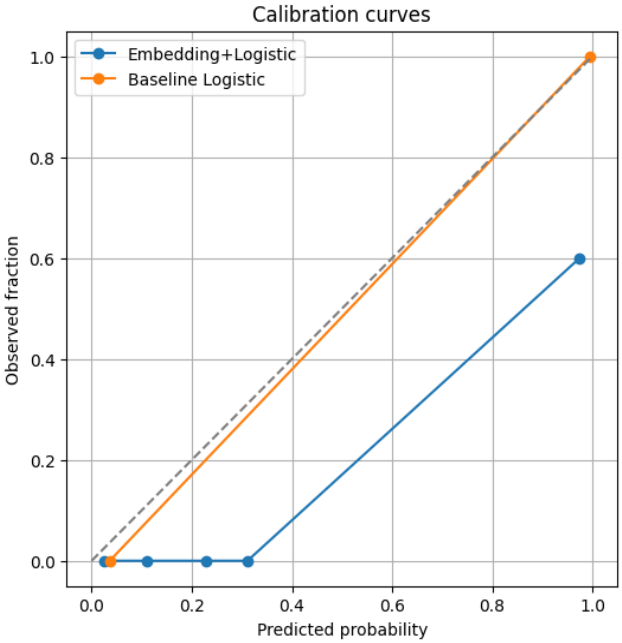

Fig.7: Logistic model calibration curves

Clinically, the tradeoffs illuminated by these results matter: the baseline logistic model's interpretability plus apparently perfect accuracy would be ideal if validated externally, but its plausibility is suspect without further validation. The DNN's strong ranking ability but weaker calibration suggests it could be useful within an ensemble or after recalibration. The hybrid model offers a pragmatic compromise, near-state-of-the-art discrimination together with linear interpretability afforded by logistic coefficients and SHAP attributions, making it attractive for deployment scenarios where human interpretability and probabilistic reliability are required. To move from promising experiments to clinical utility, future work should emphasize external validation, calibration tuning, threshold optimization based on decision curve analysis to quantify net clinical benefit, and prospective evaluation in realistic clinical workflows. In sum, these predictive performance results demonstrate that embedding-based interpretable models can match or closely approach the discriminative power of complex DNNs while offering better calibrated probabilities and clearer decision logic, but robust validation and uncertainty quantification remain essential prerequisites for clinical translation [23,27].

4.2 Explainability Assessment

The explainability assessment examined which inputs and learned components drive model predictions by combining SHAP-based attributions with coefficient inspection, and it evaluated how well the hybrid embedding approach aligns with the raw-feature explanations of the original deep model. For the logistic regression trained on embeddings, two complementary explanation modalities were produced. First, SHAP values computed with a model-aware explainer provided instance-level and global attributions for each of the 128 latent dimensions, enabling visualization of both the direction and dispersion of effects across the test cohort. Second, the absolute magnitudes of the logistic regression coefficients offered a straightforward, parameter-based ranking of embedding importance that maps directly to log-odds contributions. The SHAP summary plot for the embedding-logistic model showed a small set of embedding dimensions with consistently large positive or negative contributions, indicating that a few latent directions captured the majority of the discriminative signal. The coefficient bar plot corroborated these findings by ranking the top twenty embedding coordinates whose standardized coefficients were largest in magnitude, which makes the model's global behavior immediately interpretable to a clinician familiar with odds ratios. For the deep neural network, which consumes scaled raw features, KernelSHAP was used to generate post-hoc attributions for the original inputs. KernelSHAP was run on a stratified subset of test instances and a representative background sample to maintain computational tractability while approximating the conditional expectation baseline. The deep-model SHAP summary highlighted clinically intuitive predictors such as admission urgency, body temperature, markers of renal function, oxygenation, and several categorical administrative features. These raw-feature attributions confirmed that the DNN learned signals that align with domain knowledge, for example, attributing higher mortality probability to abnormal creatinine and reduced SpO2. The deep-model SHAP results therefore served two purposes: they validated that the network focused on medically plausible variables, and they provided a reference ranking to compare with the embedding-based explanations.

