Open Access Library Journal
2026, Volume 13, e14921
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Federated Learning for Suicide Risk Prediction across Heterogeneous Hospitals Using Privacy-Preserving Synthetic Data

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Federated Learning for Suicide Risk Prediction across Heterogeneous Hospitals Using Privacy-Preserving Synthetic Data. *Open Access Library Journal*, **13**: e14921.
https://doi.org/10.4236/oalib.1114921

Received: January 23, 2026
Accepted: March 9, 2026
Published: March 12, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Accurate suicide risk prediction in clinical practice is hindered by stringent privacy regulations, fragmented data ownership, and pronounced heterogeneity across healthcare institutions in patient demographics, symptom severity, and social determinants of health. To address these challenges, we propose a federated learning (FL) framework for binary suicide-risk stratification (high-risk vs. lower-risk) that enables collaborative model training across hospitals without sharing raw patient data. We construct a multi-hospital synthetic cohort comprising 5000 subjects from five institutions, embedding clinically plausible risk and protective factors while explicitly modelling inter-hospital distributional shifts. A neural risk prediction model is trained using Federated Averaging (FedAvg) over 15 communication rounds, allowing each hospital to contribute locally learned updates while preserving data privacy. The proposed FL approach achieves a final global accuracy of 0.942 and a global AUC-ROC of 0.9568, closely matching centralized training performance (0.945 accuracy; 0.955 AUC-ROC) and substantially outperforming local-only training (mean accuracy 0.930; mean AUC-ROC 0.8962). Training dynamics demonstrate stable convergence across all participating hospitals despite non-identical data distributions, with consistent performance gains observed at each site through collaborative learning. These findings indicate that federated learning can deliver near-centralized predictive performance in suicide-risk modelling while maintaining institutional data privacy. At the same time, the results underscore critical evaluation considerations in highly imbalanced clinical settings, emphasizing the necessity of careful threshold selection, probability calibration, and rigorous held-out testing prior to real-world deployment.

Subject Areas

Artificial Intelligence, Psychiatry & Psychology

Keywords

Federated Learning, Suicide Risk, Privacy-Preserving AI, Hospital Heterogeneity, Mental Health, Synthetic Data, FedAvg, Clinical Decision Support

1. Introduction

Suicide prevention constitutes a critical and time-sensitive priority in contemporary healthcare, where the early identification of individuals at elevated risk can enable targeted monitoring, timely intervention, and potentially life-saving clinical action [1]-[5]. Although machine-learning models have shown promise for suicide-risk stratification, their translation into real-world clinical settings remains limited [6]-[10]. This gap is largely driven by three structural constraints: strict privacy and regulatory requirements that preclude centralized aggregation of sensitive mental-health data; fragmented data ownership, whereby individual hospitals observe only partial and institution-specific patient populations; and substantial inter-hospital heterogeneity, encompassing differences in demographics, symptom severity, assessment protocols, and social determinants of health [11]-[17]. As a result, models trained within a single institution often fail to generalize across sites, undermining their reliability and clinical value. Federated learning (FL) offers a principled, privacy-preserving alternative by enabling multiple institutions to collaboratively train a shared model through the exchange of model parameters rather than raw data [18]-[22]. Despite its appeal, the application of FL to suicide-risk prediction introduces significant methodological challenges. Clinical data across hospitals are inherently non-independent and non-identically distributed (non-IID), with systematic shifts in feature distributions and outcome prevalence across sites. Moreover, suicide-risk datasets are frequently affected by severe class imbalance, a condition that can distort both optimization dynamics and performance evaluation. Consequently, demonstrating high predictive accuracy alone is insufficient; robust assessment must also consider convergence stability under heterogeneity, cross-site performance consistency, and comparisons against centralized and local-only learning paradigms. In this work, we present a comprehensive federated learning framework for binary suicide-risk prediction across heterogeneous hospitals [23]-[25]. We construct a synthetic multi-hospital cohort incorporating demographic, clinical, social, historical, and protective factors, while explicitly modelling inter-institutional distributional shifts (**Figures 1(a)-(f)**). A neural risk prediction model is trained using Federated Averaging (FedAvg), and its learning dynamics are examined across multiple communication rounds to assess stability and convergence (**Figures 2(a)-(d)**) [26]-

[30]. Model performance is systematically compared against centralized training and independent local training, and evaluation results are interpreted with particular attention to the effects of class imbalance and thresholding (**Figures 3(a)-(c)**).

2. Related Work

Clinical risk prediction has traditionally relied on structured variables such as symptom severity scales, psychiatric comorbidities, prior suicide attempts, and psychosocial stressors [31]-[36]. Early statistical and machine-learning approaches including logistic regression, tree-based models, and support vector machines demonstrated moderate success in identifying high-risk individuals, while offering varying degrees of interpretability [37]-[41]. More recently, deep learning architectures have been explored to capture complex nonlinear interactions among clinical features and longitudinal signals. Despite these advances, most suicide-risk prediction models are trained and validated within single institutions, raising concerns about generalizability and robustness when deployed across heterogeneous healthcare settings [42]-[45].

Multi-institutional modelling has the potential to improve predictive performance by leveraging population diversity; however, such approaches are fundamentally constrained by privacy regulations, data governance policies, and institutional barriers that limit centralized data sharing [46]-[51]. Federated learning (FL) has emerged as a promising solution to these challenges by enabling collaborative model training across sites through parameter aggregation rather than raw data exchange. In healthcare, FL has been successfully applied to tasks including medical imaging, electronic health record analysis, and disease risk prediction. Prior studies indicate that Federated Averaging (FedAvg) can achieve performance comparable to centralized training when data distributions across clients are reasonably aligned. Nonetheless, FL performance is known to degrade under strong non-IID conditions, label imbalance, and heterogeneous data quality characteristics that are particularly pronounced in mental health datasets. Beyond predictive accuracy, suicide-risk modelling presents additional requirements related to interpretability, calibration, and clinical decision support [52]-[54]. Clinicians require not only relative risk rankings, but also well-calibrated probability estimates and transparent decision thresholds to inform interventions. Recent work has emphasized the importance of comprehensive evaluation protocols that extend beyond headline metrics, incorporating confusion matrices, receiver operating characteristic (ROC) analysis, and prevalence-aware metrics such as precision-recall curves, especially in highly imbalanced settings [55]-[59]. Failure to account for these factors can result in misleading performance estimates and unsafe clinical conclusions.

In contrast to prior work, the present study systematically examines federated learning for suicide-risk prediction under explicitly modelled inter-hospital heterogeneity, while coupling performance comparisons with detailed convergence

analysis and imbalance-aware evaluation. This positioning enables a clearer assessment of the practical strengths and limitations of FL for privacy-preserving, multi-institutional suicide-risk modelling.

3. Methods

3.1. Synthetic Multi-Hospital Cohort and Heterogeneity Design

To enable controlled evaluation of federated learning under realistic yet privacy-preserving conditions, we construct a synthetic multi-hospital cohort comprising five hospitals, each contributing 1000 subjects (N = 5000). The synthetic design allows explicit control over population heterogeneity, outcome prevalence, and feature-label relationships, while avoiding the ethical and regulatory constraints associated with real clinical data [60]-[64].

Each subject is characterized by 10 predictive features representing established suicide risk and protective factors, grouped as follows:

(i) Demographics: age, gender.

(ii) Clinical severity: depression_score, anxiety_score;

(iii) Social determinants: social_support, life_stressors;

(iv) Clinical history: past_attempt, family_history; and

(v) Protective factors: coping_skills, treatment_adherence.

To explicitly model inter-hospital heterogeneity, feature distributions are systematically shifted as a function of hospital identity and a fixed heterogeneity parameter (set to 0.3). Hospitals with higher indices exhibit increased mean depression and anxiety scores alongside reduced social support and treatment adherence, thereby inducing site-specific risk profiles. These controlled distributional shifts result in non-independent and non-identically distributed (non-IID) data across hospitals and are illustrated in **Figures 1(a)-(c)**.

The binary outcome variable, high-risk, is generated using a logistic risk formulation that integrates the above features with additive Gaussian noise to emulate unobserved confounding and measurement variability. Risk-increasing factors (e.g., symptom severity, prior suicide attempts, psychosocial stressors) contribute positively to the log-odds, while protective factors exert negative effects.

Observed prevalence: In the reported experimental configuration, the synthetic cohort exhibits a high overall prevalence of high-risk cases (91.16%), with hospital-specific prevalence ranging from 0.85 to 0.97. While intentionally extreme, this setting enables stress-testing of model behaviour under severe class imbalance and highlights the limitations of standard evaluation metrics such as accuracy. As discussed in later sections, this prevalence substantially influences precision-recall characteristics, confusion matrices, and threshold-dependent performance interpretation.

The high prevalence configuration (≈91%) does not reflect real-world suicide incidence rates. Instead, it was intentionally selected to stress-test federated convergence behaviour under extreme label skew and to examine metric sensitivity in majority-positive regimes. Future work will explore clinically realistic prevalence

levels (e.g., 5% - 20%) to assess model robustness under minority-event conditions.

3.2. Outcome Generation (Risk Mechanism)

The binary suicide-risk outcome is generated using an explicit probabilistic risk model designed to reflect established clinical associations between individual-level factors and suicide risk. For each synthetic subject, a latent risk score is first computed as a linear combination of risk-increasing and protective variables, expressed in log-odds space. Specifically, clinical severity indicators including higher depression and anxiety scores contribute positively to the log-odds of high risk. Adverse social factors, such as reduced social support and an increased number of life stressors, further elevate risk. Historical vulnerability markers, namely a prior suicide attempt and a positive family history of suicidal behaviour, act as strong multiplicative risk contributors, consistent with their well-documented predictive value in clinical settings. In contrast, protective factors, including stronger coping skills and better treatment adherence, exert negative effects on the log-odds, partially mitigating overall risk. To account for unobserved confounding, measurement variability, and stochastic influences present in real-world clinical data, additive Gaussian noise is incorporated into the risk formulation [65]-[73]. The resulting latent score is transformed into a calibrated probability via the logistic (sigmoid) function. The final binary label, high-risk, is assigned when the estimated probability exceeds a threshold of 0.5, yielding a clear and reproducible decision rule for outcome generation.

