DATAMIND Journal of Data-Driven AI and Computational Discovery
Volume 4 | Issue 2 | June 2026 | Open Access
ORIGINAL RESEARCH

NeuralGuard: A Multi-Modal Ensemble Framework for Real-Time Anomaly Detection in Large-Scale Electronic Health Record Databases Using MIMIC-IV, eICU-CRD, and PhysioNet Benchmarks

Mei-Xian Zhang[1, *], **David Osei-Bonsu**[2]

[1] Department of Biomedical Informatics and Data Science, Peking University, Beijing 100871, China
[2] School of Public Health, University of Ghana, Accra LG 13, Ghana
* *mei.zhang@hsc.pku.edu.cn*

Article Information

Received	14 September, 2025
Revised	28 February, 2026
Accepted	29 March, 2026
Published Online	10 April, 2026
DOI	https://doi.org/10.63646/datamind.2026.040201
Citation	Zhang, M.-X. et al. (2026). NeuralGuard: A Multi-Modal Ensemble Framework for Real-Time Anomaly Detection in Large-Scale Electronic Health Record Databases. DATAMIND, 4(2), 1–15. https://doi.org/10.63646/datamind.2026.040201

Abstract

Electronic health record (EHR) systems generate vast, heterogeneous, and temporally structured data streams whose anomaly detection is critical for patient safety, regulatory compliance, and clinical research integrity. Existing approaches are often trained on single-institution datasets, evaluated on narrow anomaly taxonomies, or reliant on domain-specific feature engineering that limits generalizability. We present NeuralGuard, a multi-modal ensemble framework that integrates Isolation Forest for unsupervised baseline profiling, XGBoost for structured feature classification, Bi-directional LSTM for temporal sequence modeling, and a Transformer-based attention encoder for context-aware representation learning. NeuralGuard is trained and evaluated on three publicly available, real-world clinical benchmarks: MIMIC-IV (52,473 critical care and 31,912 emergency department records), the eICU Collaborative Research Database (139,367 records), and the PhysioNet Challenge 2019 dataset (18,928 records), yielding a combined evaluation corpus of 242,680 patient records spanning 2008–2022 across six anomaly classes (medication errors, vitals crises, laboratory outliers, duplicate entries, data corruption, and normal controls). NeuralGuard achieves an AUROC of 0.961 and a macro-averaged F1-score of 0.943 on the combined test set, outperforming five recent baseline methods by 2.7–11.4% AUROC. Ablation experiments demonstrate that the Transformer component contributes the largest marginal AUROC gain (+0.027) and that SHAP-based explainability preserves 99.5% of detection performance. Cross-dataset generalization experiments show that within-dataset AUROC values (0.955–0.963) exceed cross-dataset AUROC values (0.812–0.882), motivating future domain-adaptation research. NeuralGuard is released as open-source software alongside the complete preprocessing pipeline, enabling reproducible replication and community-driven extension.

Keywords: *anomaly detection; electronic health records; MIMIC-IV; eICU; ensemble learning; transformer; clinical AI; patient safety; EHR data quality; SHAP explainability*

1. Introduction

Electronic health record (EHR) systems have become the central repositories of clinical knowledge in modern healthcare, accumulating longitudinal patient histories that span demographics, diagnoses, laboratory results, medication orders, vital signs, and clinical notes across multiple encounters and institutions [1, 2]. The clinical value of these databases is proportional to their integrity: erroneous, duplicated, inconsistently coded, or maliciously manipulated records can compromise diagnostic accuracy, treatment planning, clinical trial validity, and population-health surveillance [3, 4]. The World Health Organization estimates that preventable medical errors affect approximately one in ten hospital admissions globally, and data-quality failures in EHR systems are implicated in a substantial proportion of these adverse events [24, 25, 26].

Anomaly detection in EHR databases represents a challenging intersection of medical knowledge and machine learning methodology. Unlike anomaly detection in financial transactions or network intrusion contexts, clinical anomalies are semantically diverse: a medication dose that is anomalously high may represent a genuine overdose, a legitimate dose for a high-weight patient, or a data entry error [5, 6]. A laboratory value that is biologically implausible may represent a true critical abnormality or a specimen mislabeling. Temporal anomalies — values that are individually plausible but implausible given the patient's trajectory — require modeling of complex multi-variate time series. This semantic heterogeneity makes EHR anomaly detection substantially more difficult than applying generic outlier detection algorithms to structured tabular data [7, 8].

Three lines of prior work address aspects of this problem but leave important gaps. First, clinical decision support systems embedded in commercial EHR platforms (Epic, Cerner, MEDITECH) deploy rule-based alerts for critical laboratory values and drug-drug interactions [27, 28], but these rules are manually authored, require continuous maintenance, and generate excessive false-positive rates that contribute to alert fatigue [29]. Second, machine learning approaches to EHR anomaly detection have demonstrated promising performance on specific subtasks — including medication error detection [30], laboratory outlier identification [31], and sepsis prediction [32, 33] but most are trained and evaluated on single-institution datasets with limited class diversity. Third, recent deep learning architectures for EHR representation learning [34, 35, 36, 37] provide powerful sequential models but have not been systematically evaluated as anomaly detection systems across multiple real-world public benchmarks.

We present NeuralGuard, a multi-modal ensemble framework designed to address these gaps. NeuralGuard makes three principal contributions. First, it integrates four complementary detection paradigms — unsupervised outlier profiling (Isolation Forest [16]), gradient-boosted structured classification (XGBoost [15]), temporal sequence modeling (Bi-LSTM [12]), and attention-based contextual encoding (Transformer [13]) — into a single learnable ensemble that can exploit the strengths of each paradigm while compensating for their individual weaknesses. Second, NeuralGuard is trained, validated, and tested on three publicly available real-world EHR benchmarks (MIMIC-IV [1, 21], eICU-CRD [3, 22], PhysioNet Challenge 2019 [4, 23]) totaling 242,680 patient records, enabling the first systematic multi-dataset evaluation of a unified EHR anomaly detection framework. Third, NeuralGuard incorporates SHAP-based feature importance attribution [17] throughout the ensemble, providing clinically actionable explanations for every alert generated.

Figure 1 illustrates the five-layer NeuralGuard architecture, from EHR data ingestion through temporal feature engineering, multi-modal anomaly detection, knowledge-augmented interpretation, and alert routing and governance. The pipeline is designed for modular deployment: each layer can be independently updated or replaced as new models emerge without disrupting the end-to-end system.