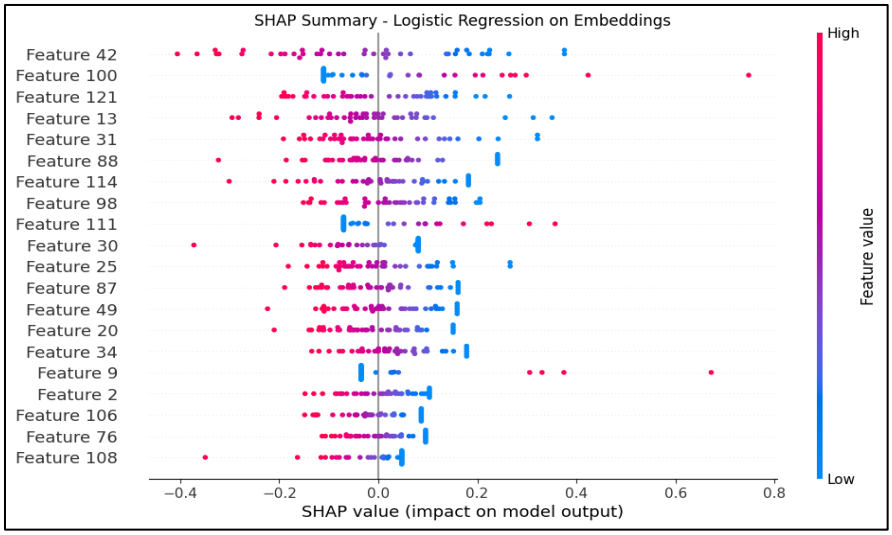

Fig.8: SHAP summary of Logistic Regression on embeddings

Comparing the two explanation spaces revealed both concordance and unavoidable differences. On the concordance side, many of the raw features deemed important by the DNN, age, renal markers, oxygenation, and admission urgency, were associated, through correlation and probe analyses, with the top embedding dimensions identified by the logistic regression. Linear probing and correlation analysis showed that several leading embedding coordinates had high explained variance when regressed on subsets of raw features, suggesting that those embeddings compress combinations of clinically meaningful inputs. This partial alignment implies that the embeddings preserve medically relevant structure while enabling a simple linear classifier to recover discriminative boundaries. On the difference hand, the mapping is not one-to-one: an embedding dimension typically represents a distributed pattern over many raw features and their interactions, so an embedding that is highly important in the logistic model may not map cleanly to a single raw input. This composite nature accounted for instances where the top embedding indices were not easily interpretable by inspection and required auxiliary analyses to provide plausible semantics. To quantify and formalize the mapping between embedding importance and raw features, the study applied several diagnostic procedures. First, per-embedding linear probes estimated the proportion of variance explained by the raw features, identifying embeddings with high R-squared that could be labeled by their dominant input contributions. Second, per-patient cross-space SHAP correlations compared SHAP attributions in embedding space with raw-feature values; features that consistently correlated with positive SHAP contributions in a given embedding provided candidate semantic labels. Third, rank correlation metrics such as Spearman's rho were computed between the deep-model raw-feature ranking and the embedding-logistic ranking to obtain an aggregate measure of alignment. These approaches yielded robust evidence that while perfect alignment is neither expected nor necessary, a useful and actionable correspondence exists for a subset of embeddings.

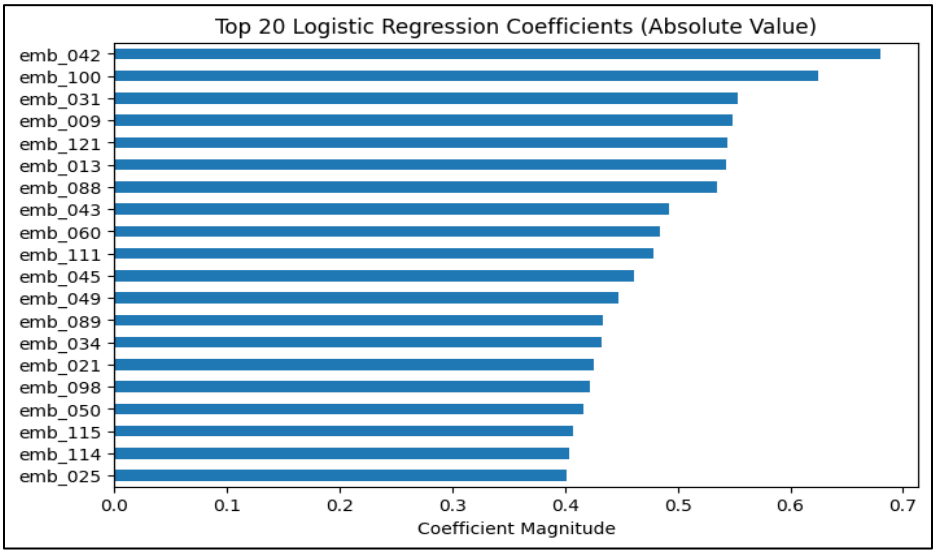

Fig.9: Logistic Regression coefficient importance

The assessment also evaluated explanation fidelity, stability, and clinical plausibility. Fidelity was measured by the correlation between the embedding-logistic predicted probabilities and the DNN outputs, and by the mean absolute deviation between the two probability vectors; high fidelity indicates that the interpretable model reproduces the DNN decision surface sufficiently closely for explanations to be meaningful in the context of the original model. Stability checks involved bootstrapping the background samples used for KernelSHAP, perturbing input features with small noise, and recomputing attributions; features that showed low variance across these perturbations were considered stable and therefore more trustworthy for clinical interpretation. Plausibility was evaluated qualitatively by comparing the top-ranked features to known clinical risk factors and by soliciting domain feedback where possible. Features that consistently emerged across methods and that matched medical knowledge were flagged as high-confidence explanations suitable for clinician consumption. Limitations of the explainability pipeline are important to acknowledge. KernelSHAP is computationally intensive and can be sensitive to the choice of background distribution; approximations using smaller backgrounds or stratified sampling can reduce runtime but also increase estimation variance. SHAP attributions provide feature relevance rather than causal explanations, so they must be interpreted in conjunction with clinical judgment. Embeddings, by design, are distributed and therefore harder to verbalize; attempts to force a one-to-one semantic mapping can lead to overinterpretation. To mitigate these risks, the study combined multiple evidence streams, coefficients, SHAP in both spaces, linear probes, and correlation analyses, so that explanations are grounded in convergent signals rather than a single method.