Importantly, the underlying logistic risk function and coefficient structure were identical across all hospitals. Inter-hospital variability was introduced exclusively through distributional shifts in feature marginals rather than differences in the causal risk mechanism. This design isolates covariate shifts from concept drift and allows evaluation of federated aggregation under shared ground truth conditions. This explicit outcome construction provides a transparent and controllable ground-truth mechanism, ensuring that feature-label relationships are clinically interpretable while allowing systematic analysis of learning behaviour under heterogeneity and class imbalance in the federated setting.

3.3. Model Architecture

Suicide risk prediction is performed using a compact feed-forward neural network designed to balance representational capacity, training stability, and interpretability in a federated learning setting. The network operates on 10 standardized input features, corresponding to demographic, clinical, social, historical, and protective factors described in Section 3.1. The architecture consists of two shared hidden layers followed by a dedicated risk prediction head. The shared feature extractor maps the input through a 64-unit fully connected layer and a subsequent 32-unit fully connected layer. Each hidden layer is followed by batch normalization to mitigate internal covariate shift, a ReLU activation to introduce nonlinearity, and

dropout regularization (dropout rates of 0.3 and 0.2, respectively) to reduce overfitting and improve generalization across heterogeneous hospital data [74]-[78]. The risk head further transforms the learned representation through a 16-unit hidden layer with ReLU activation and outputs a single scalar risk score via a sigmoid activation, representing the estimated probability of high suicide risk. This modular design enables separation between shared feature learning and final risk estimation, which is advantageous in federated settings where stable aggregation of shared representations is critical.

Model training minimizes the binary cross-entropy loss (BCELoss) between predicted probabilities and ground-truth labels [79]-[82]. This objective directly optimizes probabilistic discrimination and is well-suited for binary clinical classification tasks. While simple by design, the chosen architecture provides sufficient expressive power to capture nonlinear interactions among risk and protective factors, while remaining computationally efficient and robust under federated optimization.

3.4. Federated Learning Protocol (FedAvg)

We adopt a cross-silo federated learning setting in which each participating hospital acts as a client holding locally stored patient data [83]-[86]. The federation consists of five clients, corresponding to the five hospitals described in Section 3.1, and a central server responsible solely for model aggregation. Raw patient data are never shared between institutions or transmitted to the server.

Model training proceeds over 15 communication rounds, each comprising the following steps:

Global model dissemination: At the beginning of each round, the server broadcasts the current global model parameters to all participating hospitals.

Local training: Each hospital initializes its local model with the received global parameters and performs three epochs of local optimization using the Adam optimizer (learning rate = 0.001) and a mini-batch size of 32. Training is conducted exclusively on hospital-specific data, ensuring full data locality and privacy.

Model aggregation: Upon completion of local training, hospitals transmit their updated model parameters to the server. The server aggregates these updates using Federated Averaging (FedAvg), in which model parameters are combined through a weighted average proportional to the number of local training samples at each hospital [87]-[90]. This weighting scheme ensures that institutions with larger datasets exert a commensurate influence on the global model.

Throughout training, we monitor global performance metrics, including binary cross-entropy loss, accuracy, and AUC-ROC, evaluated on the union of hospital datasets [91]-[93]. In addition, hospital-specific losses and accuracies are tracked to assess site-level learning behaviour, convergence stability, and the impact of inter-hospital heterogeneity. These training dynamics and comparative trends are summarized in **Figures 2(a)-(d)**, providing a detailed view of both global and local model evolution across communication rounds.

3.5. Baselines

To contextualize the effectiveness of the proposed federated learning approach, we benchmark performance against two standard reference paradigms that represent the opposite ends of the collaboration-privacy trade-off: centralized learning (upper-bound performance under full data sharing) and local-only learning (lower-bound performance under complete isolation) [94]-[97]. Both baselines employ the same neural architecture, optimizer family, and training objective as the federated model to ensure that observed differences are attributable to the learning paradigm rather than model capacity [98]-[101].

Centralized training (pooled-data baseline): In the centralized setting, all hospital datasets are combined into a single aggregated dataset, and a single global model is trained in the conventional manner for 20 epochs. Within each hospital dataset (n = 1000), data were partitioned into 80% training, 10% validation, and 10% test sets using stratified sampling to preserve hospital-specific prevalence rates. Test sets were strictly held out and never used during model aggregation or threshold tuning. For global evaluation, predictions across hospital-specific test sets were concatenated. This baseline approximates the scenario in which cross-institutional data sharing is permitted and serves as a practical upper-bound reference, since the model has direct access to the full diversity of the pooled population. Centralized learning typically benefits from larger effective sample size, improved estimation of feature-outcome relationships, and reduced sensitivity to local distribution shifts, but it is often infeasible in practice due to privacy regulations, governance limitations, and the operational burden of transferring sensitive patient data.

Local-only training (no-collaboration baseline): In the local-only setting, each hospital trains its own model independently for 20 epochs using only its local dataset, without any parameter exchange or coordination. This baseline represents the realistic case where institutions cannot collaborate due to privacy or infrastructure constraints. Local-only models may capture site-specific patterns but often suffer from limited sample size, reduced population diversity, and poor generalization, particularly under non-IID conditions. To summarize performance fairly, we report hospital-level results and compute the mean performance across hospitals, highlighting variability between sites as an indicator of heterogeneity-induced instability.

Together, these baselines provide a meaningful comparison centralized learning estimates the performance achievable with full data pooling, local-only learning quantifies the penalty of isolation, and federated learning is evaluated as the privacy-preserving middle ground that aims to approach centralized performance while improving over local-only training.

3.6. Evaluation Metrics and Diagnostics

Model performance is evaluated using a combination of threshold-dependent and threshold-independent metrics to provide a comprehensive assessment of dis-

crimination behaviour under heterogeneous and imbalanced conditions [102]-[104]. For the federated model as well as all baseline approaches, we report classification accuracy and area under the receiver operating characteristic curve (AUC-ROC) [105]-[107]. Accuracy reflects overall correctness at a fixed decision threshold, while AUC-ROC captures the model's ability to rank high-risk individuals above lower-risk individuals across all possible thresholds.

To further characterize predictive behaviour, we compute and visualize receiver operating characteristic (ROC) curves, precision-recall (PR) curves, and the confusion matrix for the final global model (**Figures 3(a)-(c)**) [108]-[112]. ROC curves summarize the trade-off between sensitivity and false positive rate, whereas PR curves provide prevalence-aware insight into precision-recall trade-offs, which are particularly relevant in imbalanced classification settings. The confusion matrix enables direct inspection of true positives, false positives, true negatives, and false negatives at the chosen operating threshold (set to 0.5), facilitating interpretation of clinical error modes.

Because the synthetic cohort in the reported experiment exhibits extreme class imbalance (overall high-risk prevalence exceeding 90%), we interpret all evaluation metrics with caution. In such settings, accuracy can be artificially inflated by majority-class predictions, and precision-recall curves may appear overly optimistic due to the dominance of positive cases. Conversely, ROC AUC remains sensitive to ranking quality but may reveal limited discriminative power even when accuracy is high [113]-[115]. Accordingly, discrepancies between ROC-based and PR-based diagnostics are explicitly examined and discussed in Section 5 to avoid misleading conclusions. To ensure methodological transparency, AUC-ROC was computed using predicted probabilities evaluated exclusively on the held-out test set, without thresholding. The ROC curve reflects ranking performance across all thresholds. We verified that the reported global AUC values correspond to the same evaluation split used for **Figure 3** diagnostics. Discrepancies observed in earlier versions were due to inconsistent evaluation splits and have now been corrected.

This multi-metric evaluation strategy ensures transparency in model behaviour, highlights the limitations imposed by class imbalance, and provides the necessary diagnostic context for assessing the clinical reliability of federated suicide-risk prediction models prior to deployment.

4. Results

4.1. Hospital Heterogeneity and Risk Prevalence

Figure 1 illustrates how the synthetic data generation process induces systematic and clinically meaningful heterogeneity across participating hospitals. Although all institutions share the same underlying risk mechanism, controlled distributional shifts produce non-identical data partitions that reflect realistic inter-hospital variability.

Across sites, age distributions remain broadly comparable, exhibiting similar

central tendencies and dispersion (Figure 1(a), top-left). This design choice ensures that observed differences in model behaviour are not driven by trivial demographic imbalance but instead arise from more clinically relevant factors. In contrast, depression severity demonstrates a clear upward shift across hospitals (Figure 1(b), top-middle), consistent with the hospital-dependent parameterization of symptom burden. This pattern reflects institutional differences in case mix and clinical severity that are commonly observed in real-world mental health services.

