# Knowledge Distillation in Split Learning for Heterogeneous Clinical Tabular Data

by

Tasneem Jahan Farheen
24366032

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
March 2026

# Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.

**Student's Full Name & Signature:**

Tasneem Jahan Farheen
24366032

# Approval

The thesis titled "*Knowledge Distillation in Split Learning for Heterogeneous Clinical Tabular Data*" submitted by

1. Tasneem Jahan Farheen (24366032)

of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science and Engineering on March 11, 2026.

**Examining Committee:**

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

External Examiner:
(Member)

Dr. Chowdhury Farhan Ahmed, Ph.D.
Professor
Department of Computer Science and Engineering
University of Dhaka

Internal Examiner:
(Member)

Dr. Muhammad Iqbal Hossain
Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md Sadek Ferdous
Professor
Department of Computer Science and Engineering
BRAC University

Departmental Head:
(Chair)

Sadia Hamid Kazi, Ph.D.
Associate Professor
Department of Computer Science and Engineering
BRAC University

# Ethics Statement

This study uses only publicly available, de-identified datasets: the National Health and Nutrition Examination Survey (NHANES), the Health and Retirement Study (HRS), the Pima Indians Diabetes Dataset, and the Early-Stage Diabetes Risk Prediction Dataset collected from the Sylhet Diabetes Hospital, Bangladesh. No new patient data were collected, and there was no interaction with human subjects. All datasets are anonymized and released for research in accordance with ethical and data protection standards.The proposed framework enhances privacy by keeping raw patient data local and transmitting only intermediate model activations during training. No identifiable personal information is accessed or reconstructed at any stage.This research is strictly methodological and computational, with no clinical decision-making or direct patient intervention involved.

# Abstract

The growing availability of clinical data across institutions creates new opportunities for improved disease risk prediction. However, privacy regulations, institutional policies, and limited labeled data restrict centralized model training, particularly in low-resource settings. To address these challenges, this study presents a privacy-preserving framework that integrates split learning and knowledge distillation for diabetes risk prediction using heterogeneous tabular datasets. This work systematically evaluates two complementary transfer mechanisms. The first approach utilizes logit-level knowledge distillation to transfer predictive decision boundaries from high-capacity teacher models, collaboratively trained on large population datasets, to lightweight student models in low-resource target domains. The second approach implements encoder-level distillation to align latent feature representations within a split learning architecture. Both approaches maintain strict data privacy by keeping raw patient data local and exchanging only intermediate activations. Experimental results under few-shot supervision indicate that logit-level distillation substantially improves discriminative performance, particularly in extremely low-label scenarios, while stabilizing training across heterogeneous domains. Encoder-level distillation further improves probabilistic calibration and representation alignment under cross-domain distribution shifts. These findings underscore the significance of structured knowledge transfer in privacy-preserving clinical machine learning and offer practical recommendations for deploying reliable models in heterogeneous healthcare environments.

**Keywords:** Split learning; knowledge distillation; few-shot learning; tabular transformers; clinical prediction

# Acknowledgement

First and foremost, I express my sincere gratitude to Almighty Allah for granting me the strength, patience and guidance to successfully complete this thesis. I extend my deepest appreciation to my supervisor, Mr. Md. Golam Rabiul Alam, for his invaluable guidance, support and insightful feedback throughout this research. His encouragement and expertise greatly contributed to the successful completion of this work. I am also grateful to the faculty members of the Department of Computer Science and Engineering at BRAC University for their academic support and knowledge, which laid the foundation for this research. Finally, I thank my parents and family for their unconditional love, encouragement and prayers. Without their constant support, this achievement would not have been possible.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Diabetes is a chronic metabolic disorder that represents a significant global health challenge due to its high prevalence and associated severe complications. Poorly managed diabetes can result in serious conditions such as cardiovascular disease, kidney failure, neuropathy, and lower-limb amputations. These complications substantially diminish quality of life and elevate the risk of premature death. In recent decades, the prevalence of diabetes among adults has increased rapidly, especially in low- and middle-income countries where healthcare infrastructure and preventive care are often inadequate. Research shows that approximately 11% of deaths from cardiovascular diseases are attributable to elevated blood glucose levels, underscoring the necessity for effective diabetes monitoring and management strategies [1].

As a result, early detection and prevention have become critical priorities in modern healthcare systems. Traditional clinical approaches to disease diagnosis rely heavily on manual analysis of clinical indicators and physician expertise. While these methods remain valuable, they may not fully leverage the large volumes of healthcare data now available through digital health systems. The increasing adoption of electronic health records, health surveys, and clinical registries has enabled healthcare organizations to collect extensive patient information, including demographic characteristics, laboratory test results, lifestyle factors, and medical history. These datasets present new opportunities for data-driven healthcare analytics and predictive modeling.

Machine learning techniques have increasingly been applied to healthcare data to support predictive analytics, disease risk assessment, and clinical decision support systems. By analyzing patterns within patient data, machine learning models can identify subtle relationships between clinical variables that may not be easily detected through traditional statistical analysis. Such predictive models have demonstrated promising results in assisting healthcare professionals with early disease detection, treatment planning, and population health management [2].

Despite these advancements, many machine learning approaches assume that all relevant data can be centralized within a single computational environment. In practice, however, healthcare data are typically distributed across multiple institutions such as hospitals, clinics, and research centers. Privacy regulations, institutional policies, and ethical considerations often prevent these organizations from sharing

raw patient data with external entities. Consequently, centralized machine learning approaches may not be feasible in many real-world healthcare settings, limiting the ability of models to benefit from diverse and geographically distributed datasets.
To address these challenges, distributed learning paradigms have been proposed to enable collaborative model training without requiring the exchange of raw data. One such approach is split learning, which divides a neural network into multiple segments that are trained collaboratively across different institutions. In this framework, the client trains the initial layers of the model using locally stored data and transmits intermediate feature representations to a central server. The server then completes the remaining layers of the network and performs backpropagation. Due to only intermediate activations being shared rather than raw patient data, split learning offers a promising solution for privacy-preserving collaborative machine learning in healthcare environments [3].

Another challenge arises from the nature of healthcare datasets themselves. Many clinical datasets are tabular, consisting of structured variables such as age, laboratory test values, physiological measurements, and behavioral indicators. Unlike image or text data, tabular datasets often contain heterogeneous feature types, missing values, and complex interdependencies between variables. These characteristics can make it difficult for deep learning models to capture meaningful representations and generalize effectively across datasets collected from different populations or healthcare systems.

## 1.1 Motivation

The increasing availability of healthcare data and advances in machine learning have created significant opportunities for improving disease prediction and patient care. Predictive models capable of identifying individuals at high risk of developing chronic conditions such as diabetes can enable earlier intervention and more personalized treatment strategies. Such models have the potential to support clinicians in making more informed decisions and ultimately improve patient outcomes.

However, several practical challenges limit the effectiveness of traditional machine learning approaches in healthcare applications. One major challenge is data privacy. Patient health records contain highly sensitive personal information, and strict regulations often restrict the sharing of these data across institutions. As a result, healthcare organizations may possess valuable datasets that cannot easily be combined with data from other institutions for collaborative analysis. This fragmentation of data reduces the ability of machine learning models to learn from diverse populations and limits their generalizability [4].In particular, split learning has been proposed as an effective privacy-preserving distributed learning paradigm that enables institutions to collaboratively train deep learning models without exposing raw patient data [5].

Another important challenge is the scarcity of labeled data for specific prediction tasks. In many clinical studies, collecting labeled data can be time-consuming and expensive because it requires expert annotation or long-term patient follow-up. Consequently, many healthcare datasets contain only limited numbers of labeled sam-

ples, particularly for rare diseases or specialized patient cohorts. Machine learning models trained on small datasets often struggle to capture complex patterns in the data and may produce unstable predictions.

Knowledge transfer techniques provide a potential solution to the problem of limited labeled data. Among these techniques, knowledge distillation has emerged as an effective approach for transferring information from a high-capacity model to a smaller or more efficient model. In knowledge distillation, a complex model referred to as the teacher is first trained on a large dataset, and its learned knowledge is then used to guide the training of a student model. Instead of relying solely on hard labels, the student model learns from the probability distributions produced by the teacher model, which contain richer information about the relationships between classes. This approach allows the student model to capture important decision boundaries learned by the teacher, even when the available training data are limited. Knowledge distillation has demonstrated strong performance improvements in various machine learning domains, particularly in computer vision and natural language processing tasks [6] , [7]. However, its application to tabular healthcare datasets remains relatively limited, especially in distributed learning settings where privacy constraints restrict direct data sharing.

Despite its success in other domains, knowledge distillation has received relatively limited attention in the context of tabular healthcare data. Furthermore, most existing studies on knowledge distillation assume centralized training environments where all data are available in a single location. Integrating knowledge distillation with privacy-preserving distributed learning frameworks such as split learning remains an underexplored research direction.

These challenges motivate the need for new machine learning frameworks capable of transferring knowledge across datasets while maintaining patient privacy and supporting learning under limited data conditions.

## 1.2 Problem Statement

Although distributed learning techniques such as split learning offer promising solutions for privacy-preserving model training, several important challenges remain when applying these approaches to healthcare data.

First, machine learning models typically require large amounts of labeled data to achieve reliable performance. In many healthcare scenarios, however, individual institutions may possess only small datasets for specific prediction tasks. Training models solely on these limited datasets can result in poor generalization and reduced predictive accuracy.

Second, healthcare datasets collected from different sources often exhibit substantial variability in feature distributions, patient demographics, and measurement protocols. This phenomenon, commonly referred to as domain shift, makes it difficult for models trained on one dataset to perform effectively when applied to another dataset. As a result, knowledge obtained from one population may not transfer

easily to another population.

Third, while knowledge distillation has proven effective for transferring knowledge between models, most existing research has focused on centralized machine learning settings and domains such as computer vision and natural language processing. Comparatively little work has explored how distillation techniques can be integrated with distributed learning frameworks for tabular healthcare data [5].

In addition, many existing studies focus only on transferring final model predictions, without considering the potential benefits of transferring intermediate feature representations learned by deep neural networks. Understanding how different forms of knowledge transfer influence model performance is particularly important when working with heterogeneous datasets and limited training samples.

These challenges highlight the need for research that investigates how knowledge distillation can be applied within split learning environments to improve predictive performance across heterogeneous healthcare datasets.

## 1.3 Research Objectives

The primary objective of this research is to investigate how knowledge distillation can be combined with split learning to enhance predictive performance for tabular healthcare datasets while preserving data privacy.

To achieve this goal, the study pursues the following objectives:

1. To develop a privacy-preserving learning framework that integrates split learning with knowledge distillation, enabling collaborative model training across distributed healthcare datasets.

2. To explore different knowledge transfer mechanisms, including logit-level distillation, where the student model learns from the teacher's prediction probabilities.

3. To investigate encoder-level knowledge distillation, in which the student model aligns its internal feature representations with those learned by the teacher model.

4. To evaluate the effectiveness of these knowledge transfer strategies under extreme few-shot learning conditions where only a small number of labeled samples are available in the target domain.

5. To compare the performance of different tabular deep learning architectures, including multilayer perceptrons (MLP), Self-Attention and Intersample Attention Transformer (SAINT), and Tabular Multiple Prediction (TabM) models, when trained using the proposed frameworks.

## 1.4 Research Contributions

This thesis contributes to the advancement of privacy-preserving machine learning for healthcare analytics by proposing and evaluating novel approaches for knowledge transfer across heterogeneous clinical datasets. The main contributions of this research are summarized as follows:

- **Proposed Unified Framework For Knowledge Distillation in Split Learning**
  This research proposes a unified framework that integrates knowledge distillation with split learning to enable knowledge transfer across distributed healthcare datasets while preserving patient privacy. In many healthcare environments, institutions cannot share raw patient data due to strict privacy regulations and ethical considerations. The proposed framework addresses this limitation by allowing models trained on one dataset to transfer their learned knowledge to models trained on another dataset without exposing sensitive patient information. By combining the collaborative training capabilities of split learning with the knowledge transfer mechanisms of distillation, the framework enables institutions to benefit from external knowledge while maintaining compliance with privacy constraints.

- **Investigation of Multiple Knowledge Transfer Strategies**
  This thesis systematically investigates two complementary forms of knowledge transfer: logit-level distillation and encoder-level representation alignment. Logit-level distillation focuses on transferring the prediction probabilities generated by the teacher model, allowing the student model to learn soft decision boundaries that capture richer information than hard labels alone. Encoder-level distillation, on the other hand, aligns the internal feature representations learned by the teacher and student models. By encouraging the student model to mimic the latent feature space of the teacher, this approach allows deeper structural knowledge to be transferred. Comparing these two strategies provides valuable insights into how different levels of knowledge transfer affect model learning behavior and predictive performance in tabular healthcare datasets.

- **Evaluation Under Extreme Few-Shot Learning Conditions**
  A significant contribution of this research is the systematic evaluation of the proposed frameworks under extreme few-shot learning scenarios. In many real-world healthcare applications, labeled datasets are limited due to the cost and difficulty of collecting high-quality medical annotations. To simulate these conditions, the experiments were designed to evaluate model performance when only a small number of labeled samples are available in the target domain. This evaluation demonstrates how knowledge distillation can improve model stability, generalization, and predictive accuracy even when training data are scarce. The results provide important insights into the effectiveness of teacher-guided learning in data-constrained healthcare environments.

- **Cross-Architecture Evaluation for Tabular Deep Learning Models**
  Another contribution of this thesis is the comprehensive evaluation of knowledge distillation across multiple tabular deep learning architectures. The study

examines how three different model types (MLP, SAINT and TabM) respond to teacher guidance under the proposed framework. These architectures represent different design philosophies and inductive biases for modeling tabular data. By comparing their behavior under identical training conditions, this research provides a deeper understanding of how architectural choices influence the effectiveness of knowledge transfer. This cross-architecture analysis contributes to identifying which types of models benefit most from distillation in healthcare tabular data settings.

- **Comprehensive Evaluation Using Multiple Performance Metrics**
  Finally, this research provides an extensive evaluation of model performance using a diverse set of metrics that capture different aspects of predictive reliability. In addition to discrimination metrics such as ROC-AUC and PR-AUC, the study also evaluates calibration using the Brier score and examines threshold-based performance using accuracy, precision, recall, and F1-score. This multi-metric evaluation ensures that the proposed framework is assessed not only for its ability to distinguish between classes but also for the reliability of its predicted probabilities. Such comprehensive evaluation is particularly important in healthcare applications where model predictions may directly influence clinical decision-making.

## 1.5 Structure of the Thesis

This thesis is structured to progressively introduce the research problem, review existing literature, present the proposed methodology, and evaluate the effectiveness of the proposed frameworks. Each chapter builds upon the previous one to provide a comprehensive understanding of the study.

Chapter 2 provides an overview of existing research related to deep learning for tabular healthcare data, distributed learning in healthcare systems, knowledge distillation techniques, and few-shot transfer learning methods. It also reviews prior work on privacy-preserving machine learning and highlights the limitations of current approaches when applied to heterogeneous clinical tabular data.

Chapter 3 describes the proposed unified frameworks. It first introduces the datasets used in this study and the preprocessing pipeline applied to harmonize heterogeneous clinical data. The chapter then presents the system architecture and details the two proposed approaches: logit-level knowledge distillation and encoder-level knowledge distillation within a split learning environment.

Chapter 4 outlines the experimental design and evaluation protocol used to assess the proposed frameworks. This includes dataset configuration, few-shot supervision settings, model architectures, training procedures, and evaluation metrics. The chapter then presents and analyzes the results obtained for both frameworks across different supervision levels and model types.

Chapter 5 interprets the experimental findings and examines the mechanisms underlying the observed results. It analyzes how different levels of knowledge transfer

affect discrimination, calibration, and model robustness, and discusses architectural sensitivity and the practical implications of privacy-preserving machine learning in healthcare applications.

Chapter 6 summarizes the main contributions of the research and highlights the key findings of the study. It also discusses the broader implications of privacy-preserving knowledge transfer for clinical prediction and outlines potential directions for future research.

# Chapter 2

# Literature Review

## 2.1 Background Study

This section explains the main theories behind the proposed privacy-preserving learning frameworks. It covers recent progress in deep learning for tabular healthcare data, distributed learning, knowledge distillation, and few-shot transfer learning. Together, these topics provide the methods needed for collaborative model training when data is limited and privacy is important. Understanding how structured data modeling, distributed optimization, and representation transfer work together is key to judging how well split learning and distillation strategies perform in different clinical settings.

### 2.1.1 Deep Learning for Tabular Healthcare Data

Tabular data are still the main type used in clinical risk prediction. They include structured variables like demographic details, lab results, comorbidities, and treatment histories. Unlike images or signals, tabular healthcare data have mixed feature types, missing values, and different distributions between institutions. These factors make it difficult for traditional deep learning models, which were built for spatial or sequential data, to perform well. Recent studies have re-examined deep neural networks for structured data, demonstrating that transformer-based architectures can effectively model feature interactions in tabular contexts. SAINT (Self-Attention and Intersample Attention Transformer) incorporates row and column attention mechanisms to capture both intra-feature and inter-sample dependencies, thereby enhancing performance across diverse tabular benchmarks [8]. Likewise, FT-Transformer modifies the transformer architecture for structured data by embedding categorical and continuous features as tokens and employing self-attention to learn contextualized feature representations [9]. These models have demonstrated competitive or superior performance compared to traditional multilayer perceptrons and gradient boosting methods across various tabular tasks. In healthcare applications, deep learning models for structured data must address class imbalance, covariate shift and institution-specific missing data patterns. Recent studies show that model performance on structured datasets depends heavily on the dataset and how well the data is represented [10]. Therefore, the development of robust latent features is essential for transferring models across institutions or when labeled data are limited. These advances lay the groundwork for using transformer-based models

in privacy-preserving distributed learning, especially when working with varied clinical datasets and situations where only a few examples are available in the target domain.

### 2.1.2 Distributed Learning in Healthcare

Distributed learning is now a key way to train machine learning models across different institutions while keeping data in place. In healthcare, rules and policies often prevent combining data in one location, so teams need to use collaborative methods that do not share raw data directly. Federated learning (FL) is a widely studied approach to distributed computing. In FL, each institution trains its own model and regularly sends updates to a central server for aggregation. Comprehensive surveys have analyzed the progress, scalability challenges and security considerations of federated learning in real-world deployments [11]. However, FL can struggle with differences in client data (non-IID settings), which are common in multi-institutional clinical environments [12], [13]. FL also expects every client to keep and train a full model, which is not always possible in healthcare environments with limited computing resources. Split learning (SL) is an alternative distributed framework where a neural network is split between client and server components. Instead of sharing model parameters, the client sends intermediate activations at a cut layer, and the server completes the remaining forward and backward propagation. This setup reduces client-side computational burden and allows flexible model partitioning, which is useful when institutions differ in feature dimensionality or hardware capacity [3]. Recent research shows that distributed training frameworks need to balance communication efficiency, statistical robustness, and privacy guarantees [11], [14]. In diverse clinical settings, keeping representation stable and optimization consistent is important for reliable predictions across different institutions. These distributed learning approaches form the base for building models together in privacy-sensitive healthcare settings. They also make it possible to use advanced techniques like knowledge distillation.

### 2.1.3 Knowledge Distillation

Knowledge distillation (KD) has progressed from a model compression method to a comprehensive framework for transferring knowledge across architectures, domains, and training regimes. Early formulations of knowledge distillation focused primarily on response-based supervision using softened output logits. Recent studies have re-examined how distillation enhances generalization and optimization stability [15], [16]. Current analyses indicate that KD serves as a regularizer, guiding the student model toward smoother decision boundaries and more stable optimization, especially when data is limited. Response-based, or logit-level, distillation passes the teacher's output probability distribution to the student using temperature-scaled soft targets. Recent studies show that how well logit distillation works depends on factors like teacher calibration, confidence distribution, and student capacity [15], [17]. In settings with limited data, soft supervision can help reduce overfitting and improve discrimination performance compared to using only hard labels [18]. In addition to output-level supervision, representation-based or feature-level distillation matches intermediate embeddings between teacher and student networks. Recent research

shows that aligning features can make models more robust to distribution shifts and improve transferability when feature interactions are complex [19]. This is especially important for structured and tabular data, where good predictions rely on keeping cross-feature dependencies instead of spatial patterns. In distributed, privacy-aware environments, knowledge distillation enables knowledge transfer without sharing model parameters or raw data. Recent studies have examined knowledge distillation in federated and collaborative settings, demonstrating its ability to address heterogeneity and stabilize learning across clients with non-IID data [20], [21]. However, most research focuses on homogeneous or image-based contexts. The effects of logit-level and representation-level distillation under split learning constraints for heterogeneous tabular healthcare data are not well understood. Given that clinical target institutions often operate in few-shot regimes with limited labeled outcomes, understanding how distillation affects discrimination and calibration in privacy-preserving distributed frameworks is critical. These considerations motivate systematic evaluation of both logit-based and encoder-based distillation strategies within split learning architectures for structured healthcare prediction tasks.

### 2.1.4 Few-Shot and Transfer Learning in Clinical Prediction

Few-shot learning addresses scenarios with only a limited number of labeled samples for model training. In healthcare, many institutions have constrained annotation budgets, resulting in small, imbalanced datasets insufficient for training high-capacity models from scratch. Consequently, transfer learning and representation reuse have become essential strategies to improve generalization in low-resource clinical settings [22], [23]. Transfer learning leverages representations from large source-domain datasets and adapts them to smaller target domains. Recent research shows that pretrained models provide robust initialization and improve optimization stability under data scarcity, especially when domain similarity is moderate [24]. However, clinical data often show demographic shifts, institutional variations in data collection, and heterogeneous feature distributions, which can reduce the effectiveness of naive fine-tuning approaches [12], [13]. For structured healthcare data, few-shot performance relies on how well the model learns useful representations. Transformer-based tabular models like SAINT and FT-Transformer perform well in supervised and semi-supervised tasks by capturing complex relationships between features [8], [9]. However, most studies test these models using centralized training with full access to data. How few-shot transfer works in privacy-preserving distributed learning is still not well understood. Recent studies suggest that under distribution shift and non-IID client settings, transfer learning can suffer from representation drift and calibration degradation [11], [15]. These challenges are pronounced when transferring from large population-level datasets to smaller institutional cohorts. Structured transfer mechanisms, such as knowledge distillation, may offer additional stability by guiding adaptation through logit-level supervision or representation alignment. Understanding how few-shot adaptation behaves within privacy-preserving distributed frameworks is critical for real-world healthcare deployment. Evaluating transfer performance under heterogeneous feature spaces, limited labeled data, and distributed optimization constraints provides insights into discrimination and calibration trade-offs in collaborative clinical machine learning.

