MedFusion-LM: Explainable Large Language Model for Transforming Medical Outcomes in Federated Learning with Neural Architecture Search Blueprints

Diogen Babuc[0009−0000−5126−6480] (✉) and Teodor-Florin Fortiş[0000−0002−3143−8908]

West University of Timişoara,
blvd. V. Pârvan 4, 300223, Timişoara, Romania
{diogen.babuc,florin.fortis}@e-uvt.ro

Abstract. Federated learning (FL) has emerged as a promising approach to facilitate collaborative model training while ensuring data privacy. FL faces challenges related to heterogeneity in data distributions, interpretability of model decisions, and optimization of model architectures across decentralized nodes. This paper proposes a framework that combines FL with neural architecture search (NAS) and explainable large language model (XLLM) to overcome these issues and improve clinical outcomes. We test this approach in three medical areas. NAS is used to discover optimized model architectures tailored to heterogeneous medical data across decentralized hospitals. XLLMs are employed to interpret and communicate complex decision-making processes. Experimental validation on benchmark datasets for each clinical use case indicates improvements in predictive accuracy and clinical relevance compared to conventional federated approaches.

Keywords: privacy-preserving, neural architecture search, explainable AI, large language models, clinical decision support.

1 Introduction

The adoption of artificial intelligence (AI) in healthcare has opened new frontiers in clinical diagnostics, personalized treatment planning, and interpretation of medical images [13]. Among the many paradigms of AI, federated learning (FL) has emerged as a privacy-preserving strategy that enables collaborative model training across decentralized data silos without requiring the exchange of sensitive patient information. This is important in medical environments constrained by data protection regulations such as HIPAA [4] and GDPR [32].

Despite its promise, federated learning faces several challenges [21]. First, data heterogeneity – resulting from variations in imaging protocols, populations, and devices – can impair global model convergence and reduce generalizability. Next, the black-box nature of the deep models inhibits interpretability, which is

a key requirement in clinical settings where human oversight and explainability are critical. Finally, existing FL workflows typically use static architectures that may not adapt well to diverse institutional data distributions [34].

To overcome these limitations, our proposal is to combine FL with two core enhancements: *Neural Architecture Search (NAS)* [29] and *Explainable Large Language Model (XLLM)*. NAS allows each institution to discover optimized local architectures while contributing to a federated learning process. XLLM approaches enhance model transparency by providing contextual, human-readable explanations of decisions, bridging the gap between the black-box nature of AI and clinical reasoning.

Our approach is evaluated in three high-impact medical use cases: (1) early detection of premalignant colorectal polyps [10], (2) diagnostics for cervical cancer [5], and (3) cognitive assessment for Alzheimer's disease via dementia stage classification [8]. Through a combination of zero-shot NAS [18], knowledge distillation [9] and federated explainability [2], we show measurable improvements in predictive accuracy, interpretability, and practical deployment readiness.

This paper intends to address the following objectives: (1) the development of a privacy-preserving framework for training clinical models on decentralized, heterogeneous medical datasets; (2) improving model performance and adaptability using zero-shot NAS to enable automated, local architectural optimization; (3) integration of XLLM principles to enhance interpretability and provide clinically meaningful explanations for AI decisions.

2 Background Information

The increasing adoption of artificial intelligence (AI) in healthcare requires systems that are accurate, yet interpretable, adaptable, and privacy-preserving. In order to achieve these goals, several technologies have to be employed, such as federated learning, neural architecture search, explainable large language models, and knowledge distillation. Together, they can define a framework for trustworthy clinical AI, as depicted in Fig. 1.

2.1 Federated Learning

Federated Learning, as introduced by Mahan et al. in [23], enables a decentralized model training across multiple devices while preserving privacy. As mentioned in [34], this is a decentralized machine learning paradigm that enables multiple distributed clients (e.g., hospitals, clinics, or diagnostic centers) to collaboratively train a shared global model without exchanging raw data. In a typical FL setup, each client downloads the current international model, performs local training using its private dataset, and then uploads model updates (gradients or weights) to a central server. The server aggregates these updates to refine the global model, often using strategies such as the federated averaging (FedAvg). This cycle repeats for several rounds until convergence is reached.

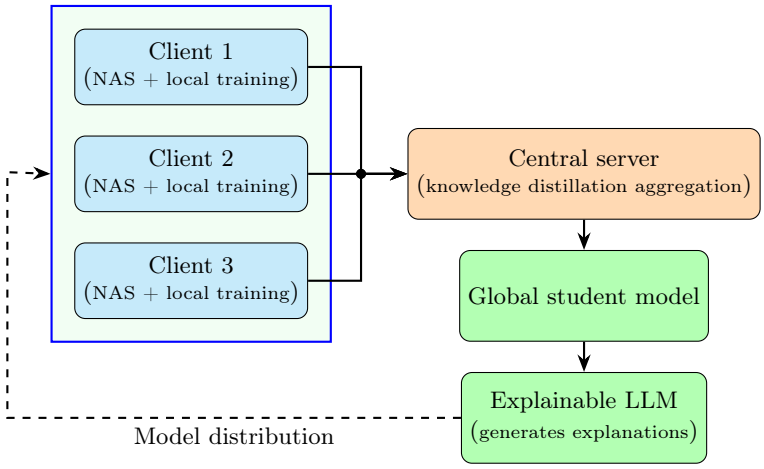

Fig. 1: Federated learning framework with NAS, knowledge distillation, and explainable large language models.