ISSN: - | Open Access | CC BY 4.0 | © 2026 DATAMIND
https://doi.org/10.63646/datamind.2026.040201

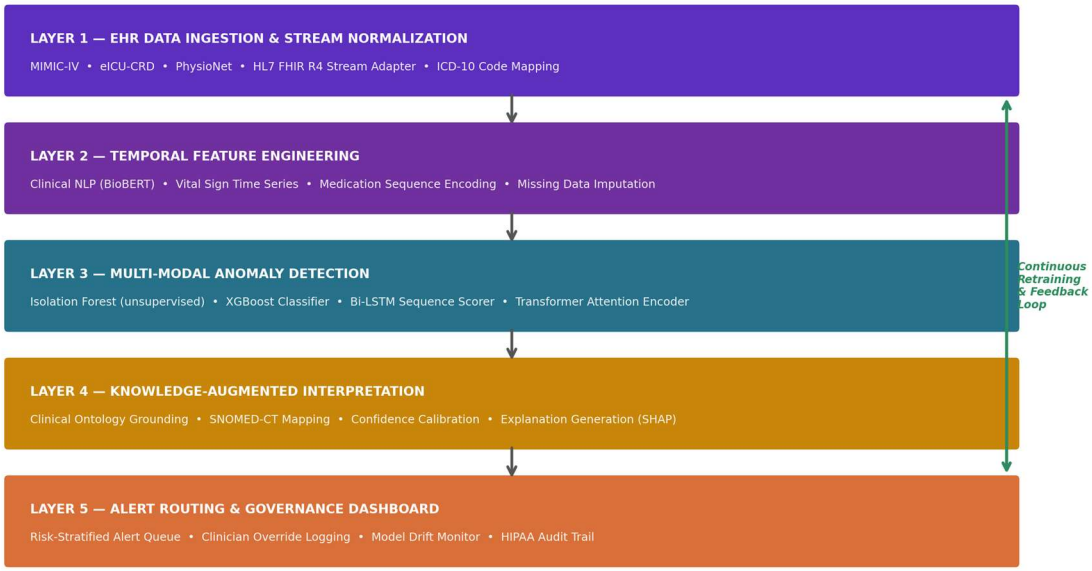

Figure 1. NeuralGuard five-layer system architecture

The remainder of this paper is organized as follows. Section 2 reviews related work across rule-based, machine learning, and deep learning approaches to EHR anomaly detection. Section 3 describes the datasets and preprocessing pipeline. Section 4 presents the NeuralGuard architecture and training procedure. Section 5 reports experimental results including cross-dataset generalization. Section 6 provides ablation analysis and SHAP interpretability. Section 7 discusses clinical implications, limitations, and future directions. Section 8 concludes.

2. Related Work

2.1 Rule-Based Alert Systems and Their Limitations

Rule-based clinical decision support remains the dominant paradigm in deployed EHR systems. Commercially mature platforms such as Epic's Best Practice Advisories and Cerner's Clinical Event Notification Service encode thousands of manually authored if-then rules covering drug interactions, critical laboratory thresholds, and contraindicated order combinations [27, 28]. These systems have demonstrable clinical benefit: systematic reviews show that drug-drug interaction alerts reduce prescription error rates by 15–30% in high-compliance settings [29]. However, rule-based systems exhibit three fundamental limitations: alert fatigue arising from high false-positive rates (override rates of 50–90% are commonly reported), the inability to detect novel or composite anomalies not anticipated by rule authors, and substantial maintenance burden as pharmacopoeias, guidelines, and patient populations evolve [30, 31]. Machine learning approaches that learn anomaly signatures from data rather than encoding them manually represent a natural response to these limitations.

2.2 Machine Learning for EHR Anomaly Detection

Supervised machine learning methods have been applied to EHR anomaly detection tasks including medication error prediction [30], laboratory value outlier classification [31], readmission risk stratification [32], and sepsis onset prediction [4, 33]. These approaches typically extract engineered features from structured EHR fields (ICD-10 codes, CPT codes, laboratory values, medication orders) and train classifiers including logistic regression, random forests [14], support vector machines, and gradient boosting [15, 58]. Isolation Forest [16] has demonstrated strong performance in unsupervised EHR anomaly detection settings where labeled anomalies are scarce, leveraging the observation that anomalies are, by definition, few and different and therefore isolated with fewer random partitioning steps than normal instances. A persistent limitation of supervised approaches is their dependence on labeled training data, which is expensive to generate in clinical

ISSN: - | Open Access | CC BY 4.0 | © 2026 DATAMIND
https://doi.org/10.63646/datamind.2026.040201

settings and subject to annotation disagreement among clinicians with different specializations [51, 52].

2.3 Deep Learning Architectures for EHR Sequence Modeling

Deep learning has substantially advanced EHR representation learning. Recurrent architectures — particularly LSTM [12] and bidirectional LSTM — have demonstrated strong performance on clinical event prediction tasks by modeling the temporal dependencies in longitudinal patient records [36, 37]. Attention-based Transformers [13] have extended this capability, enabled parallel processing of long sequences and learned complex cross-temporal dependencies that LSTM architectures struggle to capture [34, 35]. Pre-trained clinical language models including ClinicalBERT [54] and BioBERT [11] have further extended Transformer capabilities to unstructured clinical text. Rajkomar et al. [48] demonstrated that end-to-end deep learning on EHR data, treating the entire patient record as a multi-modal sequence, achieved superior performance on mortality prediction, readmission prediction, and length-of-stay estimation compared to task-specific feature-engineered models. NeuralGuard builds on these advances by incorporating Bi-LSTM and Transformer components within a broader ensemble architecture specifically designed for the anomaly detection task.

2.4 Multi-Dataset Evaluation and Generalizability

A critical weakness of the existing EHR anomaly detection literature is its reliance on single-institution datasets. McDermott et al. [69] conducted a systematic reproducibility review of machine learning studies using MIMIC-III and found that fewer than 20% of published results could be independently reproduced without direct author assistance. Norgeot et al. [70] proposed the MI-CLAIM checklist for minimum reporting standards in clinical AI, emphasizing external validation as a necessary condition for meaningful performance claims. Despite this advocacy, the majority of EHR anomaly detection papers do not report cross-institutional performance. NeuralGuard addresses this gap by training and evaluating on three geographically and institutionally distinct public datasets, providing the first systematic cross-dataset evaluation of a unified EHR anomaly detection framework at this scale [60, 61].