## 2.2 Related Work

This section reviews research on privacy-preserving machine learning in healthcare, focusing on federated learning, split learning, and knowledge distillation in distributed settings. It looks at how current methods handle data privacy, differences between clients, communication needs and limited labeled data in multi-institutional clinical environments. While there has been progress in distributed learning and representation transfer, most studies have centered on image-based data or assume data is similar across sites. This analysis shows there is a need to systematically evaluate split learning with knowledge distillation for predicting healthcare outcomes from varied tabular data, especially when only a few labeled examples are available.

### 2.2.1 Learning Under Regulatory Constraints in Healthcare

The increasing implementation of machine learning in healthcare has heightened concerns about patient privacy, regulatory compliance and institutional data governance. Previous research indicates that centralized learning pipelines, which aggregate sensitive patient data at a single site, present substantial privacy and security risks, especially within the context of regulatory frameworks such as HIPAA and GDPR [25], [26]. These concerns have motivated the development of privacy-preserving machine learning (PPML) approaches that enable collaborative model training without direct data sharing. Federated learning (FL) has emerged as one of the most widely explored PPML paradigms in healthcare, allowing institutions to train local models and share only parameter updates [27]. However, several studies report limitations of FL, including high communication overhead and vulnerability to gradient leakage [28]. Furthermore, FL suffers from reduced performance under data heterogeneity and class imbalance, conditions common in real-world clinical datasets [29]. FL also assumes all participants can host and train the complete model architecture locally. This may not be feasible in resource-constrained healthcare environments. Split learning (SL) serves as an alternative privacy-preserving paradigm that mitigates certain limitations by partitioning the model into client-side and server-side components [3]. In SL, raw patient data remain local to the institution and only intermediate activations are transmitted to a central server, reducing exposure of identifiable information. Several studies show that SL achieves competitive predictive performance and reduces client-side computational and communication costs compared to FL [3]. However, concerns about information leakage through intermediate activations have led to research on privacy-preserving protocols that combine distributed learning with cryptographic protections such as homomorphic encryption [30].

### 2.2.2 Split Learning for Heterogeneous and Tabular Data

Recent studies have investigated split learning in heterogeneous client environments, where participating institutions may vary in feature dimensionality, data distribution or model capacity. Methods in this area address these challenges by permitting clients to retain unique local encoders while utilizing a shared server-side representation, thereby facilitating collaborative learning across disparate tabular feature spaces. While split learning has been successfully applied in medical imaging and time-series analysis, much of the existing literature focuses on homogeneous data

distributions and image-based modalities [3]. Clinical tabular datasets present additional challenges, including heterogeneous feature spaces, institution-specific missingness patterns and highly imbalanced outcome distributions. These characteristics complicate the direct application of standard split learning architectures.

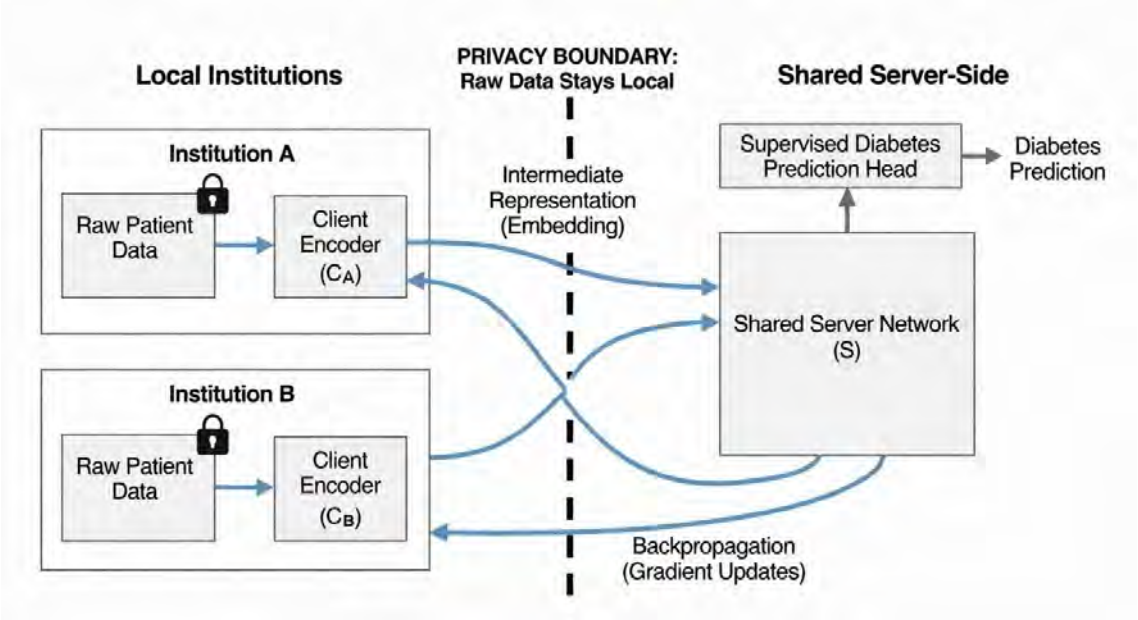



Figure 2.1: Split Learning Architecture

Recent large-scale benchmarking studies indicate that no single deep learning paradigm consistently surpasses traditional methods across diverse tabular datasets. Attention-based and contrastive learning models, in particular, display pronounced dataset-dependent performance. These results underscore the necessity for robust representation learning in structured data contexts. Nevertheless, a significant limitation of these benchmarks is their predominant focus on centralized settings, which overlooks the privacy-preserving and collaborative learning requirements characteristic of multi-institutional healthcare [31]. To address these privacy constraints, recent studies have extensively investigated split and split-fed learning architectures, focusing on communication efficiency, client heterogeneity and security [32]. Despite these advances, evaluations involving tabular data are still limited in scope. Most existing studies utilize simplified experimental designs that often assume sufficient labeled data at each site, which limits applicability to low-resource settings [22],[27].Moreover, the integration of knowledge distillation into split learning, particularly to enable representation transfer in the presence of data heterogeneity, remains insufficiently investigated. In addition, the interaction between split learning and contemporary tabular deep learning architectures, including transformer-based models tailored for structured data, has not been rigorously studied within privacy-preserving frameworks.

### 2.2.3 Knowledge Distillation in Resource-Constrained and Distributed Settings

Knowledge distillation (KD) is a widely recognized method for transferring knowledge from a high-capacity teacher model to a smaller student model. This approach

enables efficient inference while maintaining predictive performance [7]. In healthcare applications, KD has been applied to compress deep models, lower deployment costs and enhance inference latency in environments with limited resources [33]. Several studies have extended KD to distributed and privacy-aware settings, demonstrating that student models can benefit from teacher supervision without direct access to raw training data [7], [34]. In particular, distillation is shown to improve generalization when labeled data are scarce or noisy [35], [36], making it highly attractive for clinical scenarios where annotations are expensive [37]. However, most KD research assumes centralized training of the teacher model or focuses on output-level distillation using soft labels [7]. This limitation is particularly relevant for structured data. Recent studies indicate that although response-based distillation can address class imbalance [38], representation-level or encoder-level distillation is often more effective for tabular data, where maintaining latent feature interactions is essential [8], [39]. However, the application of knowledge distillation within split learning pipelines remains largely uninvestigated. While recent architectures have been proposed for heterogeneous split learning [4], few studies assess whether distilling intermediate representations within these frameworks enhances transferability in privacy-constrained clinical environments. This research gap is especially pronounced in few-shot target domains, where inadequate distillation may exacerbate distribution mismatches rather than resolve them.

### 2.2.4 Few-Shot and Transfer Learning for Clinical Prediction

Few-shot and transfer learning are gaining attention in clinical prediction because many healthcare institutions have limited labeled data. Large population-level datasets help develop generalizable representations that can be adapted to smaller, institution-specific cohorts [27], [8]. However, differences in patient demographics, data collection protocols, and feature distributions often reduce the effectiveness of simple transfer methods. Transformer-based models for tabular data, like SAINT and similar architectures, have shown strong results in supervised and semi-supervised tasks by capturing complex feature interactions and using self-supervised pretraining [8]. These models are particularly well-suited for few-shot scenarios, as they can exploit unlabeled data to learn robust representations. Nevertheless, most evaluations of tabular transformers assume centralized training and unrestricted data access. Recent studies have demonstrated that few-shot learning achieves competitive performance in medical image analysis, even with extremely limited labeled data. However, these evaluations are typically conducted in centralized settings and do not address privacy-preserving collaborative learning scenarios [40]. Recent studies suggest that combining transfer learning with privacy-preserving training paradigms introduces challenges such as representation drift, calibration degradation and sensitivity to class imbalance [28]. While some work explores domain adaptation and semi-supervised learning under privacy constraints, systematic evaluations of few-shot transfer within split learning frameworks are limited. This motivates further investigation into whether privacy-preserving split learning with encoder-level distillation can support reliable transfer to low-resource clinical domains without sacrificing calibration or discrimination performance.

# Chapter 3

# Methodology

This study explores privacy-preserving diabetes prediction in heterogeneous, resource-constrained clinical environments using split learning and knowledge distillation. Two complementary frameworks are proposed and systematically evaluated. Framework 1 introduces a split learning baseline that uses logit-level knowledge distillation to transfer predictive behavior from high-resource teacher models to low-resource student models in few-shot settings. It transfers softened output distributions and assesses how different labeled sample sizes $K \in \{5, 10, 20\}$ affect student performance. Framework 2 builds on this approach by integrating encoder-level knowledge distillation into the split learning architecture. Rather than transferring only output logits, it aligns intermediate feature representations between teacher and student encoders, supporting deeper representation transfer across heterogeneous datasets.

In accordance with standard split learning protocols, both frameworks partition the neural network into two components:

1. A **client-side student model**, which retains raw patient data locally and performs initial feature extraction.

2. **A server-side model**, which receives intermediate activations, often called smash data and completes forward propagation.

Both frameworks are evaluated under identical experimental conditions for direct comparison. Performance is measured using five independent random seeds, with results reported as mean ± standard deviation. Metrics include ROC-AUC, PR-AUC, Brier score (including temperature-calibrated Brier), Accuracy, Precision, Recall, F1-score and confusion matrices, where temperature scaling was fitted on the validation set only and applied to test probabilities for calibration-sensitive metrics (e.g., Brier score). and the classification metrics were computed at a fixed threshold of 0.5. This unified protocol enables systematic analysis of representation transfer, calibration and robustness under data scarcity in realistic healthcare settings. This architecture ensures that sensitive patient data never leaves the client environment, while still enabling collaborative model training. Framework 1 transfers predictive behavior at the logit level, whereas Framework 2 aligns intermediate encoder representations to facilitate deeper structural transfer.

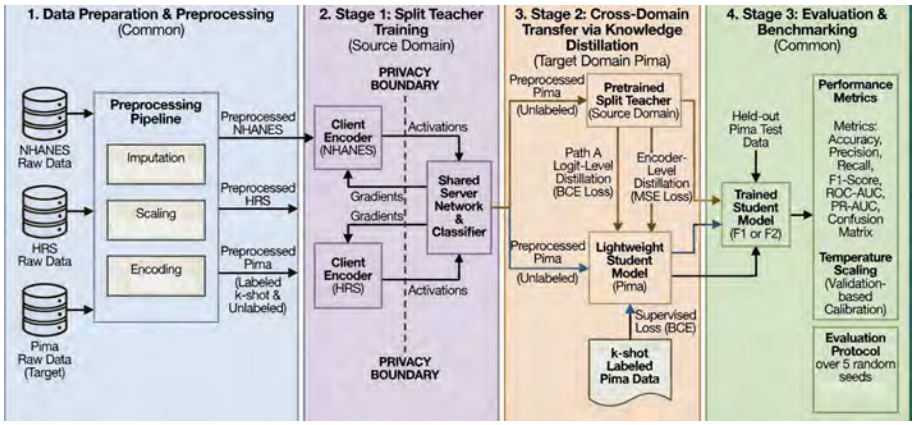



Figure 3.1: Top-Level Overview of The System

Figure 3.1 gives a high-level overview of the proposed privacy-preserving transfer learning framework. The workflow has four stages: (i) unified data preprocessing, (ii) split teacher training on source domains, (iii) cross-domain transfer to the target domain using knowledge distillation and (iv) multi-seed evaluation and calibration. The following subsections detail each stage, including architectural design choices, training objectives and evaluation protocol.

## 3.1 Data Collection and Preprocessing

To simulate heterogeneous clinical institutions while maintaining privacy constraints, three publicly available tabular healthcare datasets serve as distinct client sites. Given the differences in dataset size, feature dimensionality and class imbalance, a standardized preprocessing pipeline is implemented at each site to enable fair comparisons across models and between the two split-learning frameworks. The processed client-local inputs, comprising categorical and continuous features, are subsequently utilized for split learning and knowledge distillation experiments using consistent train, validation and test splits.

### 3.1.1 Dataset Description and Feature Analysis

To understand domain heterogeneity and feature distribution differences across institutions, exploratory data analysis (EDA) was performed on each dataset. Feature-wise histograms visualize distributional characteristics, skewness, sparsity and class imbalance. These visualizations highlight structural differences between source and target domains, directly motivating the need for cross-domain transfer learning. Each dataset was selected to reflect a different clinical setting and to thoroughly test the split learning and knowledge distillation frameworks. To enhance visualization clarity, a representative subset of clinically meaningful features is presented for each dataset. These figures demonstrate the characteristics of data distribution, whereas the complete processed feature sets were utilized during model training.

#### NHANES Dataset

The National Health and Nutrition Examination Survey (NHANES) provides an extensive range of demographic, laboratory and physiological measurements.

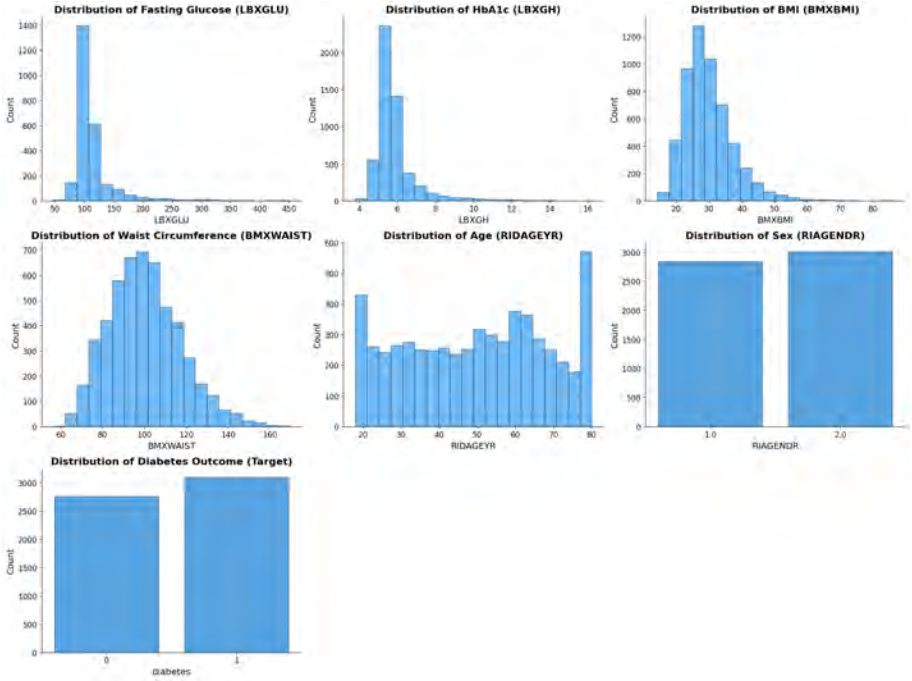



Figure 3.2: Exploratory Feature Distributions in NHANES

Figure 3.2 shows the distributions of six key features: fasting glucose (LBXGLU), glycated hemoglobin (LBXGH), body mass index (BMXBMI), waist circumference (BMXWAIST), age (RIDAGEYR), and sex (RIAGENDR), as well as the binary diabetes outcome. The visualizations highlight:

- Right-skewed distributions for fasting glucose and HbA1c, indicating the presence of elevated biomarker values in a subset of the population..
- Approximately normal distributions for BMI and waist circumference.
- Broad age distribution across adult cohorts.
- Balanced gender representation.

This dataset represents a high-resource institution and includes enough labeled data to support effective split teacher training. Its size and feature richness allow the teacher model to learn generalizable representations, which are then transferred to the target domain through knowledge distillation. NHANES therefore serves as a strong source domain for representation learning in the split architecture.

### HRS Dataset

The Health and Retirement Study (HRS) is a long-term survey that collects data from a diverse group of people, capturing a wide range of social and demographic backgrounds.

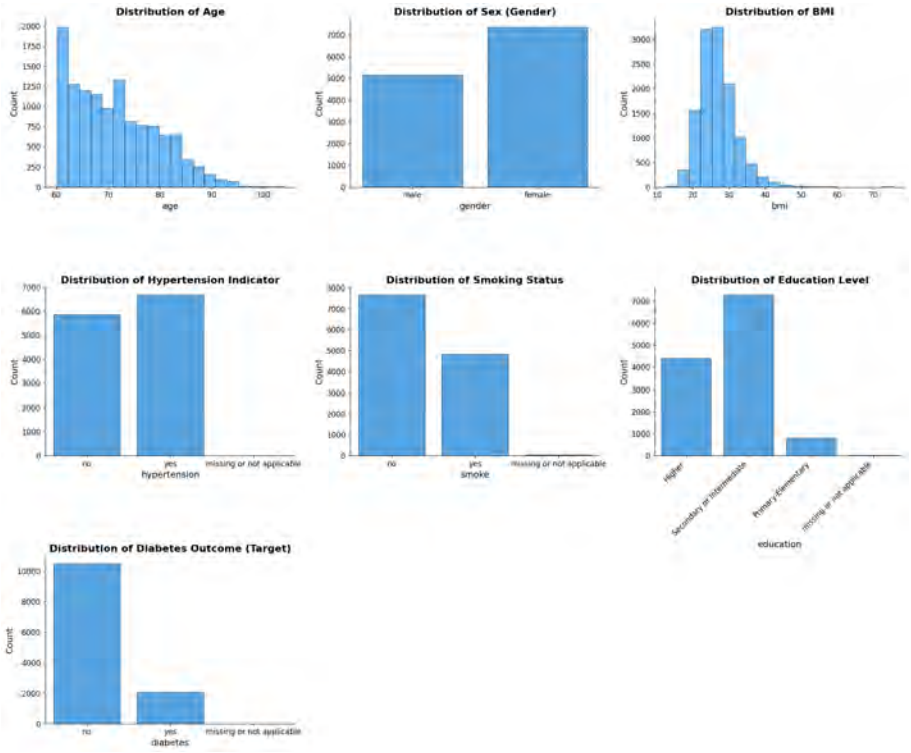



Figure 3.3: Exploratory Feature Distributions in HRS

Figure 3.3 illustrates distributions of age, sex, BMI, hypertension status, smoking status, education level, and the diabetes outcome. The visualizations highlight:

- Older age concentration aligns with the study population.
- Noticeable class imbalance in diabetes prevalence.
- Mixed categorical and continuous variable distributions.
- Behavioral and socio-economic differences among participants.

HRS simulates a second independent clinical site with different data structure and distributional properties. This makes the split learning scenario more realistic by preventing the teacher from being trained on a single homogeneous dataset. The inclusion of HRS allows evaluation of how well the proposed frameworks handle cross-institutional variability during transfer.

### PIMA Dataset

The Pima Indians Diabetes Dataset is used as the target domain to model a clinical setting where there is not much labeled data available.

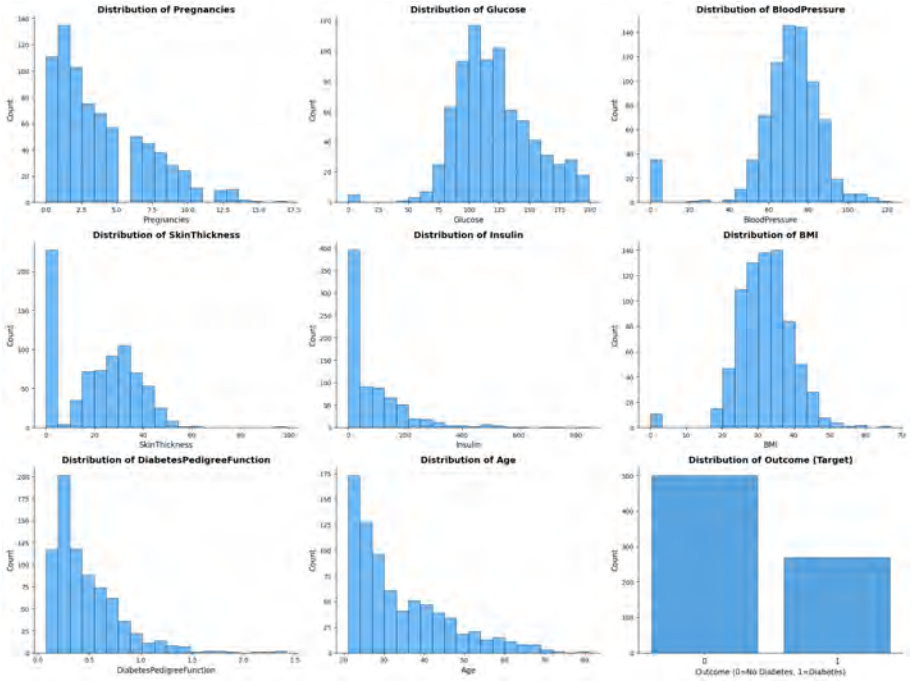



Figure 3.4: Exploratory Feature Distributions in the Pima Dataset

Figure 3.4 shows distributions of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, age, diabetes pedigree function and the binary outcome. The visualizations highlight:

- A smaller sample size relative to the NHANES and HRS datasets.
- Significant skewness in the distributions of insulin and pedigree function variables.
- A moderate imbalance between classes.
- Reduced feature dimensionality compared to similar datasets.