The primary motivation behind FL is data privacy. In sensitive healthcare domains, patient data is protected by stringent regulations, and ethical considerations often prohibit centralized data sharing. FL addresses this by ensuring that patient data remain localized to the institution where it was generated, while still contributing to a broader training objective. This capability is particularly beneficial when the pooling of diverse datasets can significantly enhance model generalization, but it is often infeasible due to legal, institutional or technical constraints.

Beyond privacy, federated learning also brings practical advantages benefits in terms of scalability and data utilization. Institutions with limited data sets or rare diseases cases can participate in global training and benefit from the collective model improvements, which is usually not possible in traditional centralized learning paradigms. Moreover, FL can easily adapt to edge computing environments, where models are trained directly on medical imaging devices or wearable sensors, and updates are shared asynchronously [26].

However, the advantages of FL are counterbalanced by several key challenges, such as statistical heterogeneity. Data distributions often vary across clients due to differences in patient demographics, imaging devices, annotation practices, or diagnostic criteria. This non-independent and identically distributed (non-IID) nature of client data can degrade convergence and cause the global model to underperform on outlier distributions. Another major challenge is the heterogeneity of the system, where clients have varying computational resources, bandwidth, and availability, leading to unequal participation and possible biases in model training.

Standard FL setups typically assume a fixed architecture model shared across all clients. While this simplifies the aggregation process, it limits the systemțs

ability to adapt to local data characteristics, especially when clients differ in data modalities or complexity. This rigidity can lead to underfitting on certain clients and overfitting on others. Additionally, model interpretability remains a major concern. Since FL is typically based on deep neural networks, the resulting models often operate as black boxes, making it difficult for clinicians to understand or trust their decisions, especially in high-stakes environments for the diagnosis of cancer or neurodegenerative diseases [7].

2.2 Neural Architecture Search

Neural architecture search (NAS) is a subfield of automated machine learning (AutoML) focused on the automated discovery of optimal neural network architectures for a given task and dataset [3]. Traditionally, neural network design has relied heavily on expert intuition and manual experimentation, which can be time-consuming, suboptimal, and infeasible in highly variable environments such as decentralized healthcare systems. NAS addresses this limitation by automating the architecture design, enabling more efficient and often higher-performing model development.

A typical NAS framework consists of three key components: a) a search space – the set of all possible architectures, b) a search strategy – reinforcement learning, evolutionary algorithms, or gradient-based methods, c) and a performance estimation strategy – early stopping or proxy tasks. The goal is to explore this space efficiently and identify architectures that optimize a target objective, such as classification accuracy, latency, or model size [33].

In the context of federated learning, NAS offers a particularly compelling solution to statistical heterogeneity, one of the key challenges of FL. When data distributions vary across clients – as is common in healthcare due to different populations, imaging modalities, and institutional protocols – a single shared model architecture may not perform uniformly well. NAS allows each client to tailor its model architecture to local data characteristics, thereby improving representation capacity and model fit. For example, a smaller convolutional architecture might suffice for one hospital with simple, high-quality data, while another (with more complex cases) may benefit from a deeper or hybrid model.

Integrating NAS into a federated framework is not an easy task. One of the primary challenges is the inconsistency of architectures across clients. Traditional FL methods, such as FedAvg [17], require all clients to share an identical architecture to aggregate weights. With NAS, clients may end up with structurally different models. This makes direct parameter aggregation infeasible. To address this, we adopt a knowledge distillation-based aggregation strategy. Instead of sharing model weights, each client shares soft predictions (logits) on a shared public or synthetic dataset. A centralized student model then distills these predictions into a unified architecture, effectively learning a consensus model that captures the knowledge from heterogeneous client models [6].

To further enhance scalability, we use zero-shot NAS, a lightweight and efficient variant that evaluates candidate architectures without full training. Zero-shot NAS drastically reduces computational overhead. This makes it feasible

to deploy across multiple FL clients with limited hardware resources. As a result, the framework is suitable for both academic-scale datasets and real-world deployments in distributed hospital networks.

Moreover, our use of NAS is closely coupled with clinical objectives. Rather than optimizing solely for accuracy, we incorporate multiple constraints and evaluation criteria, including inference latency (for specific settings), parameter count (for embedded devices), and interpretability (through attention mechanisms and hierarchical layers).

2.3 Explainable Large Language Models

Large language model (LLM) approaches have transformed natural language processing (NLP) [22,30], demonstrating strong performance on complex tasks such as question answering, summarization, and contextual reasoning, often in zero- or few-shot settings. In the medical domain, specialized LLMs like BioBERT [14], BioGPT [20], and Med-PaLM [27] have shown great promise in understanding and generating clinically relevant text. However, their opaque decision-making process limit their practical adoption in safety-critical environments such as healthcare.

To address this, recent advances in explainable artificial intelligence (XAI) have been extended to LLMs, resulting in what we refer to as explainable large language models [2]. XLLMs integrate mechanisms for generating interpretable outputs alongside predictions. These can take the form of attention heatmaps, token-level rationales, natural language justifications, or evidence-based answer highlights. In contrast to black-box output, XLLMs offer the transparency that is essential for clinical trust and regulatory acceptance.

In our approach, XLLMs serve as interpretable interfaces between model outputs and human users, particularly clinical staff. For example, given a diagnostic input, such as pathology image labels or patient notes, the XLLM can provide also an explanation. This supports human-in-the-loop workflows, where clinicians can validate or challenge model decisions based on the provided justifications.