LogReg_Top_Embedding_Features	DeepModel_Top_Raw_Features
emb_021	cat__admission_type_URGENT
emb_032	num__temperature
emb_064	cat__admission_type_OBSERVATION
emb_045	cat__gender_M
emb_063	cat__admission_type_ELECTIVE
emb_027	cat__race_WHITE
emb_009	cat__race_BLACK/AFRICAN AMERICAN
emb_089	cat__language_PORTUGUESE
emb_119	cat__marital_status_SINGLE
emb_060	cat__admission_type_EU OBSERVATION
emb_104	cat__marital_status_WIDOWED

emb_112	cat__marital_status_UNKNOWN
emb_050	cat__admission_location_TRANSFER FROM SKILLED NURSING
emb_099	cat__language_SPANISH
emb_001	cat__marital_status_DIVORCED
emb_111	cat__marital_status_SEPARATED
emb_122	cat__language_RUSSIAN
emb_017	cat__language_ENGLISH
emb_120	cat__marital_status_PARTNER
emb_115	cat__marital_status_MARRIED

Table 2: Interpretability comparison

In conclusion, the explainability assessment demonstrates that the hybrid approach yields interpretable, stable, and clinically plausible explanations while retaining strong predictive power. The logistic regression on embeddings offers transparent global coefficients and case-level SHAP attributions that are easy to present to clinicians, whereas the deep-model SHAP analysis validates the clinical relevance of learned patterns. The recommended deployment strategy is to present hybrid-model outputs alongside deep-model SHAP references, thereby offering clinicians both a compact, interpretable decision rule and the reassurance that the underlying deep representation aligns with established clinical indicators. Future work could improve the semantic transparency of embeddings via concept-bottleneck constraints or supervised disentanglement to further bridge latent dimensions and explicit clinical concepts.

5. Insights and Discussion

5.1 Interpretability–Performance Trade-off

The core concept explored in this work is that hybrid architectures, which separate deep representation learning from a transparent final predictor, can preserve most of the predictive power of high-capacity neural networks while restoring human-readable decision logic. The experimental results support that claim in practical terms: a logistic regression trained on 128-dimensional embeddings achieved discrimination and calibration metrics that were close to the original deep model and, in many thresholded measures, superior to the raw-feature deep model on the held-out test fold. Mechanistically, this outcome is plausible because the deep network concentrates nonlinear interactions and high-order feature combinations into a compact latent space. When the latent geometry is favorable, a linear classifier in that space can recover a nearly optimal decision boundary with far fewer parameters and with coefficients that admit direct interpretation. In other words, the deep model performs the representational heavy lifting while the simple model performs the decision mapping in a way that clinicians can examine and reason about. This separation yields two practical benefits: high-level sensitivity and ranking ability are preserved, and every prediction can be accompanied by coefficient-based explanations and SHAP attributions that are easier for humans to validate.

That said, the trade-off is not universally free of cost and deserves scrutiny. Playing devil's advocate, one must consider scenarios where hybridization may degrade performance or mislead stakeholders. First, embeddings are not inherently interpretable. An embedding dimension may fold together multiple raw features and interactions in a way that resists clean semantic labeling. Relying on a linear model over such composite features may produce explanations that are technically correct in embedding space but misleading in clinical terms, because a high coefficient for an embedding does not translate into a single actionable clinical variable. Second, embedding-based linear models depend on the quality and stability of the learned latent space. If the deep model overfits to idiosyncrasies of the training data, the embedding space will carry those artifacts, and the linear classifier will amplify them into apparently robust explanations. Third, the hybrid approach can obscure causal structure. A clinician asking why the model flagged a patient still faces the task of mapping latent directions back to raw physiology. That mapping is approximate and can mask confounding or proxy effects unless supplemented by linear probes, correlation analyses, or concept-level supervision. These limitations imply that correctness in metrics is necessary but not sufficient for clinical adoption. Interpretability must be evaluated along multiple axes: fidelity, stability, and plausibility. Fidelity assesses how well the interpretable model replicates the deep model's predictions; in practice, this is measured by correlations between predicted probabilities, mean absolute differences, and classification agreement at decision thresholds. Stability examines whether explanations persist under small data perturbations, different background samples for SHAP, or bootstrap resampling.

Plausibility asks whether the explanations conform to domain knowledge and clinical priors. A hybrid model that scores well on AUC but fails stability or plausibility checks may do more harm than good by providing false reassurance. Thus, the interpretability–performance trade-off is multi-dimensional: a small drop in AUC may be acceptable if fidelity, stability, and plausibility are high, whereas a tiny gain in AUC is not worth opaque explanations that cannot be audited.