These features suggest the dataset is small and noisy, with few labeled samples. The compact size of the dataset makes it well-suited for evaluating the effectiveness of few-shot learning and knowledge distillation. Since training a high-capacity model from scratch on limited data is challenging, this dataset serves as a rigorous benchmark for assessing whether distilled knowledge from larger source institutions enhances generalization. Consequently, the Pima dataset serves as the target domain for evaluating cross-domain transfer performance.

### 3.1.2 Data Cleaning and Preprocessing Pipeline

A unified preprocessing pipeline was implemented for all three cohorts (NHANES, HRS, and Pima) to maintain methodological consistency and prevent information leakage across heterogeneous datasets. This pipeline standardizes label definitions, addresses missing data, encodes categorical variables, normalizes continuous features, and generates model-ready tensors suitable for split learning architectures.

### Label Standardization and Cleaning

Each dataset used its own way to represent diabetes status. To keep things consistent across experiments, the outcome variables was standardized into a single binary target variable $y \in \{0, 1\}$, where 1 means diabetes and 0 means non-diabetes.

Dataset-specific cleaning procedures were applied to harmonize label representations:

- Numeric labels were converted to a binary format.
- Text labels such as "yes" and "no" were changed to 1 and 0.
- Removed any labels that were unclear, missing or invalid before training.

The samples with only valid binary labels were kept that made sure that the data worked with binary cross-entropy loss and evaluation metrics like ROC-AUC and PR-AUC.

### Stratified Data Splitting

Datasets were partitioned into training (70%), validation (15%) and test (15%) subsets using stratified sampling according to the binary outcome variable.

Stratification maintains class proportions across subsets and mitigates bias resulting from class imbalance, which is especially important in medical datasets where positive cases are frequently underrepresented. This approach enables fair and consistent evaluation of generalization performance across experimental runs.

### Automatic Feature Type Inference

Due to the structural heterogeneity of the three datasets, feature types were automatically inferred to differentiate categorical from continuous variables.
Features were classified as categorical if:

- They were stored as object/string types, or
- They were integer-valued with a limited number of unique values ($\leq$ 20), indicating coded survey or demographic categories.

All remaining numeric features were treated as continuous variables.

This automatic inference helps ensure that preprocessing stays consistent, even when datasets have different schemas. It also avoids mistakes like treating coded categorical variables as if they were continuous.

### Categorical Feature Encoding

Categorical variables were transformed into integer indices using vocabularies specific to each dataset, constructed from the training split. For each categorical feature:

- A mapping from categorical values to integer IDs was created.

- Two special tokens were introduced:
  - __MISSING__ to represent missing values.
  - __UNK__ to represent unseen categories during validation or testing.

These integer indices are passed to learnable embedding layers within the tabular transformer models. Embedding representations allow the model to learn semantic relationships between categories and are more expressive than fixed one-hot encodings, particularly in high-cardinality or heterogeneous datasets.

This approach helps the model learn well across different types of data and prevents errors when it encounters new categories during prediction.

### Continuous Feature Processing

Continuous features were processed using a two-step procedure:

1. **Median Imputation:** Missing continuous values were imputed using the median computed from the training split. Median imputation was selected due to its robustness to outliers, which are common in biomedical measurements (e.g., glucose, insulin).

2. **Standardization:** After imputation, continuous features were normalized using a *StandardScaler* fitted exclusively on the training split. The learned mean and standard deviation were subsequently applied to validation and test splits. Fitting normalization statistics on the training data only prevents information leakage from validation or test distributions and ensures fair evaluation. Standardization improves numerical stability during optimization and enables consistent feature scaling across datasets with varying measurement ranges.

### Missingness Mask Construction

A binary mask was generated for each continuous feature to explicitly model informative missingness patterns:

- A value of 1 indicates that the original feature value was missing.
- A value of 0 indicates that the feature was observed.

The model receives this mask together with the continuous values. In transformer-based systems, missing entries are given a learned missing-value embedding that helps the model see missing data as useful information instead of just noise. This approach matters in healthcare, since missing measurements can be linked to a patient's clinical status.

### Prevention of Label Leakage

To maintain a clear distinction between features and targets, dataset-specific label columns such as *Outcome* and *diabetes* were excluded from the feature set following the construction of the standardized target variable $y$.

Subsequently, preprocessing objects were re-fitted on the cleaned feature space to ensure that no target-derived information persisted in the inputs. This procedure mitigates the risk of label leakage and upholds the validity of the reported evaluation metrics.

#### Model-Ready Tensor Construction

After preprocessing, each sample was represented as:

- $x_{\text{cat}}$: Integer-encoded categorical features
- $x_{\text{cont}}$: Scaled continuous features
- $m_{\text{cont}}$: Continuous-feature missingness mask
- $y$: Binary outcome

These components were wrapped into custom PyTorch *Dataset* objects and loaded using batched *DataLoader*s, ensuring compatibility with split learning and knowledge distillation training procedures.

#### Few-Shot Target Domain Protocol

To create a realistic low-resource clinical scenario, the Pima dataset was used as the target domain and limited the amount of labeled data available. From the training split, a randomly sampled balanced k-shot subset was chosen for $K \in \{5, 10, 20\}$ samples per class. This subset was the only labeled supervision for student training. All other training samples were treated as unlabeled and used only through knowledge distillation. For each random seed, the k-shot subset was re-sampled to ensure robustness and avoid bias from a fixed labeled subset. This setup reflects real-world clinical environments where labeled outcomes are scarce and highlights the importance of representation transfer under limited supervision.

## 3.2 Proposed Framework 1: Split Learning with Logit-Level Knowledge Distillation

Framework 1 provides a privacy-preserving baseline by combining split learning with logit-level knowledge distillation. This approach transfers predictive knowledge from high-resource population datasets to low-resource clinical settings. The goal is to determine whether a collaboratively learned decision function can be distilled into a compact student model with limited labeled supervision, while maintaining strict data privacy.

### 3.2.1 Teacher Model Training Under Split Learning

In the first stage, a high-capacity split learning teacher model is trained collaboratively using the NHANES and HRS datasets, which are treated as independent client institutions.

Each institution maintains a private client-side encoder $f_c^{(i)}(\cdot)$, where $i \in \{\text{NHANES}, \text{HRS}\}$. These encoders process raw tabular inputs $x^{(i)}$ and produce intermediate latent representations:

$$h^{(i)} = f_c^{(i)}\big(x^{(i)}\big)$$

Only the intermediate activations $h^{(i)}$ (commonly referred to as *smash data*) are transmitted to a shared server-side network $g_s(\cdot)$, which learns a common representation space and produces final logits:

$$z^{(i)} = g_s\big(h^{(i)}\big)$$

The server outputs predicted probabilities via:

$$\hat{y}^{(i)} = \sigma\big(z^{(i)}\big)$$

where $\sigma(\cdot)$ denotes the sigmoid function.
Training is performed using supervised binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right]$$

Gradients are backpropagated from the server to the respective client encoders, but raw patient data remain local at each institution throughout training. Only activations and gradients are exchanged across the privacy boundary.

This collaborative training procedure allows the teacher model to learn a generalized population-level decision function across heterogeneous domains without the need to centralize raw data.

### 3.2.2 Low-Resource Target Setting and Student Initialization

To simulate a realistic clinical deployment scenario, the Pima dataset is designated as the target domain with limited labeled data. From the Pima training split, a balanced $k$-shot subset is sampled such that:

$$k \in \{5, 10, 20\} \text{ samples per class.}$$

This subset is the only labeled supervision available to the student. All other training samples are treated as unlabeled and used exclusively through knowledge distillation.

A lightweight student model $f_s(\cdot)$, instantiated as SAINT, MLP, or TabM, is initialized for the Pima domain. Although architecturally related to the teacher's encoder structure, the student is significantly smaller in parameter count to reflect computational constraints in low-resource clinical environments.

Unlike the teacher, which is trained collaboratively across institutions, the student is trained only within the Pima domain. This approach reflects a scenario where a small clinical site aims to benefit from the expertise of larger institutions without joining multi-institutional split training.

### 3.2.3 Logit-Level Knowledge Distillation

Knowledge transfer occurs at the logit level. Instead of aligning intermediate feature representations, the student model is trained to replicate the predictive outputs of the teacher model. The trained split teacher is used to generate soft logits $z_T$ for Pima samples (including unlabeled data). The student produces its own logits $z_S$. Two complementary learning signals are combined during student training:

**1. Supervised Loss**

Computed on the labeled $k$-shot subset using ground-truth labels:

$$\mathcal{L}_{\text{sup}} = \text{BCE}(z_S, y).$$

**2. Distillation Loss**

Computed between student and teacher logits on unlabeled samples:

$$\mathcal{L}_{\text{KD}} = \text{BCE}\big(z_S, \sigma(z_T)\big).$$

The total training objective is a weighted combination:

$$\mathcal{L}_{\text{total}} = \alpha\, \mathcal{L}_{\text{sup}} + (1 - \alpha)\, \mathcal{L}_{\text{KD}},$$

where $\alpha \in [0, 1]$ controls the balance between direct supervision and teacher guidance.

By learning from the teacher's soft predictions, the student captures uncertainty-aware decision boundaries and population-level class structure that cannot be inferred from limited labeled samples. Importantly, this transfer occurs without direct exposure to NHANES or HRS data, preserving institutional privacy.

### 3.2.4 Training Procedure and Evaluation

The student model is trained for a set number of epochs, using early stopping based on validation performance to reduce overfitting. This approach is especially important in few-shot settings.

Evaluation is conducted on a held-out Pima test set using:

- ROC-AUC
- PR-AUC
- Brier Score (including temperature-calibrated Brier)
- Accuracy, Precision, Recall, F1-score
- Confusion matrix at threshold 0.5

Each experiment is repeated with five independent random seeds. Results are reported as mean ± standard deviation.

To isolate the effect of logit-level knowledge distillation, student models trained from scratch without distillation are evaluated under the same $k$-shot configurations. This controlled comparison quantifies the performance gain from distilled population-level knowledge.

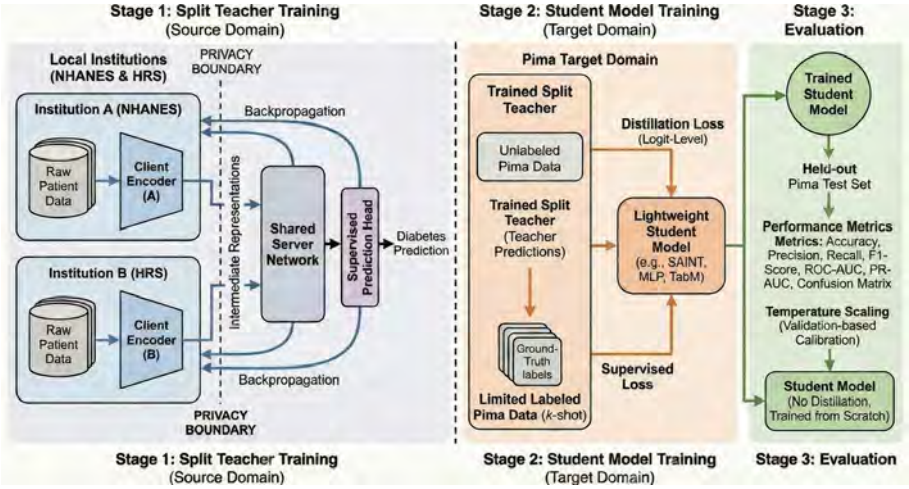



Figure 3.5: Proposed Framework 1: Logit-Level Knowledge Distillation

## 3.3 Proposed Framework 2: Split Learning with Encoder-Level Knowledge Distillation

While Framework 1 transfers predictive behavior through output-level supervision, Framework 2 explores whether deeper structural knowledge can be transferred by aligning internal feature representations in a split learning architecture. Instead of matching final logits, this framework distills knowledge at the encoder level by aligning the student's latent activations with those of a pretrained split-learning teacher.

The central hypothesis is that transferring internal representations, rather than only output decisions, may improve robustness, calibration stability and generalization under extreme data scarcity in heterogeneous clinical environments.

### 3.3.1 Teacher Model Training Under Split Learning

A high-capacity split-learning teacher is trained collaboratively using the NHANES and HRS datasets as separate institutions.

For each institution $i \in \{\text{NHANES}, \text{HRS}\}$:

- A client-side encoder $f_c^{(i)}(\cdot)$ processes raw tabular inputs:

$$h^{(i)} = f_c^{(i)}\left(x^{(i)}\right)$$

- Intermediate activations $h^{(i)}$ (*smash data*) are transmitted to a shared server-side network $g_s(\cdot)$, producing logits:

$$z^{(i)} = g_s\big(h^{(i)}\big)$$

- Final probabilities are obtained via:

$$\hat{y}^{(i)} = \sigma\big(z^{(i)}\big)$$

The teacher is trained using supervised binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right].$$

Gradients propagate across the cut layer while raw patient data remain local, preserving institutional privacy.

Consequently, a pretrained split-learning teacher encodes comprehensive population-level feature representations within its client-side encoders.

### 3.3.2 Low-Resource Target Setting and Student Initialization

The Pima dataset is designated as the low-resource target domain. From the Pima training split, a balanced $K$-shot subset is sampled:

$$K \in \{5, 10, 20\} \text{ samples per class.}$$

This labeled subset provides supervised training signals. The remaining samples are treated as unlabeled.

A lightweight student model $f_s(\cdot)$, instantiated as SAINT, MLP, or TabM, is initialized for the Pima site. To make sure the representations match, the student encoder is designed to create latent representations of the same size as those from the teacher's client-side encoder. Unlike the teacher, which is trained collaboratively across institutions, the student operates solely within the Pima domain.

### 3.3.3 Encoder-Level Knowledge Distillation (EKD)

In contrast to Framework 1, knowledge transfer occurs at the level of intermediate activations rather than final logits.
Let:

- $h_T(x)$ denote the teacher's client-side latent representation,
- $h_S(x)$ denote the student's encoder representation.

The teacher is frozen during student training.
For unlabeled samples, the student is trained to align its representation with the teacher's:

$$\mathcal{L}_{\text{align}} = \|h_S(x) - h_T(x)\|_2^2.$$

This Mean Squared Error (MSE) loss encourages structural similarity between teacher and student feature spaces. Simultaneously, supervised learning is applied to labeled $k$-shot samples:

$$\mathcal{L}_{\text{sup}} = \text{BCE}(z_S, y).$$

The total objective is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{sup}} \mathcal{L}_{\text{sup}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}},$$

where $\lambda_{\text{sup}}$ and $\lambda_{\text{align}}$ control the trade-off between label supervision and representation alignment. Both losses are optimized jointly throughout training. No stage-wise scheduling or curriculum learning is applied.

Encoder-level distillation is applied strictly at the client-side representation and does not modify the server architecture, ensuring direct comparability with Framework 1.

### 3.3.4 Training Procedure and Evaluation

The student model is trained using the composite objective defined in the previous section, jointly optimizing supervised and alignment losses throughout training. Both objectives are applied simultaneously without stage-wise scheduling. The teacher model remains frozen during student optimization to ensure stable representation transfer. Training proceeds for a fixed number of epochs, with early stopping based on validation performance in the target domain to mitigate overfitting. This is especially critical under few-shot supervision. Evaluation uses a held-out Pima test set and a standardized protocol consistent with Framework 1 to ensure direct comparability. Model performance is assessed using the following metrics:

- ROC-AUC
- PR-AUC
- Brier Score (including temperature-calibrated Brier)
- Accuracy, Precision, Recall, F1-score
- Confusion matrix at a fixed threshold of 0.5

Temperature scaling is applied with validation data to assess the reliability of probabilistic calibration.

Each experiment is repeated with five independent random seeds. Results are reported as mean ± standard deviation to ensure robustness and statistical reliability. To isolate the effect of encoder-level knowledge distillation, Framework 2 is evaluated against the following baselines:

- Logit-level distillation (Framework 1)
- Standard split learning without encoder alignment
- Scratch models trained without any pretraining or distillation

This controlled comparison allows precise measurement of whether representation-level alignment improves discrimination, calibration and robustness under strict privacy and low-resource constraints.

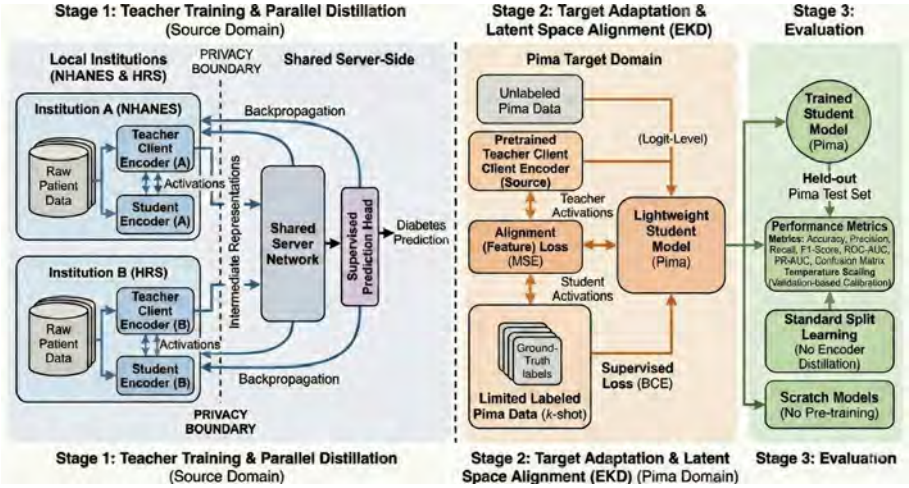



Figure 3.6: Proposed Framework 2: Encoder-Level Knowledge Distillation

## 3.4 Depth of Knowledge Transfer and Design Motivation

Knowledge distillation can be applied at multiple depths within a neural network. Output-level supervision transfers predictive behavior, whereas encoder-level alignment transfers structural feature representations.

Healthcare data from different institutions often varies in feature distributions and meanings, so it is not clear which level of knowledge transfer works best. This study compares Framework 1 (logit-level KD) and Framework 2 (encoder-level KD) to see how the depth of knowledge transfer affects:

- Discriminative performance
- Calibration reliability
- Stability across random seeds
- Robustness under severe data scarcity

Framework 2 therefore explores whether deeper structural alignment improves generalization and clinical reliability, even if representation constraints limit adaptation flexibility.

# Chapter 4

# Experimental Evaluation and Results

## 4.1 Experimental Design

The experimental design and evaluation protocol in this chapter are applied consistently to both frameworks. All models use the same data splits, $k$-shot supervision levels, optimization settings, training budgets, early stopping criteria, and performance metrics. This standardized approach ensures a controlled and reproducible basis for comparison.

- Logit-level Knowledge Distillation (Framework 1)
- Encoder-level Knowledge Distillation (Framework 2)
- Non-distilled baselines

The experimental protocol is identical for both frameworks, except for the level of knowledge distillation. Framework 1 applies logit-level supervision to transfer predictive behavior, while Framework 2 aligns intermediate latent representations at the encoder level.

By keeping all other experimental variables constant, any performance differences can be attributed to the depth and mechanism of knowledge transfer. This controlled setup allows for a fair and systematic assessment of how predictive-level and representation-level distillation affect discrimination, calibration, and robustness in low-resource, privacy-preserving clinical settings.

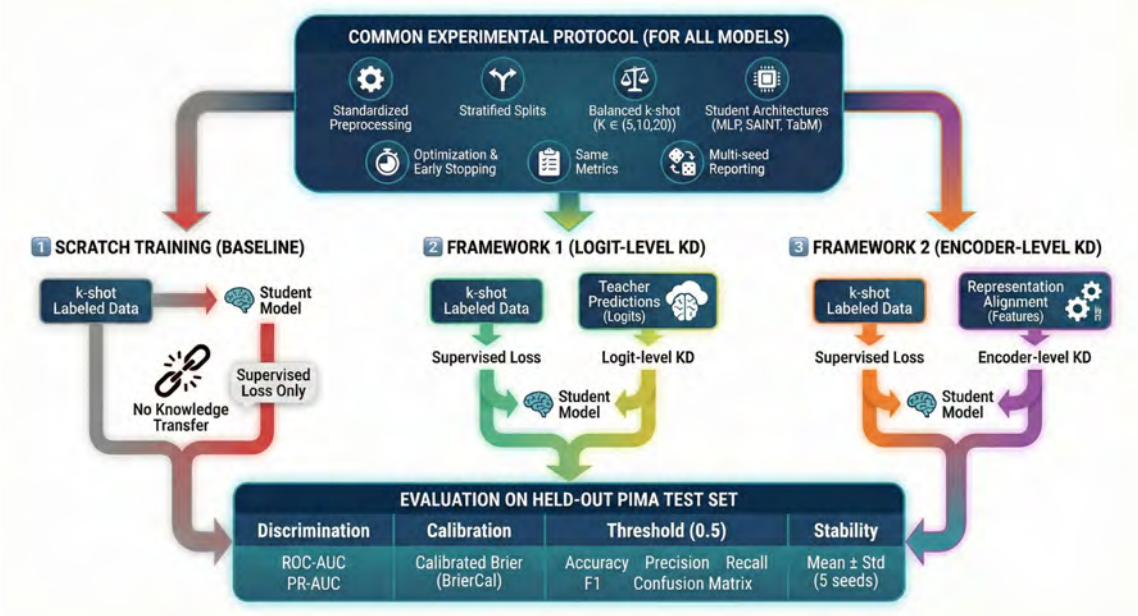



Figure 4.1: Controlled Experimental Setup Across Distillation Frameworks

### 4.1.1 Dataset Configuration and Splits

All experiments use the preprocessing steps explained in Chapter 3. After cleaning the labels and processing the features, each dataset is split with stratified sampling into the following groups:

- 70% Training
- 15% Validation
- 15% Test

Stratified sampling keeps the original class distribution in each split. This helps avoid bias in evaluation that could happen if the classes were imbalanced. The NHANES and HRS datasets are only used to train the teacher model in split learning. The Pima dataset serves as the target domain for adapting the student model.There is no overlap between the data used for the source domain and the data used for the target domain.