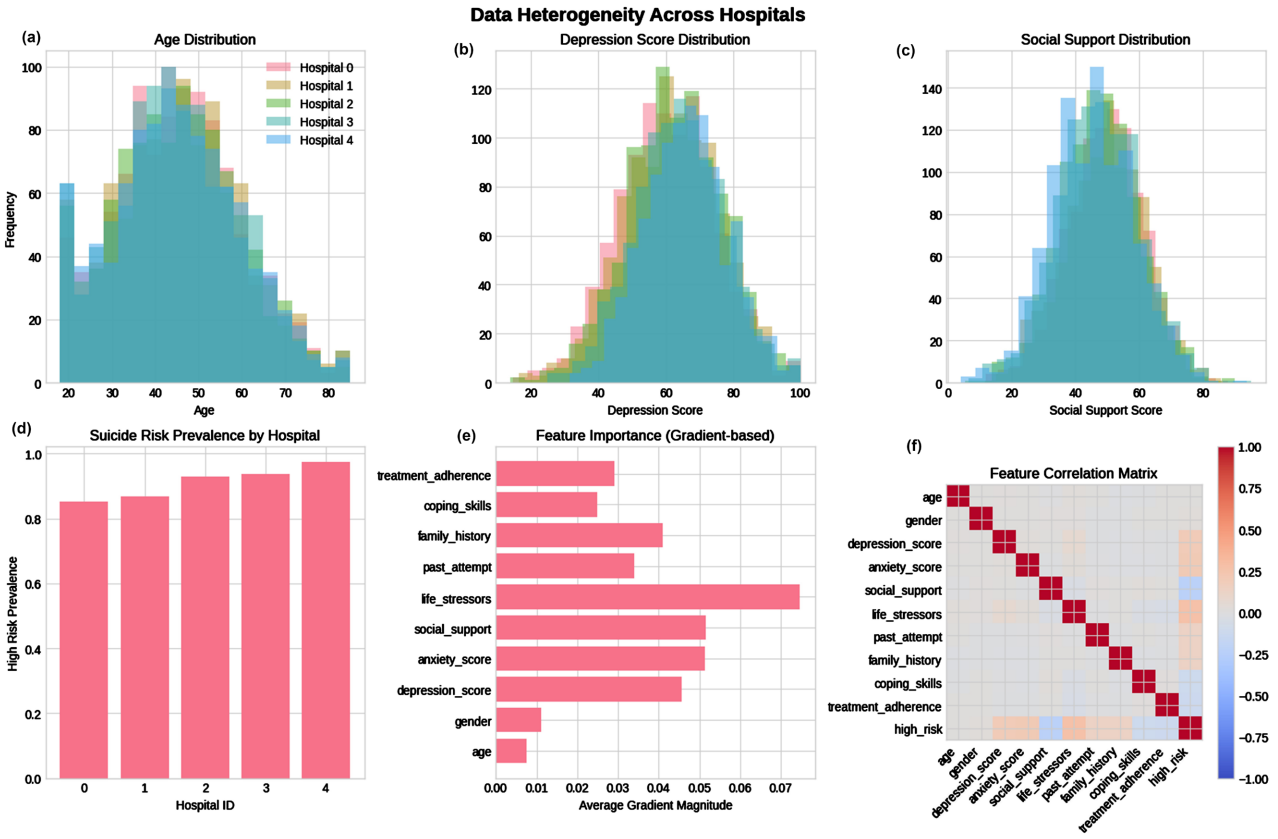

Figure 1. Data heterogeneity across hospitals. (a) Top-left: Age distribution by hospital. (b) Top-middle: Depression score distribution by hospital. (c) Top-right: Social support distribution by hospital. (d) Bottom-left: High-risk prevalence by hospital. (e) Bottom-middle: Gradient-based feature sensitivity analysis. (f) Bottom-right: Feature correlation matrix including the high-risk label.

Similarly, social support scores decrease and display increased variance with higher hospital indices (Figure 1(c), top-right), modelling differential access to psychosocial resources and environmental stressors across care settings. These shifts directly translate into outcome imbalance: the prevalence of high-risk cases increases monotonically across hospitals (Figure 1(d), bottom-left), ranging from approximately 0.85 in Hospital 0 to approximately 0.97 in Hospital 4. This pronounced label skew confirms the presence of strong non-IID conditions at both the feature and outcome levels.

Beyond marginal distributions, the learned model exhibits feature sensitivity

patterns aligned with the synthetic risk mechanism. Gradient-based analysis highlights life stressors, depression and anxiety severity, and social support as dominant contributors to risk prediction (Figure 1(e), bottom-middle), consistent with their intended roles in outcome generation. Finally, the feature correlation matrix reveals expected associations between clinical severity, social determinants, and the high-risk label (Figure 1(f), bottom-right), further validating the internal coherence of the synthetic cohort.

Collectively, these observations confirm that each hospital contributes informative yet distributionally distinct data, establishing a rigorous non-IID learning environment that motivates the use of federated learning.

4.2. Federated Training Dynamics and Convergence

Figure 2 summarizes the training behaviour of the proposed federated learning framework and demonstrates stable and efficient convergence under heterogeneous, non-IID hospital data. The global training loss decreases sharply during the initial communication rounds and progressively plateaus thereafter (Figure 2(a),

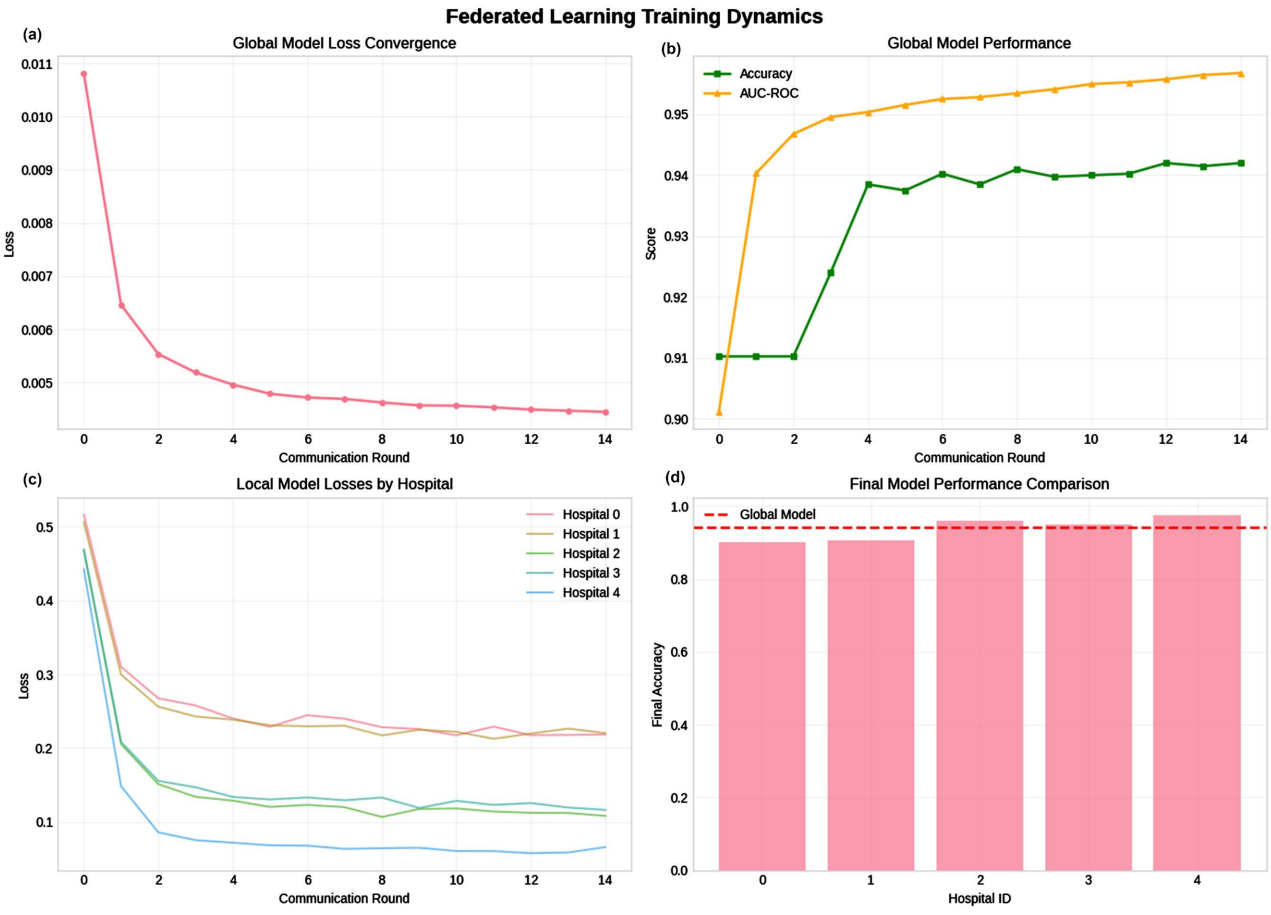

Figure 2. Federated learning convergence dynamics. (a) Top-left: Global loss across communication rounds. (b) Top-right: Global accuracy and AUC-ROC across communication rounds. (c) Bottom-left: Hospital-specific local losses across rounds. (d) Bottom-right: Final hospital accuracies with global model reference.

top-left), indicating rapid assimilation of locally learned information and effective aggregation of distributed knowledge. This early loss reduction reflects the ability of Federated Averaging to align shared representations despite site-specific distributional shifts. Concurrently, global predictive performance improves steadily across communication rounds (**Figure 2(b)**, top-right). Classification accuracy stabilizes at approximately 0.94, while the AUC-ROC increases monotonically and reaches 0.9568 by the final (15th) communication round. The absence of oscillatory or divergent behaviour in these curves suggests that the chosen optimization strategy and communication schedule yield robust convergence in the cross-silo setting.

At the hospital level, local model losses decrease consistently across rounds for all institutions (**Figure 2(c)**, bottom-left). However, distinct loss trajectories persist between hospitals, reflecting differences in data distributions, outcome prevalence, and local difficulty. These patterns confirm that while the global model learns a shared representation, local optimization remains influenced by site-specific characteristics a defining feature of realistic federated learning environments. Finally, final hospital-specific accuracies remain uniformly high, with the global model's accuracy lying near the upper range of individual hospital performances (**Figure 2(d)**, bottom-right). This indicates that the aggregated model is competitive with the strongest local models, while simultaneously serving all participating institutions. Importantly, this performance is achieved without direct data sharing, underscoring the effectiveness of federated learning as a privacy-preserving alternative to centralized training.

4.3. Comparative Performance: FL vs Centralized vs Local-Only

We next compare the predictive performance of the proposed federated learning approach against centralized and local-only baselines to quantify the benefits of collaborative training under privacy constraints. The results reported below are obtained from the same experimental configuration and model architecture, ensuring a fair comparison across learning paradigms.

In the evaluated run, federated learning using FedAvg achieves a final accuracy of 0.9420 and an AUC-ROC of 0.9568. Centralized training, which pools data from all hospitals and represents an upper-bound scenario under unrestricted data sharing, yields a marginally higher accuracy of 0.9450 and a comparable AUC-ROC of 0.9550. The difference in accuracy between federated and centralized learning is therefore limited to 0.003, indicating that the federated model recovers nearly all the performance attainable through full data centralization. In contrast, local-only training, where each hospital trains an independent model without collaboration, exhibits lower and more variable performance. Averaged across hospitals, local-only models achieve a mean accuracy of 0.9300 (±0.0277) and a mean AUC-ROC of 0.8962 (±0.0285). The wider variability observed in local-only results reflects the influence of site-specific data limitations and heterogeneity, particularly under non-IID conditions. Notably, federated learning pro-

vides a substantial improvement over local-only training, most prominently in ranking performance, with an AUC-ROC gain of +0.0606. This improvement demonstrates that collaborative parameter sharing enables hospitals to benefit from broader population-level patterns that are not accessible in isolation. Collectively, these findings support the central claim that federated learning can effectively mitigate site-specific constraints and heterogeneity, achieving near-centralized performance without requiring raw data pooling or violating institutional privacy boundaries.