Moreover, XLLMs enhance model transparency in a federated setting, where understanding what models have learned across diverse institutions is challenging. Since client models may be architecturally and data-wise diverse (especially under NAS), generating consistent, interpretable outputs via XLLMs creates a shared language for explaining predictions. This not only increases user trust but also facilitates collaborative validation and feedback cycles between institutions.

Our proposed model, MedFusion-LM, incorporates explainability natively within the language modeling component, enabling both predictive reasoning and explanation generation as parallel outputs. This design aligns with the broader vision of human-centric AI and makes the framework more suitable for deployment in real-world clinical environments.

2.4 Knowledge Distillation

Knowledge distillation (KD) [24] is a model compression and transfer learning technique in which a smaller, compact student model learns to imitate the behavior of a larger, more complex teacher model or ensemble. Originally proposed for resource-efficient deployment, KD has since evolved into a powerful tool for aggregating knowledge from heterogeneous sources. This makes it especially suitable for federated learning configurations.

Traditional FL approaches rely on weight aggregation, assuming that all clients use the same architecture. However, this assumption becomes limiting when combined with NAS, where each client can learn an optimal but unique architecture tailored to its data. To bridge this architectural heterogeneity, we use a KD-based aggregation approach. Rather than sharing model parameters, each client generates logits or probabilities on a dataset [16]. These outputs are then aggregated by a central server to train a unified student model, which absorbs the collective knowledge of the diverse local models.

This strategy is inspired by frameworks such as FedMD [15] (heterogeneous federated learning through model distillation) and FedDF [19] (ensemble distillation for robust model fusion in federated learning), which demonstrate that distillation-based FL can outperform weight averaging. For current implementation, the student model becomes the shared global model that is redistributed to clients in the next training round. It is important to mention that knowledge distillation also serves as a privacy-preserving mechanism. Since only logits are shared, not raw data or gradients, the risk of data leakage is minimized.

3 The Proposed Framework

Our proposed framework integrates XLLMs within a federating learning (FL) system enhanced by zero-shot NAS (ZS-NAS) and logit-based knowledge distillation. The core innovation lies in enabling LLMs to operate as interpretable, adaptive, and privacy-preserving diagnostic agents across distributed clinical environments. Each component of the framework empowers LLMs with personalization, explainability, and scalability in the federated setting.

The medical AI pipeline that fuses large models and multimodal learning (MedFusion-LM) (Fig. 2) begins with decentralized clinical data (EHRs, notes, scans) at client institutions, ensuring privacy preservation. Each client performs a zero-shot neural architecture search to adaptively select lightweight LLM adapters suitable for local data characteristics. These models are fine-tuned and produce soft predictions and rationales which are centrally distilled into a unified, interpretable student model. The resulting global student XLLM is redistributed to clients, supporting continuous, explainable, and privacy-compliant learning rounds.

3.1 Zero-Shot Architecture Search for LLM Adapters

To accommodate diverse local data distributions and clinical formats, each FL client performs a lightweight, training-free ZS-NAS procedure to identify optimal

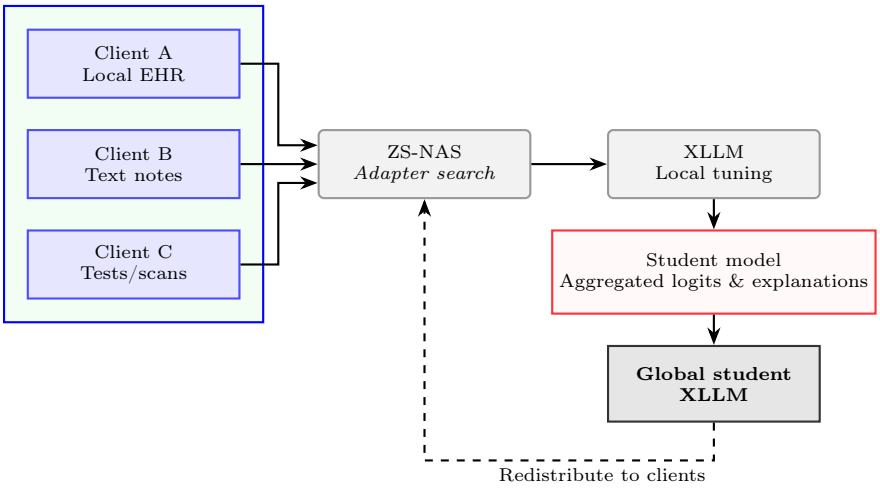

Fig. 2: Layout of the **MedFusion-LM** framework. Clients send data to a shared adapter selection and tuning pipeline. Aggregated outputs form a unified student LLM redistributed for future rounds.

architectural components, such as attention-enhanced LLM adapters, convolutional front-ends, or domain-specific encoders, that plug into a shared pre-trained LLM backbone.

ZS-NAS allows clients to evaluate architectural variants without full training, by using low-cost performance estimators such as gradient norms, parameter entropy, or Jacobian sensitivity. This ensures that each client discovers an architecture that maximizes the utility of the LLM for its specific clinical modality (colonoscopy analysis and results, cytology sequences, and cognitive assessment).