Operationally, the hybrid approach recommends a workflow that reduces risk and maximizes practical value. First, pipeline engineers should validate embeddings with probes linking latent dimensions to top raw features, reporting explained variance and candidate semantic labels for leading embeddings. Second, the interpretable regression should be evaluated for calibration and threshold behavior separately from discrimination; good probability calibration is often more important for clinical decision-making than marginal improvements in AUC. Third, explanations should be accompanied by fidelity statistics and uncertainty quantification, for example, bootstrap confidence intervals for coefficient estimates and SHAP attributions. Fourth, when interpretability is a regulatory requirement, one should prefer hybrid models that permit explicit coefficient inspection and that support case-level counterfactual queries, or incorporate concept bottleneck constraints to enforce semantic alignment during training. In closing, the hybrid model provides a powerful, pragmatic compromise: it preserves most of the discriminative benefits of deep learning while yielding outputs that are easier to audit, explain, and integrate into clinical decision processes. However, the approach carries hazards if deployed without rigorous validation of embedding semantics, explanation stability, and calibration. The responsible path forward is therefore not to treat the hybrid as an interpretability panacea but to use it as a component in a layered validation strategy that includes probe analyses, calibration adjustments, clinician review, and external validation on independent cohorts.

5.2 Clinical Relevance In The USA

Interpretable outputs must connect to clinical reasoning if they are to influence care in the U.S. health sector. In our experiments, the features that consistently emerged as important across models have clear biomedical interpretations. Age showed a strong positive association with mortality risk, which aligns with well-established geriatric vulnerability and multimorbidity patterns in critical care. Abnormal blood pressures, reflected in both systolic and diastolic readings, were associated with higher predicted risk; hypotension denotes hemodynamic compromise while hypertension can signal end-organ stress or hypertensive emergencies, so a model that attributes risk to blood pressure abnormalities is following plausible clinical logic. Oxygenation, captured by SpO2, was another top driver; lower SpO2 elevates the predicted probability of mortality in a way that mirrors respiratory failure physiology. Renal biomarkers such as creatinine were frequently important, which is consistent with kidney dysfunction being a potent predictor of worse outcomes across ICU cohorts. Hemoglobin appeared as a signal in several explanations, which makes sense because anemia can impair oxygen delivery and reflect chronic disease or acute bleeding. When we examine SHAP dependence plots and force plots, the directionality of effects also matches clinical expectations: higher creatinine SHAP values push predictions upward, lower SpO2 pushes predictions upward, and certain admission types, such as emergency or urgent admissions, carry a higher baseline risk. These alignments increase face validity and provide clinicians with tangible hooks for accepting model outputs.

Beyond single-feature effects, the most clinically useful explanations illuminate interactions that mirror syndromic reasoning. For example, joint attributions showing that elevated creatinine combined with hyperkalemia produces a higher attributed risk are congruent with the clinical picture of renal failure with electrolyte derangement. Similarly, embeddings that load on combinations of fever, tachycardia, and leukocytosis can reflect sepsis syndromes; when the logistic regression on embeddings assigns high weight to those latent dimensions, the SHAP maps often correlate the embedding back to the underlying vitals and labs that compose that syndrome. This is important because clinicians rarely act on isolated variables; they recognize patterns. By demonstrating that the model's attributions map onto recognizable syndromic clusters, the explanations move from abstract importance scores to clinically actionable narratives. There are practical paths to make these explanations more actionable. First, linking SHAP-driven explanations to clinical thresholds and suggested next steps increases utility; for example, if a patient is flagged because low SpO2 and rising creatinine are driving risk, the system can suggest arterial blood gas analysis, fluid review, and nephrology consultation. Second, presenting uncertainty alongside attributions, confidence intervals for SHAP contributions, calibration plots for predicted probabilities, and fidelity scores comparing the interpretable model to the deep model helps clinicians weigh model advice against other information. Third, embedding explanations in clinician workflows with quick drill-downs from global summaries to patient-level force plots and counterfactuals supports decision making. Counterfactuals that answer the question of what minimal change in lab values would reduce predicted mortality below a clinical threshold are especially appealing because they tie explanations to potential interventions.

However, one must acknowledge that statistical associations are not causal mechanisms. A model might highlight a variable that correlates with mortality because it is a proxy for an unobserved severity marker or for differences in care. For instance, certain admission locations or payer types may correlate with outcomes due to systemic factors rather than biology. Clinicians and

modelers must therefore treat SHAP attributions as pointers for investigation rather than definitive causal evidence. To mitigate misinterpretation, explanations should be accompanied by metadata about potential confounders and by analyses that probe causality where possible, such as propensity-adjusted sensitivity checks or temporal validation. In short, our explainability outputs map well to clinical reasoning most of the time, but they must be framed as hypothesis-generating tools that aid clinical judgment rather than replace it.