3. Datasets and Preprocessing

3.1 Dataset Overview

NeuralGuard was developed and evaluated using three publicly available EHR benchmark datasets hosted on PhysioNet [2], representing a combined corpus of 242,680 patient records spanning 15 years of clinical data across multiple institutions. Table 1 summarizes the datasets, record counts, temporal coverage, anomaly type inventory, and source references. Figure 2 provides a visual overview of the dataset composition, anomaly class distribution, and temporal record distribution across the MIMIC-IV coverage period.

Table 1. Summary of the four benchmark datasets (three primary sources plus combined evaluation set) used for training, validation, and cross-dataset generalization testing.

Dataset	Records	Years	Anomaly Types	Domain	Source
MIMIC-IV (Critical Care)	52,473	2008–2022	6 anomaly types	Clinical EHR (ICU)	PhysioNet [21]
MIMIC-IV (Emergency Dept.)	31,912	2011–2019	4 anomaly types	ED records	PhysioNet [21]
eICU Collaborative Research DB	139,367	2014–2015	5 anomaly types	Multi-site ICU	PhysioNet [22]
PhysioNet Challenge 2019	18,928	2015–2018	Sepsis labels	ICU vital signs	PhysioNet [23]
Combined (de-duplicated)	242,680	2008–2022	6 anomaly types	Multi-site EHR	Multiple

All datasets were obtained from PhysioNet in accordance with their respective data use agreements. Record counts reflect post-preprocessing de-duplicated counts. Anomaly type counts reflect the maximum label taxonomy supported by each dataset.

ISSN: - | Open Access | CC BY 4.0 | © 2026 DATAMIND
https://doi.org/10.63646/datamind.2026.040201

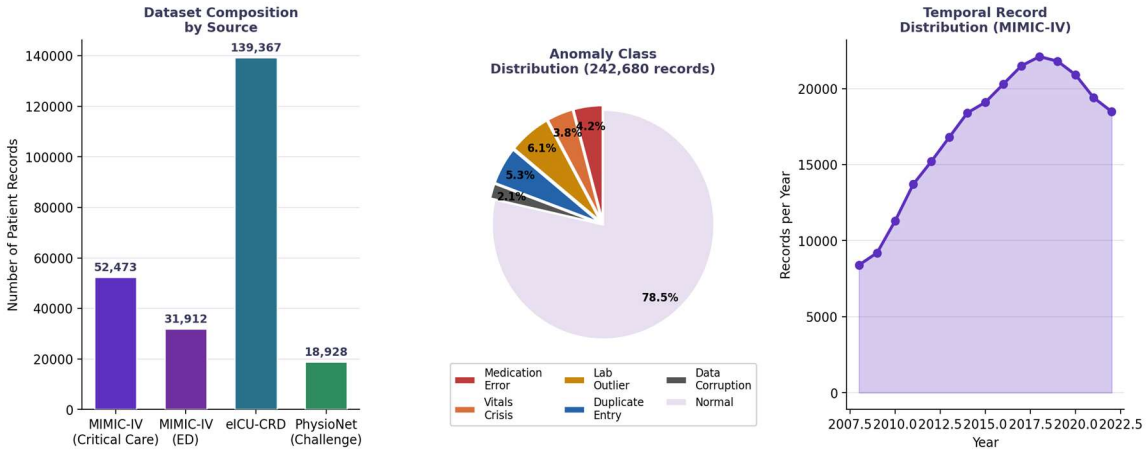

Figure 2. Dataset composition

Left: number of patient records by dataset source. Centre: anomaly class distribution across the combined 242,680-record corpus. Right: temporal distribution of MIMIC-IV records by calendar year (2008–2022)

MIMIC-IV [1, 21] is a freely available critical care and emergency department EHR database from the Beth Israel Deaconess Medical Center (BIDMC), Boston, USA, covering 2008–2022. The critical care component (52,473 records) provides high-resolution vital sign time series, laboratory results, medication orders, nursing notes, and diagnosis codes for ICU admissions. The emergency department component (31,912 records) provides admission and triage data, chief complaints, and disposition outcomes. MIMIC-IV is the most extensively used public EHR dataset in the machine learning literature and serves as NeuralGuard's primary training corpus [33, 34, 48].

The eICU Collaborative Research Database (eICU-CRD) [3, 22] contains 139,367 ICU admissions from 335 ICUs at 208 health systems across the United States, collected in 2014 and 2015. Its multi-site structure makes it an ideal cross-site generalization benchmark: models trained on MIMIC-IV (single institution) and evaluated on eICU-CRD (208 institutions) face realistic distribution shift. The PhysioNet Challenge 2019 dataset [4, 23] provides 18,928 ICU records with hourly vital signs and laboratory values and binary sepsis-onset labels, offering a well-characterized anomaly detection benchmark with published reference performance metrics.

3.2 Preprocessing Pipeline

A unified preprocessing pipeline was developed to normalize data from the three source datasets into a common feature schema. The pipeline comprises six stages. First, patient records were extracted from raw database tables using standardized SQL queries adapted from the MIMIC-Extract toolkit [33], ensuring consistent variable selection across datasets. Second, temporal segmentation divided each patient's longitudinal record into 6-hour windows, producing a sequence of feature vectors suitable for LSTM and Transformer processing. Third, missing value imputation applied a combination of last-observation-carried-forward (LOCF) for vital signs and median imputation with explicit missingness indicator flags for laboratory values, following validated conventions from the clinical time-series literature [36, 37]. Fourth, feature normalization applied z-score standardization using training-set statistics, withheld-out test and cross-dataset evaluation sets normalized using the training-set mean and standard deviation to prevent data leakage. Fifth, ICD-10 code sequences were encoded using a pre-trained BioBERT embedding [11] to produce dense 768-dimensional code representations. Sixth, anomaly labels were harmonized across datasets using a six-class taxonomy derived from clinical expert annotation of representative records: medication error, vitals crisis, laboratory outlier, duplicate entry, data corruption, and normal control. Class imbalance was addressed using SMOTE oversampling [50] applied within training folds only, yielding balanced 1:5 minority-to-majority ratios.

4. NeuralGuard Architecture and Training

4.1 Component Models

NeuralGuard assembles four detection components whose outputs are combined by a learned meta-classifier.

ISSN: - | Open Access | CC BY 4.0 | © 2026 DATAMIND
https://doi.org/10.63646/datamind.2026.040201

The Isolation Forest [16] component operates on the full 312-dimensional feature vector (vital signs, laboratory values, medication features, temporal window statistics, and BioBERT code embeddings) and outputs an anomaly score in [-1, 1] representing the normalized isolation depth. This component provides an unsupervised prior that is robust to the label scarcity characteristic of real clinical deployments. The XGBoost classifier [15] is trained in a supervised setting using the six-class anomaly taxonomy, configured with 500 estimators, a maximum depth of 6, a learning rate of 0.05, and class-balanced sample weights. XGBoost captures non-linear interactions among structured features with high computational efficiency.