These LLM-tuned adapters are then used to locally fine-tune or prompt the shared LLM on private data, while preserving the base model's alignment. Importantly, this setup enables model flexibility at the edge while maintaining a common language-based representation space across all clients (see Fig. 1).

$$\mathcal{S}_{\text{ZS-NAS}}(a) = \lambda_1 \cdot \text{Entropy}(a) + \lambda_2 \cdot \|\nabla \mathcal{L}(a)\| + \lambda_3 \cdot \text{Jacobian}(a) \tag{1}$$

In Eq. (1), $\mathcal{S}_{\text{ZS-NAS}}(a)$ represents the zero-shot architecture score for candidate a, computed without full model training. The terms $\text{Entropy}(a)$, $\|\nabla \mathcal{L}(a)\|$, and $\text{Jacobian}(a)$ quantify architectural uncertainty, gradient strength, and sensitivity to input perturbations, respectively. The coefficients λ_1, λ_2, and λ_3 are tunable weights that balance the contribution of each proxy metric to the final score.

3.2 Logit-Based Knowledge Distillation for LLM Consensus

Because architectures vary across clients, traditional weight aggregation is infeasible. Instead, our system employs logit-based knowledge distillation [19] to unify the distributed reasoning of multiple XLLMs.

Each client generates soft label predictions and textual rationales from their LLM on a shared public or synthetic dataset. These logits are aggregated on a central server to train a student model that mimics the collective outputs (see Fig. 2). The student is designed to retain both the classification accuracy and the explanation fidelity by learning the predicted class probabilities and the associated explanation embeddings from the XLLMs.

This process allows the federated system to: (a) preserve architectural heterogeneity while maintaining semantic alignment; (b) distill diverse clinical reasoning into a unified, interpretable LLM-based student model; (c) enable explanation-level aggregation, where the quality of generated rationales improves via ensemble distillation.

$$\mathcal{L}_{\text{KD}} = \sum_{i=1}^{N} \text{KL}\left(\mathbf{z}_i^{\text{client}} \parallel \mathbf{z}_i^{\text{student}}\right) \tag{2}$$

In Eq. (2), $\mathcal{L}_{\text{KD}}$ denotes the knowledge distillation loss used to train the student model by imitating the soft outputs of the client models. Here, $\mathbf{z}_i^{\text{client}}$ and $\mathbf{z}_i^{\text{student}}$ are the logit vectors (soft predictions) from the i-th client model and the central student model, respectively. The KL divergence $\text{KL}(\cdot\|\cdot)$ measures the difference between these distributions over N shared input samples, encouraging the student to align its predictions with those of the clients.

3.3 Federated Explainability via LLM Rationales

Unlike traditional federated models that produce raw scores, our framework is designed for human-AI cooperation. The distilled student model is capable of generating clinically grounded rationales for each prediction in natural language. These explanations can be reviewed, validated, or contested by human experts, enhancing transparency, trust, and safety.

Furthermore, because LLM outputs can include references to clinical evidence, observed features, and diagnostic criteria, the framework supports accurate predictions and also aligned justifications across federated institutions.

The training cycle proceeds as follows: (i) clients perform ZS-NAS to select lightweight, explainability-enhancing components for their local XLLM adapter; (ii) each client fine-tunes the LLM locally and shares soft logits and textual explanations on public prompts; (iii) a centralized student model learns to replicate both decision and explanation patterns; (iv) the distilled global XLLM is redistributed to clients, where the process repeats iteratively (see Fig. 2).

This design allows our system to scale across diverse healthcare environments, delivering LLM-powered, transparent, and adaptive AI models that adhere to both data privacy and clinical accountability.

Explainable Federated Inference. Traditional federated learning systems primarily output numerical predictions, offering limited transparency into how those predictions are derived. In contrast, our framework enables *explainable federated inference* [25], in which each model prediction is accompanied by a natural language rationale, attention visualizations or feature highlights generated by the XLLM. These explanations are embedded into the model output pipeline and are derived during both training and inference (see Fig. 1).

This approach ensures that predictions are not only accurate but also interpretable, fulfilling a critical requirement in clinical environments where decisions must be auditable and justifiable. By incorporating explanation tokens into the training objective, we encourage the LLM to co-learn the diagnostic decision and its reasoning trace, resulting in outputs that are both performance-optimized and human-verifiable.

Furthermore, the explainable output can be aggregated during distillation. The student model not only learns the soft labels from the teacher models, but also aligns with their generated rationales. This multiobjective learning helps enforce consistency in explanation quality across clients, even when they differ in data domains and architecture design.

$$\mathcal{L}_{\text{expl}} = \sum_{i=1}^{N} \|\mathbf{r}_i^{\text{client}} - \mathbf{r}_i^{\text{student}}\|_2^2 \tag{3}$$

In Eq. 3, $\mathcal{L}_{\text{expl}}$ denotes the alignment loss of the explanation, which ensures that the student model replicates the explanatory rationales of the client models. The vectors $\mathbf{r}_i^{\text{client}}$ and $\mathbf{r}_i^{\text{student}}$ represent the explanation embeddings (e.g., attention weights or textual rationale features) produced by the i-th client and the student model, respectively. The loss is computed as the squared Euclidean distance (L2) over N samples, encouraging the student to learn both the decisions and their interpretive justifications.

Client-Aware Model Fusion Conventional federated systems aggregate client models, assuming uniformity in architecture and training conditions. This assumption breaks down in real-world clinical settings, where datasets, modalities, and patient populations vary widely. To overcome this, we propose a *client-aware model fusion* [12] strategy that leverages both knowledge distillation and metalearning principles to synthesize a global model from heterogeneous clients.

Rather than treating all client contributions equally, our framework evaluates the reliability of the client model based on factors such as performance confidence, consistency of explanations, and domain alignment. These factors are encoded into the distillation process by weighting the logits of each client and the rationale embeddings when training the student model. As a result, the student model is not a naive average but a contextually fused global model that reflects client diversity while maintaining strong generalization.