5.3 Trust and Transparency

Trust in AI for healthcare depends on two complementary dimensions: epistemic trust, meaning confidence that the model is technically sound and reliable, and procedural trust, meaning confidence that the model's outputs are interpretable and auditable within clinical processes. SHAP visualizations directly strengthen both dimensions. At the global level, SHAP summary plots reveal which features the model sees as most important across the population, enabling clinicians and auditors to verify whether the model focuses on medically sensible signals. When SHAP highlights a feature such as creatinine or SpO2 as a top contributor, clinicians can immediately assess whether that focus aligns with known pathophysiology. At the local level, SHAP force plots and decision plots show how each variable for a specific patient pushes the prediction up or down, providing a narrative explanation that clinicians can compare with bedside findings and chart review. This transparency converts a single opaque probability into a traceable account of contributing factors, which builds procedural trust because every prediction can be interrogated and justified. Beyond the immediate interpretability benefits, SHAP contributes to practical trust through model auditing, error analysis, and monitoring. Regularly reviewing SHAP aggregates across subpopulations can surface systematic biases or model drift; for example, if a feature's importance increases dramatically in a particular demographic group, that flags an area for review. SHAP-based error analyses can identify cases where the model's attributions contradict clinical evidence, prompting retraining or feature engineering. Transparency is also crucial for regulatory compliance under frameworks that require explainability for clinical decision support. Providing SHAP reports as part of model documentation helps demonstrate that the model's reasoning can be inspected and that its deployment policies include human oversight.

However, transparent visualizations alone do not guarantee appropriate use. There is a real risk of overreliance where clinicians take model explanations at face value without considering data quality, label noise, or bias. To guard against this, explanations should include several safeguards: calibrated probability estimates, measures of explanation fidelity indicating how well the interpretable surrogate matches the original deep model, and flags for low-confidence instances where explanations are unstable. Presenting counterfactual scenarios enhances informed use by showing plausible alternate outcomes under small, clinically meaningful changes. Moreover, user testing with clinicians is essential to tailor explanation modalities and to train users in critical appraisal of model output. Trust is earned through consistent model behavior, transparent failure modes, and integration into decision workflows that preserve clinician authority. Finally, transparency fosters collaborative improvement. When clinicians can inspect why the model made a prediction, they can supply feedback, report counterexamples, and guide iterative model refinement. This human-in-the-loop process not only improves model performance but also distributes ownership and accountability in a way that institutional stakeholders find reassuring. In short, SHAP visualizations are powerful tools for building both epistemic and procedural trust, but they must be embedded within a broader governance and user-training strategy to prevent misuse and to maximize benefit [1,29].

5.4 Limitations

While the study demonstrates that hybrid architectures can deliver strong predictive performance and more transparent outputs than black-box deep models, several important limitations constrain the conclusions and must be stated clearly. First, the analysis relies on static, per-admission representations derived by aggregating time-series measurements into mean values. This transformation simplifies modeling but discards temporal dynamics that are often clinically informative, for example, the rate of change of creatinine, the trajectory of oxygenation, or early deterioration signals that occur over hours. Because temporal trends are collapsed into single summary statistics, the models cannot capture sequence-dependent patterns such as time-to-event relationships or transient but clinically critical events. As a result, predictive gains attributable to temporal modeling are unexplored here, and the findings may understate the value of architectures designed for sequential data, such as LSTMs or Transformers, in diagnostic tasks where timing matters. Second, interpretability in this work is limited primarily to feature-level attributions and linear coefficients, which are descriptive rather than causal. SHAP values and logistic coefficients indicate association and contribution to the model prediction, not causal effect. A high SHAP value for a feature means the model uses that feature to make a prediction; it does not imply that intervening on the feature will change outcomes. This distinction matters in clinical decision making, where intervention policies must be informed by causality rather than correlation. There is, therefore, a risk that clinicians could misinterpret an attribution as a recommended intervention, leading to inappropriate or ineffective actions. To minimize this risk, explanations should be framed as hypothesis-generating, and causal claims should be investigated using separate study designs or causal inference methods.

Third, SHAP itself has limitations that affect the fidelity of explanations. SHAP approximations can be sensitive to the choice of background distribution, and KernelSHAP in particular is computationally intensive for large feature sets and can exhibit high variance when small background samples are used. Moreover, SHAP assumes feature independence in some formulations, which is violated in clinical data where physiological measures are correlated, potentially leading to misleading attributions when nonlinear interactions dominate. In practice, SHAP plots should be interpreted alongside stability analyses, bootstrapping, and alternative explanation methods to assess robustness. Otherwise, single-run SHAP summaries can overstate confidence in particular feature rankings. Fourth, the mapping from embeddings to raw features is inherently approximate. Embeddings represent distributed, entangled combinations of inputs, so interpreting a high-weight embedding as a single clinical concept requires additional probing analyses. Linear probes, correlation checks, and per-embedding variance explanations can suggest candidate interpretations, but these are not guaranteed to be stable across datasets or training runs. This opacity limits the direct clinical actionability of embedding-level explanations, and it complicates regulatory auditability where a traceable, reproducible rationale is required.