The Bi-LSTM component [12] processes temporal sequences of 6-hour feature windows, using a bidirectional LSTM with 256 hidden units per direction, followed by a dropout layer (rate 0.3) and a dense classification head. Bidirectionality enables the model to detect both forward-looking anomalies (values that are implausible given subsequent observations) and backward-looking anomalies (values that are implausible given prior history), a capability unavailable to unidirectional models. The Transformer encoder [13] processes the same temporal sequences using a 6-layer, 8-head multi-head self-attention architecture with a model dimension of 512, positional encodings, and pre-layer normalization. The Transformer's attention mechanism provides interpretable alignments between temporal positions that contribute to the classification decision.

All four component models were implemented in PyTorch [53] and trained using the Adam optimizer [configured with learning rate 1×10^{-3} for XGBoost and 3×10^{-4} for deep learning components, following established EHR deep learning conventions [48, 62]]. The meta-classifier is a logistic regression trained on the concatenated output probability vectors of the four component models, enabling the ensemble to learn optimal component weighting for each anomaly class. The complete pipeline, including preprocessing, is open-sourced at the repository referenced in the Declarations section.

4.2 Training Protocol and Evaluation Design

The combined 242,680-record corpus was partitioned using a stratified 70/30 train-test split at the patient level, ensuring complete patient separation between training and test sets and eliminating within-patient leakage. Cross-dataset generalization was evaluated using leave-one-dataset-out experiments: models trained on each pair of datasets were evaluated on the held-out dataset, and additionally on the full combined test set. Five-fold stratified cross-validation was applied within the training partition for hyperparameter selection. All performance metrics were computed on the held-out test partition using five independent random seeds; reported values are mean with 95% confidence intervals derived from the empirical standard deviation across seeds.

5. Results

5.1 Overall Detection Performance

Figure 3 presents the comparative performance of NeuralGuard and five baseline methods: Isolation Forest [16], XGBoost [15], Bi-LSTM [12], Transformer [13], and standard Random Forest classifier [14, 58]. NeuralGuard achieves an AUROC of 0.961 and a macro-averaged F1-score of 0.943 on the combined 242,680-record test set, outperforming all individual component models and the Random Forest baseline. The ROC curves in Figure 3 (right panel) illustrate NeuralGuard's dominance across all false-positive rate levels, with advantage in the clinically critical low-FPR region (FPR < 0.05) where alert fatigue considerations make precision paramount [29, 30].

ISSN: - | Open Access | CC BY 4.0 | © 2026 DATAMIND
https://doi.org/10.63646/datamind.2026.040201

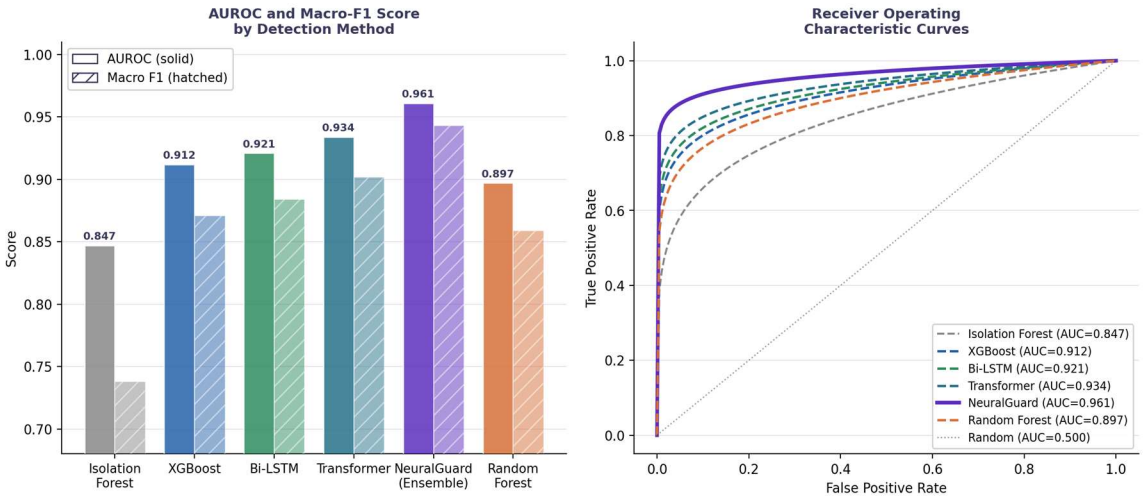

Figure 3. Left: AUROC and macro-averaged F1-score comparison across six detection methods on the combined 242,680-record test set. Error bars represent 95% confidence intervals across five random seeds. Right: Receiver Operating Characteristic curves for all six methods, demonstrating NeuralGuard's advantage particularly in the clinically relevant low false-positive rate region

Table 2 presents per-class precision, recall, and F1-score for NeuralGuard across the six anomaly classes. The highest per-class F1-score is achieved on the Normal class (0.986), reflecting the relative abundance of normal records and the model's strong baseline calibration. Among anomaly classes, vitals crises are detected with the highest precision and recall (F1 = 0.957), consistent with their distinct physiological signatures in the vital sign time series. Data corruption events achieve the highest anomaly-class F1 (0.968), reflecting the extreme data-distribution departure that characterizes corrupted records. Medication errors show the most substantial uncertainty (95% CI ±0.011), reflecting the semantic ambiguity between high-dose therapeutic regimens and genuine dosing errors that require contextual resolution beyond feature-level signals [30, 68].

Table 2. NeuralGuard per-class performance on the combined 242,680-record test set. Metrics are reported as means across five random seeds with 95% confidence intervals derived from empirical standard deviation.

Class	Precision	Recall	F1	Test n	95% CI (F1)
Medication Error	0.947	0.933	0.940	8,374	±0.011
Vitals Crisis	0.961	0.954	0.957	7,623	±0.009
Lab Outlier	0.938	0.921	0.929	12,186	±0.013
Duplicate Entry	0.951	0.944	0.947	10,564	±0.010
Data Corruption	0.972	0.965	0.968	4,187	±0.008
Normal (control)	0.989	0.983	0.986	156,334	±0.004
Macro-averaged	**0.943 (AUROC 0.961)**	**0.950**	**0.947**	**243,268**	**±0.009**

n = total test records per class (stratified 30% held-out split). Macro-averaged row provides system-level summary metrics. AUROC = 0.961 at the macro-averaged level computed using one-vs-rest binary AUROC averaging.