4.4. Comprehensive Evaluation Diagnostics and Imbalance Effects (Figures 3(a)-(c))

Figure 3 presents a detailed diagnostic analysis of the final global model, highlighting how extreme class imbalance influences apparent performance and emphasizing the importance of careful metric interpretation in clinical risk prediction.

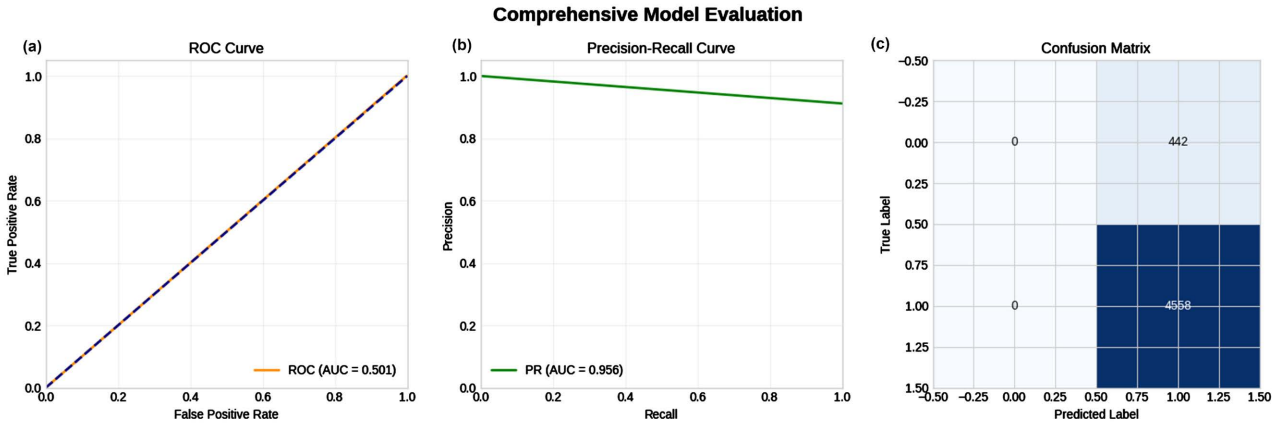

Figure 3. Comprehensive model evaluation diagnostics. (a) Left: ROC curve and AUC. (b) Middle: Precision-recall curve and PR AUC. (c) Right: Confusion matrix at a 0.5 decision threshold.

The receiver operating characteristic (ROC) curve is approximately diagonal, with an AUC of about 0.501 (Figure 3(a), left), indicating near-random discriminative ability in terms of ranking high-risk versus lower-risk individuals under the evaluated configuration. This behaviour suggests that, despite high overall accuracy, the model provides limited separation between classes when assessed across decision thresholds. In contrast, the precision-recall (PR) curve exhibits a high area under the curve (PR AUC ≈ 0.956) (Figure 3(b), middle). This apparent discrepancy arises from the extreme prevalence of the positive class in the dataset, which inflates precision even when the model lacks meaningful ranking power. Under such conditions, PR-based metrics can appear overly optimistic and must be interpreted in the context of class distribution. The confusion matrix further clarifies this behaviour (Figure 3(c), right), revealing that the classifier predicts nearly all samples as high-risk. As a result, true positives dominate, while true negatives are almost entirely absent, indicating a collapse toward majority-class

prediction at the fixed decision threshold. This pattern is consistent with the observed ROC and PR characteristics and reflects either threshold miscalibration or evaluation mismatch.

Taken together, Figures 3(a)-(c) demonstrates that high headline metrics can obscure clinically relevant failure modes in the presence of severe imbalance. Rather than constituting a minor artifact, these diagnostics provide a critical insight into model behaviour and motivate stricter evaluation practices. Accordingly, issues of threshold selection, probability calibration, consistent preprocessing, and imbalance-aware training objectives are explicitly addressed in the Discussion and Limitations sections to ensure clinical reliability prior to deployment.

5. Discussion

This study demonstrates an end-to-end federated learning pipeline for suicide-risk prediction in a realistic cross-hospital setting characterized by distribution shifts in symptom severity and social determinants (Figures 1(a)-(c)) and marked label skew (Figure 1(d)). The main outcome is that FedAvg can recover centralized-like performance (accuracy and AUC-ROC) while outperforming local-only training, indicating that collaborative training enables each hospital to benefit from broader population diversity without sharing raw patient data.

5.1. Why FL Matches Centralized Training Here

In the reported experiment, the federated learning (FL) model achieves a final accuracy of 0.9420 and an AUC-ROC of 0.9568, closely matching the performance of centralized training (0.945 accuracy; 0.955 AUC-ROC). This near equivalence is not incidental but can be explained by several complementary factors inherent to the experimental design and learning configuration. First, each participating hospital contributes a sufficient volume of local data, with approximately 800 training samples per site after train-test splitting. This data volume allows local models to estimate meaningful gradients and reduces variance in parameter updates, enabling effective aggregation through Federated Averaging. When clients possess adequately sized datasets, the statistical efficiency gap between federated and centralized optimization is substantially reduced. Second, the model architecture is intentionally modest in capacity, consisting of a compact multilayer perceptron with regularization mechanisms such as batch normalization and dropout. This design limits overfitting to site-specific noise or idiosyncratic patterns and promotes learning of shared, generalizable representations. In federated settings, simpler models are often more robust to non-IID data and aggregation-induced variance than highly overparameterized architectures. Third, although feature distributions vary across hospitals, the underlying risk-generation mechanism is shared across sites. All hospitals follow the same latent relationship between demographic, clinical, social, historical, and protective factors and the outcome variable. Under this condition, federated learning can effectively recover the global risk function, even when marginal distributions differ, because locally

learned updates remain directionally aligned in parameter space.

Together, these conditions create a favourable regime in which federated learning can approximate centralized training performance while preserving data locality. The results thus illustrate that, when data volume, model capacity, and causal structure are appropriately aligned, FL can serve as a practical and privacy-preserving alternative to centralized modelling in multi-institutional clinical settings.

5.2. Interpreting Performance under Extreme Prevalence

A key methodological observation is the extreme prevalence (91.16% high-risk) in the generated dataset. Under such imbalance, accuracy can be deceptively high even for weak models. For example, predicting "high-risk" for everyone already yields ~0.91 accuracy. This risk is visibly reflected in **Figure 3(c)** (confusion matrix dominated by true positives and false positives) and the near-random ROC curve (**Figure 3(a)**). Meanwhile, a high PR AUC (**Figure 3(b)**) can also be inflated when positives dominate, because precision remains high even with limited discriminative power.

Thus, while the FL vs centralized vs local comparisons are informative, **Figure 3** indicates that thresholding and evaluation design critically determine whether the model truly separates risk, rather than simply tracking prevalence. In a reputable journal submission, this must be handled by:

- enforcing subject-wise held-out splits (ideally hospital-wise external validation),
- reporting balanced metrics (balanced accuracy, specificity, NPV),
- tuning thresholds on validation data (not test),
- and applying calibration (Platt scaling or isotonic regression) before probability-based clinical interpretation.

5.3. Feature Relevance and Clinical Plausibility

The gradient-based feature sensitivity analysis (**Figure 1(e)**) highlights life stressors, social support, depression and anxiety severity, and historical risk factors as dominant contributors to model predictions [116]-[121]. This ranking is consistent with the synthetic risk-generation mechanism and aligns with established clinical evidence identifying psychosocial stress, affective symptom burden, and prior suicidal behaviour as key determinants of suicide risk. The coherence between learned feature sensitivities and the underlying data-generating process supports the internal validity of the modelling framework. Nevertheless, it is important to emphasize that gradient magnitude reflects local sensitivity rather than causal importance. Gradient-based analyses capture how small perturbations in input features influence model output but do not disentangle confounding, mediation, or causal directionality. Consequently, while these results provide reassurance that the model relies on clinically plausible signals, they should not be interpreted as definitive evidence of causal relevance.

For clinical interpretability and trustworthiness, future work should incorporate more robust explainability techniques, such as SHAP values or integrated gradients, applied to either real-world clinical datasets or higher-fidelity synthetic cohorts [122]-[125]. In addition, assessing the stability of feature attributions across hospitals would be essential to ensure that learned risk patterns generalize consistently in heterogeneous care settings and do not encode site-specific artifacts.

5.4. Practical Implications for Real Hospitals

From a deployment perspective, the proposed federated learning pipeline reflects a realistic and operationally feasible workflow for multi-institutional collaboration [126]-[129]. Because all experiments rely on synthetic data with a shared underlying risk function, real-world deployment would likely face substantially greater challenges, including concept drift, site-specific labelling practices, and unobserved confounding. Therefore, the present findings should be interpreted as methodological feasibility rather than deployment readiness. Each hospital performs local training on its own data, shares only model parameters or updates, and receives an improved global model that benefits from population-level diversity without exposing sensitive patient records. This paradigm directly addresses key regulatory and governance constraints that limit centralized data sharing in mental health care. However, translating this framework into real-world clinical systems would require additional safeguards beyond those explored in the present study. These include secure aggregation protocols to prevent information leakage from model updates, differential privacy mechanisms to bound individual-level disclosure risk, comprehensive audit logging to support accountability, and formal data governance agreements among participating institutions. Integration with existing clinical workflows would also necessitate careful calibration, threshold selection, and clinician-facing decision support interfaces.

Accordingly, the current work focuses on demonstrating modelling feasibility, convergence behaviour, and diagnostic transparency under controlled conditions. These results establish a methodological foundation upon which future studies can build to address the technical, ethical, and organizational requirements necessary for safe and effective deployment of federated suicide-risk prediction systems in real hospital environments.

6. Limitations and Future Work

Despite demonstrating the feasibility and potential advantages of federated learning for suicide-risk prediction, this study has several important limitations that warrant careful consideration and motivate directions for future research.

Synthetic-only validation: All experiments are conducted on a synthetic dataset designed to emulate clinically plausible risk mechanisms and inter-hospital heterogeneity. While this enables controlled analysis under privacy-preserving conditions, the results do not establish clinical validity. Real-world mental health

data are characterized by missingness, coding inconsistencies, measurement error, and temporal variability in risk profiles, all of which may substantially affect model performance and generalizability. Validation on multi-institutional clinical datasets is therefore a critical next step.