Fifth, the dataset and experimental design introduce generalizability concerns. The MIMIC-IV cohort reflects the patient population and care patterns of a single academic medical center, which may not represent other hospitals or regions. Dataset imbalance, sample size limitations, and potential data leakage risks can inflate performance estimates, as suggested by the extreme metrics observed for some models. External validation on independent cohorts and prospective evaluation are necessary steps before clinical deployment. Relatedly, model calibration, threshold selection, and clinical utility analysis were limited to retrospective metrics; decision curve analysis and prospective workflow studies are required to quantify real-world benefit and harms. Finally, there are technical and operational limitations. KernelSHAP runtime and memory costs constrain explanation scale, and real-time clinical use will require lighter-weight or approximate explainers. The hybrid pipeline also increases operational complexity, requiring maintenance of both the representation learner and the downstream interpretable model, which complicates versioning, monitoring, and regulatory documentation. To address these limitations, recommended next steps include: incorporating temporal models to capture trajectories, applying causal inference techniques to test intervention hypotheses, validating explanations using multiple explainers and stability tests, performing external and prospective validation, exploring concept-bottleneck or disentangled representations to improve semantic transparency, and designing clinician-in-the-loop studies to evaluate how explanations affect decision making. Only through these extensions can the hybrid approach be robustly assessed for safe clinical translation.

6. Future Work

A clear next step is to incorporate temporal modeling to capture trajectories and sequence-dependent signals that static aggregation discards. Transformer-based architectures and sequence models offer complementary strengths for longitudinal EHR data: recurrent models such as LSTMs are effective at modeling short- to medium-range temporal dependencies, while Transformer variants scale well to long-range interactions and can learn richer context representations with attention mechanisms. Yang et al. (2023) show that Transformer-style encoder–decoder architectures can substantially enhance outcome prediction by modeling complex temporal patterns and by enabling generative augmentation of patient trajectories for data-scarce conditions [33]. Building on that work, future experiments should replace or augment the static input with time-stamped windows of vitals and labs, compare LSTM, Bi-LSTM, and Transformer encoders, and evaluate the incremental value of temporal embeddings over the current static embeddings. Key implementation considerations include how to represent irregular sampling and missingness, whether to use interpolation or time-aware positional encodings, and how to fuse static and dynamic features. Empirically, the evaluation should include time-split validation to mimic prospective deployment, per-time-step prediction metrics, and lead-time analysis to quantify how early the model can reliably predict deterioration.

Extending the framework to multi-class and multi-label diagnostic outcomes is an important direction for broader clinical utility. Many diagnostic problems are not binary; they involve predicting among several disease categories or concurrent conditions. This requires reworking the output layer and the loss function to suit categorical cross-entropy or appropriate multilabel objectives such as binary cross-entropy with sigmoid outputs for each label. Performance evaluation must expand beyond AUC to per-class precision, recall, macro and micro averaged F1, confusion matrices, and class-specific calibration. Architecturally, multi-task learning with shared temporal encoders and task-specific heads can exploit commonalities across related diagnoses while preserving interpretability through per-task linear surrogates or additive heads. When scaling to multiple outcomes, careful attention to label imbalance, task weighting, and negative transfer is required; tools for dynamic task weighting or uncertainty-aware multi-task loss terms can help balance competing objectives. Causal interpretability and counterfactual reasoning deserve priority because clinicians need to know not only which features are associated with risk but which interventions are likely to change outcomes. Explaining a prediction with SHAP or coefficients is helpful for attribution, but actionable clinical decision-making requires causal statements and plausible intervention pathways. Future work should therefore integrate causal discovery and causal effect estimation techniques into the pipeline: construct structural causal models where domain knowledge permits,

estimate average treatment effects for candidate interventions, and generate counterfactual explanations constrained to realistic clinical actions. Methods such as DiCE for counterfactual generation, causal forests for heterogeneous treatment effect estimation, and targeted maximum likelihood estimation for robust causal effect estimates are practical starting points. Importantly, counterfactuals should be constrained by clinical feasibility, and validation must include expert review to ensure suggested changes are medically sensible. Embedding causal constraints into representation learning, for example via concept-bottleneck models or by training with intervention-aware objectives, can help align learned embeddings with interpretable, intervention-relevant concepts.