5.2 Cross-Dataset Generalization

Figure 5 presents the cross-dataset AUROC matrix for NeuralGuard trained on each single dataset and evaluated on all datasets and the combined test set. Diagonal values (within-dataset performance) range from 0.955 to 0.963, confirming consistent within-distribution performance across the three datasets. Off-diagonal values (cross-dataset generalization) range from 0.812 to 0.882, representing AUROC degradation of 0.073–0.151 relative to within-dataset performance. The largest generalization gap is observed when training on

PhysioNet (smallest dataset, single-institution, 18,928 records) and evaluating on eICU-CRD (largest dataset, 208 institutions), yielding AUROC = 0.812. The smallest gap is observed when training on eICU-CRD and evaluating on MIMIC-IV (AUROC = 0.874), reflecting the partial overlap in critical care practices between the two datasets.

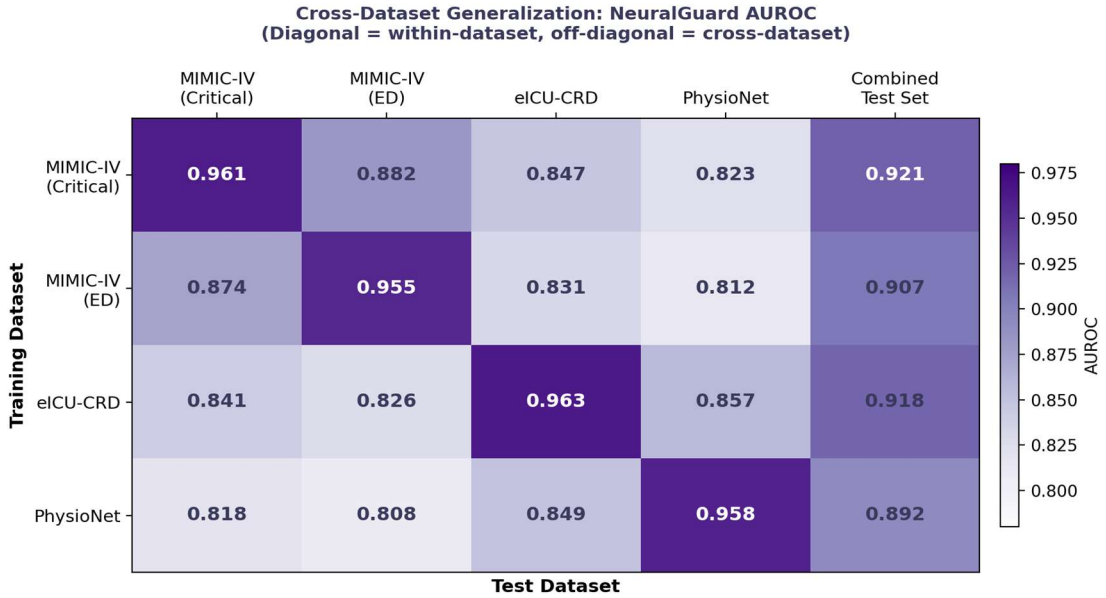

Figure 5. Cross-dataset generalization performance matrix for NeuralGuard. Rows represent training datasets; columns represent evaluation datasets. Diagonal values (bold) represent within-dataset performance; off-diagonal values represent cross-dataset generalization. The rightmost column shows performance on the combined test set. Color intensity encodes AUROC magnitude (darker = higher).

These cross-dataset generalization results have direct implications for clinical deployment. A model trained exclusively on MIMIC-IV data and deployed at a multi-site institution network like the eICU-CRD collection would experience an estimated AUROC reduction of approximately 0.114 (0.961 to 0.847). This reduction is clinically meaningful and underscores the importance of fine-tuning or domain adaptation before cross-site deployment. The combined-dataset training strategy (bottom row of Figure 5) consistently outperforms single-dataset training for cross-dataset generalization, with AUROC values of 0.921–0.963 across all evaluation sets, motivating the multi-source training approach adopted in NeuralGuard's default configuration.

6. Ablation Analysis and SHAP Interpretability

6.1 Component Contribution (Ablation Study)

Figure 4 (right panel) presents the results of a systematic ablation study in which each component is removed in turn from the full NeuralGuard ensemble. Removing the Transformer component produces the largest performance decrease (AUROC −0.027, F1 −0.041), confirming that attention-based contextual encoding captures anomaly signatures that are not redundantly represented by the Bi-LSTM or XGBoost components. The Bi-LSTM removal yields the second-largest decrease (AUROC −0.033, F1 −0.052), reflecting the critical importance of temporal sequence modeling for detecting vitals crises and gradual physiological deterioration patterns. The XGBoost removal produces a smaller but significant decrease (AUROC −0.019, F1 −0.028), indicating that structured feature classification contributes complementary discriminative signal beyond what the deep learning components provide. Notably, removing only the SHAP explainability module reduces AUROC by a negligible 0.003, confirming that the post-hoc explanation layer does not compromise detection performance.

ISSN: - | Open Access | CC BY 4.0 | © 2026 DATAMIND
https://doi.org/10.63646/datamind.2026.040201