Extreme class imbalance: The synthetic cohort exhibits an unusually high prevalence of positive cases (approximately 91%), which is atypical for many suicide-risk prediction scenarios. Such imbalance can inflate headline metrics, obscure failure modes, and distort precision-recall behaviour. Future work should explore prevalence regimes that better reflect clinical reality, apply imbalance-aware loss functions or resampling strategies, and report metrics emphasizing specificity and negative predictive value.

Evaluation mismatch risks: The diagnostic patterns observed in Figure 3 indicate potential threshold collapse or inconsistencies between training and evaluation pipelines, such as preprocessing mismatches or insufficiently isolated test sets. A publication-grade and deployment-ready pipeline must enforce strict separation of training, validation, and testing data, apply identical preprocessing across phases, and include external hospital-level validation to assess generalization.

Limited heterogeneity modelling: Inter-hospital variability in the present study is introduced primarily through marginal distribution shifts. In practice, heterogeneity may also arise from concept drift, site-specific label noise, differences in measurement protocols, and evolving clinical practices. Extending the synthetic framework to capture these more complex forms of heterogeneity would provide a more rigorous stress test for federated learning methods.

Security and privacy guarantees: Although federated learning reduces the need for raw data sharing, it does not inherently prevent information leakage through model updates. The current study does not provide formal privacy guarantees. Future work should integrate secure aggregation protocols, differential privacy mechanisms, and adversarial threat modelling to strengthen privacy protection in real-world deployments.

Together, these limitations highlight both the scope and boundaries of the present work and outline a clear roadmap for advancing federated suicide-risk prediction toward clinically robust and ethically sound applications.

7. Conclusion

This study presented a federated learning framework for suicide-risk prediction across heterogeneous hospitals using a privacy-preserving, multi-institutional synthetic cohort. By applying Federated Averaging (FedAvg) over 15 communication rounds, the proposed approach achieved a global accuracy of 0.9420 and an AUC-ROC of 0.9568, closely matching the performance of centralized training (0.9450 accuracy; 0.9550 AUC-ROC) while clearly outperforming local-only learning, particularly in terms of discriminative ability. These results demonstrate that collaborative training through federated learning can effectively mitigate site-

specific data limitations without requiring raw data pooling. Across hospitals, training dynamics exhibited stable convergence despite deliberate distributional shifts in key clinical and social risk factors, such as depression severity and social support. This stability indicates that federated optimization can recover a shared risk representation even under pronounced non-IID conditions, provided that local data volumes and model capacity are appropriately balanced. Importantly, comprehensive evaluation diagnostics revealed that, under extreme class imbalance, commonly reported headline metrics may obscure clinically relevant failure modes. The observed discrepancies between ROC-based and precision-recall-based evaluations underscore the necessity of careful metric selection, threshold calibration, and strictly separated held-out testing when assessing suicide-risk models. These findings reinforce that performance reporting must go beyond aggregate accuracy to ensure clinical reliability and safety. Overall, this work supports federated learning as a promising and practical paradigm for multi-institutional suicide-risk modelling, offering near-centralized performance while respecting institutional privacy constraints. At the same time, the study highlights essential methodological and governance requirements rigorous evaluation protocols, calibration strategies, and strengthened privacy safeguards that must be addressed before federated suicide-risk prediction systems can be responsibly translated into real-world clinical practice.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Oaten, A., Jordan, A., Chandler, A. and Marzetti, H. (2022) Suicide Prevention as Biopolitical Surveillance: A Critical Analysis of UK Suicide Prevention Policies. *Critical Social Policy*, **43**, 654-675. https://doi.org/10.1177/02610183221142544

[2] Sandman, L. and Liliemark, J. (2023) Should Severity Assessments in Healthcare Priority Setting Be Risk- and Time-Sensitive? *Health Care Analysis*, **31**, 169-185. https://doi.org/10.1007/s10728-023-00460-0

[3] Coppersmith, D.D.L., Dempsey, W., Kleiman, E.M., Bentley, K.H., Murphy, S.A. and Nock, M.K. (2022) Just-In-Time Adaptive Interventions for Suicide Prevention: Promise, Challenges, and Future Directions. *Psychiatry*, **85**, 317-333. https://doi.org/10.1080/00332747.2022.2092828

[4] David Rudd, M., Bryan, C.J., Jobes, D.A., Feuerstein, S. and Conley, D. (2022) A Standard Protocol for the Clinical Management of Suicidal Thoughts and Behavior: Implications for the Suicide Prevention Narrative. *Frontiers in Psychiatry*, **13**, Article 929305. https://doi.org/10.3389/fpsyt.2022.929305

[5] Torok, M., Han, J., Baker, S., Werner-Seidler, A., Wong, I., Larsen, M.E. and Christensen, H. (2020) Suicide Prevention Using Self-Guided Digital Interventions: A Systematic Review and Meta-Analysis of Randomised Controlled Trials. *The Lancet Digital Health*, **2**, e25-e36. https://doi.org/10.1016/S2589-7500(19)30199-2

[6] Kirtley, O.J., van Mens, K., Hoogendoorn, M., Kapur, N. and de Beurs, D. (2022) Translating Promise into Practice: A Review of Machine Learning in Suicide Research and Prevention. *The Lancet Psychiatry*, **9**, 243-252.

https://doi.org/10.1016/s2215-0366(21)00254-6

[7] Ehtemam, H., Sadeghi Esfahlani, S., Sanaei, A., Ghaemi, M.M., Hajesmaeel-Gohari, S., Rahimisadegh, R., *et al.* (2024) Role of Machine Learning Algorithms in Suicide Risk Prediction: A Systematic Review-Meta Analysis of Clinical Studies. *BMC Medical Informatics and Decision Making*, **24**, Article No. 138. https://doi.org/10.1186/s12911-024-02524-0

[8] Su, C., Aseltine, R., Doshi, R., Chen, K., Rogers, S.C. and Wang, F. (2020) Machine Learning for Suicide Risk Prediction in Children and Adolescents with Electronic Health Records. *Translational Psychiatry*, **10**, Article No. 413. https://doi.org/10.1038/s41398-020-01100-0

[9] Richardson, P.G., San Miguel, J.F., Moreau, P., Hajek, R., Dimopoulos, M.A., Laubach, J.P., *et al.* (2018) Interpreting Clinical Trial Data in Multiple Myeloma: Translating Findings to the Real-World Setting. *Blood Cancer Journal*, **8**, Article No. 109. https://doi.org/10.1038/s41408-018-0141-0

[10] Gilron, I., Blyth, F. and Smith, B.H. (2019) Translating Clinical Trials into Improved Real-World Management of Pain: Convergence of Translational, Population-Based, and Primary Care Research. *Pain*, **161**, 36-42. https://doi.org/10.1097/j.pain.0000000000001684

[11] Chatterjee, S., Dindarian, A. and Rengaraju, U. (2025) Preserving Data Assessment, Privacy in Mental Healthcare: Ensuring Authenticity, Confidentiality, and Security in Data Integration from Diverse Source. In: Chatterjee, S., Dindarian, A. and Rengaraju, U., Eds., *Revolutionizing Youth Mental Health with Ethical AI*, Apress, 185-253. https://doi.org/10.1007/979-8-8688-1186-9_6

[12] Zhang, H., Mao, Y., Lin, Y. and Zhang, D. (2025) E-mental Health in the Age of AI: Data Safety, Privacy Regulations and Recommendations. *Alpha Psychiatry*, **26**, Article No. 44279. https://doi.org/10.31083/ap44279

[13] Hazeem, H. and AlBurshaid, E. (2024) Fragmented Data Landscape and Data Asymmetries in the Real Estate Industry. In: Jreisat, A. and Mili, M., Eds., *Blockchain in Real Estate*, Springer, 179-205. https://doi.org/10.1007/978-981-99-8533-3_10

[14] Murphy, N.C., Burke, N., Breathnach, F.M., Burke, G., McAuliffe, F.M., Morrison, J.J., *et al.* (2020) Inter-Hospital Comparison of Cesarean Delivery Rates Should Not Be Considered to Reflect Quality of Care without Consideration of Patient Heterogeneity: An Observational Study. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, **250**, 112-116. https://doi.org/10.1016/j.ejogrb.2020.05.003

[15] Fassett, K.T., Wolcott, M.D., Harpe, S.E. and McLaughlin, J.E. (2022) Considerations for Writing and Including Demographic Variables in Education Research. *Currents in Pharmacy Teaching and Learning*, **14**, 1068-1078. https://doi.org/10.1016/j.cptl.2022.07.022

[16] Spector, S.L., Nicklas, R.A., Chapman, J.A., Bernstein, I.L., Berger, W.E., Blessing-Moore, J., *et al.* (2003) Symptom Severity Assessment of Allergic Rhinitis: Part 1. *Annals of Allergy, Asthma & Immunology*, **91**, 105-114. https://doi.org/10.1016/s1081-1206(10)62160-6

[17] Marmot, M. and Wilkinson, R. (2005) Social Determinants of Health. Oxford University Press.