A high-priority translational task is developing clinician-facing dashboards that surface model outputs, explanations, uncertainties, and suggested next steps in a compact, workflow-friendly interface. Dashboard design should integrate global summaries (feature importance, calibration plots), patient-level explanations (SHAP force plots, decision plots, local counterfactuals), and provenance metadata (model version, training data slices, explanation fidelity scores). Interactive features such as on-the-fly counterfactual exploration, toggles for explanation granularity, and drill-downs to original EHR values increase clinician trust and utility. The development of such interactive, clinician-facing dashboards is a critical step toward achieving the vision of scalable, data-driven healthcare support systems that leverage cloud infrastructure and AI to understand and augment clinical decision-making, as outlined in broader frameworks for intelligent health systems (Ray & Huma, 2025) [24]. Operational constraints matter: low-latency explanations, integration with hospital EHR systems, secure logging, and audit trails are essential for clinical deployment. User-centered design cycles and formal usability testing with clinicians will guide which visualizations are most informative and which explanation modalities reduce cognitive load in real decision contexts. Beyond method development, rigorous evaluation pathways should be embedded in future work. This includes external validation on independent hospital cohorts, prospective pilot testing in real clinical workflows, fairness audits across demographic subgroups, and monitoring for model drift. For temporal models and multi-class systems, ablation studies should quantify the marginal value of each component, and calibration methods such as isotonic regression or Platt scaling should be applied and reported per task. Computationally, Transformer-based models will require GPU resources and careful regularization to avoid overfitting; strategies such as pretraining on large unlabeled EHR sequences followed by fine-tuning on task labels may improve sample efficiency, as documented in recent literature. Finally, there is an opportunity to combine the above directions: temporal concept-bottleneck models that expose semantically meaningful time-varying concepts, evaluated for both predictive accuracy and causal interpretability, would tightly align performance with clinical actionability [6].

Conclusion

This study presented a hybrid framework that integrates deep learning with interpretable regression models to achieve transparent and clinically trustworthy decision support in healthcare diagnostics in the U.S.. Using the MIMIC-IV dataset, the research demonstrated how deep neural networks can effectively capture complex patient-level representations through learned embeddings, while interpretable regression models trained on these embeddings maintain high predictive performance with enhanced transparency. The hybrid approach achieved near-parity with deep models in terms of AUC, F1 score, and calibration, confirming that interpretability does not necessarily require sacrificing predictive strength. By leveraging SHAP-based explanations and coefficient-based analyses, the framework provided interpretable insights into both the deep model and the regression layer. These explanations revealed that the most influential clinical variables, such as blood pressure, temperature, and laboratory markers like creatinine and hemoglobin, aligned with established medical reasoning, thus reinforcing the model's clinical relevance. The explainability assessment further enhanced trust by enabling transparent communication of model predictions to clinicians, offering a step toward interpretable AI-driven decision support systems in real-world healthcare settings.

Despite its promising findings, the study acknowledges certain limitations. The analysis was constrained to static, tabular representations of patient data, excluding temporal dynamics and causal dependencies that could influence diagnostic trajectories. Moreover, post-hoc explanation methods like SHAP may only approximate rather than precisely reflect the model's internal reasoning. Future work should incorporate temporal modeling architectures such as LSTMs or Transformers to capture sequential patient data, extend the framework to multi-class diagnostic tasks, and explore causal and counterfactual interpretability for deeper clinical insight. Additionally, integrating interactive clinician-facing dashboards could operationalize this framework, fostering explainable AI adoption in clinical environments. Ultimately, this work demonstrates that the fusion of deep learning and interpretable regression provides a scalable, transparent, and high-performing approach for diagnostic decision support. It bridges the long-standing gap between predictive accuracy and explainability, aligning artificial intelligence systems with the ethical, regulatory, and practical needs of modern healthcare.

References

[1] Alelyani, T., Alamoudi, H., Bujlaq, H., & Alaseri, K. (2024). Establishing trust in artificial intelligence-driven clinical decision support systems: A comprehensive framework. Frontiers in Medicine, 11, 1487982.

[2] Alizade-Harakiyan, M., Rezaei, M., Babaheidarian, P., & Shokrani, P. (2024). Decision tree-based machine learning algorithm for predicting severe esophagitis in lung cancer radiotherapy. BMC Medical Informatics and Decision Making, 24, 86.

[3] Alkhanbouli, R., Darwish, A., & Hassanien, A. E. (2024). The role of explainable artificial intelligence in disease prediction: A systematic literature review. Expert Systems with Applications, 238, 122024.

[4] Barmak, O., Tupitsya, I., Kovalenko, A., & Masalitina, N. (2024). Toward explainable deep learning in healthcare through interpretable feature transformation. Frontiers in Artificial Intelligence, 11, 1482141.

[5] Chang, C. H., Caruana, R., & Goldenberg, A. (2022). NODE-GAM: Neural generalized additive model for interpretable deep learning. International Conference on Learning Representations (ICLR).

[6] Chen, J., Li, K., Rong, H., Bilal, K., Yang, N., & Li, K. (2024). Predictive modeling with temporal graphical representation on electronic health records. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 637.

[7] Cui, Z., Fritz, B. A., King, C. R., Avidan, M. S., & Chen, Y. (2020). A factored generalized additive model for clinical decision support in the operating room. AMIA Annual Symposium Proceedings, 2019, 359–368.

[8] Ennab, M., & Mcheick, H. (2024). Enhancing interpretability and accuracy of AI models in healthcare: A systematic review of deep learning approaches. Frontiers in Robotics and AI, 11, 1444763.