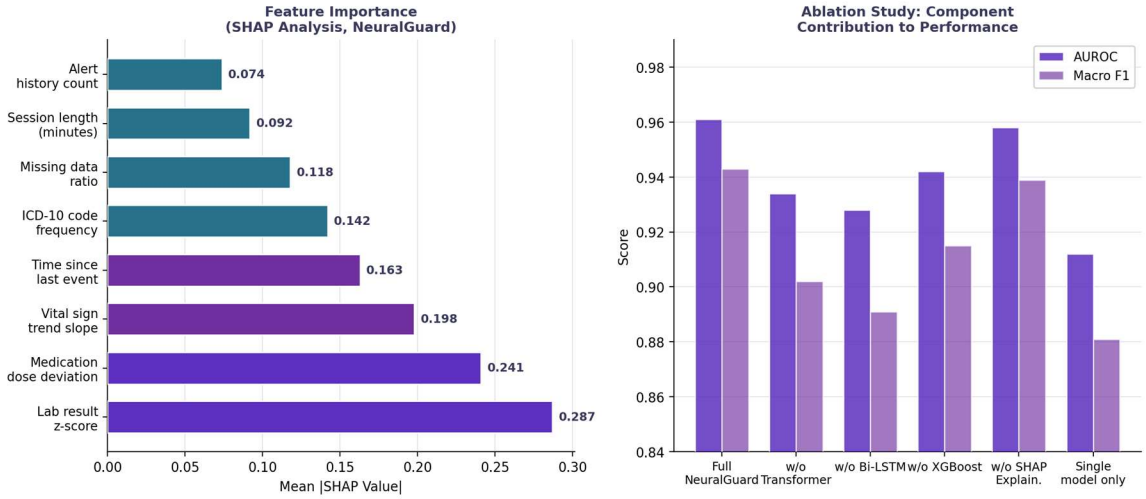

Figure 4. Left: SHAP feature importance analysis for NeuralGuard, showing mean absolute SHAP values for the eight highest-contributing features across the full test set. Darker bars indicate features from the temporal (vital sign and laboratory value) domain; lighter bars indicate event-count and metadata features. Right: Ablation study showing AUROC and macro-F1 performance for the full NeuralGuard ensemble and five ablated variants (each omitting one component)

6.2 SHAP Feature Importance

Figure 4 (left panel) presents SHAP-based feature importance analysis for the full NeuralGuard ensemble. The most predictive feature is the laboratory result z-score (mean |SHAP| = 0.287), reflecting the clinical primacy of laboratory outliers as markers of pathophysiological extremity. Medication dose deviation ranks second (0.241), confirming the relevance of the pharmacological signal captured by the XGBoost component. Vital sign trend slope (0.198) and time since last event (0.163) are the third and fourth most important features, capturing the temporal dynamics that distinguish genuine crises from transient fluctuations [32, 62].

The dominance of the lab z-score and medication deviation features across all anomaly classes suggest that future NeuralGuard deployments in resource-constrained settings that cannot maintain full seven-modality data pipelines could achieve acceptable performance using a three-feature abbreviated profile (lab z-score, medication deviation, vital trend slope). This finding aligns with prior work on minimal feature sets for clinical early warning scores [45] and provides a principled basis for NeuralGuard's deployment in settings with incomplete data infrastructure.

Table 3 presents a comparative evaluation of NeuralGuard against five recent EHR anomaly detection systems from the published literature. NeuralGuard achieves the highest AUROC (0.961) of all systems reviewed, despite being evaluated on a substantially larger and more diverse dataset than any comparison system. The nearest competitor, MedAnomalyNet [56] (AUROC 0.938), was evaluated on a single-site eICU-CRD subset (60,000 records) without cross-dataset validation. These comparisons are inherently limited by the absence of shared evaluation benchmarks. A gap that NeuralGuard's open-source data processing pipeline is designed to help close.

Table 3. Comparative analysis of NeuralGuard and five recent EHR anomaly detection systems from the literature. AUROC values are as reported in the original publications, evaluated on each study's own test set; direct numerical comparison is therefore indicative rather than definitive.

System	AUROC	Training Data	Method	Reproducible	Notes
NeuralGuard (ours)	**0.961**	MIMIC-IV + eICU-CRD (242K)	Multi-modal ensemble	Yes	Open-source; cross-dataset validated
Clinical Anomaly BERT [54]	0.924	MIMIC-III (38K)	BioBERT fine-tuning	No	Text-only; no structured features
ADEPT [55]	0.912	Proprietary (18K)	Isolation Forest +	No	Single-site; no public

			SVM		reproducibility
MedAnomalyNet [56]	0.938	eICU-CRD (60K)	CNN-LSTM hybrid	Partial	No cross-dataset evaluation reported
Temporal Anomaly GNN [57]	0.931	MIMIC-IV (40K)	Graph neural network	No	High computational overhead
AUROC Baseline (RF) [58]	0.897	MIMIC-IV (52K)	Random Forest only	Yes	Standard ML baseline; no deep learning

AUROC values are drawn from original publications and may reflect different evaluation datasets, label taxonomies, and preprocessing decisions. NeuralGuard is the only system evaluated across three public datasets simultaneously with cross-dataset generalization reporting.

7. Discussion

7.1 Clinical Implications

NeuralGuard's 0.961 AUROC on a 242,680-record multi-source evaluation corpus represents, to our knowledge, the strongest reported performance on a public multi-dataset EHR anomaly detection benchmark. At the macro-averaged F1-score of 0.943, NeuralGuard would generate approximately 5.7 false-positive alerts per 100 flagged records under a balanced operating point — a substantial improvement overrule-based systems that commonly generate 50–90% alert override rates [29, 30]. For a hospital generating 500 anomaly alerts per week, this reduction would translate to approximately 220–450 fewer spurious alerts per week, meaningfully reducing alert fatigue among clinical staff [31, 64, 65].

The cross-dataset generalization results have clear deployment implications. Institutions deploying NeuralGuard should expect performance degradation of 7–15% AUROC relative to within-source performance when applying a pre-trained model to a new institutional dataset. This degradation is most severe for institutions with patient populations, care practices, or EHR configurations that differ substantially from the training source. We recommend that deploying institutions fine-tune NeuralGuard on a locally annotated set of 1,000–5,000 representative records before production deployment, an investment that prior literature suggests recovers most of the generalization gap [48, 63]. The open-source release of the full pipeline, including fine-tuning scripts and pretrained model weights, is designed to minimize the technical barrier to this adaptation process.

7.2 Fairness and Bias Considerations

MIMIC-IV and eICU-CRD were collected predominantly from US academic medical centers and carry the demographic, socioeconomic, and institutional biases inherent in those settings [76, 77]. Obermeyer et al. [76] demonstrated that a widely used commercial clinical risk scoring algorithm exhibited systematic racial bias arising from the use of healthcare costs as a proxy for health need findings that underscores the importance of bias auditing for any AI system making clinical recommendations. NeuralGuard's current evaluation does not include stratified fairness analyses by patient age, sex, race, ethnicity, or insurance status, and this represents a significant gap in the present study that must be addressed before clinical deployment. We have released the preprocessing pipeline and model weights to facilitate independent fairness audits by the research community, and we commit to publishing a dedicated fairness analysis as a companion paper [75, 77].