This strategy ensures robustness to noisy or biased clients and allows the global model to prioritize high-fidelity contributions, especially in low-data or

high-variability domains. It also lays the groundwork for future adaptive weighting schemes that incorporate human feedback, model uncertainty, or explanation quality as part of the fusion logic.

Together, explainable federated inference and client-aware model fusion elevate the capabilities of federated LLM systems, enabling them to deliver high predictive accuracy and meaningful, aligned, and trustworthy decision support across institutional boundaries.

$$\alpha_i = \frac{\exp(\gamma_1 c_i + \gamma_2 e_i + \gamma_3 d_i)}{\sum_{j=1}^{N} \exp(\gamma_1 c_j + \gamma_2 e_j + \gamma_3 d_j)} \tag{4}$$

In Eq. (4), α_i represents the normalized weight assigned to the client i during the fusion of the client-aware model. The terms c_i, e_i, and d_i correspond to the client's confidence score, explanation consistency, and domain alignment, respectively, while γ_1, γ_2, and γ_3 are scaling factors that control the influence of each metric. The softmax formulation ensures that the weights α_i sum to 1 across all N clients, allowing the global model to prioritize high-quality and trustworthy client contributions during aggregation.

4 Results and Discussion

This section presents the empirical results of our proposed framework in three medical tasks: detection of colorectal polyps' premalignancy (using the PolyDB dataset [11]), cervical cancer prediagnosis (based on the SIPakMeD dataset [28]), and assessment of Alzheimer's disease [31]. We evaluated and compared four state-of-the-art language models – BioBERT, BioGPT, Med-PaLM, and the proposed MedFusion-LM – across multiple dimensions, including the effectiveness of knowledge distillation in heterogeneous environments.

Additionally, we benchmark the performance of representative models from diverse architectural families [1], including CNNs, vision transformers (ViT), autoencoders, and hybrid models that combine CNNs with shallow classifiers, such as random forests. This holistic comparison enables us to assess raw accuracy and model adaptability, explainability, and cross-domain transferability in federated clinical settings.

As illustrated in Fig. 3, MedFusion-LM outperforms all baseline models – BioBERT, BioGPT, and Med-PaLM – across the three evaluated clinical tasks. Although Med-PaLM demonstrates strong performance in the Alzheimer's domain, MedFusion-LM exceeds it by a margin of over 4 percentage points, achieving the highest precision (96.61%). The gains are even more pronounced in the colorectal and cervical cancer tasks, where MedFusion-LM attains accuracies of 92.92% and 96.04%, respectively, suggesting superior adaptability and generalization. These improvements validate the effectiveness of integrating zero-shot NAS, explainable LLMs, and logit-based knowledge distillation in federated clinical settings.

Fig. 4 illustrates the mean accuracy of four models across 10-, 15-, and 30-fold cross-validation settings on the colorectal polyps dataset. MedFusion-LM

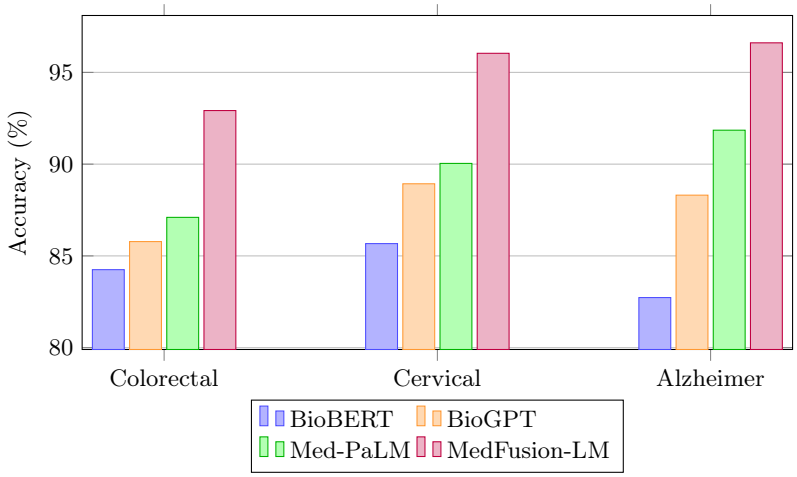

Fig. 3: Accuracy of four large language models evaluated on three clinical tasks. MedFusion-LM consistently outperforms others across all use cases.

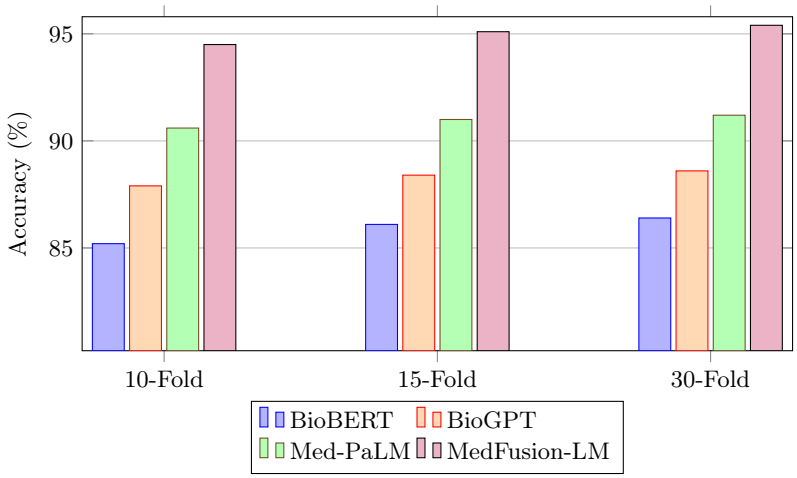

Fig. 4: Mean accuracy for each model across different k-fold cross-validation settings (colorectal polyps samples).

consistently achieves the highest accuracy in all settings, surpassing the second-best model, Med-PaLM, by approximately 4 percentage points on average. In particular, MedFusion-LM demonstrates both high performance and stability, with minimal variance as the fold count increases, indicating robustness to partitioning strategies. These results highlight the framework's strong generalization capability, driven by federated NAS tuning and explanation-guided knowledge distillation.