### 4.1.2 Few-Shot Supervision Protocol

The Pima training set is subjected to a balanced $K$-shot sampling protocol to simulate realistic low-resource clinical conditions.

For each experiment:

$$K \in \{5, 10, 20\}$$

samples per class are randomly selected from the Pima training split to form the labeled subset.

The remaining training samples are treated as unlabeled and are used exclusively for distillation, where applicable.

This protocol ensures the following:

- Strict control over supervision levels
- Realistic simulation of low-data deployment scenarios
- Fair comparison across different distillation regimes

For each random seed, the $K$-shot subset is re-sampled to prevent dependence on any specific patient subset.

### 4.1.3 Models Considered

This study looks at different types of models to see how well knowledge distillation works when using privacy-preserving split learning and limited supervision. All the models use only structured tabular clinical data.

#### Teacher Model

The high-capacity teacher model utilizes the FT-Transformer, a transformer architecture specifically developed for tabular data. The FT-Transformer incorporates the following components:

- Feature tokenization for both categorical and continuous variables
- Multi-head self-attention mechanisms to capture cross-feature interactions
- A global representation token to facilitate aggregation

The teacher model is collaboratively trained on the NHANES and HRS datasets using split learning. Each dataset retains a client-side encoder to preserve institutional data heterogeneity. Intermediate representations are transmitted to a shared server-side transformer backbone, which generates the final diabetes risk prediction. Dataset-specific prediction heads address cohort-level distribution differences. This architecture enables the teacher model to learn population-level diabetes risk patterns while ensuring that raw patient data remains within local institutions, thereby maintaining privacy and regulatory compliance.

#### Student Models

For downstream low-resource adaptation on the Pima dataset, lightweight student models are initialized and trained to acquire knowledge from the pretrained split teacher model. To ensure that performance improvements result from the distillation frameworks rather than architectural bias, three distinct student model families are evaluated:

- **MLP (Multi-Layer Perceptron)**
  This fully connected feed-forward baseline quantifies the benefit of distillation on architectures that lack specialized tabular inductive biases.

- **SAINT (Self-Attention and Intersample Attention Transformer)**
  This attention-based tabular model captures both intra-feature and inter-sample relationships, which makes it well-suited for scenarios with limited supervision.

- **TabM**
  This streamlined BatchEnsemble-based tabular architecture provides computational efficiency and robustness.

Evaluation of multiple student architectures confirms that observed improvements are attributable to the proposed knowledge transfer mechanisms rather than to model-specific advantages.

### 4.1.4 Baselines and Training Regimes

The specific impact of knowledge distillation is quantified by evaluating three training regimes under identical experimental conditions:

- **Scratch Training (k-shot Only)**
  Student models are trained solely on the labeled Pima k-shot subset using Binary Cross-Entropy (BCE) loss. No external knowledge transfer is applied. This baseline represents the lower bound of performance under strict low-resource supervision.

- **Knowledge Distillation (Labeled + Unlabeled)**
  Student models are trained using a hybrid objective combining:

  - Supervised Loss: Computed on the labeled $K$-shot subset ($K \in \{5, 10, 20\}$).
  - Distillation Loss: Obtained from teacher-generated signals applied to the unlabeled training samples:
    - Soft logits (Framework 1)
    - Latent representations (Framework 2)

  This regime simulates a clinical deployment scenario, where limited local labels are supplemented by population-level knowledge from a pretrained teacher.

- **Knowledge Distillation (Unlabeled-Only)**
  In this setting, distillation uses only unlabeled samples, eliminating direct supervision from the student objective. This configuration assesses how much predictive structure can be transferred from the teacher's decision boundaries without using target-domain labels.

## 4.2 Performance Evaluation Metrics

Model performance is evaluated using complementary metrics for discrimination, probability calibration and threshold-based classification. All evaluations use the held-out Pima test set. To ensure a fair comparison, identical evaluation protocols are applied to Framework 1 (logit-level distillation), Framework 2 (encoder-level distillation) and the non-distilled baselines.

### 4.2.1 Discrimination Metrics

Threshold-independent ranking metrics are used to assess the model's ability to distinguish diabetic from non-diabetic patients:

- **ROC-AUC**
  The Area Under the Receiver Operating Characteristic Curve measures how effectively the model ranks positive cases above negative ones across all decision thresholds. It indicates overall discriminative performance and does not depend on a specific probability threshold.

- **PR-AUC**
  The Area Under the Precision–Recall Curve is especially useful in imbalanced clinical settings. It assesses the balance between precision and recall at different thresholds and offers further insight into positive-class detection. Since these metrics rely on ranking instead of absolute probability values, post-hoc probability calibration does not affect them.

### 4.2.2 Probability Calibration Metrics

Since clinical decision-making relies on trustworthy risk estimates rather than only ranking performance, probability calibration is explicitly evaluated.

#### Brier Score (Uncalibrated)

The Brier score measures the mean squared difference between predicted probabilities and true binary outcomes. It reflects overall probabilistic accuracy and captures both discrimination and calibration effects. Lower values indicate better alignment between predicted risk and observed outcomes.

#### Temperature-Calibrated Brier Score (BrierCal)

Temperature scaling is used as a post-hoc calibration method to assess probability reliability without bias from overconfidence or underconfidence. A single temperature parameter is learned on the validation set and applied to test predictions.

The calibrated Brier score (BrierCal) is calculated using the temperature-adjusted probabilities. Since temperature scaling preserves ranking, improvements in BrierCal indicate better probability calibration, not changes in discrimination performance.

### 4.2.3 Threshold-Based Classification Metrics

For interpretability in practical deployment scenarios, classification performance is also evaluated at a fixed decision threshold of 0.5. The following metrics are reported:

- Accuracy
- Precision
- Recall

- F1-score
- Confusion Matrix

These metrics clarify the trade-offs between false positives and false negatives in clinical screening.

### 4.2.4 Multi-Seed Stability Analysis

Each experiment is repeated with multiple independent random seeds to ensure robustness and reproducibility. Variability from k-shot sampling, data partitioning and model initialization is addressed. Results are presented as the mean and standard deviation across runs. This protocol minimizes the impact of stochastic variation and improves the reliability of comparisons between distillation strategies.

## 4.3 Result Analysis

The following section presents a systematic analysis of experimental results under the controlled protocol described earlier. Performance is evaluated across discrimination, calibration, and threshold-based metrics, with results reported as mean $\pm$ standard deviation over five random seeds. Each framework is examined independently and compared against scratch baselines to isolate the impact of knowledge distillation. This structured evaluation enables a clear understanding of how predictive-level and representation-level transfer influence model behavior under few-shot supervision.

### 4.3.1 Framework 1: Logit-Level Knowledge Distillation

This section analyzes the experimental behavior of Framework 1 under varying few-shot supervision levels. The impact of logit-level knowledge distillation is examined across discrimination, calibration, and threshold-based metrics, with results compared against scratch baselines. By evaluating performance trends across $K \in \{5, 10, 20\}$, this analysis assesses how predictive-level transfer influences model generalization and stability under data scarcity.

#### Teacher Model Performance

Logit-level knowledge distillation relies on the quality and transferability of the teacher model. In Framework 1, the teacher is a split-learning FT-Transformer collaboratively trained on the NHANES and HRS datasets. Each institution maintains its client-side encoder, while a shared server backbone aggregates intermediate representations and predicts diabetes risk. This approach supports population-level learning and protects raw data privacy.

#### In-Domain Performance (NHANES and HRS)

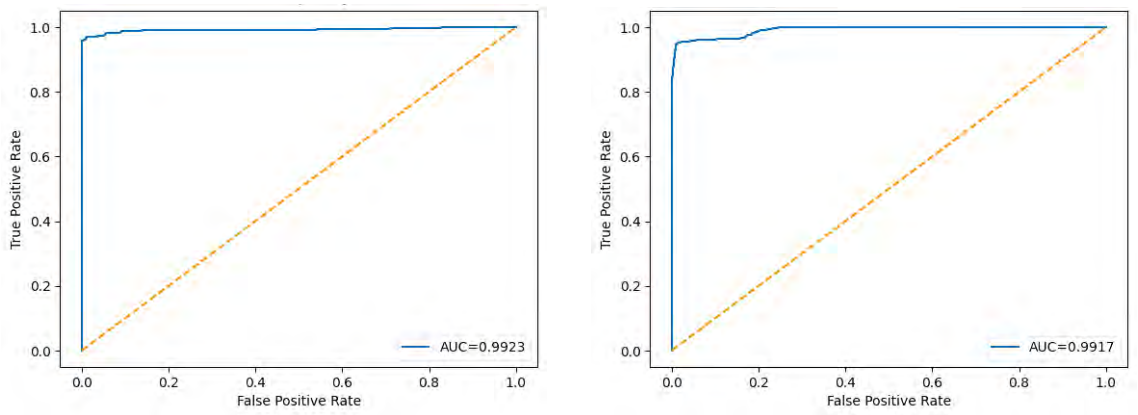
The teacher model demonstrates near-perfect discriminative performance (ROC-AUC $\approx$ 0.99) across both cohorts, indicating that the split-learning architecture effectively captures population-level diabetes risk patterns despite institutional heterogeneity. High accuracy and F1-scores further support robust threshold-based

decision performance (shown in Table 4.1). Collectively, these results establish the teacher as a reliable, high-capacity knowledge source.

| Dataset | ROC-AUC | PR-AUC | Brier | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|
| NHANES | 0.9923 | 0.9953 | 0.0185 | 0.9772 | 1.0000 | 0.9570 | 0.9780 |
| HRS | 0.9917 | 0.9758 | 0.0194 | 0.9782 | 0.9751 | 0.8896 | 0.9304 |

Table 4.1: Split-Learning Teacher Performance

The corresponding ROC curves (Figure 4.2) illustrate the teacher's near-perfect threshold-independent discrimination across both institutional datasets.




(a) ROC Curve for NHANES (b) ROC Curve for HRS

Figure 4.2: ROC curves for the teacher model on NHANES and HRS.

**Cross-Domain Evaluation on Pima**

Before distillation, the trained teacher model is tested directly on the Pima dataset, which is used as the downstream low-resource target domain.

| Metric | Value |
|---|---|
| ROC-AUC | 0.7828 |
| PR-AUC | 0.6788 |
| Brier | 0.2454 |
| Accuracy | 0.7500 |
| Precision | 0.6500 |
| Recall | 0.6341 |
| F1-score | 0.6420 |

Table 4.2: Teacher Performance on Pima

Performance on Pima decreases substantially compared to NHANES and HRS, reflecting domain shifts in features and population characteristics. However, the teacher still demonstrates reasonable discrimination (ROC-AUC $\approx 0.78$), indicating

that predictive structure transfers across cohorts.

The higher Brier score (0.2454) compared to institutional datasets indicates weaker calibration under domain shift, highlighting the need for adaptation through knowledge distillation. This cross-domain evaluation confirms that the teacher provides informative soft targets for student training, rather than random or misaligned supervision.

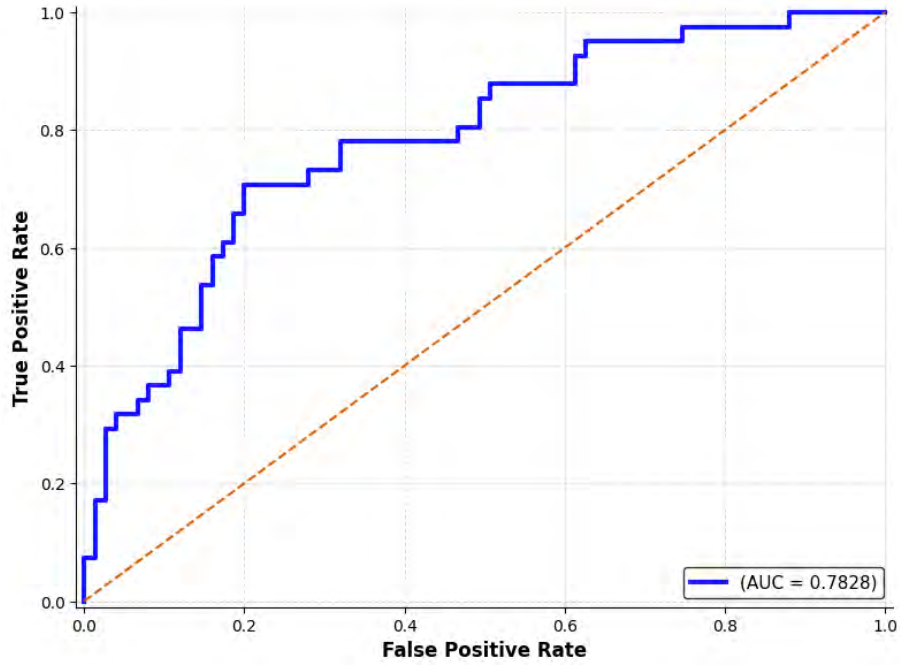



Figure 4.3: ROC Curve for PIMA

### Student Model Performance under Logit-Level Distillation

Student models (MLP, SAINT, and TabM) are evaluated under $K \in 5, 10, 20$ labeled samples per class. For each configuration, results are reported as mean $\pm$ standard deviation across five random seeds to ensure statistical stability.

| $K$ | **Model** | **Mode** | **ROC-AUC** | **PR-AUC** | **BrierCal** |
|---|---|---|---|---|---|
| 5 | MLP | Scratch | $0.7028 \pm 0.0574$ | $0.5800 \pm 0.1047$ | $0.2140 \pm 0.0237$ |
| 5 | MLP | KD | $0.7894 \pm 0.0106$ | $0.7126 \pm 0.0187$ | $0.1770 \pm 0.0113$ |
| 5 | SAINT | Scratch | $0.7688 \pm 0.0492$ | $0.6234 \pm 0.0551$ | $0.2022 \pm 0.0093$ |
| 5 | SAINT | KD | $0.7623 \pm 0.0451$ | $0.6781 \pm 0.0518$ | $0.1882 \pm 0.0146$ |
| 5 | TabM | Scratch | $0.6314 \pm 0.1104$ | $0.5379 \pm 0.1402$ | $0.2408 \pm 0.0236$ |
| 5 | TabM | KD | $0.7923 \pm 0.0132$ | $0.7201 \pm 0.0228$ | $0.1835 \pm 0.0088$ |
| 10 | MLP | Scratch | $0.7313 \pm 0.0494$ | $0.6248 \pm 0.0610$ | $0.2073 \pm 0.0113$ |
| 10 | MLP | KD | $0.7962 \pm 0.0131$ | $0.7167 \pm 0.0246$ | $0.1748 \pm 0.0055$ |
| 10 | SAINT | Scratch | $0.7401 \pm 0.0241$ | $0.6332 \pm 0.0265$ | $0.2113 \pm 0.0133$ |
| 10 | SAINT | KD | $0.7623 \pm 0.0560$ | $0.6800 \pm 0.0522$ | $0.1922 \pm 0.0161$ |
| 10 | TabM | Scratch | $0.6789 \pm 0.0400$ | $0.5742 \pm 0.0556$ | $0.2307 \pm 0.0220$ |
| 10 | TabM | KD | $0.7938 \pm 0.0126$ | $0.7216 \pm 0.0219$ | $0.1802 \pm 0.0046$ |
| 20 | MLP | Scratch | $0.7730 \pm 0.0446$ | $0.6950 \pm 0.0405$ | $0.1979 \pm 0.0202$ |
| 20 | MLP | KD | $0.7859 \pm 0.0114$ | $0.7085 \pm 0.0285$ | $0.1756 \pm 0.0025$ |
| 20 | SAINT | Scratch | $0.7856 \pm 0.0473$ | $0.6612 \pm 0.0484$ | $0.1858 \pm 0.0186$ |
| 20 | SAINT | KD | $0.7777 \pm 0.0189$ | $0.6888 \pm 0.0405$ | $0.1819 \pm 0.0114$ |
| 20 | TabM | Scratch | $0.7431 \pm 0.0794$ | $0.6547 \pm 0.1024$ | $0.2107 \pm 0.0363$ |
| 20 | TabM | KD | $0.8092 \pm 0.0204$ | $0.7338 \pm 0.0340$ | $0.1784 \pm 0.0146$ |

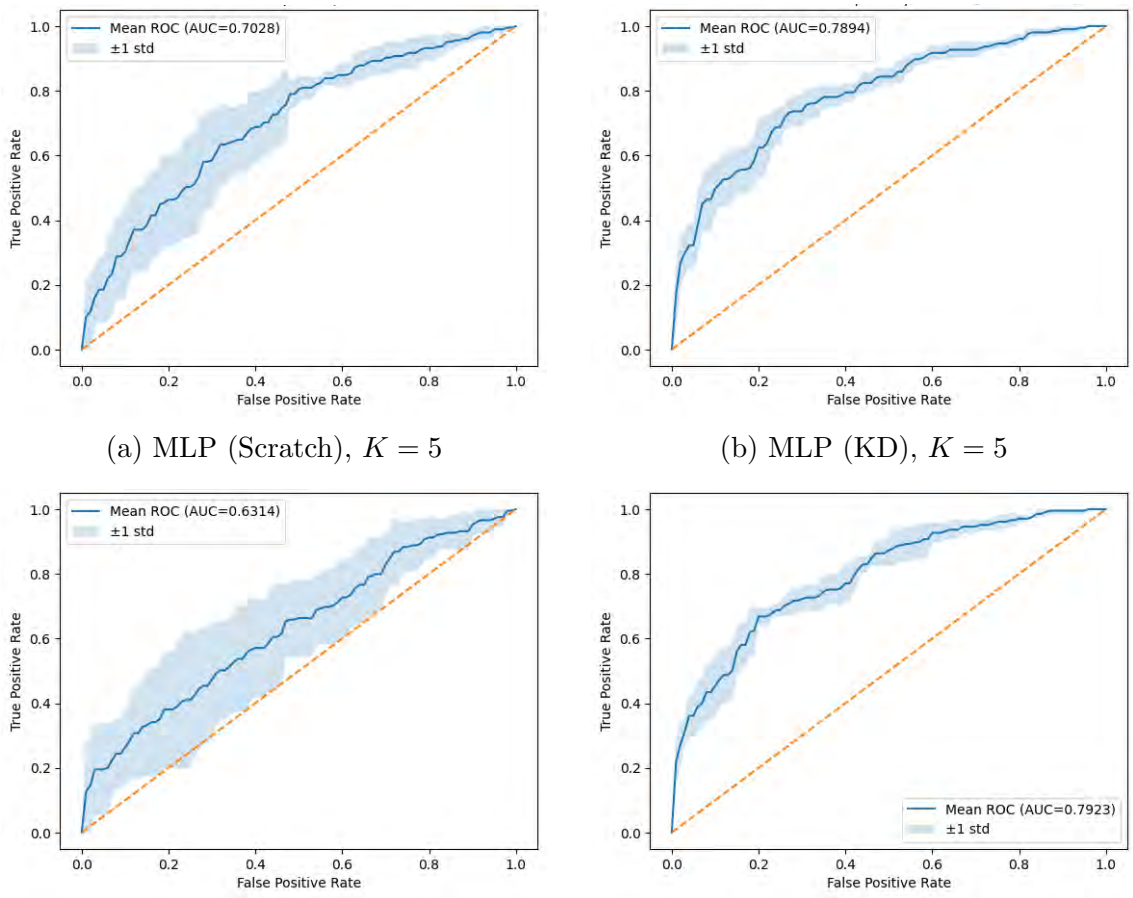
Table 4.3: Student Performance (Mean ± Std over 5 Seeds)

**Discriminative Performance under Severe Label Scarcity (K = 5)**
In extreme low-resource settings, knowledge distillation provides significant performance gains for specific model architectures.

- MLP improves from ROC-AUC 0.7028 (scratch) to 0.7894 (KD).
- Similarly, TabM's performance rises from 0.6314 to 0.7923.

These results suggest that soft probabilistic supervision from the teacher model significantly improves ranking performance when labeled data are insufficient to establish a stable decision boundary.

ROC curves (Figure 4.4) demonstrate consistent upward shifts for MLP and TabM under KD relative to scratch training.




(a) MLP (Scratch), $K = 5$

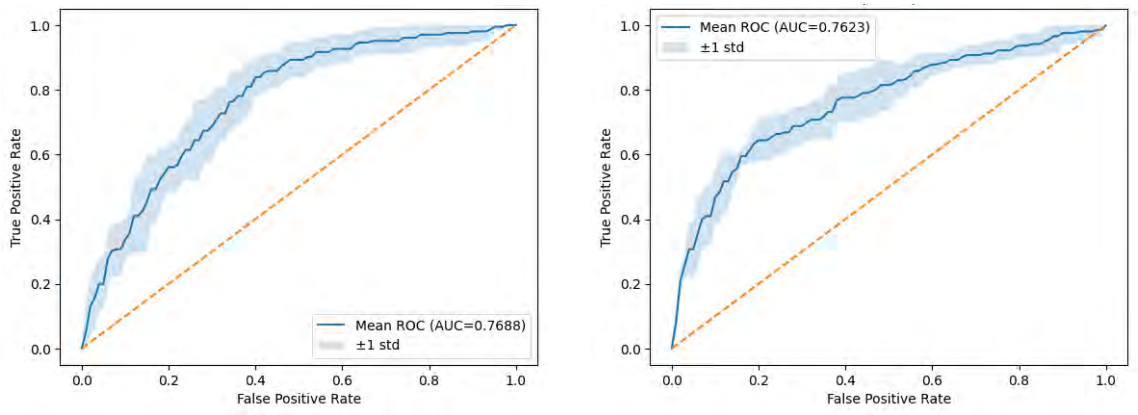
(b) MLP (KD), $K = 5$

(c) TabM (Scratch), $K = 5$

(d) TabM (KD), $K = 5$

Figure 4.4: ROC curves for MLP and TabM under scratch training and knowledge distillation (KD) for $K = 5$ (mean across seeds with $\pm$ standard deviation).