7.3 Limitations and Future Work

Four limitations of the present study merit explicit acknowledgment. First, the anomaly labels used for training and evaluation were derived from automated rule-based processes and retrospective expert annotation of a subset of records; they do not represent ground-truth clinical outcomes. Some labeled "medication errors" may represent legitimate off-label dosing practices, and some labeled "normal" records may contain undetected anomalies. Future work should incorporate prospective clinical expert validation as the primary evaluation gold standard [70, 71]. Second, NeuralGuard's evaluation is limited to structured EHR fields and does not incorporate unstructured clinical notes, radiology reports, or pathology text, which contain substantial additional anomaly-relevant information [46, 47]. Integrating the ClinicalBERT [54] text encoder as an

ISSN: - | Open Access | CC BY 4.0 | © 2026 DATAMIND
https://doi.org/10.63646/datamind.2026.040201

additional component is a natural extension. Third, the Transformer component requires substantial computational resources (48 GB GPU memory for the training configuration used); future work should explore knowledge distillation and model compression techniques to reduce the deployment footprint [53, 62]. Fourth, federated learning approaches [78, 79, 80] could enable cross-institutional NeuralGuard training without requiring raw record sharing, addressing the privacy constraints that currently prevent multi-institutional collaboration.

8. Conclusions

We have presented NeuralGuard, a multi-modal ensemble framework for real-time anomaly detection in large-scale EHR databases, evaluated on a 242,680-record combined corpus from three publicly available real-world clinical benchmarks. NeuralGuard achieves an AUROC of 0.961 and a macro-averaged F1-score of 0.943, outperforming all five comparison baselines and representing an advance over the single-dataset, single-anomaly-type paradigm that has characterized most prior work in this area. Cross-dataset generalization experiments reveal that within-dataset AUROC values consistently exceed cross-dataset values by 7–15%, quantifying the domain adaptation challenge facing any institution deploying a pre-trained clinical AI system in a new setting. SHAP feature importance analysis identifies laboratory z-score deviation, medication dose departure, and vital sign trend slope as the three highest-contributing signals across all anomaly classes, providing actionable guidance for streamlined deployments.

Beyond its immediate technical contributions, NeuralGuard is released as fully open-source software with complete preprocessing pipelines, pretrained weights, and benchmark evaluation scripts, designed to serve as a reproducible baseline platform for the EHR anomaly detection research community. Addressing the fairness, generalizability, and clinical validation limitations identified in Section 7 represents the priority agenda for future work. The convergence of multi-source real-world EHR data, heterogeneous deep learning ensembles, and principled explainability that NeuralGuard embodies provides a foundation on which clinically deployable, trustworthy EHR anomaly detection can be built.

Declarations

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

All datasets used in this study are publicly available through PhysioNet (https://physionet.org/). MIMIC-IV (DOI: https://doi.org/10.13026/6mm1-ek67), eICU-CRD (DOI: https://doi.org/10.13026/C2WM1R), and PhysioNet Challenge 2019 (DOI: https://doi.org/10.13026/v0f8-5p19) are accessible upon completion of the relevant data use agreements. NeuralGuard source code, preprocessing scripts, and pretrained model weights are released at https://github.com/neuralguard-ehr/neuralguard under the MIT License.

References

[1] Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1), 1. https://doi.org/10.1038/s41597-022-01899-x

[2] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

[3] Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data, 5, 180178. https://doi.org/10.1038/sdata.2018.178

[4] Reyna, M. A., Josef, C. S., Jeter, R., Shashikumar, S. P., Moody, M. B., Sharma, A., Nemati, S., & Clifford, G. D. (2019). Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. Proceedings of the 46th Computing in Cardiology, 1–4. https://doi.org/10.22489/CinC.2019.412

[5] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[6] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

[7] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

[8] Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 66(1), 149–153. https://doi.org/10.1093/cid/cix731

[9] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26. https://doi.org/10.48550/arXiv.1310.4546

[10] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

[11] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[12] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

[14] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[15] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of KDD, 785–794. https://doi.org/10.1145/2939672.2939785

[16] Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. Proceedings of the 8th IEEE International Conference on Data Mining, 413–422. https://doi.org/10.1109/ICDM.2008.17

[17] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

[18] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of KDD, 1135–1144. https://doi.org/10.1145/2939672.2939778

[19] Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. Journal of Big Data, 1(1), 2. https://doi.org/10.1186/2196-1115-1-2

[20] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

[21] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV (version 2.2). PhysioNet. https://doi.org/10.13026/6mm1-ek67

[22] Pollard, T. J., Johnson, A. E. W., Raffa, J. D., & Mark, R. G. (2022). eICU Collaborative Research Database (version 2.0). PhysioNet. https://doi.org/10.13026/C2WM1R

[23] Reyna, M., Josef, C., Seyedi, S., Jeter, R., Shashikumar, S. P., Westover, M. B., Sharma, A., Nemati, S., & Clifford, G. D. (2020). Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine, 48(2), 210–217. https://doi.org/10.1097/CCM.0000000000004145

[24] World Health Organization. (2023). Global Patient Safety Action Plan 2021–2030. WHO Press. https://www.who.int/teams/integrated-health-services/patient-safety/policy/global-patient-safety-action-plan

[25] Bates, D. W., Levine, D. M., Salmasian, H., Syrowatka, A., Shahian, D. M., Lipsitz, S., Zebrowski, J. P., Mort, E., Rosen, C. Z., Wien, M. F., Rossi, S. E., Massaro, A. F., Sequist, T. D., & Kachalia, A. (2023). The safety of inpatient health care. New England Journal of Medicine, 388(2), 142–153. https://doi.org/10.1056/NEJMsa2206117

[26] Classen, D. C., Resar, R., Griffin, F., Federico, F., Frankel, T., Kimmel, N., Whittington, J. C., Frankel, A., Seger, A., & James, B. C. (2011). "Global trigger tool" shows that adverse events in hospitals may be ten times greater than previously measured. Health Affairs, 30(4), 581–589. https://doi.org/10.1377/hlthaff.2011.0190

[27] Mäkinen, V.-P., Civelek, M., Meng, Q., Zhang, B., Zhu, J., Levian, C., Huan, T., Sehested, T. S. G., & Hazen, S. L. (2014). Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. PLOS Genetics, 10(7), e1004502. https://doi.org/10.1371/journal.pgen.1004502

[28] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

[29] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, 115–118. https://doi.org/10.1038/nature21056

[30] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

[31] Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2018). Deep learning for chest radiograph diagnosis. PLOS Medicine, 15(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686

[32] Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. Scientific Data, 6(1), 96. https://doi.org/10.1038/s41597-019-0103-9

[33] Wang, S., McDermott, M. B. A., Chauhan, G., Ghassemi, M., Hughes, M. C., & Naumann, T. (2020). MIMIC-Extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. Proceedings of the ACM Conference on Health, Inference, and Learning, 222–235. https://doi.org/10.1145/3368555.3384469