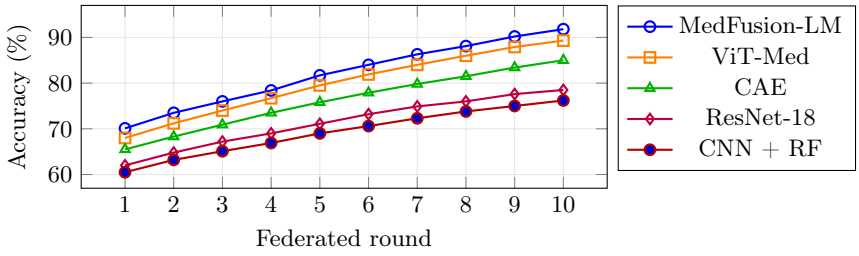

Fig. 5: Average accuracy progression of five models' quintessence (families) over 10 federated training rounds on the Alzheimer's dataset. LLMs show steeper and higher convergence in the Flower federated learning environment.

Fig. 5 illustrates the average accuracy progression of five model families over 10 federated training rounds on the cervical cancer datasets. MedFusion-LM (LLM-based) demonstrates the steepest and highest convergence, achieving over 91% accuracy by round 10, significantly outperforming ViT-Med, the convolutional autoencoder (CAE) model, and CNN-based (ResNet-18 and CNN + RF) baselines. The shallow hybrid model (CNN + RF) shows the slowest improvement and the lowest final accuracy, underscoring its limitations in the capture of complex medical features. These results emphasize the superior representational capacity and learning efficiency of language model-driven architectures in federated clinical environments.

Table 1: Simulated transfer performance across medical domains.

Source	**Target**	**Model**	**Target acc.**	**Fine-tuning drop**
Colorectal	Cervical	MedFusion-LM	86.7%	5.1%
		ViT-Med	85.6%	6.2%
		ResNet-18	84.4%	7.4%
		CAE	83.2%	8.6%
Cervical	Alzheimer	MedFusion-LM	85.3%	6.6%
		ViT-Med	84.0%	7.9%
		ResNet-18	83.1%	8.8%
		CAE	82.3%	9.6%
Alzheimer	Colorectal	MedFusion-LM	88.1%	6.9%
		ViT-Med	86.4%	8.6%
		ResNet-18	85.3%	9.7%
		CAE	84.5%	10.5%

Table 1 summarizes the simulated cross-domain transferability of four representative models across colorectal, cervical, and Alzheimer datasets. MedFusion-LM consistently achieves the highest target accuracy and the lowest fine-tuning

Table 2: Scientific comparison of model families in federated clinical AI systems.

Aspect	Metric	LLM	ViT	AE	CNN
Interpretability	Qualitative rating	★★★★★	★★★★☆	★★☆☆☆	★★★☆☆
Transferability	Cross-domain acc.	★★★★★	★★★★☆	★★★☆☆	★★☆☆☆
Training efficiency	Resource usage	★★☆☆☆	★★☆☆☆	★★★☆☆	★★★★☆
FL compatibility	Arch. support	★★★★★	★★★☆☆	★★★☆☆	★★☆☆☆
Multimodal capability	Input flexibility	★★★★☆	★★★☆☆	★★☆☆☆	★☆☆☆☆
Robustness to noise	Tolerance score	★★★☆☆	★★★☆☆	★★☆☆☆	★★☆☆☆
Explanation alignment	Alignment support	Yes	Yes	No	No

drop in all source-target combinations, demonstrating its strong generalization and adaptation capabilities. Compared to CNNs, autoencoder (AE) approaches, and ViT-based models, it exhibits greater resilience to domain shifts, which is critical for real-world deployments where disease distributions vary.

The star ratings reported in Table 2 reflect a structured qualitative assessment derived from empirical results, architectural analysis, and published benchmarks. Each of the analyzed aspects, like interpretability, transferability, and robustness, was scored on a scale of one to five stars based on diverse criteria: rationales in LLMs, attention heatmaps in ViTs, and post hoc visualizations in CNNs; inferred transferability from cross-domain performance drops (Table 1); convergence speed and computational overhead during training; architectures that can flexibly support adapter tuning or knowledge distillation; ingestion of diverse input modalities (images, text, structured data); data variability and label noise; explanation-level alignment across clients.

These results support the robustness of our LLM-driven framework in heterogeneous clinical environments and highlight its suitability for transfer learning under federated constraints.

5 Generalizable Insights about Responsible Application of Machine Learning in Healthcare

The development and deployment of machine learning systems in healthcare require a balance between predictive performance, ethical safeguards, and clinical usability. Our work with MedFusion-LM illustrates several principles that extend beyond the specific medical domains studied.