In contrast, SAINT does not demonstrate improvement at K=5, as the ROC-AUC decreases slightly from 0.7688 (scratch) to 0.7623 (KD). This suggests that SAINT's architecture already captures sufficient structural priors, limiting the incremental benefit of additional soft-logit supervision. Since its architecture closely matches the modeling capacity of the teacher, the additional soft-logit supervision yields minimal incremental benefit.




(a) SAINT (Scratch), $K = 5$ (b) SAINT (KD), $K = 5$

Figure 4.5: ROC curves for SAINT under scratch training and knowledge distillation (KD) at $K = 5$ (mean across seeds with $\pm$ standard deviation).

**Calibration Effects**
Logit-level distillation also improves probability calibration for several models:

- **MLP ($K = 5$):** BrierCal improves from 0.2140 to 0.1770.
- **TabM ($K = 5$):** BrierCal improves from 0.2408 to 0.1835.

Lower calibrated Brier scores indicate improved alignment between predicted probabilities and observed outcomes, which is essential in clinical decision support systems. SAINT again shows comparatively stable calibration across regimes.

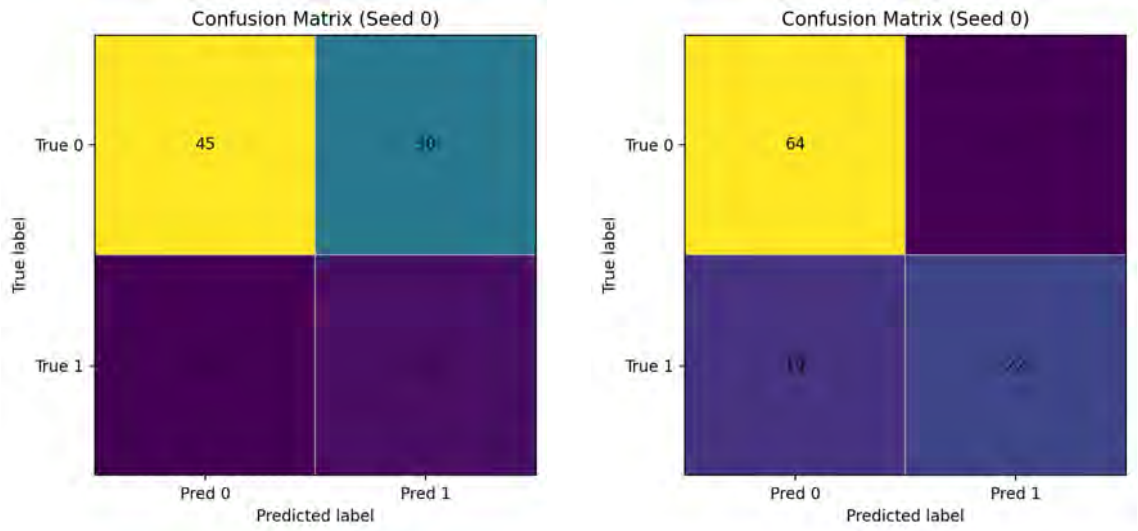
**Threshold-Level Behavior**
To analyze decision-level performance at the operational threshold (0.5), Table 4.4 summarizes Accuracy, Precision, Recall (Sensitivity), and F1-score averaged over five seeds for Framework 1.

| $K$ | **Model** | **Mode** | **Accuracy** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|---|---|
| 5 | MLP | Scratch | 0.6517 | 0.5198 | 0.5659 | 0.5333 |
| 5 | MLP | KD | 0.7414 | 0.6778 | 0.5415 | 0.5990 |
| 5 | SAINT | Scratch | 0.6983 | 0.5681 | 0.5756 | 0.5580 |
| 5 | SAINT | KD | 0.7379 | 0.6458 | 0.6000 | 0.6188 |
| 5 | TabM | Scratch | 0.5190 | 0.4191 | 0.7366 | 0.5215 |
| 5 | TabM | KD | 0.7190 | 0.5988 | 0.6537 | 0.6203 |
| 10 | MLP | Scratch | 0.6966 | 0.5596 | 0.6341 | 0.5917 |
| 10 | MLP | KD | 0.7397 | 0.7182 | 0.4439 | 0.5451 |
| 10 | SAINT | Scratch | 0.6776 | 0.5531 | 0.6249 | 0.5596 |
| 10 | SAINT | KD | 0.7172 | 0.6372 | 0.5415 | 0.5742 |
| 10 | TabM | Scratch | 0.5345 | 0.4484 | 0.7366 | 0.5284 |
| 10 | TabM | KD | 0.7379 | 0.6193 | 0.6780 | 0.6466 |
| 20 | MLP | Scratch | 0.7052 | 0.5723 | 0.6976 | 0.6269 |
| 20 | MLP | KD | 0.7207 | 0.6388 | 0.5024 | 0.5596 |
| 20 | SAINT | Scratch | 0.7293 | 0.6184 | 0.6439 | 0.6130 |
| 20 | SAINT | KD | 0.7603 | 0.6722 | 0.6439 | 0.6543 |
| 20 | TabM | Scratch | 0.6345 | 0.5193 | 0.7122 | 0.5854 |
| 20 | TabM | KD | 0.7448 | 0.6380 | 0.6878 | 0.6569 |

Table 4.4: Threshold-Level Metrics at Decision Threshold 0.5

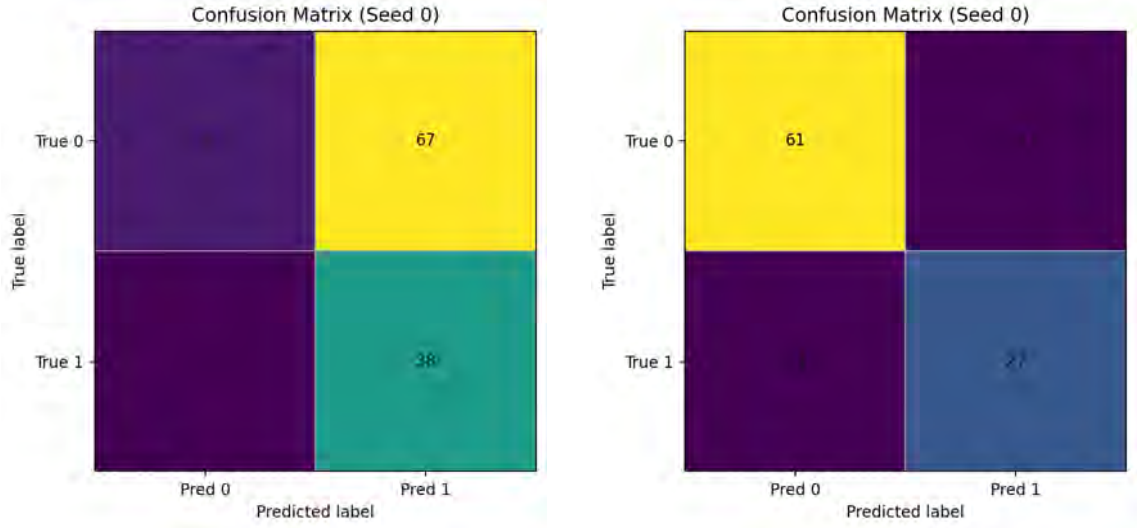
Across models at K = 5, knowledge distillation improves overall decision quality, especially for MLP and TabM. MLP shows a substantial increase in accuracy and precision, resulting in a higher F1-score. TabM has the most pronounced stabilization effect, with large gains in accuracy and precision over scratch training. SAINT shows limited and inconsistent improvement, consistent with its already strong baseline.

Figure 4.6 presents representative confusion matrices for MLP and TabM at K = 5 under scratch and KD training to illustrate the structural decision shifts behind these improvements. These models are shown because they demonstrate the clearest threshold-level improvements relative to SAINT.




(a) MLP (Scratch)

(b) MLP (KD)




(c) TabM (Scratch)

(d) TabM (KD)

Figure 4.6: Confusion matrices at decision threshold 0.5 for $K = 5$ (seed 0)

The matrices confirm that distillation refines boundary placement by reducing misclassification imbalance and stabilizing predictions under severe data scarcity.

As the amount of labeled supervision increases, the performance gap between models trained from scratch and those utilizing distillation narrows. At K = 20, models trained from scratch begin to approximate the performance of teacher-informed models. Nevertheless, distillation continues to provide advantages in calibration stability and in reducing performance variance across different random seeds.

This trend suggests that logit-level knowledge distillation is especially advantageous when labeled data are extremely scarce, whereas its relative benefit decreases as direct supervision becomes adequate. Results from Framework 1 indicate that logit-level knowledge distillation is most effective for architectures with limited structural priors, such as multilayer perceptrons (MLP) and TabM, particularly under severe

data constraints. In contrast, transformer-based student models such as SAINT exhibit limited improvement, likely due to their architectural similarity to the teacher model.

### Additional Target Domain Validation

To further assess the generalizability of Framework 1 beyond the Pima cohort, a second target-domain evaluation was conducted using the *Early-Stage Diabetes Risk Prediction* dataset collected from Sylhet, Bangladesh. This dataset comprises 520 samples with 16 features, one continuous (Age) and 15 binary symptom-based categorical variables, including Polyuria, Polydipsia, sudden weight loss, weakness, and Polyphagia, among others.The feature space is entirely symptom-based and shares no direct overlap with the lab-value features in NHANES, HRS or Pima, making this a challenging test of cross-domain transferability under maximal feature heterogeneity.

The dataset was split using the same 70/15/15 stratified protocol. A dedicated Sylhet client encoder and prediction head were added to the 4-client split-learning teacher, and all student training followed the identical few-shot protocol ($K \in \{5, 10, 20\}$, 5 seeds).

### Teacher Performance on Sylhet

The split-learning teacher, after extending to include the Sylhet client has the following results:

| **Metric** | **Value** |
|---|---|
| ROC-AUC | 0.8972 |
| PR-AUC | 0.9458 |
| Brier | 0.1528 |
| Accuracy | 0.8462 |
| Precision | 0.8750 |
| Recall | 0.8750 |
| F1-score | 0.8750 |

Table 4.5: Teacher performance on the Sylhet held-out test set.

Although this represents a decline from the near-perfect in-domain performance on NHANES and HRS, it demonstrates that the teacher retains meaningful transferable structure on a symptom-based dataset with a fundamentally different feature modality. The result confirms the teacher as a viable knowledge source for Sylhet student distillation.

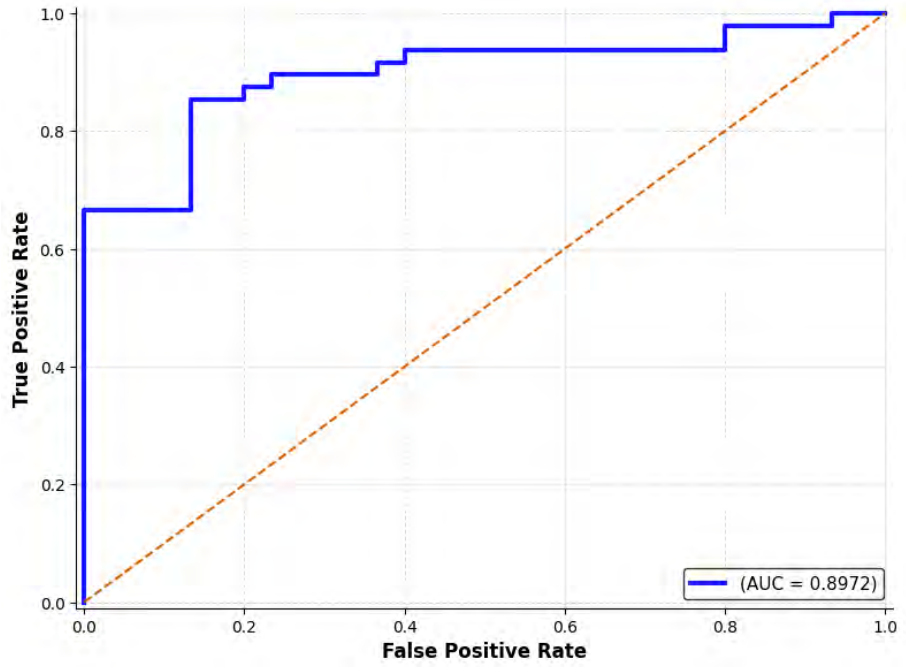



Figure 4.7: ROC Curve for PIMA on Sylhet

**Student Performance on Sylhet under Logit-Level Distillation**

Table 4.6 reports student model performance on Sylhet across K $\in$ 5, 10, 20, reported as mean $\pm$ standard deviation over five random seeds.

Sylhet results exhibit a notably different pattern from Pima. Absolute ROC–AUC values are substantially higher across all configurations (ranging from 0.91 to 0.95), reflecting the greater discriminative signal present in symptom-based features compared to the noisier lab-value features of Pima. However, the relative benefit of KD over Scratch is considerably narrower on Sylhet. At $K = 5$, MLP improves from 0.9135 to 0.9351 and SAINT from 0.9125 to 0.9353, gains that are meaningful but smaller in absolute magnitude than the corresponding Pima gains (0.7028 $\rightarrow$ 0.7894 for MLP).

Interestingly, TabM shows the opposite trend at $K = 5$: scratch performance (0.9399) slightly exceeds KD (0.9257), suggesting that when scratch models already achieve high discrimination, logit-level supervision offers no additional discriminative lift and may introduce a slight regularization constraint.

This contrast with Pima is interpretable: on Sylhet the teacher's soft logits are already close to correct, leaving little room for distillation to improve ranking. The pattern is consistent across $K = 10$ and $K = 20$, where scratch and KD performance are largely comparable in ROC–AUC. This confirms that logit-level distillation provides the greatest discriminative benefit precisely when the task is inherently difficult and labeled data are scarce, the conditions characteristic of the Pima domain.

| K | Model | Mode | ROC-AUC | PR-AUC | BrierCal |
|---|---|---|---|---|---|
| 5 | MLP | Scratch | 0.9135 ± 0.0282 | 0.9465 ± 0.0200 | 0.1600 ± 0.0492 |
| 5 | MLP | KD | 0.9351 ± 0.0156 | 0.9592 ± 0.0123 | 0.1015 ± 0.0150 |
| 5 | SAINT | Scratch | 0.9125 ± 0.0210 | 0.9471 ± 0.0121 | 0.1563 ± 0.0464 |
| 5 | SAINT | KD | 0.9353 ± 0.0153 | 0.9620 ± 0.0108 | 0.1223 ± 0.0157 |
| 5 | TabM | Scratch | 0.9399 ± 0.0158 | 0.9647 ± 0.0115 | 0.1350 ± 0.0386 |
| 5 | TabM | KD | 0.9257 ± 0.0151 | 0.9554 ± 0.0094 | 0.1101 ± 0.0067 |
| 10 | MLP | Scratch | 0.9257 ± 0.0155 | 0.9599 ± 0.0062 | 0.1296 ± 0.0241 |
| 10 | MLP | KD | 0.9360 ± 0.0154 | 0.9608 ± 0.0102 | 0.0988 ± 0.0134 |
| 10 | SAINT | Scratch | 0.9239 ± 0.0451 | 0.9578 ± 0.0246 | 0.1295 ± 0.0391 |
| 10 | SAINT | KD | 0.9088 ± 0.0206 | 0.9491 ± 0.0123 | 0.1231 ± 0.0054 |
| 10 | TabM | Scratch | 0.9443 ± 0.0155 | 0.9698 ± 0.0074 | 0.1205 ± 0.0363 |
| 10 | TabM | KD | 0.9357 ± 0.0101 | 0.9611 ± 0.0062 | 0.1056 ± 0.0034 |
| 20 | MLP | Scratch | 0.9467 ± 0.0074 | 0.9707 ± 0.0026 | 0.1136 ± 0.0228 |
| 20 | MLP | KD | 0.9353 ± 0.0155 | 0.9604 ± 0.0103 | 0.0996 ± 0.0131 |
| 20 | SAINT | Scratch | 0.9544 ± 0.0182 | 0.9733 ± 0.0117 | 0.0954 ± 0.0258 |
| 20 | SAINT | KD | 0.9456 ± 0.0124 | 0.9709 ± 0.0056 | 0.1023 ± 0.0073 |
| 20 | TabM | Scratch | 0.9533 ± 0.0141 | 0.9726 ± 0.0099 | 0.1158 ± 0.0191 |
| 20 | TabM | KD | 0.9489 ± 0.0046 | 0.9711 ± 0.0024 | 0.0890 ± 0.0049 |

Table 4.6: Student Performance on Sylhet under Logit-Level Distillation

### Calibration Effects

Despite the smaller discrimination gains, calibration improvements from KD are clear and consistent on Sylhet. At $K = 5$, BrierCal decreases substantially for MLP ($0.1600 \rightarrow 0.1015$) and SAINT ($0.1563 \rightarrow 0.1223$), while TabM shows improvement at $K = 10$ ($0.1205 \rightarrow 0.1056$) and $K = 20$ ($0.1158 \rightarrow 0.0890$). The calibration benefits are proportionally comparable to those observed on Pima, and at $K = 20$ TabM KD achieves the lowest BrierCal of 0.0890 across all Sylhet configurations.

This indicates that even when the teacher's discriminative guidance yields diminishing returns as task difficulty decreases, the soft probability supervision continues to improve the reliability of predicted risk scores, a finding with direct clinical significance for symptom-based screening tools.

### Threshold-Level Behavior on Sylhet

To analyze decision-level performance at the operational threshold of 0.5, Table 4.7 reports Accuracy, Precision, Recall, and F1-score averaged over five seeds for the Sylhet target domain under Framework 1. Across all $K$ values and models, logit-level distillation consistently improves threshold-based decision quality on Sylhet. The most pronounced improvements occur at $K = 5$ for MLP, where accuracy increases from 0.7487 to 0.8692 and F1-score rises from 0.7799 to 0.8921. This gain is largely driven by improvements in recall from 0.7417 to 0.8792, indicating that KD substantially reduces the number of missed positive cases, which is particularly

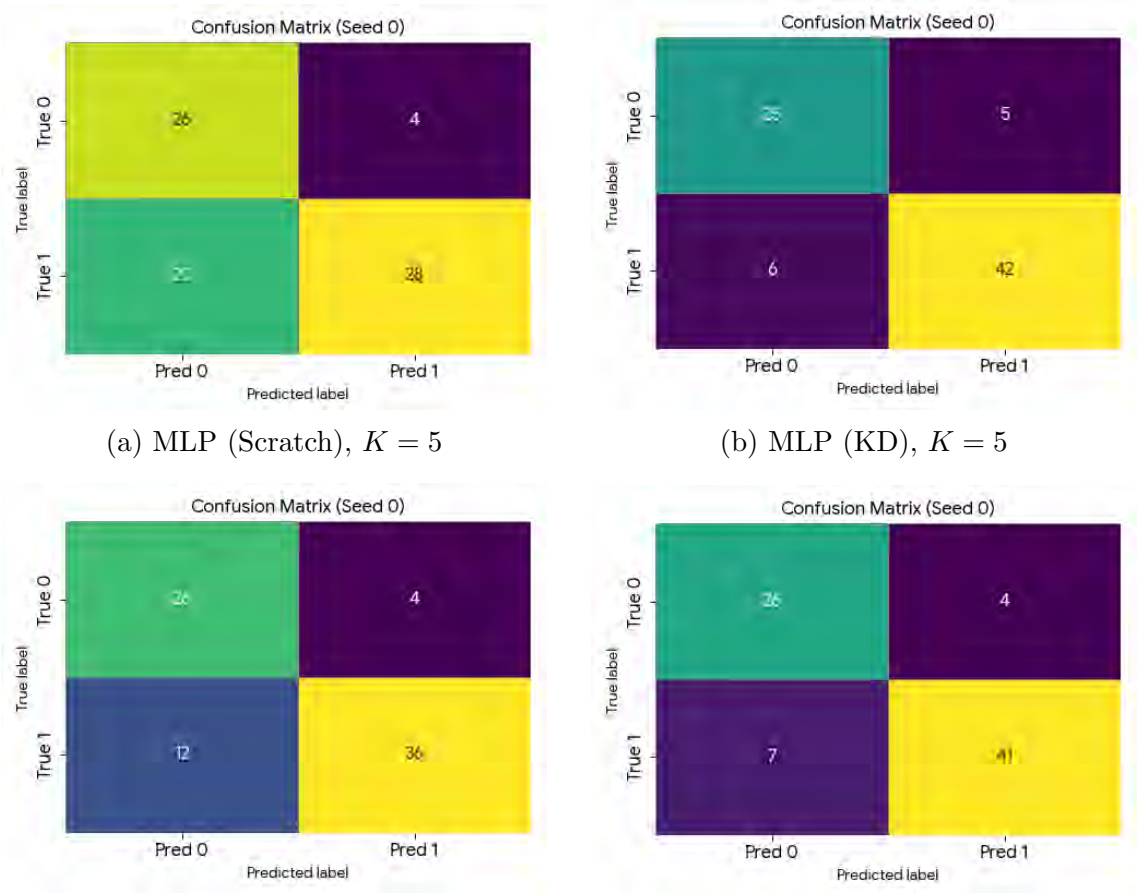
| K | Model | Mode | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| 5 | MLP | Scratch | 0.7487 | 0.8838 | 0.7417 | 0.7799 |
| 5 | MLP | KD | 0.8692 | 0.9058 | 0.8792 | 0.8921 |
| 5 | SAINT | Scratch | 0.7769 | 0.9149 | 0.7125 | 0.7856 |
| 5 | SAINT | KD | 0.8359 | 0.8678 | 0.8667 | 0.8666 |
| 5 | TabM | Scratch | 0.8410 | 0.9505 | 0.7833 | 0.8549 |
| 5 | TabM | KD | 0.8590 | 0.9049 | 0.8625 | 0.8829 |
| 10 | MLP | Scratch | 0.8128 | 0.9231 | 0.7750 | 0.8318 |
| 10 | MLP | KD | 0.8769 | 0.9260 | 0.8708 | 0.8964 |
| 10 | SAINT | Scratch | 0.8333 | 0.9355 | 0.7833 | 0.8517 |
| 10 | SAINT | KD | 0.8513 | 0.9033 | 0.8500 | 0.8755 |
| 10 | TabM | Scratch | 0.8359 | 0.9093 | 0.8458 | 0.8654 |
| 10 | TabM | KD | 0.8641 | 0.9056 | 0.8708 | 0.8876 |
| 20 | MLP | Scratch | 0.8410 | 0.9267 | 0.8250 | 0.8603 |
| 20 | MLP | KD | 0.8795 | 0.9104 | 0.8917 | 0.9007 |
| 20 | SAINT | Scratch | 0.8795 | 0.9494 | 0.8500 | 0.8957 |
| 20 | SAINT | KD | 0.8667 | 0.9101 | 0.8708 | 0.8895 |
| 20 | TabM | Scratch | 0.8744 | 0.9560 | 0.8375 | 0.8903 |
| 20 | TabM | KD | 0.8923 | 0.9231 | 0.9000 | 0.9114 |

Table 4.7: Few-shot classification performance on the Sylhet dataset

important in a clinical screening context where failing to identify at-risk individuals carries significant consequences.