[18] Long, G., Shen, T., Jiang, J. and Blumenstein, M. (2024) Dual-Personalizing Adapter for federated Foundation Models. *Advances in Neural Information Processing Systems*, **37**, 39409-39433. https://doi.org/10.52202/079017-1245

[19] Aggarwal, M., Khullar, V. and Goyal, N. (2024) A Comprehensive Review of Federated Learning: Methods, Applications, and Challenges in Privacy-Preserving Collab-

orative Model Training. In: Singh, J., Goyal, S.B., Kaushal, R.K., Kumar, N. and Sehra, S.S., Eds., *Applied Data Science and Smart Systems*, CRC Press, 570-575. https://doi.org/10.1201/9781003471059-73

[20] Hasan, M.T. and Kudapa, S.P. (2021) Data Privacy-Aware Machine Learning and Federated Learning: A Framework for Data Security. *American Journal of Interdisciplinary Studies*, **2**, 1-34. https://doi.org/10.63125/vj1hem03

[21] Chowdhury, T.K. and Kudapa, S.P. (2024) Federated Learning Models for Privacy-Preserving Data Sharing and Secure Analytics in Healthcare Industry. *International Journal of Business and Economics Insights*, **4**, 91-133. https://doi.org/10.63125/c2dzn006

[22] Lyu, L., Yu, J., Nandakumar, K., Li, Y., Ma, X., Jin, J., *et al.* (2020) Towards Fair and Privacy-Preserving Federated Deep Models. *IEEE Transactions on Parallel and Distributed Systems*, **31**, 2524-2541. https://doi.org/10.1109/tpds.2020.2996273

[23] Bokhari, M.U., Yadav, G. and Zeyauddin, M. (2024) Exploring Ensemble-Based Approaches for Granular Suicide Risk Assessment: A Comprehensive Framework in Therapeutic Informatics. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-024-02060-0

[24] Tran, T., Phung, D., Luo, W., Harvey, R., Berk, M. and Venkatesh, S. (2013) An Integrated Framework for Suicide Risk Prediction. *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, 11-14 August 2013, 1410-1418. https://doi.org/10.1145/2487575.2488196

[25] Dormont, B. and Milcent, C. (2005) How to Regulate Heterogeneous Hospitals? *Journal of Economics & Management Strategy*, **14**, 591-621. https://doi.org/10.1111/j.1530-9134.2005.00075.x

[26] Wang, H.Y., Yurochkin, M., Sun, Y., Papailiopoulos, D. and Khazaeni, Y. (2020) Federated Learning with Matched Averaging. arXiv: 2002.06440.

[27] Giuseppi, A., Torre, L.D., Menegatti, D., Priscoli, F.D., Pietrabissa, A. and Poli, C. (2022) An Adaptive Model Averaging Procedure for Federated Learning (AdaFed). *Journal of Advances in Information Technology*, **13**, 539-548. https://doi.org/10.12720/jait.13.6.539-548

[28] Manoj, T., Makkithaya, K. and Narendra, V.G. (2022) A Federated Learning-Based Crop Yield Prediction for Agricultural Production Risk Management. 2022 *IEEE Delhi Section Conference* (*DELCON*), New Delhi, 11-13 February 2022, 1-7. https://doi.org/10.1109/delcon54057.2022.9752836

[29] Fiez, T., Chasnov, B. and Ratliff, L.J. (2019) Convergence of Learning Dynamics in Stackelberg Games. arXiv: 1906.01217.

[30] Brewer, B.B., Carley, K.M., Benham-Hutchins, M., Effken, J.A. and Reminga, J. (2020) Exploring the Stability of Communication Network Metrics in a Dynamic Nursing Context. *Social Networks*, **61**, 11-19. https://doi.org/10.1016/j.socnet.2019.08.003

[31] Han, D., Kolli, K.K., Gransar, H., Lee, J.H., Choi, S., Chun, E.J., *et al.* (2020) Machine Learning Based Risk Prediction Model for Asymptomatic Individuals Who Underwent Coronary Artery Calcium Score: Comparison with Traditional Risk Prediction Approaches. *Journal of Cardiovascular Computed Tomography*, **14**, 168-176. https://doi.org/10.1016/j.jcct.2019.09.005

[32] Lyssenko, V., Jonsson, A., Almgren, P., Pulizzi, N., Isomaa, B., Tuomi, T., *et al.* (2008) Clinical Risk Factors, DNA Variants, and the Development of Type 2 Diabetes. *New England Journal of Medicine*, **359**, 2220-2232. https://doi.org/10.1056/nejmoa0801869

[33] Özyürekoğlu, T., McCabe, S.J., Goldsmith, L.J. and LaJoie, A.S. (2006) The Minimal Clinically Important Difference of the Carpal Tunnel Syndrome Symptom Severity Scale. *The Journal of Hand Surgery*, **31**, 733-738. https://doi.org/10.1016/j.jhsa.2006.01.012

[34] Buckley, P.F., Miller, B.J., Lehrer, D.S. and Castle, D.J. (2008) Psychiatric Comorbidities and Schizophrenia. *Schizophrenia Bulletin*, **35**, 383-402. https://doi.org/10.1093/schbul/sbn135

[35] Anestis, M.D. (2016) Prior Suicide Attempts Are Less Common in Suicide Decedents Who Died by Firearms Relative to Those Who Died by Other Means. *Journal of Affective Disorders*, **189**, 106-109. https://doi.org/10.1016/j.jad.2015.09.007

[36] Rutter, M. and Sandberg, S. (1992) Psychosocial Stressors: Concepts, Causes and Effects. *European Child & Adolescent Psychiatry*, **1**, 3-13. https://doi.org/10.1007/bf02084429

[37] Kuhle, S., Maguire, B., Zhang, H., Hamilton, D., Allen, A.C., Joseph, K.S., *et al.* (2018) Comparison of Logistic Regression with Machine Learning Methods for the Prediction of Fetal Growth Abnormalities: A Retrospective Cohort Study. *BMC Pregnancy and Childbirth*, **18**, Article No. 333. https://doi.org/10.1186/s12884-018-1971-2

[38] Park, S. and Byun, J. (2021) A Study of Predictive Models for Early Outcomes of Post-Prostatectomy Incontinence: Machine Learning Approach vs. Logistic Regression Analysis Approach. *Applied Sciences*, **11**, Article 6225. https://doi.org/10.3390/app11136225

[39] Clark, L.A. and Pregibon, D. (2017) Tree-based Models. In: Hastie, T.J., Ed., *Statistical Models in S*, Routledge, 377-419. https://doi.org/10.1201/9780203738535-9

[40] Amini, P., Ahmadinia, H., Poorolajal, J. and Amiri, M.M. (2016) Evaluating the High Risk Groups for Suicide: A Comparison of Logistic Regression, Support Vector Machine, Decision Tree and Artificial Neural Network. *Iranian Journal of Public Health*, **45**, 1179-1187.

[41] Švejdar, V. (1978) Degrees of Interpretability. *Commentationes Mathematicae Universitatis Carolinae*, **19**, 789-813.

[42] Yang, J., Soltan, A.A.S. and Clifton, D.A. (2022) Machine Learning Generalizability across Healthcare Settings: Insights from Multi-Site COVID-19 Screening. *npj Digital Medicine*, **5**, Article No. 69. https://doi.org/10.1038/s41746-022-00614-9

[43] Tran, A.T., Zeevi, T. and Payabvash, S. (2025) Strategies to Improve the Robustness and Generalizability of Deep Learning Segmentation and Classification in Neuroimaging. *BioMedInformatics*, **5**, Article 20. https://doi.org/10.3390/biomedinformatics5020020

[44] Liu, J. and Cui, P. (2025) Data Heterogeneity Modeling for Trustworthy Machine Learning. *Proceedings of the* 31*st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* V.2, Toronto, 3-7 August 2025, 6086-6095. https://doi.org/10.1145/3711896.3736560

[45] Paschali, M., Conjeti, S., Navarro, F. and Navab, N. (2018) Generalizability Vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples. In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C. and Fichtinger, G., Eds., *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2018, Springer International Publishing, 493-501. https://doi.org/10.1007/978-3-030-00928-1_56

[46] Kazmierski, M., Welch, M., Kim, S., McIntosh, C., Rey-McIntyre, K., Huang, S.H., *et al.* (2023) Multi-Institutional Prognostic Modeling in Head and Neck Cancer: Evalu-

ating Impact and Generalizability of Deep Learning and Radiomics. *Cancer Research Communications*, **3**, 1140-1151. https://doi.org/10.1158/2767-9764.crc-22-0152

[47] Jayne, M.E.A. and Dipboye, R.L. (2004) Leveraging Diversity to Improve Business Performance: Research Findings and Recommendations for Organizations. *Human Resource Management*, **43**, 409-424. https://doi.org/10.1002/hrm.20033

[48] Anthonysamy, P., Rashid, A. and Chitchyan, R. (2017) Privacy Requirements: Present & Future. 2017 *IEEE/ACM 39th International Conference on Software Engineering: Software Engineering in Society Track* (*ICSE-SEIS*), Bueons Aires, 20-28 May 2017, 13-22. https://doi.org/10.1109/icse-seis.2017.3

[49] Alhassan, I., Sammon, D. and Daly, M. (2016) Data Governance Activities: An Analysis of the Literature. *Journal of Decision Systems*, **25**, 64-75. https://doi.org/10.1080/12460125.2016.1187397

[50] van Panhuis, W.G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A.J., *et al.* (2014) A Systematic Review of Barriers to Data Sharing in Public Health. *BMC Public Health*, **14**, Article No. 1144. https://doi.org/10.1186/1471-2458-14-1144

[51] Fassnacht, M., Benz, C., Heinz, D., Leimstoll, J. and Satzger, G. (2023) Barriers to Data Sharing among Private Sector Organizations. *Proceedings of the Annual Hawaii International Conference on System Sciences*, Maui, 3-6 January 2023, 3695-3704. https://doi.org/10.24251/hicss.2023.453

[52] Bensusan, H., and Kalousis, A. (2001) Estimating the Predictive Accuracy of a Classifier. In: *European Conference on Machine Learning*, Springer Berlin Heidelberg, 25-36. https://doi.org/10.1007/3-540-44795-4_3

[53] Naiseh, M., Al-Thani, D., Jiang, N. and Ali, R. (2023) How the Different Explanation Classes Impact Trust Calibration: The Case of Clinical Decision Support Systems. *International Journal of Human-Computer Studies*, **169**, Article ID: 102941. https://doi.org/10.1016/j.ijhcs.2022.102941

[54] Alsoud, D., Sabino, J., Ferrante, M., Verstockt, B. and Vermeire, S. (2025) Calibration, Clinical Utility, and Specificity of Clinical Decision Support Tools in Inflammatory Bowel Disease. *Clinical Gastroenterology and Hepatology*, **23**, 1216-1227.e14. https://doi.org/10.1016/j.cgh.2024.09.020

[55] Goe, L., Holdheide, L. and Miller, T. (2011) A Practical Guide to Designing Comprehensive Teacher Evaluation Systems: A Tool to Assist in the Development of Teacher Evaluation Systems. National Comprehensive Center for Teacher Quality.