Federated learning, by design, enables training on data from multiple institutions without centralizing sensitive records. This diversity can reduce the risk of demographic or institutional bias, but fairness audits and bias-aware evaluation protocols must complement it to address disparities in model performance. Preserving patient confidentiality is paramount under regulations such as GDPR. FL inherently mitigates privacy risks by keeping data local, and the use of logit-based aggregation further limits leakage. Nevertheless, integrating secure aggregation and differential privacy can provide stronger guarantees against

adversarial inference. Clinical adoption depends on transparent reasoning. The integration of XLLMs into the predictive pipeline enables models to produce both decisions and interpretable rationales. This principle, aligning outputs with human-readable justifications, applies to a wide range of diagnostic and prognostic systems.

In healthcare, models must remain reliable across heterogeneous settings, modalities, and patient populations. Techniques such as neural architecture search for local adaptation and knowledge distillation for consensus learning can enhance robustness, ensuring stable performance despite domain shifts or noisy inputs. Deployment should be guided by multidisciplinary oversight, incorporating clinical expertise, ethical review, and compliance with evolving AI governance standards. Embedding feedback loops and human-in-the-loop mechanisms fosters accountability and supports continuous improvement.

By embedding these principles into the design and lifecycle of ML systems, healthcare applications can achieve technical excellence and sustainable impact.

6 Conclusions

This paper introduced MedFusion-LM, an integrated framework that unifies federated learning, neural architecture search, and explainable large language models to address core challenges in clinical AI, namely data heterogeneity, architectural rigidity, and decision opacity. By leveraging zero-shot NAS, client-side architectures were dynamically tailored to local data distributions without the need for exhaustive training. Simultaneously, logit-based knowledge distillation enabled the fusion of diverse client models into a cohesive global student model without requiring architectural uniformity.

Experimental results across three medical domains: colorectal polyps, cervical cancer, and Alzheimer's disease, demonstrated that MedFusion-LM outperforms state-of-the-art baselines such as BioBERT, BioGPT, and Med-PaLM in both predictive accuracy and cross-validation robustness. The integration of LLM-generated rationales supports explainable federated inference, a critical feature for clinical adoption and trust.

Our findings suggest that combining adaptive architecture search with explainable, language-based representations paves the way toward a new class of federated AI systems—those that are not only privacy-preserving and performant, but also interpretable, resilient, and clinically aligned. Future work will explore real-time deployment in hospital environments, extension to multimodal data, and clinician-in-the-loop feedback mechanisms to further close the gap between AI prediction and medical reasoning.

Acknowledgments. The research conducted in this paper was partially supported by project *Romanian Hub for Artificial Intelligence*, Smart Growth, Digitization, and Financial Instruments Program, 2021-2027, MySMIS no. 334906 and the UVT 1000 Develop Fund of the West University of Timișoara.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bootun, D., Auzine, M.M., Ayesha, N., Idris, S., Saba, T., Khan, M.H.M.: Adamaex—alzheimer's disease classification via attention-enhanced autoencoders and xai. Egyptian Informatics Journal **30**, 100688 (2025). https://doi.org/10.1016/j.eij.2025.100688
2. Chaddad, A., Lu, Q., Li, J., Katib, Y., Kateb, R., Tanougast, C., Bouridane, A., Abdulkadir, A.: Explainable, domain-adaptive, and federated artificial intelligence in medicine. IEEE/CAA Journal of Automatica Sinica **10**(4), 859–876 (Apr 2023). https://doi.org/10.1109/jas.2023.123123
3. Chaiyarin, S., Rojbundit, N., Piyabenjarad, P., Limpitigranon, P., Wisitthipakdeekul, S., Nonthasaen, P., Achararit, P.: Neural architecture search for medicine: A survey. Informatics in Medicine Unlocked **50**, 101565 (2024). https://doi.org/10.1016/j.imu.2024.101565
4. Cohen, I.G., Mello, M.M.: HIPAA and protecting health information in the 21st century. Jama **320**(3), 231–232 (Jul 2018). https://doi.org/10.1001/jama.2018.5630
5. Cohen, P.A., Jhingran, A., Oaknin, A., Denny, L.: Cervical cancer. The Lancet **393**(10167), 169–182 (2019). https://doi.org/10.1016/s0140-6736(18)32470-x
6. Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: FedAvg with fine tuning: local updates lead to representation learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022). https://doi.org/10.5555/3600270.3601038
7. Crowson, M.G., Moukheiber, D., Arévalo, A.R., Lam, B.D., Mantena, S., Rana, A., Goss, D., Bates, D.W., Celi, L.A.: A systematic review of federated learning applications for biomedical data. PLOS Digital Health **1**(5), e0000033 (May 2022). https://doi.org/10.1371/journal.pdig.0000033
8. DeTure, M.A., Dickson, D.W.: The neuropathological diagnosis of Alzheimer's disease. Molecular Neurodegeneration **14**(1), 32 (Aug 2019). https://doi.org/10.1186/s13024-019-0333-5
9. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (Mar 2021). https://doi.org/10.1007/s11263-021-01453-z
10. Granados-Romero, J.J., Valderrama-Treviño, A.I., Contreras-Flores, E.H., Barrera-Mera, B., Herrera Enríquez, M., Uriarte-Ruíz, K., Ceballos-Villalba, J.C., Estrada-Mata, A.G., Alvarado Rodríguez, C., Arauz-Peña, G.: Colorectal cancer: a review. International Journal of Research in Medical Sciences **5**(11), 4667 (Oct 2017). https://doi.org/10.18203/2320-6012.ijrms20174914
11. Jha, D., Tomar, N.K., Sharma, V., Trinh, Q., Biswas, K., Pan, H., Jha, R.K., Durak, G., Hann, A., Varkey, J., Dao, H.V., Dao, L.V., Nguyen, B.P., Pham, K.C., Tran, Q.T., Papachrysos, N., Rieders, B., Schmidt, P.T., Geissler, E., Berzin, T.M., Halvorsen, P., Riegler, M.A., de Lange, T., Bagci, U.: PolypDB: A curated multi-center dataset for development of AI algorithms in colonoscopy. ArXiv **abs/2409.00045** (2024). https://doi.org/10.48550/ARXIV.2409.00045