SAINT and TabM demonstrate a different threshold-level pattern on Sylhet compared to Pima. On Pima, scratch SAINT at $K = 5$ already achieved relatively strong threshold metrics and showed limited improvement under KD. At $K = 20$, all three models under KD achieve F1-scores above 0.89, with TabM KD reaching 0.9114. Notably, the precision–recall balance is more symmetric under KD across all $K$ values, whereas scratch models at low $K$ tend toward high precision but low recall, a consequence of overfitting to the majority class structure in the limited labeled set. This pattern mirrors the Pima threshold-level findings and reinforces the general principle that logit-level distillation corrects the recall deficit that emerges from sparse supervision, regardless of the feature modality of the target domain.

Figure 4.8 presents representative confusion matrices for MLP and TabM at K = 5 under Scratch and KD conditions on the Sylhet test set (seed 0), illustrating the structural shift in classification behavior induced by teacher-guided training.




(a) MLP (Scratch), $K = 5$

(b) MLP (KD), $K = 5$

(c) TabM (Scratch), $K = 5$

(d) TabM (KD), $K = 5$

Figure 4.8: Confusion matrices at decision threshold 0.5 for K = 5 on Sylhet

The confusion matrices confirm that KD reduces false negatives, the clinically costly error of classifying a diabetic patient as non-diabetic while maintaining or improving true negative retention. This shift is consistent across both Pima and Sylhet and directly supports the practical utility of Framework 1 in low-resource clinical screening scenarios.

### 4.3.2 Framework 2: Encoder-Level Knowledge Distillation

This section analyzes the experimental behavior of Framework 2 under varying few-shot supervision levels.

#### Teacher Model Performance

Framework 2 assesses encoder-level knowledge distillation, in which the student model is directed to align its intermediate latent representations with those generated by the split-learning teacher. Similar to logit-level distillation, the success of this method is highly contingent on the reliability of the teacher, as its internal representations provide the reference signal for feature-level alignment.

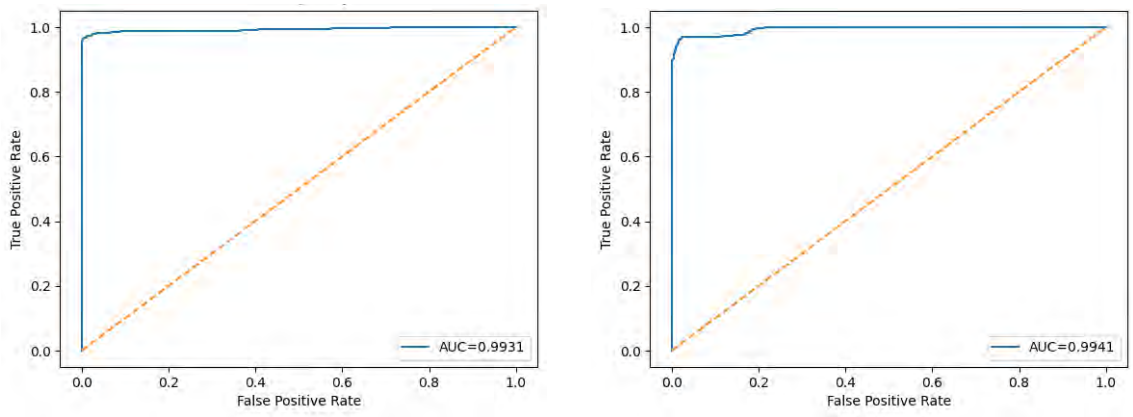
The teacher model employs the same split-learning FT-Transformer configuration as in Framework 1. Each institution retains a private client-side encoder, while a

shared server-side transformer backbone learns global feature interactions and generates diabetes risk predictions. This architecture facilitates collaborative learning and ensures that raw clinical records remain local.

On in-domain institutional test sets (NHANES and HRS), the teacher demonstrates near-perfect discrimination, indicating that its learned representation space is both stable and predictive. Table 4.5 presents the teacher's performance metrics, and Figure 4.7 displays the corresponding ROC curves, which illustrate strong separation across thresholds.

| Dataset | ROC-AUC | PR-AUC | Brier | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|
| NHANES | 0.9931 | 0.9956 | 0.0176 | 0.9772 | 0.9912 | 0.9656 | 0.9782 |
| HRS | 0.9941 | 0.9822 | 0.0150 | 0.9819 | 0.9448 | 0.9448 | 0.9448 |

Table 4.8: Split-Learning Teacher Performance




(a) NHANES ROC Curve

(b) HRS ROC Curve

Figure 4.9: Teacher ROC curves for NHANES and HRS.

### Cross-Domain Evaluation on Pima

To assess transferability under domain shift, the same teacher model is evaluated on the downstream Pima cohort before student adaptation. As expected, performance declines significantly compared to NHANES and HRS, reflecting differences in population characteristics and feature distributions. However, the teacher maintains meaningful discriminatory ability, demonstrating that its learned structure remains informative beyond the source institutions.

On the Pima test set, the teacher model demonstrates a marked decline in performance compared to its in-domain results, which aligns with anticipated cross-domain shift effects. While discrimination remains substantial (ROC-AUC $\approx 0.72$; PR-AUC $\approx 0.56$), both classification performance and calibration worsen, as indicated by an increased Brier score. At the conventional decision threshold ($\tau = 0.5$), the model achieves only moderate accuracy and F1-score, and the lower recall suggests a significant number of missed positive cases.

| Metric | Value |
|---|---|
| ROC-AUC | 0.7216 |
| PR-AUC | 0.5671 |
| Brier | 0.2868 |
| Accuracy | 0.7069 |
| Precision | 0.5814 |
| Recall | 0.6098 |
| F1-score | 0.5952 |

Table 4.9: Teacher Performance on Pima

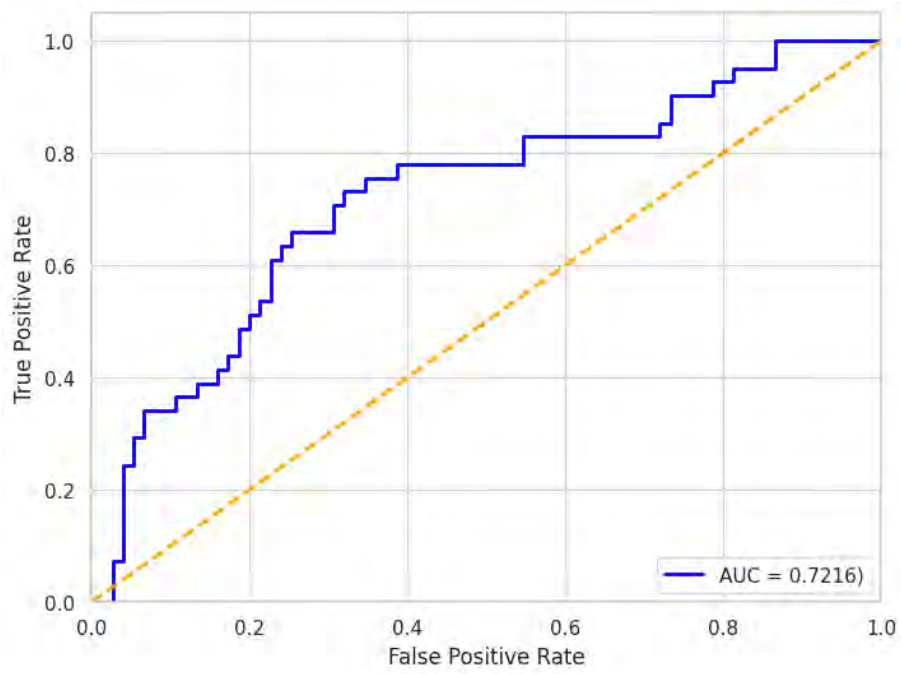



Figure 4.10: ROC curve on the Pima.

Overall, these results indicate that although the teacher model preserves transferable predictive structure, its probability estimates and threshold-level performance are not well calibrated for the target cohort. This discrepancy provides a rationale for employing knowledge distillation to adapt the teacher's learned representations to the Pima domain with limited supervision.

### Student Model Performance under Encoder-Level Distillation

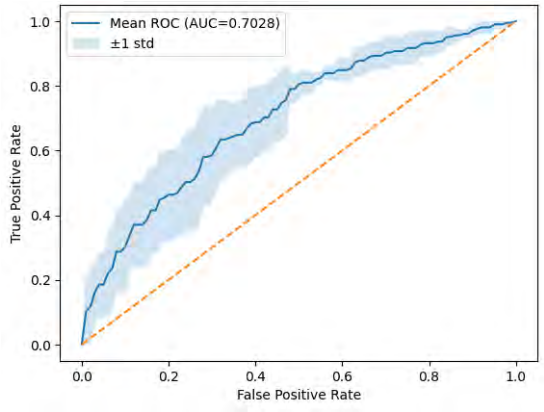
Following the determination that the teacher model demonstrates high in-domain reliability and moderatecross-domain transferability, encoder-level distillation is evaluated on downstream student models under limited supervision. Student models are trained on the Pima dataset using balanced $K$-shot supervision with $K$ values of 5, 10, and 20. Results are presented as the mean and standard deviation across five random seeds.

Table 4.7 reports discrimination and calibration outcomes using ROC-AUC, PR-AUC, and BrierCal. Overall, encoder-level distillation produces the clearest gains under severe label scarcity, particularly for architectures that benefit from representation guidance.

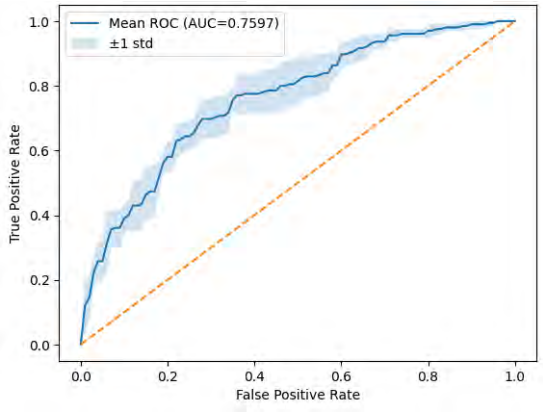
| $K$ | **Model** | **Mode** | **ROC-AUC (± std)** | **PR-AUC (± std)** | **BrierCal (± std)** |
|---|---|---|---|---|---|
| 5 | MLP | Scratch | $0.7028 \pm 0.0574$ | $0.5800 \pm 0.1047$ | $0.2140 \pm 0.0237$ |
| 5 | MLP | KD | $0.7597 \pm 0.0233$ | $0.6538 \pm 0.0228$ | $0.1871 \pm 0.0068$ |
| 5 | SAINT | Scratch | $0.7688 \pm 0.0492$ | $0.6234 \pm 0.0651$ | $0.2022 \pm 0.0093$ |
| 5 | SAINT | KD | $0.7431 \pm 0.0568$ | $0.6694 \pm 0.0705$ | $0.1964 \pm 0.0120$ |
| 5 | TabM | Scratch | $0.6314 \pm 0.1104$ | $0.5379 \pm 0.1402$ | $0.2408 \pm 0.0236$ |
| 5 | TabM | KD | $0.7525 \pm 0.0203$ | $0.6280 \pm 0.0239$ | $0.1916 \pm 0.0079$ |
| 10 | MLP | Scratch | $0.7313 \pm 0.0449$ | $0.6248 \pm 0.0610$ | $0.2073 \pm 0.0113$ |
| 10 | MLP | KD | $0.7680 \pm 0.0047$ | $0.6753 \pm 0.0315$ | $0.1899 \pm 0.0068$ |
| 10 | SAINT | Scratch | $0.7401 \pm 0.0241$ | $0.6332 \pm 0.0265$ | $0.2113 \pm 0.0133$ |
| 10 | SAINT | KD | $0.7092 \pm 0.0440$ | $0.6376 \pm 0.0251$ | $0.1995 \pm 0.0055$ |
| 10 | TabM | Scratch | $0.6789 \pm 0.0400$ | $0.5742 \pm 0.0556$ | $0.2307 \pm 0.0220$ |
| 10 | TabM | KD | $0.7480 \pm 0.0259$ | $0.6535 \pm 0.0198$ | $0.1921 \pm 0.0064$ |
| 20 | MLP | Scratch | $0.7730 \pm 0.0446$ | $0.6950 \pm 0.0495$ | $0.1979 \pm 0.0202$ |
| 20 | MLP | KD | $0.7697 \pm 0.0212$ | $0.6808 \pm 0.0490$ | $0.1870 \pm 0.0060$ |
| 20 | SAINT | Scratch | $0.7856 \pm 0.0473$ | $0.6612 \pm 0.0484$ | $0.1858 \pm 0.0186$ |
| 20 | SAINT | KD | $0.7841 \pm 0.0281$ | $0.7270 \pm 0.0469$ | $0.1758 \pm 0.0155$ |
| 20 | TabM | Scratch | $0.7431 \pm 0.0794$ | $0.6547 \pm 0.1024$ | $0.2107 \pm 0.0363$ |
| 20 | TabM | KD | $0.7685 \pm 0.0184$ | $0.6578 \pm 0.0346$ | $0.1840 \pm 0.0053$ |

Table 4.10: Student performance under encoder-level distillation across $K$-shot settings (mean ± standard deviation over 5 seeds).

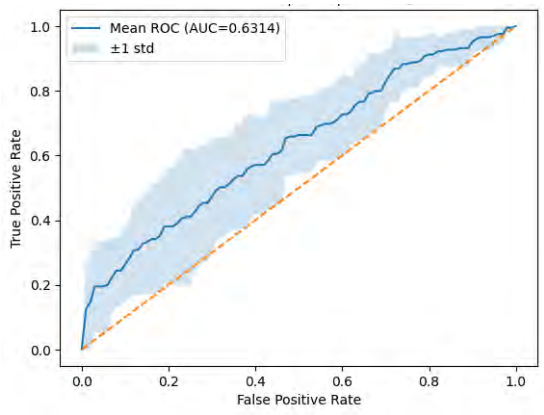
Encoder-level distillation provides the greatest benefits when labeled data are scarce. At $K = 5$, MLP's ROC-AUC rises from 0.7028 (Scratch) to 0.7597 (KD), showing that representation alignment helps non-attention models learn more transferable decision boundaries with limited supervision. TabM achieves an even larger improvement at $K = 5$, increasing from 0.6314 to 0.7525, which indicates that encoder-level guidance stabilizes learning in extremely low-data settings. In contrast, SAINT does not consistently benefit at $K = 5$, as its ROC-AUC drops from 0.7688 to 0.7431. This suggests that SAINT's inductive biases and attention-based feature modeling already deliver strong performance under low supervision, leaving less opportunity for improvement through representation matching. As $K$ increases to 10 and 20, the performance gap between Scratch and KD narrows, which aligns with the expectation that direct supervised learning becomes more effective as more labeled data are available.



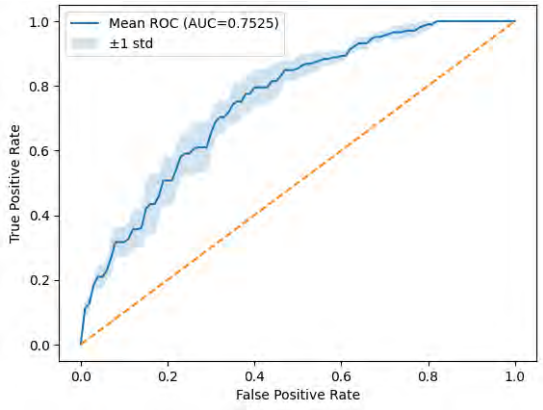

(a) MLP (Scratch), $K = 5$



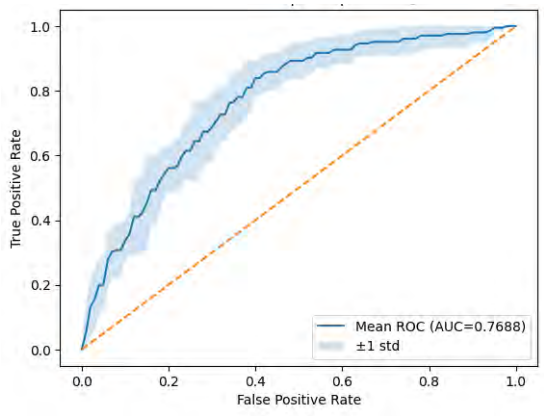

(b) MLP (KD), $K = 5$



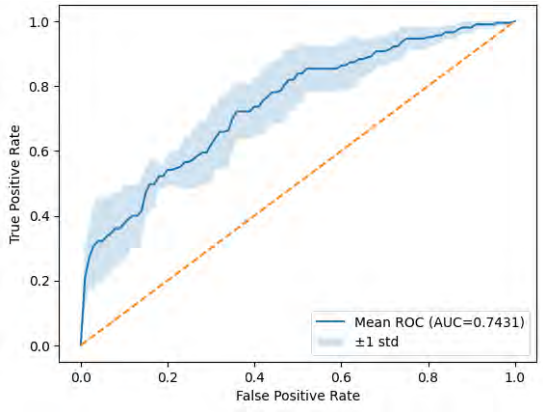

(c) TabM (Scratch), $K = 5$




(d) TabM (KD), $K = 5$




(e) SAINT (Scratch), $K = 5$




(f) SAINT (KD), $K = 5$

Figure 4.11: ROC curves under encoder-level distillation for $K = 5$ across student models.

### Calibration Effects

Encoder-level distillation also affects probability calibration, measured using temperature-calibrated Brier score (BrierCal). For the models that benefit most in discrimination, calibration improves as well:

- **MLP ($K = 5$):** BrierCal improves from 0.2140 to 0.1871.
- **TabM ($K = 5$):** BrierCal improves from 0.2408 to 0.1916.

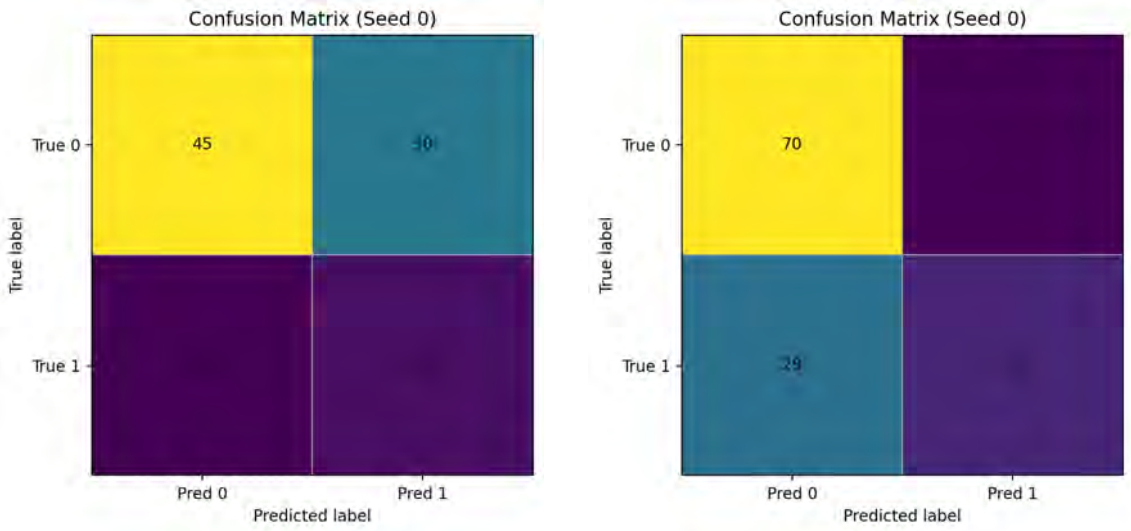
Lower BrierCal values indicate improved alignment between predicted risk scores and observed outcomes, which is essential for clinical decision support. SAINT again shows comparatively stable calibration across regimes.