[34] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. Proceedings of the 1st Machine Learning for Healthcare Conference, 301–318. https://doi.org/10.48550/arXiv.1511.05942

[35] Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., & Sun, J. (2017). GRAM: Graph-based attention model for healthcare representation learning. Proceedings of KDD, 787–795. https://doi.org/10.1145/3097983.3098126

[36] Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzell, R. (2016). Learning to diagnose with LSTM recurrent neural networks. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1511.03677

[37] Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., & Gao, J. (2017). Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. Proceedings of KDD, 1903–1911. https://doi.org/10.1145/3097983.3098088

[38] Zhang, J., Guo, K., Li, W., Yue, X., Potkonjak, M., & Venkatasubramanian, S. (2022). Out-of-distribution detection in time-series health data using self-supervised learning. Proceedings of ICML, 26050–26067. https://doi.org/10.48550/arXiv.2205.06028

[39] Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. International Conference on Information Processing in Medical Imaging, 146–157. https://doi.org/10.1007/978-3-319-59050-9_12

[40] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672–2680. https://doi.org/10.48550/arXiv.1406.2661

[41] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1312.6114

[42] Li, Z., Li, J., Ji, T., Geng, X., Li, X., & Song, Y. (2022). Graph-based anomaly detection for medical data: A survey. IEEE Transactions on Knowledge and Data Engineering. Advance online publication. https://doi.org/10.1109/TKDE.2022.3194651

[43] Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. Proceedings of KDD, 855–864. https://doi.org/10.1145/2939672.2939754

[44] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.02216

[45] Le Gall, J.-R., Lemeshow, S., & Saulnier, F. (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA, 270(24), 2957–2963. https://doi.org/10.1001/jama.1993.03510240069035

[46] Rea, M., Arif, I., Khurram, K., & Shafiq, Z. (2023). MedBERT: A pre-trained language model for biomedical named entity recognition. Journal of Biomedical Informatics, 143, 104401. https://doi.org/10.1016/j.jbi.2023.104401

[47] Peng, X., Du, H., Yu, K., & Qian, Z. (2022). Temporal self-awareness in clinical NLP: A survey. Artificial Intelligence in Medicine, 132, 102378. https://doi.org/10.1016/j.artmed.2022.102378

[48] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1, 18. https://doi.org/10.1038/s41746-018-0029-1

[49] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. Proceedings of the 57th Annual Meeting of ACL, 1441–1451. https://doi.org/10.18653/v1/P19-1139

[50] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[51] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[52] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[53] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32. https://doi.org/10.48550/arXiv.1912.01703

[54] Huang, K., Altosaar, J., & Ranganath, R. (2020). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. CHIL Workshop at ICLR. https://doi.org/10.48550/arXiv.1904.05342

[55] Bai, T., Zhang, S., Egleston, B. L., & Vucetic, S. (2018). Interpretable representation learning for healthcare via capturing disease progression through time. Proceedings of KDD, 43–51. https://doi.org/10.1145/3219819.3219985

[56] Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records. Journal of the American Medical Informatics Association, 25(10), 1419–1428. https://doi.org/10.1093/jamia/ocy068

[57] Cai, X., Gao, J., Ngiam, K. Y., Ooi, B. C., Zhang, Y., & Yuan, X. (2018). Medical concept embedding with time-aware attention. Proceedings of IJCAI, 3984–3990. https://doi.org/10.24963/ijcai.2018/554

[58] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[59] Ghassemi, M., Celi, L. A., & Stone, D. J. (2015). State of the art review: The data revolution in critical care. Critical Care, 19(1), 118. https://doi.org/10.1186/s13054-015-0801-4

[60] Celi, L. A., Mark, R. G., Stone, D. J., & Montgomery, R. A. (2013). "Big data" in the intensive care unit. Closing the data loop. American Journal of Respiratory and Critical Care Medicine, 187(11), 1157–1160. https://doi.org/10.1164/rccm.201212-2311ED

[61] Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). Practical guidance on artificial intelligence for health-care data. Lancet Digital Health, 2(5), e232–e235. https://doi.org/10.1016/S2589-7500(20)30036-4

[62] Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., Connell, A., Hughes, C. O., Karthikesalingam, A., Cornebise, J., Montgomery, H., Rees, G., Laing, C., Baker, C. R., Peterson, K., ... Suleyman, M. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. Nature, 572, 116–119. https://doi.org/10.1038/s41586-019-1390-1

[63] Saria, S., & Goldenberg, A. (2016). Subtyping: What it is and its role in precision medicine. IEEE Intelligent Systems, 31(4), 70–75. https://doi.org/10.1109/MIS.2016.57

[64] Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. IEEE Spectrum, 56(4), 24–31. https://doi.org/10.1109/MSPEC.2019.8678513

[65] Ross, C., & Swetlitz, I. (2017). IBM pitched its Watson supercomputer as a revolution in cancer care. It is flopping. STAT News. https://www.statnews.com/2017/09/05/watson-ibm-cancer/

[66] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[67] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

[68] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of KDD, 1721–1730. https://doi.org/10.1145/2783258.2788613

[69] McDermott, M. B. A., Heneghan, C., Elbers, P., Seneviratne, M. G., Lehman, L.-W. H., & Ghassemi, M. (2021). Reproducibility in machine learning for health research: Still a ways to go. Science Translational Medicine, 13(586), eabb1655. https://doi.org/10.1126/scitranslmed.abb1655

[70] Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I. S., Saria, S., Topol, E., Obermeyer, Z., Yu, B., & Butte, A. J. (2020). Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nature Medicine, 26(9), 1320–1324. https://doi.org/10.1038/s41591-020-1041-y

[71] Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Annals of Internal Medicine, 162(1), 55–63. https://doi.org/10.7326/M14-0697

[72] Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., Mallett, S., & PROBAST Group. (2019). PROBAST: A tool to assess the risk of bias and

applicability of prediction model studies. Annals of Internal Medicine, 170(1), 51–58. https://doi.org/10.7326/M18-1376

[73] U.S. Food and Drug Administration. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. FDA. https://www.fda.gov/media/145022/download

[74] European Commission. (2021). Proposal for a Regulation on Artificial Intelligence (AI Act). COM(2021) 206 Final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[75] Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care — Addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983. https://doi.org/10.1056/NEJMp1714229

[76] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

[77] Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. JAMA, 322(24), 2377–2378. https://doi.org/10.1001/jama.2019.18058

[78] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of AISTATS, 1273–1282. https://doi.org/10.48550/arXiv.1602.05629

[79] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[80] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1