12. Jin, R., Tong, B., Yang, S., Hou, B., Shen, L.: ICAFS: Inter-client-aware feature selection for vertical federated learning (2025). https://doi.org/10.48550/ARXIV.2504.10851
13. Krishnan, G., Singh, S., Pathania, M., Gosavi, S., Abhishek, S., Parchani, A., Dhar, M.: Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm. Frontiers in artificial intelligence **6**, 1227091 (Aug 2023). https://doi.org/10.3389/frai.2023.1227091
14. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (09 2019). https://doi.org/10.1093/bioinformatics/btz682
15. Li, D., Wang, J.: Fedmd: Heterogenous federated learning via model distillation. arXiv preprint arXiv:1910.03581 (2019)
16. Li, K., Yu, L., Wang, S., Heng, P.A.: Towards cross-modality medical image segmentation with online mutual knowledge distillation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 775–783. Association for the Advancement of Artificial Intelligence (AAAI) (2020). https://doi.org/10.1609/aaai.v34i01.5421
17. Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the convergence of FedAVG on non-IID data. arXiv preprint arXiv:1907.02189 **34**(04), 4723–4730 (Apr 2019). https://doi.org/10.1609/aaai.v34i04.5905
18. Lin, M., Wang, P., Sun, Z., Chen, H., Sun, X., Qian, Q., Li, H., Jin, R.: Zen-NAS: A zero-shot NAS for high-performance image recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 347–356. IEEE (Oct 2021). https://doi.org/10.1109/iccv48922.2021.00040
19. Lin, T., Kong, L., Stich, S.U., Jaggi, M.: Ensemble distillation for robust model fusion in federated learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
20. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics **23**(6), 1–11 (2022). https://doi.org/10.1093/bib/bbac409
21. Mammen, P.M.: Federated learning: Opportunities and challenges. arXiv preprint arXiv:2101.05428 (2021). https://doi.org/10.48550/arXiv.2101.05428
22. Masoumi, S., Amirkhani, H., Sadeghian, N., Shahraz, S.: Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research. Systematic Reviews **13**(1), 107 (2024). https://doi.org/10.1186/s13643-024-02470-y
23. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR: Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282. PMLR (2017)
24. Meng, H., Lin, Z., Yang, F., Xu, Y., Cui, L.: Knowledge distillation in medical data mining: A survey. In: 5th International Conference on Crowd Science and Engineering. p. 175–182. ICCSE '21, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503181.3503211
25. Mu, J., Kadoch, M., Yuan, T., Lv, W., Liu, Q., Li, B.: Explainable federated medical image analysis through causal learning and blockchain. IEEE Journal of Biomedical and Health Informatics **28**(6), 3206–3218 (2024). https://doi.org/10.1109/jbhi.2024.3375894

26. Oh, W., Nadkarni, G.N.: Federated learning in health care using structured medical data. Advances in kidney disease and health **30**(1), 4–16 (2023). https://doi.org/10.1053/j.akdh.2022.11.007
27. Park, K., Sayres, R., Sellergren, A., Pollard, T., Jamil, F., Kohlberger, T., Lau, C., Kiraly, A.: Application of Med-PaLM 2 in the refinement of MIMIC-CXR labels (2025). https://doi.org/10.13026/7WMP-JX90
28. Plissiti, M.E., Dimitrakopoulos, P., Sfikas, G., Nikou, C., Krikoni, O., Charchanti, A.: Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE (Oct 2018). https://doi.org/10.1109/icip.2018.8451588
29. Ren, P., Xiao, Y., Chang, X., Huang, P.y., Li, Z., Chen, X., Wang, X.: A comprehensive survey of neural architecture search: Challenges and solutions. ACM Computing Surveys (CSUR) **54**(4), 1–34 (May 2021). https://doi.org/10.1145/3447582
30. Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., Tara, M.: A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Medical Informatics and Decision Making **25**(1), 117 (Mar 2025). https://doi.org/10.1186/s12911-025-02954-4
31. Subramoniam, M., Aparna, T.R., Anurenjan, P.R., Sreeni, K.G.: Deep Learning-Based Prediction of Alzheimer's Disease from Magnetic Resonance Images, pp. 145–151. Springer Nature Singapore (2022). https://doi.org/10.1007/978-981-16-7771-7_12
32. Voigt, P., von dem Bussche, A.: The EU General Data Protection Regulation (GDPR): A Practical Guide, vol. 10. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-62328-8
33. Wang, Y., Zhen, L., Zhang, J., Li, M., Zhang, L., Wang, Z., Feng, Y., Xue, Y., Wang, X., Chen, Z., Luo, T., Goh, R.S.M., Liu, Y.: MedNAS: Multiscale training-free neural architecture search for medical image analysis. IEEE Transactions on Evolutionary Computation **28**(3), 668–681 (Jun 2024). https://doi.org/10.1109/tevc.2024.3352641
34. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y.: A survey on federated learning. Knowledge-Based Systems **216**, 106775 (Mar 2021). https://doi.org/10.1016/j.knosys.2021.106775