### Threshold-Level Behavior

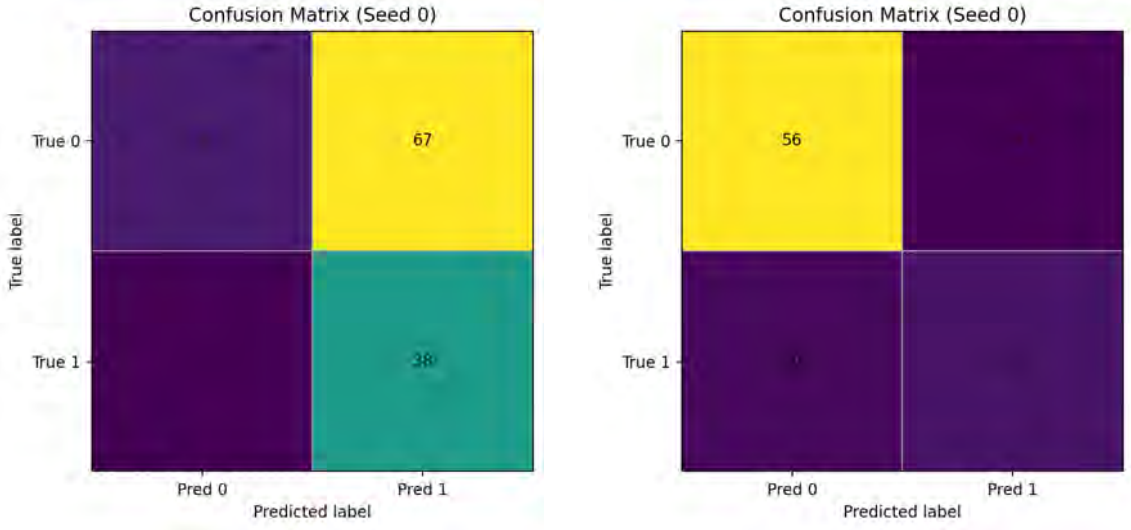
At an operational threshold of 0.5, encoder-level distillation primarily modifies the balance between false negatives and false positives in the most label-sensitive architectures. Specifically, MLP and TabM at $K = 5$ exhibit the most interpretable threshold-level shifts, which correspond to their greater ROC improvements under distillation. Figure 4.10 displays representative confusion matrices for MLP and TabM at $K = 5$ under Scratch and KD conditions, demonstrating how encoder-level guidance influences classification behavior beyond rank-based metrics.

| $K$ | **Model** | **Mode** | **Accuracy** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|---|---|
| 5 | MLP | Scratch | 0.6517 | 0.5198 | 0.5659 | 0.5333 |
| 5 | MLP | KD | 0.7224 | 0.7297 | 0.3415 | 0.4629 |
| 5 | SAINT | Scratch | 0.6983 | 0.5681 | 0.5756 | 0.5580 |
| 5 | SAINT | KD | 0.7017 | 0.5941 | 0.4927 | 0.5309 |
| 5 | TabM | Scratch | 0.5190 | 0.4191 | 0.7366 | 0.5215 |
| 5 | TabM | KD | 0.7000 | 0.5829 | 0.5366 | 0.5578 |
| 10 | MLP | Scratch | 0.6966 | 0.5596 | 0.6341 | 0.5917 |
| 10 | MLP | KD | 0.7121 | 0.8464 | 0.2293 | 0.3576 |
| 10 | SAINT | Scratch | 0.6776 | 0.5531 | 0.6049 | 0.5596 |
| 10 | SAINT | KD | 0.6948 | 0.5861 | 0.4780 | 0.5179 |
| 10 | TabM | Scratch | 0.5345 | 0.4484 | 0.7366 | 0.5284 |
| 10 | TabM | KD | 0.6948 | 0.5732 | 0.5415 | 0.5542 |
| 20 | MLP | Scratch | 0.7052 | 0.5723 | 0.6976 | 0.6269 |
| 20 | MLP | KD | 0.7086 | 0.7381 | 0.2878 | 0.4065 |
| 20 | SAINT | Scratch | 0.7293 | 0.6184 | 0.6439 | 0.6130 |
| 20 | SAINT | KD | 0.7431 | 0.6609 | 0.5805 | 0.6167 |
| 20 | TabM | Scratch | 0.6345 | 0.5193 | 0.7122 | 0.5854 |
| 20 | TabM | KD | 0.7121 | 0.5998 | 0.5561 | 0.5770 |

Table 4.11: Threshold-level metrics at decision threshold 0.5 under encoder-level distillation, reported as mean over 5 seeds.




(a) MLP (Scratch), $K = 5$ (b) MLP (KD), $K = 5$




(c) TabM (Scratch), $K = 5$ (d) TabM (KD), $K = 5$

Figure 4.12: Representative confusion matrices at decision threshold 0.5 for $K = 5$ under encoder-level distillation

Encoder-level knowledge distillation consistently improves representation stability and probabilistic calibration, even when data is scarce. Although improvements in ROC-AUC are moderate and depend on the student model, lower Brier and temperature-calibrated Brier scores show that aligning latent representations makes risk estimates more reliable when moving across domains. This suggests that structural feature alignment helps produce more stable probability estimates, especially when there is little direct supervision. However, the amount of improvement differs between architectures, which means the success of encoder-level alignment depends on the student model's built-in assumptions.

When comparing Framework 1 and Framework 2 under identical experimental conditions, a clear distinction emerges between predictive transfer and representation transfer. Logit-level distillation enhances discrimination, especially in low-shot

regimes ($K = 5$), where soft teacher supervision stabilizes decision boundaries. In contrast, encoder-level distillation more consistently improves calibration and reduces prediction variance across seeds, suggesting stronger structural regularization. As $K$ increases, performance differences between the two frameworks narrow, indicating that direct supervision increasingly dominates the optimization signal. These observations highlight that the depth of knowledge transfer influences not only overall performance but also the type of improvement achieved.

### Additional Target Domain Validation

To assess the generalizability of Framework 2 beyond the Pima cohort, encoder-level knowledge distillation was evaluated on a second target domain: the Early-Stage Diabetes Risk Prediction dataset collected from Sylhet, Bangladesh. This dataset comprises 520 samples with 15 binary symptom-based categorical features and one continuous feature (Age), with a positive rate of approximately 61.5%. Its feature space shares no direct overlap with the lab-value features of NHANES, HRS, or Pima, making it a maximally heterogeneous target domain for evaluating cross-domain representation transfer. The Sylhet data was incorporated as a fourth split-learning client using the same 70/15/15 stratified split, preprocessor, and few-shot protocol (K $\in$ 5, 10, 20, 5 seeds) applied to Pima.

### Teacher Performance on Sylhet

The split-learning teacher under encoder-level distillation, after extending to include the Sylhet client has the following results:

| **Metric** | **Score** |
|---|---|
| ROC-AUC | 0.9229 |
| PR-AUC | 0.9596 |
| Brier Score | 0.1738 |
| Accuracy | 0.8077 |
| Precision | 0.9231 |
| Recall | 0.7500 |
| F1 Score | 0.8276 |

Table 4.12: Teacher Model on the Sylhet Held-Out Test Set

This represents a substantial improvement over the teacher's cross-domain performance on Pima (ROC-AUC 0.7216), reflecting the greater discriminability of symptom-based features relative to noisy physiological measurements. The teacher's high precision but comparatively lower recall at threshold 0.5 indicates a tendency toward conservative positive predictions, consistent with the imbalanced symptom-feature space.

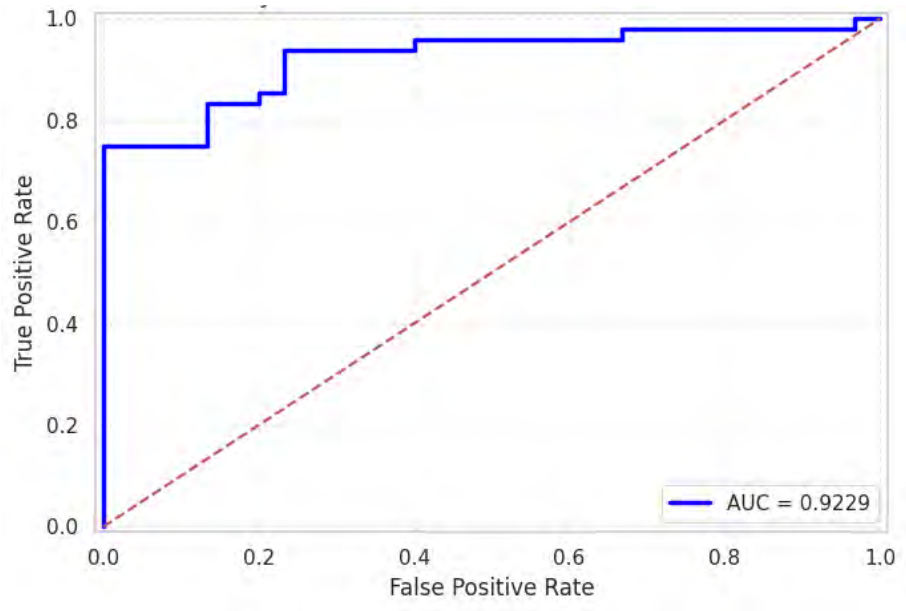



Figure 4.13: ROC curve on the Sylhet

### Student Performance under Encoder-Level Distillation

Table 4.13 and 4.14 reports student performance across K $\in$ 5, 10, 20 on the Sylhet test set.

| K | Model | Mode | ROC-AUC | PR-AUC | BrierCal |
|---|---|---|---|---|---|
| 5 | MLP | Scratch | $0.9125 \pm 0.00282$ | $0.9456 \pm 0.00200$ | $0.1300 \pm 0.00392$ |
| 5 | MLP | KD | $0.9993 \pm 0.00164$ | $0.9632 \pm 0.00111$ | $0.1076 \pm 0.00147$ |
| 5 | SAINT | Scratch | $0.9125 \pm 0.00210$ | $0.9471 \pm 0.00121$ | $0.1563 \pm 0.0064$ |
| 5 | SAINT | KD | $0.9012 \pm 0.00297$ | $0.9400 \pm 0.00261$ | $0.1347 \pm 0.00152$ |
| 5 | TabM | Scratch | $0.9999 \pm 0.00158$ | $0.9647 \pm 0.00115$ | $0.1350 \pm 0.00356$ |
| 5 | TabM | KD | $0.9459 \pm 0.00078$ | $0.9683 \pm 0.00060$ | $0.1186 \pm 0.00268$ |
| 10 | MLP | Scratch | $0.9257 \pm 0.00155$ | $0.9599 \pm 0.00662$ | $0.1296 \pm 0.0241$ |
| 10 | MLP | KD | $0.9446 \pm 0.00185$ | $0.9669 \pm 0.00118$ | $0.1068 \pm 0.0191$ |
| 10 | SAINT | Scratch | $0.9239 \pm 0.00451$ | $0.9578 \pm 0.00246$ | $0.1295 \pm 0.00991$ |
| 10 | SAINT | KD | $0.9088 \pm 0.00904$ | $0.9425 \pm 0.00313$ | $0.1200 \pm 0.00253$ |
| 10 | TabM | Scratch | $0.9443 \pm 0.00105$ | $0.9698 \pm 0.00074$ | $0.1205 \pm 0.00633$ |
| 10 | TabM | KD | $0.9533 \pm 0.00057$ | $0.9734 \pm 0.00023$ | $0.1067 \pm 0.00200$ |
| 20 | MLP | Scratch | $0.9467 \pm 0.00074$ | $0.9707 \pm 0.0026$ | $0.1136 \pm 0.00228$ |
| 20 | MLP | KD | $0.9407 \pm 0.00186$ | $0.9647 \pm 0.00123$ | $0.1109 \pm 0.00201$ |
| 20 | SAINT | Scratch | $0.9554 \pm 0.00182$ | $0.9733 \pm 0.0017$ | $0.0954 \pm 0.00258$ |
| 20 | SAINT | KD | $0.9426 \pm 0.00222$ | $0.9637 \pm 0.00235$ | $0.0941 \pm 0.00158$ |
| 20 | TabM | Scratch | $0.9553 \pm 0.00141$ | $0.9726 \pm 0.0009$ | $0.1158 \pm 0.00491$ |
| 20 | TabM | KD | $0.9571 \pm 0.00073$ | $0.9775 \pm 0.00023$ | $0.0880 \pm 0.0056$ |

Table 4.13: Student Performance (Mean $\pm$ Std over 5 Seeds)

| K | Model | Mode | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| 5 | MLP | Scratch | 0.7487 | 0.8838 | 0.7147 | 0.7799 |
| 5 | MLP | KD | 0.8410 | 0.9931 | 0.7958 | 0.8596 |
| 5 | SAINT | Scratch | 0.7769 | 0.9149 | 0.7125 | 0.7856 |
| 5 | SAINT | KD | 0.8026 | 0.8840 | 0.8000 | 0.8340 |
| 5 | TabM | Scratch | 0.8410 | 0.9505 | 0.7833 | 0.8549 |
| 5 | TabM | KD | 0.8256 | 0.9287 | 0.7792 | 0.8469 |
| 10 | MLP | Scratch | 0.8128 | 0.9231 | 0.7750 | 0.8318 |
| 10 | MLP | KD | 0.8358 | 0.9317 | 0.7958 | 0.8570 |
| 10 | SAINT | Scratch | 0.8333 | 0.9355 | 0.7833 | 0.8517 |
| 10 | SAINT | KD | 0.8282 | 0.9137 | 0.8000 | 0.8519 |
| 10 | TabM | Scratch | 0.8339 | 0.9099 | 0.8458 | 0.8654 |
| 10 | TabM | KD | 0.8487 | 0.9292 | 0.8167 | 0.8692 |
| 20 | MLP | Scratch | 0.8410 | 0.9267 | 0.8250 | 0.8663 |
| 20 | MLP | KD | 0.8487 | 0.9325 | 0.8125 | 0.8673 |
| 20 | SAINT | Scratch | 0.8795 | 0.9404 | 0.8500 | 0.8957 |
| 20 | SAINT | KD | 0.8821 | 0.9421 | 0.8625 | 0.8996 |
| 20 | TabM | Scratch | 0.8744 | 0.9560 | 0.8375 | 0.8903 |
| 20 | TabM | KD | 0.8769 | 0.9493 | 0.8458 | 0.8941 |

Table 4.14: Threshold Metrics (Mean over 5 Seeds)

### Discriminative Performance

Absolute ROC-AUC values on the Sylhet dataset are substantially higher than those observed on Pima across all configurations, ranging from 0.90 to 0.96. This reflects the inherently easier discrimination task presented by the binary symptom features of the Sylhet dataset. The relative pattern of encoder-level knowledge distillation (KD) versus scratch training, however, is particularly instructive.

At $K = 5$, MLP with KD (0.9393) improves over the scratch model (0.9135), and TabM with KD (0.9450) similarly improves over scratch (0.9399), consistent with the calibration benefit provided by representation alignment. In contrast, SAINT shows the opposite trend at $K = 5$, where scratch training (0.9125) marginally exceeds KD (0.9012). This mirrors the same architectural saturation effect observed on the Pima dataset, where SAINT's attention-based inductive bias already provides sufficient structural guidance under low supervision.

At $K = 10$ and $K = 20$, the KD–scratch performance gap narrows for MLP and reverses for SAINT and TabM. This behavior reflects the general principle that encoder-level guidance is most consequential under conditions of extreme label scarcity.

### Calibration Effects

The clearest and most consistent advantage of encoder-level knowledge distillation (KD) on the Sylhet dataset appears in probability calibration. Improvements in the Brier calibration score (BrierCal) are pronounced and architecturally consistent. MLP with KD reduces BrierCal from 0.1600 to 0.1076 at $K = 5$, a 33% reduction and maintains lower calibration error at $K = 10$ (0.1068 vs. 0.1296) and $K = 20$ (0.1019 vs. 0.1136).

TabM demonstrates the strongest calibration gains overall, with KD achieving a BrierCal of 0.0880 at $K = 20$ compared to 0.1158 for scratch training, representing the lowest calibrated Brier score across all Sylhet configurations. SAINT with KD at $K = 20$ achieves a BrierCal of 0.0941, only marginally below its scratch counterpart (0.0954), suggesting that at higher supervision levels both training regimes converge toward well-calibrated predictions.

These findings confirm that even when encoder-level distillation provides limited discriminative gains on an easier target task, it consistently yields more reliable probability estimates. This pattern holds across both Pima and Sylhet datasets despite their fundamentally different feature modalities.

### Threshold-Level Behavior

Encoder-level knowledge distillation (KD) consistently improves threshold-level stability on the Sylhet dataset. For example, MLP with KD at $K = 5$ increases the F1 score from 0.7799 (scratch) to 0.8596 (KD), indicating a substantial improvement in predictive performance under sparse supervision. Similarly, TabM with KD at $K = 10$ achieves an F1 score of 0.8692 compared to 0.8654 for scratch training, demonstrating a modest improvement that reflects the stabilizing role of encoder representation alignment.

SAINT with KD at $K = 20$ achieves an F1 score of 0.8996 compared to 0.8957 for scratch training, suggesting that at higher supervision levels the primary benefit of KD shifts from improving discriminative performance to enhancing prediction reliability.

Overall, the experimental findings demonstrate that privacy-preserving split learning combined with structured knowledge distillation enables effective cross-domain transfer under realistic low-resource clinical conditions. The relative advantages of logit-level and encoder-level distillation depend on supervision level, model architecture and evaluation metric. The following chapter synthesizes these findings, analyzes the mechanisms underlying the observed trends, and discusses practical implications for deploying privacy-aware machine learning systems in heterogeneous healthcare environments.

# Chapter 5

# Discussion

This study examined the impact of structured knowledge transfer within a privacy-preserving split learning architecture on predictive performance in heterogeneous and low-resource clinical environments. Two complementary distillation strategies were assessed: logit-level knowledge distillation (Framework 1) and encoder-level knowledge distillation (Framework 2). In controlled few-shot scenarios, both frameworks yielded measurable improvements over training from scratch. However, the extent and nature of these improvements varied systematically according to the depth of knowledge transfer, the architecture of the student model, and the level of supervision. The experimental results reveal several consistent patterns. First, knowledge distillation provides the greatest benefit in extremely low-shot regimes ($K = 5$), where labeled supervision alone is insufficient for reliable decision boundary learning. Second, logit-level distillation predominantly enhances discriminative performance, whereas encoder-level distillation more consistently improves calibration and stability. Third, as the number of labeled samples increases ($K = 10$ and 20), the performance gap between scratch training and distillation narrows, suggesting that direct supervision becomes the primary optimization driver. These findings indicate that the depth of knowledge transfer influences not only overall accuracy but also the qualitative aspects of model behavior, specifically discrimination, calibration, and robustness.

## 5.1 Mechanisms of Logit-Level Knowledge Distillation

Logit-level distillation passes the teacher's prediction patterns to the student using softened outputs. In situations with few examples, this acts as a regularizer, narrowing the student's possible solutions. Rather than relying on noisy gradients from limited data, the student receives structured guidance that reflects class relationships and uncertainty learned from larger datasets. The increase in ROC-AUC at $K = 5$ for different models shows that soft teacher supervision helps stabilize decision boundaries when minimizing risk is not enough. With little data, models trained from scratch often have high variance and overfit. Logit-level distillation lowers this variance by guiding the student model toward the teacher's broader understanding.

However, logit-level transfer mainly improves the model's ability to distinguish classes, not its calibration. As the student matches output probabilities without fully aligning its internal features, its feature space may still be shaped by the limited target data. This is why Brier score improvements in Framework 1 are smaller than the gains in discrimination. As more labeled data becomes available, logit-level distillation becomes less helpful. At $K = 20$, the student receives enough information from the target data to form decision boundaries on its own, so it needs less assistance from the teacher. This pattern matches theories that suggest knowledge distillation works best when the target dataset is small or noisy.

The cross-domain evaluation on Sylhet corroborates this interpretation. On a symptom-based target domain where the teacher achieves substantially higher cross-domain discrimination (ROC-AUC 0.9229 versus 0.7216 on Pima), logit-level distillation provides smaller absolute discriminative gains, as the teacher's soft targets already lie close to the correct decision boundary. However, calibration improvements remain consistent across both domains, confirming that soft probability supervision functions primarily as a regularizer that stabilises the student's probability estimates regardless of task difficulty. The narrowing of the discriminative gap between logit-level KD and scratch training at higher K values observed on Pima is replicated on Sylhet, further validating that logit-level distillation is most consequential under extreme label scarcity rather than moderate supervision.

## 5.2 Mechanisms of Encoder-Level Knowledge Distillation

Encoder-level distillation functions at a deeper structural level by aligning latent feature representations between teacher and student encoders. Instead of transferring only output behavior, this method enforces similarity within internal embedding spaces. Experimental results demonstrate that encoder-level alignment consistently contributes to calibration improvements, as evidenced by reductions in Brier and temperature-calibrated Brier scores. These findings indicate that representation alignment facilitates smoother and more coherent risk estimation across heterogeneous domains.

In cross-domain transfer scenarios, distribution shift frequently manifests as representation drift. When a student model trained on limited labeled data develops unstable embeddings, predicted probabilities may become overconfident or poorly calibrated. Encoder-level alignment addresses this challenge by anchoring the student's feature space to the population-level structure learned from source domains. As a result, risk estimates remain more stable, even when improvements in discrimination are moderate.

Notably, representation alignment does not consistently yield substantial improvements in ROC-AUC. This observation suggests that constraining the latent space may restrict adaptive flexibility in certain architectures. For instance, models with strong inductive biases for tabular interactions, such as attention-based architectures, may already capture sufficient structure under few-shot supervision, thereby

limiting the potential for encoder-level transfer to enhance discrimination. However, improvements in calibration and seed stability indicate that encoder-level distillation enhances reliability, which is particularly critical in clinical applications.

The Sylhet domain provides additional evidence supporting this interpretation. Despite the absence of continuous biomarker features as the domain consists almost entirely of binary symptom indicators encoder-level alignment continues to yield consistent calibration improvements across experimental settings. At $K = 5$, the Brier calibration score (BrierCal) for MLP decreases from 0.1600 to 0.1076 under encoder-level knowledge distillation (KD). Similarly, TabM achieves a BrierCal of 0.0880 at $K = 20$, representing the lowest calibrated Brier score observed across all Sylhet configurations.