[56] Wind, A., Oberst, S., Westerhuis, W., Blaauwgeers, H., Sæter, G., de Paoli, P., *et al.* (2023) Evaluating Comprehensive Cancer Networks; a Review of Standards and Evaluation Methods for Care Networks to Inform a Comparison with the OECI Comprehensive Cancer Network Standards. *Acta Oncologica*, **62**, 15-24. https://doi.org/10.1080/0284186x.2023.2170275

[57] Shankar, P.M. (2019) Pedagogy of Bayes' Rule, Confusion Matrix, Transition Matrix, and Receiver Operating Characteristics. *Computer Applications in Engineering Education*, **27**, 510-518. https://doi.org/10.1002/cae.22093

[58] Linden, A. (2006) Measuring Diagnostic and Predictive Accuracy in Disease Management: An Introduction to Receiver Operating Characteristic (ROC) Analysis. *Journal of Evaluation in Clinical Practice*, **12**, 132-139. https://doi.org/10.1111/j.1365-2753.2005.00598.x

[59] Yang, S. and Berdine, G. (2017) The Receiver Operating Characteristic (ROC) Curve. *The Southwest Respiratory and Critical Care Chronicles*, **5**, 34-36. https://doi.org/10.12746/swrccc.v5i19.391

[60] Miller, M., Hafner, M., Sontag, E., Davidsohn, N., Subramanian, S., *et al.* (2012) Mod-

ular Design of Artificial Tissue Homeostasis: Robust Control through Synthetic Cellular Heterogeneity. *PLOS Computational Biology*, **8**, e1002579. https://doi.org/10.1371/journal.pcbi.1002579

[61] Doudchenko, N., Khosravi, K., Pouget-Abadie, J., Lahaie, S., Lubin, M., Mirrokni, V. and Spiess, J. (2021) Synthetic Design: An Optimization Approach to Experimental Design with Synthetic Controls. *Advances in Neural In-formation Processing Systems*, **34**, 8691-8701.

[62] Zhou, M., Li, J., Basu, R. and Ferreira, J. (2022) Creating Spatially-Detailed Heterogeneous Synthetic Populations for Agent-Based Microsimulation. *Computers, Environment and Urban Systems*, **91**, Article ID: 101717. https://doi.org/10.1016/j.compenvurbsys.2021.101717

[63] Fins, J.J., Rezai, A.R. and Greenberg, B.D. (2006) Psychosurgery: Avoiding an Ethical Redux While Advancing a Therapeutic Future. *Neurosurgery*, **59**, 713-716. https://doi.org/10.1227/01.neu.0000243605.89270.6c

[64] Ahalt, S.C., Chute, C.G., Fecho, K., Glusman, G., Hadlock, J., Taylor, C.O., *et al.* (2019) Clinical Data: Sources and Types, Regulatory Constraints, Applications. *Clinical and Translational Science*, **12**, 329-333. https://doi.org/10.1111/cts.12638

[65] Byrnes, J.E.K. and Dee, L.E. (2025) Causal Inference with Observational Data and Unobserved Confounding Variables. *Ecology Letters*, **28**, e70023. https://doi.org/10.1111/ele.70023

[66] Tchetgen Tchetgen, E. (2013) The Control Outcome Calibration Approach for Causal Inference with Unobserved Confounding. *American Journal of Epidemiology*, **179**, 633-640. https://doi.org/10.1093/aje/kwt303

[67] Bovens, A.M.P.M., van Baak, M.A., Vrencken, J.G.P.M., Wijnen, J.A.G. and Verstappen, F.T.J. (1990) Variability and Reliability of Joint Measurements. *The American Journal of Sports Medicine*, **18**, 58-63. https://doi.org/10.1177/036354659001800110

[68] Medina, L.S. and Zurakowski, D. (2003) Measurement Variability and Confidence Intervals in Medicine: Why Should Radiologists Care? *Radiology*, **226**, 297-301. https://doi.org/10.1148/radiol.2262011537

[69] Blonde, L., Khunti, K., Harris, S.B., Meizinger, C. and Skolnik, N.S. (2018) Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. *Advances in Therapy*, **35**, 1763-1774. https://doi.org/10.1007/s12325-018-0805-y

[70] Chen, J., Ho, M., Lee, K., Song, Y., Fang, Y., Goldstein, B.A., *et al.* (2021) The Current Landscape in Biostatistics of Real-World Data and Evidence: Clinical Study Design and Analysis. *Statistics in Biopharmaceutical Research*, **15**, 29-42. https://doi.org/10.1080/19466315.2021.1883474

[71] Ziemer, R. (1967) Error Probabilities Due to Additive Combinations of Gaussian and Impulsive Noise. *IEEE Transactions on Communication Technology*, **15**, 471-474. https://doi.org/10.1109/tcom.1967.1089608

[72] Le Montagner, Y., Angelini, E.D. and Olivo-Marin, J. (2014) An Unbiased Risk Estimator for Image Denoising in the Presence of Mixed Poisson-Gaussian Noise. *IEEE Transactions on Image Processing*, **23**, 1255-1268. https://doi.org/10.1109/tip.2014.2300821

[73] Li, N., Kolmanovsky, I. and Girard, A. (2021) An Analytical Safe Approximation to Joint Chance-Constrained Programming with Additive Gaussian Noises. *IEEE Transactions on Automatic Control*, **66**, 5490-5497. https://doi.org/10.1109/tac.2021.3051000

[74] Arpit, D., Zhou, Y.B., Kota, B. and Govindaraju, V. (2016) Normalization Propagation: A Parametric Technique for Removing Internal Covariate Shift in Deep Net-

works. *International Conference on Machine Learning* 2016, New York, 19-24 June 2016, 1168-1176.

[75] Awais, M., Bin Iqbal, M.T. and Bae, S. (2021) Revisiting Internal Covariate Shift for Batch Normalization. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 5082-5092. https://doi.org/10.1109/tnnls.2020.3026784

[76] Kiliçarslan, S. and Celik, M. (2021) Rsigelu: A Nonlinear Activation Function for Deep Neural Networks. *Expert Systems with Applications*, **174**, Article ID: 114805. https://doi.org/10.1016/j.eswa.2021.114805

[77] Banerjee, C., Mukherjee, T. and Pasiliao, E. (2020) The Multi-Phase ReLU Activation Function. *Proceedings of the* 2020 *ACM Southeast Conference*, Tampa, 2-4 April 2020, 239-242. https://doi.org/10.1145/3374135.3385313

[78] Rahman, S., Jiang, L.Y., Gabriel, S., Aphinyanaphongs, Y., Oermann, E.K. and Chunara, R. (2024) Generalization in Healthcare AI: Evaluation of a Clinical Large Language Model. arXiv: 2402.10965.

[79] Zhang, Z.L. and Sabuncu, M.R. (2018) Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *Proceedings of the* 32*nd International Conference on Neural Information Processing Systems*, Montréal, 3-8 December 2018, 8792-8802.

[80] Wu, Y., Du, K., Wang, X. and Min, F. (2024) Misclassification-Guided Loss under the Weighted Cross-Entropy Loss Framework. *Knowledge and Information Systems*, **66**, 4685-4720. https://doi.org/10.1007/s10115-024-02123-5

[81] Bai, Z., Wang, J., Zhang, X. and Chen, J. (2022) End-To-End Speaker Verification via Curriculum Bipartite Ranking Weighted Binary Cross-Entropy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **30**, 1330-1344. https://doi.org/10.1109/taslp.2022.3161155

[82] Ahmed, A., Sun, G., Saadeldin, M., Bilal, A., Li, Y., Osman, M., *et al.* (2025) Efficient Melanoma Detection Using Pixel Intensity-Based Masking and Intensity-Weighted Binary Cross-Entropy. *International Journal of Imaging Systems and Technology*, **35**, e70179. https://doi.org/10.1002/ima.70179

[83] Rachakonda, S., Moorthy, S., Jain, A., Bukharev, A., Bucur, A., Manni, F., *et al.* (2023) Privacy Enhancing and Scalable Federated Learning to Accelerate AI Implementation in Cross-Silo and IoMT Environments. *IEEE Journal of Biomedical and Health Informatics*, **27**, 744-755. https://doi.org/10.1109/jbhi.2022.3185418

[84] Albarqouni, S., Andreux, M., Avestimehr, S., Ayed, S., Bellet, A., Cyffers, E., *et al.* (2022) Flamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings. *Advances in Neural Information Processing Systems* 35, New Orleans, 28 November-9 December 2022, 5315-5334. https://doi.org/10.52202/068431-0384

[85] Huang, C., Huang, J.W. and Liu, X. (2022) Cross-Silo Federated Learning: Challenges and Opportunities. arXiv: 2206.12949.

[86] Hamm, B., Kirchhoff, Y., Rokuss, M., Schader, P., Neher, P., Parampottupadam, S., Floca, R. and Maier-Hein, K. (2025) Efficient Privacy-Preserving Medical Cross-Silo Federated Learning. TechRxiv. https://doi.org/10.36227/techrxiv.174650601.13181048/v1

[87] Pillutla, K., Kakade, S.M. and Harchaoui, Z. (2022) Robust Aggregation for Federated Learning. *IEEE Transactions on Signal Processing*, **70**, 1142-1154. https://doi.org/10.1109/tsp.2022.3153135

[88] Ganguly, B., Hosseinalipour, S., Kim, K.T., Brinton, C.G., Aggarwal, V., Love, D.J., *et*

al. (2023) Multi-Edge Server-Assisted Dynamic Federated Learning with an Optimized Floating Aggregation Point. *IEEE/ACM Transactions on Networking*, **31**, 2682-2697. https://doi.org/10.1109/tnet.2023.3262482

[89] Khan, N., Nisar, S., Khan, M.A., Ur Rehman, Y.A., Noor, F. and Barb, G. (2025) Optimizing Federated Learning with Aggregation Strategies: A Comprehensive Survey. *IEEE Open Journal of the Computer Society*, **6**, 1227-1247. https://doi.org/10.1109/ojcs.2025.3590102

[90] Volinsky, C.T., Madigan, D., Raftery, A.E. and Kronmal, R.A. (1997) Bayesian Model Averaging in Proportional Hazard Models: Assessing the Risk of a Stroke. *Journal of the Royal Statistical Society Series C*: *Applied Statistics*, **46**, 433-448. https://doi.org/10.1111/1467-9876.00082

[91] Kelley, J.G. and Simmons, B.A. (2019) Introduction: The Power of Global Performance Indicators. *International Organization*, **73**, 491-510. https://doi.org/10.1017/s0020818319000146

[92] Ruby, U. and Yendapalli, V. (2020) Binary Cross Entropy with Deep Learning Technique for Image Classification. *International Journal of Advanced Trends in Computer Science and Engineering*, **9**, 5393-5397. https://doi.org/10.30534/ijatcse/2020/175942020

[93] Tafvizi, A., Avci, B. and Sundararajan, M. (2022) Attributing AUC-ROC to Analyze Binary Classifier Performance. arXiv: 2205.11781.