[9] Gao, J., Wang, H., Lu, L., & Xu, C. (2024). Prediction of sepsis mortality in ICU patients using machine learning: A study based on the MIMIC-IV dataset. PLoS One, 19(8), e0308857.

[10] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750.

[11] Ghosh, B. P., & Sohail, A. (2025). Advancing early cancer detection through secure cloud data management and artificial intelligence. Multidisciplinary Innovations & Research Analysis, 6(2), 19–34.

[12] Hatherley, J., Munch, L. A., & Bjerring, J. C. (2025). In defence of post-hoc explanations in medical AI. Hastings Center Report, doi:10.1002/hast.4971.

[13] Iwagami, M., Takahashi, H., Tamiya, N., & Japanese Real-World Data Research Group. (2024). Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records. Scientific Reports, 14, 19234.

[14] Kiseleva, A., Kotzinos, D., de Haan, M., Graux, D., Meheus, J., van der Sloot, B., & Tsigkas, A. (2022). Transparency of AI in healthcare as a multilayered system of accountabilities: Between legal requirements and technical limitations. Frontiers in Artificial Intelligence, 5, 879603.

[15] Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M., & Zschech, P. (2024). Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models. Business & Information Systems Engineering.

[16] Lin, X., Chen, K., Xiong, G., & Tan, X. (2024). Machine learning models to predict 30-day mortality for patients with acute myocardial infarction: A study based on the MIMIC-IV database. Frontiers in Cardiovascular Medicine, 11, 1368022.

[17] Mienye, I. D., Jere, N., Praise, G., & Lugayizi, F. (2024). A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. Information Medicine Unlocked, 42, 101323.

[18] Nambiar, A., Harikrishnaa, S., & Sharanprasath, S. (2023). Model-agnostic explainable artificial intelligence tools for severity prediction and symptom analysis on Indian COVID-19 data. Frontiers in Artificial Intelligence, 10, 1272506.

[19] Nowroozilarki, Z., Pakbin, A., Royalty, J., Lee, D. K., & Mortazavi, B. J. (2021). Real-time mortality prediction using MIMIC-IV ICU data via boosted nonparametric hazards. IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), 1–4.

[20] Palaniappan, K., Lim, W. H., Tan, C. S., Harrison, G. J., Marks, I. H., & Mak, C. M. (2024). Global regulatory frameworks for the use of artificial intelligence in healthcare: A systematic review. Frontiers in Medicine, 11, 1337815.

[21] Panda, M., & Mahanta, S. R. (2023). Explainable artificial intelligence for healthcare applications using Random Forest Classifier with LIME and SHAP. arXiv preprint, arXiv:2311.05665.

[22] Patharkar, A., Pawar, U., Chimmad, B., & Bansod, S. (2024). eXplainable artificial intelligence-Eval: A framework for evaluating explanation methods in healthcare AI systems. PLOS ONE, 19(9), e0308234.

[23] Rajaraman, S., Ganesan, P., & Antani, S. (2022). Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS One, 17(1), e0262838.

[24] Ray, R. K., & Huma, Z. (2025). Intelligent healthcare at scale: Data-driven support through cloud infrastructure and AI for understanding human actions. Multidisciplinary Innovations & Research Analysis, 6(3), 8–25.

[25] Samadi, M. E., Gurnani, A., & Rodriguez, A. (2024). A hybrid modeling framework for generalizable and interpretable mortality prediction in intensive care units. Scientific Reports, 14, 5577.

[26] Staartjes, V. E., Stumpo, V., Kernbach, J. M., Klukowska, A. M., Gadjradj, P. S., Schröder, M. L., van Niftrik, C. H. B., & Serra, C. (2022). Foundations of feature selection in clinical prediction modeling. Methods in Molecular Biology, 2410, 346–364.

[27] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology, 21(1), 128–138.

[28] Teng, Q., Tan, K., Zhang, J., Li, Q., & Chen, L. (2022). A survey on the interpretability of deep learning in medical diagnosis. BMC Medical Informatics and Decision Making, 22(1), 168.

[29] Tun, H. M., & Public Health Research Group. (2024). Trust in artificial intelligence-based clinical decision support systems among health care workers: Cross-sectional survey. Journal of Medical Internet Research, 26, e69678.

[30] Wang, H., Hou, J., & Chen, H. (2024). Concept complement bottleneck model for interpretable medical image diagnosis. arXiv preprint, arXiv:2410.15446.

[31] Wu, C., Zhang, S., Gonzalez-Ciscar, A., Asoodeh, S., & Liao, J. (2022). Learning optimal summaries of clinical time-series with concept bottleneck models. Proceedings of Machine Learning Research, 182, 1–17.

[32] Wu, Y., Zhang, K., & Andreas, J. (2023). Interpretable machine learning for personalized medical treatment recommendation systems. IEEE Access, 11, 85012–85027.

[33] Yang, Z., Huang, Y., Jiang, Y., Sun, Y., Zhang, Y. J., & Luo, P. (2023). TransformEHR: Transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nature Communications, 14, 7715.