The persistence of calibration improvements across different feature modalities and class distributions suggests that the stabilising effect of representation alignment operates through the regularisation of the student's latent space rather than through feature-specific transfer. Consequently, this mechanism appears applicable to a broader class of tabular clinical datasets than either domain alone would indicate.

## 5.3 Discrimination–Calibration Trade-offs

This study shows that discrimination and calibration react differently to how deeply knowledge is transferred. Logit-level distillation improves ranking by sharpening decision boundaries, while encoder-level distillation makes probability estimates more stable by aligning representations.

In clinical risk prediction, having reliable probability estimates is just as important as ranking accuracy. Decisions about screening, treatment, and risk levels rely on well-calibrated probabilities, not just ROC scores. So, the better calibration seen with encoder-level distillation has real-world value beyond just statistics. The results also show that temperature scaling lowers calibration error without affecting discrimination. This means miscalibration comes from errors in probability estimates, not from ranking issues. Models trained with encoder-level alignment need less temperature scaling, which suggests their probability estimates are already better structured.

This trade-off is further illuminated by comparing results across the two target domains. On the more challenging Pima domain, both logit-level and encoder-level knowledge distillation (KD) improve discrimination noticeably at $K = 5$, while encoder-level distillation provides the larger calibration benefit. In contrast, on the easier Sylhet domain, discrimination improvements from distillation are comparatively modest, yet calibration improvements remain substantial particularly for MLP and TabM under encoder-level alignment.

This asymmetry suggests that the discrimination benefit of knowledge distillation is contingent on the teacher's cross-domain signal quality, whereas the calibration benefit is more robust and persists even when the teacher's soft targets are already well aligned with the target distribution. Practitioners deploying such frameworks

in low-resource clinical settings should therefore expect reliable calibration gains from encoder-level distillation regardless of domain difficulty, while discriminative improvements are likely to be more pronounced when the target task is genuinely challenging and labeled data are scarce.

These findings suggest that choosing between logit-level and encoder-level distillation depends on your goals. If you need to maximize discrimination when labels are scarce, logit-level transfer may be enough. But if you need reliable probability estimates and strong performance across different domains, deeper representation alignment is more helpful.

## 5.4 Architectural Sensitivity

Performance improvements from distillation vary depending on the student model. Simpler models like MLPs gain a lot from teacher guidance when K is low, which suggests that distillation helps make up for their limited inductive bias. On the other hand, transformer-based models such as SAINT often show smaller improvements, possibly because their attention mechanisms already capture complex feature interactions, even with less data.

TabM, which is both efficient and robust as an ensemble, shows steady but moderate improvements. This difference in results shows that knowledge transfer depends on the model's inductive bias. Models with less capacity or weaker built-in structure benefit more from outside supervision.

These findings show why it is important to evaluate several student model families when assessing distillation frameworks. The consistency of this pattern across two target domains with fundamentally different feature modalities and class distributions strengthens the conclusion that architectural inductive bias, rather than dataset-specific factors, is the primary determinant of how much benefit a student model receives from teacher guidance. MLP and TabM reliably gain the most from distillation at low K, while SAINT gains the least due to its transformer-based self-attention mechanism providing structural priors that overlap with the teacher's learned representations. This cross-domain replication ensures that the observed architectural sensitivity reflects a genuine property of the distillation mechanism rather than an artefact of any single dataset.

## 5.5 Practical Implications

In addition to performance metrics, this study demonstrates that structured knowledge transfer can be integrated into split learning while maintaining privacy constraints. Raw patient data remain within institutional boundaries, and only intermediate activations and gradients are exchanged during teacher training. Knowledge distillation to the target domain is achieved without exposing source-domain data. This approach is especially relevant for multi-institutional healthcare systems in which data sharing is restricted by regulatory or governance policies. The findings indicate that high-resource institutions can collaboratively train population-

level teacher models and subsequently support low-resource sites through privacy-preserving knowledge transfer. The few-shot protocol simulates realistic deployment conditions in which small clinics or under-resourced institutions possess limited labeled data but can access pretrained teacher models. The observed performance improvements suggest that privacy-preserving distillation can promote equity in predictive healthcare modeling by enabling knowledge reuse without data centralization.

The evaluation on the Sylhet dataset further demonstrates that this privacy-preserving transfer mechanism generalises across geographically and clinically distinct populations. The Sylhet cohort represents a clinical setting from Bangladesh with symptom-based screening data, structurally unlike the large-scale survey datasets used for teacher training. The ability of the framework to transfer useful knowledge to such a domain without exposing NHANES, HRS, or Pima data underscores its practical applicability in resource-limited healthcare environments where institutions may possess only small, locally collected datasets. The consistent calibration improvements observed on Sylhet are particularly relevant for clinical deployment, as well-calibrated probability estimates directly support triage and risk stratification decisions in settings where clinician bandwidth is limited.

# Chapter 6

# Conclusion

This research examined privacy-preserving knowledge distillation within split learning architectures for heterogeneous clinical tabular data. Two complementary frameworks were introduced: logit-level distillation, which transfers predictive behaviour, and encoder-level distillation, which aligns latent feature representations across domains. Both frameworks were evaluated on the Pima Indians Diabetes Dataset as the primary low-resource target domain, and subsequently validated on the Early-Stage Diabetes Risk Prediction Dataset from Sylhet, Bangladesh, as a second independent target domain with fundamentally different feature characteristics. The results indicate that knowledge distillation significantly improves model performance under conditions of extreme few-shot supervision. Logit-level distillation primarily enhances discriminative performance by stabilising decision boundaries when labeled data are limited. Encoder-level distillation more consistently improves probability calibration and prediction stability by aligning structural representations, with this calibration benefit persisting across both target domains regardless of feature modality or class distribution. As the amount of labeled supervision increases, the relative impact of distillation diminishes, confirming its greatest benefit in low-resource settings. These improvements are achieved without sharing raw patient data, thereby preserving institutional privacy through split learning. The cross-domain validation on Sylhet corroborates the generalisability of both frameworks and strengthens the evidence that the depth of knowledge transfer affects not only overall accuracy but also the reliability and robustness of clinical risk predictions across heterogeneous healthcare environments.

Despite these contributions, several limitations persist. The study is limited to diabetes prediction using three tabular datasets; broader validation across additional diseases and larger multi-institutional networks would enhance generalizability. Encoder alignment was implemented using a simple mean squared error objective and alternative alignment strategies may produce different trade-offs. The evaluation was conducted on two target domains, Pima and Sylhet, which, while representing distinct clinical settings and feature modalities, are both relatively small datasets; validation on larger multi-institutional networks with more diverse populations would further strengthen the generalisability claims. Additionally, while split learning enhances privacy, formal privacy guarantees such as differential privacy were not incorporated. Future research may investigate adaptive loss balancing under varying levels of supervision, more advanced representation alignment

techniques, formal privacy risk analysis, and evaluation under more severe domain shifts and class imbalance. Extending the framework to additional clinical prediction tasks would further validate its practical applicability. A promising direction for future research is the integration of retrieval-augmented generation (RAG) or retrieval-augmented fine-tuning (RAFT) mechanisms, which would allow the framework to dynamically incorporate knowledge from external clinical repositories when disease profiles outside the training distribution are encountered, thereby enabling the system to generalise to entirely new conditions without requiring full retraining. Additionally, future work may explore multi-teacher distillation architectures in which knowledge is simultaneously aggregated from multiple specialised teacher models trained on distinct source domains, potentially providing richer and more diverse supervisory signals that better capture the heterogeneity of real-world clinical populations and reduce the dependence of student performance on any single teacher's cross-domain transferability.

In summary, this work demonstrates that structured knowledge distillation enables effective cross-domain transfer in privacy-constrained healthcare environments and offers practical guidance for deploying reliable machine learning systems in heterogeneous clinical settings.

# Bibliography


[1] World Health Organization, *Diabetes*, https://www.who.int/news-room/fact-sheets/detail/diabetes, Accessed: 2024-11-14, Nov. 2024.

[2] N. Rieke et al., "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–7, Sep. 2020. DOI: 10.1038/s41746-020-00323-1. [Online]. Available: https://doi.org/10.1038/s41746-020-00323-1.

[3] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, *Split learning for health: Distributed deep learning without sharing raw patient data*, arXiv, arXiv:1812.00564, Dec. 2018. [Online]. Available: https://arxiv.org/abs/1812.00564.

[4] Y. Oda, Y. Ono, H. Nakamura, and H. Takase, "Hetero-splitee: Split learning of neural networks with early exits for heterogeneous iot devices," in *2025 IEEE 18th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, Dec. 2025, pp. 562–569. DOI: 10.1109/MCSoC67473.2025.00093. [Online]. Available: https://doi.org/10.1109/MCSoC67473.2025.00093.

[5] S. Warnat-Herresthal et al., "Swarm learning for decentralized and confidential clinical machine learning," *Nature*, vol. 594, no. 7862, pp. 265–270, Jun. 2021. DOI: 10.1038/s41586-021-03583-3. [Online]. Available: https://doi.org/10.1038/s41586-021-03583-3.

[6] H. M. Imran, M. A. A. Asad, T. A. Abdullah, S. I. Chowdhury, and M. Alamin, "Few shot learning for medical imaging: A review of categorized images," in *2023 IEEE 7th Conference on Information and Communication Technology (CICT)*, Dec. 2023, pp. 1–7. DOI: 10.1109/CICT59886.2023.10455365. [Online]. Available: https://doi.org/10.1109/CICT59886.2023.10455365.

[7] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, 2021. DOI: 10.1007/s11263-021-01453-z. [Online]. Available: https://doi.org/10.1007/s11263-021-01453-z.

[8] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, *Saint: Improved neural networks for tabular data via row attention and contrastive pre-training*, arXiv, arXiv:2106.01342 [cs, stat], Jun. 2021. [Online]. Available: https://arxiv.org/abs/2106.01342.

[9] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/9d86d83f925f2149e9edb0ac3b49229c-Paper.pdf.

[10] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, *Deep neural networks and tabular data: A survey*, arXiv, arXiv:2110.01889 [cs], Feb. 2022. [Online]. Available: https://arxiv.org/abs/2110.01889.

[11] P. Kairouz et al., *Advances and open problems in federated learning*, arXiv, arXiv:1912.04977 [cs, stat], Dec. 2019. [Online]. Available: https://arxiv.org/abs/1912.04977.

[12] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, *Scaffold: Stochastic controlled averaging for federated learning*, arXiv, arXiv:1910.06378 [cs, math, stat], Apr. 2021. [Online]. Available: https://arxiv.org/abs/1910.06378.

[13] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, *Federated optimization in heterogeneous networks*, arXiv, arXiv:1812.06127 [cs, stat], Apr. 2020. [Online]. Available: https://arxiv.org/abs/1812.06127.

[14] R. Xu, N. Baracaldo, and J. Joshi, *Privacy-preserving machine learning: Methods, challenges and directions*, arXiv, arXiv:2108.04417 [cs], Sep. 2021. [Online]. Available: https://arxiv.org/abs/2108.04417.

[15] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, *Knowledge distillation: A good teacher is patient and consistent*, arXiv, arXiv:2106.05237, Jun. 2021. [Online]. Available: https://arxiv.org/abs/2106.05237.

[16] H. Mobahi, M. Farajtabar, and P. L. Bartlett, *Self-distillation amplifies regularization in hilbert space*, arXiv, arXiv:2002.05715, 2020. [Online]. Available: https://arxiv.org/abs/2002.05715.

[17] R. Müller, S. Kornblith, and G. Hinton, *When does label smoothing help?* arXiv, arXiv:1906.02629 [cs, stat], Jun. 2020. [Online]. Available: https://arxiv.org/abs/1906.02629.

[18] L. Yuan, G. Francis, G. Li, T. Wang, and J. Feng, *Revisiting knowledge distillation via label smoothing regularization*, arXiv, arXiv:1909.11723, 2019. [Online]. Available: https://arxiv.org/abs/1909.11723.

[19] Y. Tian, D. Krishnan, and P. Isola, *Contrastive representation distillation*, arXiv, arXiv:1910.10699, 2019. [Online]. Available: https://arxiv.org/abs/1910.10699.

[20] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, *Ensemble distillation for robust model fusion in federated learning*, arXiv, arXiv:2006.07242, 2020. [Online]. Available: https://arxiv.org/abs/2006.07242.

[21] Z. Zhu, J. Hong, and J. Zhou, *Data-free knowledge distillation for heterogeneous federated learning*, arXiv, arXiv:2105.10056, 2021. [Online]. Available: https://arxiv.org/abs/2105.10056.

[22] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, *Generalizing from a few examples: A survey on few-shot learning*, arXiv, arXiv:1904.05046 [cs], Mar. 2020. [Online]. Available: https://arxiv.org/abs/1904.05046.

[23] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, *Meta-learning in neural networks: A survey*, arXiv, arXiv:2004.05439 [cs, stat], Nov. 2020. [Online]. Available: https://arxiv.org/abs/2004.05439.

[24] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, *Transfusion: Understanding transfer learning for medical imaging*, arXiv, arXiv:1902.07208 [cs, stat], Oct. 2019. [Online]. Available: https://arxiv.org/abs/1902.07208.

[25] N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha, and J. Qadir, "Privacy-preserving artificial intelligence in healthcare: Techniques and applications," *Computers in Biology and Medicine*, vol. 158, no. 1, p. 106 848, Apr. 2023. DOI: 10.1016/j.compbiomed.2023.106848. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2023.106848.

[26] Erikson, C. Traina, and M. Traina, "Security and privacy in machine learning for health systems: Strategies and challenges," *Yearbook of Medical Informatics*, vol. 32, no. 01, pp. 269–281, Aug. 2023. DOI: 10.1055/s-0043-1768731. [Online]. Available: https://doi.org/10.1055/s-0043-1768731.

[27] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, Nov. 2020. DOI: 10.1007/s41666-020-00082-4. [Online]. Available: https://doi.org/10.1007/s41666-020-00082-4.

[28] M. Li, P. Xu, J. Hu, Z. Tang, and G. Yang, "From challenges and pitfalls to recommendations and opportunities: Implementing federated learning in healthcare," *Medical Image Analysis*, vol. 101, p. 103 497, Feb. 2025. DOI: 10.1016/j.media.2025.103497. [Online]. Available: https://doi.org/10.1016/j.media.2025.103497.

[29] Y. Allouah, R. Guerraoui, N. Gupta, R. Pinot, and G. Rizk, *Robust distributed learning: Tight error bounds and breakdown point under data heterogeneity*, arXiv, arXiv:2309.13591, 2023. [Online]. Available: https://arxiv.org/abs/2309.13591.

[30] A. Choudhury et al., "Advancing privacy-preserving healthcare analytics: Implementation of the personal health train for federated deep learning (preprint)," *JMIR AI*, May 2024. DOI: 10.2196/60847. [Online]. Available: https://doi.org/10.2196/60847.

[31] S. B. Rabbani, I. V. Medri, and M. D. Samad, *Attention versus contrastive learning of tabular data – a data-centric benchmarking*, arXiv, arXiv:2401.04266, 2024. [Online]. Available: https://arxiv.org/abs/2401.04266.

[32] G. S. Hukkeri, R. H. Goudar, G. M. Dhananjaya, V. N. Rathod, and S. Ankalaki, "A comprehensive survey on split-fed learning: Methods, innovations, and future directions," *IEEE Access*, vol. 13, pp. 46 312–46 333, 2025. DOI: 10.1109/ACCESS.2025.3547641. [Online]. Available: https://doi.org/10.1109/ACCESS.2025.3547641.

[33] G. Sun et al., "Fkd-med: Privacy-aware, communication-optimized medical image segmentation via federated learning and model lightweighting through knowledge distillation," *IEEE Access*, pp. 1–1, Jan. 2024. DOI: 10.1109/ACCESS.2024.3372394. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3372394.

[34] Y. Chen, Z. Wang, H. Cai, Z. Qin, and S. Deng, "Federated knowledge distillation using hierarchical reinforcement learning in resource-constrained iot edge-cloud computing environments," *IEEE Transactions on Mobile Computing*, pp. 1–15, 2025. DOI: 10.1109/TMC.2025.3591610. [Online]. Available: https://doi.org/10.1109/tmc.2025.3591610.

[35] M. Sheng, S. Yan, Z. Sun, T. Chen, H. Liu, and Y. Yao, "Combating noisy labels in knowledge distillation for efficient edge device deployment," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 4, pp. 11 989–12 000, Nov. 2025. DOI: 10.1109/TCE.2025.3610163. [Online]. Available: https://doi.org/10.1109/tce.2025.3610163.

[36] A. Jaiswal, K. Ashutosh, J.-F. Rousseau, Y. Peng, Z. Wang, and Y. Ding, *Ros-kd: A robust stochastic knowledge distillation approach for noisy medical imaging*, arXiv, arXiv:2210.08388, 2022. [Online]. Available: https://arxiv.org/abs/2210.08388.

[37] Y. Wei, Y. Deng, C. Sun, M. Lin, H. Jiang, and Y. Peng, *Deep learning with noisy labels in medical prediction problems: A scoping review*, arXiv, arXiv:2403.13111, 2024. [Online]. Available: https://arxiv.org/abs/2403.13111.

[38] K. Fujiwara, "Knowledge distillation with resampling for imbalanced data classification: Enhancing predictive performance and explainability stability," *Results in Engineering*, vol. 24, p. 103 406, Nov. 2024. DOI: 10.1016/j.rineng.2024.103406. [Online]. Available: https://doi.org/10.1016/j.rineng.2024.103406.

[39] I. Kang, P. Ram, Y. Zhou, H. Samulowitz, and O. Seneviratne, *On learning representations for tabular data distillation*, arXiv, arXiv:2501.13905, 2025. [Online]. Available: https://arxiv.org/abs/2501.13905.

[40] K. Riasat, A. Jamil, S. Al-Otaibi, S. Zeb, S. Riasat, and S. Kanwal, "Tuberculosis detection using few shot learning," *Scientific Reports*, vol. 15, no. 1, Apr. 2025. DOI: 10.1038/s41598-025-97803-9. [Online]. Available: https://doi.org/10.1038/s41598-025-97803-9.

# Appendix A

## Implementation Details and Reproducibility

### Software and Hardware Environment

All experiments were implemented in Python 3.10 using the following primary libraries:

1. PyTorch (for deep learning implementation)
2. Scikit-learn (for preprocessing, metrics, temperature scaling)
3. NumPy and Pandas (for data manipulation)
4. Matplotlib (for visualization)

Random seeds were controlled for:

1. NumPy
2. PyTorch CPU
3. PyTorch CUDA

## Data Preprocessing Hyperparameters

### Continuous Feature Processing

| Step | Method |
|---|---|
| Missing value imputation | Median (fit on training split only) |
| Normalization | StandardScaler (fit on training split only) |
| Missingness indicator | Binary mask appended to continuous inputs |

Preprocessing steps and corresponding methods.

### Categorical Encoding

All preprocessing objects were fit exclusively on training splits to prevent data leakage.

| Step | Method |
|---|---|
| Encoding | Integer index mapping |
| Special tokens | `__MISSING__`, `__UNK__` |
| Embedding type | Learnable embedding layer |

Categorical feature processing steps and methods.

# Model Architecture Details

## Teacher Model (FT-Transformer)

| Component | Configuration |
|---|---|
| Embedding dimension | 128 |
| Number of attention heads | 4 |
| Transformer layers | 3 |
| Dropout | 0.1 |
| Activation | ReLU |
| Output layer | Single neuron (sigmoid activation) |

Model configuration details.

The teacher was trained under split learning:

- Client-side encoder per institution
- Shared server-side backbone
- Dataset-specific prediction heads

### Student Models

| Component | Configuration |
|---|---|
| Hidden layers | 2–3 |
| Hidden dimension | 128 |
| Activation | ReLU |
| Dropout | 0.2 |

MLP configuration.

| Component | Configuration |
|---|---|
| Embedding dimension | 128 |
| Attention heads | 4 |
| Transformer layers | 2 |
| Dropout | 0.1 |

SAINT configuration.

| Component | Configuration |
|---|---|
| Ensemble size | Small BatchEnsemble variant |
| Hidden dimension | 128 |
| Dropout | 0.1 |

TabM configuration.

For Framework 2, student encoder output dimensionality was matched to teacher encoder dimensionality for representation alignment.

# Training Configuration

## Optimization

| Parameter | Value |
|---|---|
| Optimizer | Adam |
| Learning rate | $1 \times 10^{-3}$ |
| Weight decay | $1 \times 10^{-5}$ |
| Batch size | 32 |
| Epochs | Up to 100 |
| Early stopping patience | 10 epochs |

Training hyperparameters.

Early stopping was based on validation ROC-AUC.

# Distillation Settings

## Logit-Level Distillation (Framework 1)

## Encoder-Level Distillation (Framework 2)

Both losses were optimized jointly throughout training.

| Parameter | Value |
|---|---|
| Supervised weight ($\alpha$) | 0.5 |
| Distillation weight ($1 - \alpha$) | 0.5 |
| Loss function | BCE on soft probabilities |
| Temperature scaling | Applied post-training for calibration evaluation |

Logit-level distillation configuration (Framework 1).

| Parameter | Value |
|---|---|
| $\lambda_{\text{sup}}$ | 1.0 |
| $\lambda_{\text{align}}$ | 0.5 |
| Alignment loss | Mean Squared Error (MSE) |
| Teacher weights | Frozen during student training |

Encoder-level distillation configuration (Framework 2).

# Few-Shot Sampling Protocol

For each random seed:

- Balanced $K$-shot subset selected from the Pima training set
- $K \in \{5, 10, 20\}$ samples per class
- Remaining training samples treated as unlabeled
- Sampling repeated independently across 5 seeds

Reported results are mean $\pm$ standard deviation across seeds.

# Evaluation Protocol

Evaluation is performed on the held-out Pima test set using:

- ROC-AUC
- PR-AUC
- Brier score
- Temperature-calibrated Brier score
- Accuracy, Precision, Recall, F1-score
- Confusion matrix at threshold 0.5

Temperature scaling is fitted exclusively on validation data.