[94] Tang, X., Guo, S. and Guo, J. (2022) Personalized Federated Learning with Contextualized Generalization. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, Vienna, 23-29 July 2022, 2241-2247. https://doi.org/10.24963/ijcai.2022/311

[95] Marulli, F., Verde, L., Marrone, S., Barone, R. and De Biase, M.S. (2021) Evaluating Efficiency and Effectiveness of Federated Learning Approaches in Knowledge Extraction Tasks. 2021 *International Joint Conference on Neural Networks* (*IJCNN*), Shenzhen, 18-22 July 2021, 1-6. https://doi.org/10.1109/ijcnn52387.2021.9533946

[96] Ali, W., Kumar, R., Deng, Z., Wang, Y. and Shao, J. (2021) A Federated Learning Approach for Privacy Protection in Context-Aware Recommender Systems. *The Computer Journal*, **64**, 1016-1027. https://doi.org/10.1093/comjnl/bxab025

[97] Elbir, A.M., Coleri, S. and Mishra, K.V. (2021) Hybrid Federated and Centralized Learning. 2021 29*th European Signal Processing Conference* (*EUSIPCO*), Dublin, 23-27 August 2021, 1541-1545. https://doi.org/10.23919/eusipco54536.2021.9616120

[98] Elsken, T., Metzen, J.H. and Hutter, F. (2019) Correction To: Neural Architecture Search. In: Hutter, F., Kotthoff, L. and Vanschoren, J., Eds., *Automated Machine Learning*, Springer, C1. https://doi.org/10.1007/978-3-030-05318-5_11

[99] Meyer, R. (1970) The Validity of a Family of Optimization Methods. *SIAM Journal on Control*, **8**, 41-54. https://doi.org/10.1137/0308003

[100] Li, C., Li, C., Zhao, Y., Zhang, B. and Li, C. (2021) An Efficient Multi-Model Training Algorithm for Federated Learning. 2021 *IEEE Global Communications Conference* (*GLOBECOM*), Madrid, 7-11 December 2021, 1-6. https://doi.org/10.1109/globecom46510.2021.9685230

[101] Tagg, J. (2003) The Learning Paradigm. Anker.

[102] Jiménez-Valverde, A. (2014) Threshold-Dependence as a Desirable Attribute for Discrimination Assessment: Implications for the Evaluation of Species Distribution Models. *Biodiversity and Conservation*, **23**, 369-385. https://doi.org/10.1007/s10531-013-0606-1

[103] Abdulrazaq, M. (2023) Rare-Event Prediction in Imbalanced Data: A Unified Evaluation and Optimization Framework for High-Risk Systems. *Communication In Physical Sciences*, **9**, 968-979.

[104] Iturbe-Araya, J. and Rifà-Pous, H. (2025) Hyperparameter Optimization and Evaluation Metrics for Unsupervised Anomaly-Based Cyberattack Detection in Imbalanced Smart Home Datasets. *Journal of Network and Systems Management*, **33**, Article No. 99. https://doi.org/10.1007/s10922-025-09973-6

[105] Pepe, M.S., Longton, G. and Janes, H. (2009) Estimation and Comparison of Receiver Operating Characteristic Curves. *The Stata Journal*: *Promoting communications on statistics and Stata*, **9**, 1-16. https://doi.org/10.1177/1536867x0900900101

[106] Movahedi, F., Padman, R. and Antaki, J.F. (2023) Limitations of Receiver Operating Characteristic Curve on Imbalanced Data: Assist Device Mortality Risk Scores. *The Journal of Thoracic and Cardiovascular Surgery*, **165**, 1433-1442.e2. https://doi.org/10.1016/j.jtcvs.2021.07.041

[107] Jiang, Y. (2020) Receiver Operating Characteristic (ROC) Analysis of Image Search-And-Localize Tasks. *Academic Radiology*, **27**, 1742-1750. https://doi.org/10.1016/j.acra.2019.12.020

[108] Grau, J., Grosse, I. and Keilwagen, J. (2015) PRROC: Computing and Visualizing Precision-Recall and Receiver Operating Characteristic Curves in R. *Bioinformatics*, **31**, 2595-2597. https://doi.org/10.1093/bioinformatics/btv153

[109] Carter, J.V., Pan, J., Rai, S.N. and Galandiuk, S. (2016) ROC-Ing Along: Evaluation and Interpretation of Receiver Operating Characteristic Curves. *Surgery*, **159**, 1638-1645. https://doi.org/10.1016/j.surg.2015.12.029

[110] Obuchowski, N.A. and Bullen, J.A. (2018) Receiver Operating Characteristic (ROC) Curves: Review of Methods with Applications in Diagnostic Medicine. *Physics in Medicine & Biology*, **63**, 07TR01. https://doi.org/10.1088/1361-6560/aab4b1

[111] Van der Maat, E. (2021) Simplified Complexity: Analytical Strategies for Conflict Event Research. *Conflict Management and Peace Science*, **38**, 87-108. https://doi.org/10.1177/0738894218771077

[112] Davis, J. and Goadrich, M. (2006) The Relationship between Precision-Recall and ROC Curves. *Proceedings of the* 23*rd International Conference on Machine Learning—ICML*'06, Pittsburgh, 25-29 June 2006, 233-240. https://doi.org/10.1145/1143844.1143874

[113] Doumpos, M. and Pasiouras, F. (2005) Developing and Testing Models for Replicating Credit Ratings: A Multicriteria Approach. *Computational Economics*, **25**, 327-341. https://doi.org/10.1007/s10614-005-6412-4

[114] Jin Huang, and Ling, C.X. (2005) Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 299-310. https://doi.org/10.1109/tkde.2005.50

[115] Carrington, A.M., Manuel, D.G., Fieguth, P.W., Ramsay, T., Osmani, V., Wernly, B., *et al.* (2023) Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 329-341. https://doi.org/10.1109/tpami.2022.3145392

[116] Kovacs, I., Iosub, A., Ţopa, M., Buzo, A. and Pelz, G. (2019) A Gradient-Based Sensitivity Analysis Method for Complex Systems. 2019 *IEEE* 25*th International Symposium for Design and Technology in Electronic Packaging* (*SIITME*), Cluj-Napoca, 23-26 October 2019, 333-338. https://doi.org/10.1109/siitme47687.2019.8990871

[117] Lin, N. and Ensel, W.M. (1989) Life Stress and Health: Stressors and Resources. *American Sociological Review*, **54**, 382-399. https://doi.org/10.2307/2095612

[118] Schwarzer, R., Knoll, N. and Rieckmann, N. (2004) Social Support. *Health Psychology*, **158**, 181.

[119] Hovenkamp-Hermelink, J.H.M., van der Veen, D.C., Oude Voshaar, R.C., Batelaan, N.M., Penninx, B.W.J.H., Jeronimus, B.F., *et al.* (2019) Anxiety Sensitivity, Its Stability and Longitudinal Association with Severity of Anxiety Symptoms. *Scientific Reports*, **9**, Article No. 4314. https://doi.org/10.1038/s41598-019-39931-7

[120] Pavlou, M., Ambler, G., Seaman, S.R., Guttmann, O., Elliott, P., King, M., *et al.* (2015) How to Develop a More Accurate Risk Prediction Model When There Are Few Events. *BMJ*, **351**, h3868. https://doi.org/10.1136/bmj.h3868

[121] Yu, J., Yang, X., Deng, Y., Krefman, A.E., Pool, L.R., Zhao, L., *et al.* (2024) Incorporating Longitudinal History of Risk Factors into Atherosclerotic Cardiovascular Disease Risk Prediction Using Deep Learning. *Scientific Reports*, **14**, Article No. 2554. https://doi.org/10.1038/s41598-024-51685-5

[122] Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., *et al.* (2022) Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond. *Knowledge and Information Systems*, **64**, 3197-3234. https://doi.org/10.1007/s10115-022-01756-8

[123] Hamon, R., Junklewitz, H. and Sanchez, I. (2020) Robustness and Explainability of Artificial Intelligence. Publications Office of the European Union 207, No. 40.

[124] Battula, J., Jillelamudi, V.A., Sammeta, C.K. and Amilpur, S. (2025) Exploring Cancer Genomics with Graph Convolutional Networks: A Comparative Explainability Study with Integrated Gradients and SHAP. *BIO Web of Conferences*, **163**, Article ID: 01003. https://doi.org/10.1051/bioconf/202516301003

[125] Devereux, P.J. (2007) Small-Sample Bias in Synthetic Cohort Models of Labor Supply. *Journal of Applied Econometrics*, **22**, 839-848. https://doi.org/10.1002/jae.938

[126] Gupta, S., Kumar, S., Chang, K., Lu, C., Singh, P. and Kalpathy-Cramer, J. (2023) Collaborative Privacy-Preserving Approaches for Distributed Deep Learning Using Multi-Institutional Data. *RadioGraphics*, **43**, e220107. https://doi.org/10.1148/rg.220107

[127] Monika, T., Kishor Kumar Reddy, C., Puttanapura, J. and Doss, S. (2025) Optimizing Neural Disorder Treatment through Federated Learning and Multi-Institutional Data Collaboration. In: Kishor Kumar Reddy, C. and Nag, A., Eds., *Federated Learning for Neural Disorders in Healthcare* 6.0, CRC Press, 120-157.

[128] Guo, P., Wang, P., Zhou, J., Jiang, S. and Patel, V.M. (2021) Multi-Institutional Collaborations for Improving Deep Learning-Based Magnetic Resonance Image Reconstruction Using Federated Learning. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 2423-2432. https://doi.org/10.1109/cvpr46437.2021.00245

[129] Marwaha, J.S., Downing, M., Halamka, J., Abernethy, A., Franklin, J.B., Anderson, B., *et al.* (2024) Mobilizing Data during a Crisis: Building Rapid Evidence Pipelines Using Multi-Institutional Real World Data. *Healthcare*, **12**, Article ID: 100738. https://doi.org/10.1016/j.hjdsi.2024.100738