Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 2 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i02.58737

PAPER

Federated Learning with Adaptive Intermediate Model Selection for Predicting IVIG Resistance in Kawasaki Disease

Namitha T N[1] (✉), Raghavendra S[1], Vinith R[2]

[1]Christ (Deemed to be University), Bangalore, India

[2]Amrita Vishwa Vidyapeetham, Coimbatore, India

namitha.tn@res.christuniversity.in

ABSTRACT

Kawasaki disease (KD), a rare pediatric illness affecting children under five, is treated with intravenous immunoglobulin (IVIG). But 10–20% of patients are resistant to IVIG, and these resistant kids face a higher risk of coronary artery abnormalities. Identifying resistance early is vital, yet data scarcity, class imbalance, and the disease's rarity necessitate nationwide collaboration, which is often hindered by country-specific privacy policies. Federated learning (FL) provides a practical way for different parties to collaborate on training a model while keeping their raw data private and secure. To enhance model adaptability across diverse clinical populations, we propose an adaptive intermediate model selection strategy in federated learning. Each client retains the version—global or locally fine-tuned—that performs best on its own data, using customizable performance metrics such as F1-score or recall. The system was implemented using the Flower FL framework, with three simulated clients and a shared convolutional neural network (CNN) architecture. Experiments demonstrated that the global model achieved stronger performance than conventional models, and several clients obtained further gains by selecting intermediate models aligned with their data. This approach introduces a novel balance between worldwide collaboration and local personalization in FL, offering a flexible and clinically meaningful solution for IVIG resistance prediction.

KEYWORDS

federated learning (FL), Kawasaki disease (KD), intravenous immunoglobulin (IVIG) resistance, ADASYN, flower framework, convolutional neural network (CNN), adaptive model selection

1 INTRODUCTION

Artificial intelligence (AI) is transforming the healthcare sector nowadays, and it helps the clinicians in the early disease detection, risk assessment, and better treatment planning. Machine learning (ML) algorithms assist medical decision-making by analyzing the available clinical and laboratory data. Kawasaki disease (KD) is

Namitha, T. N., Raghavendra, S., Vinith, R. (2026). Federated Learning with Adaptive Intermediate Model Selection for Predicting IVIG Resistance in Kawasaki Disease. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(2), pp. 109–123. https://doi.org/10.3991/ijoe.v22i02.58737

Article submitted 2025-09-19. Revision uploaded 2025-11-15. Final acceptance 2025-11-15.

an acute, self-limiting inflammation of blood vessels that mainly affects children below five years. It is considered the leading cause of acquired heart disease in children worldwide. The standard treatment includes a single high dose of intravenous immunoglobulin (IVIG), which greatly reduces the risk of coronary artery abnormalities (CAAs). But, around 10–20% of treated patients do not respond to IVIG therapy and continue to have fever and inflammation. This resistance-showing group faces a ninefold higher risk of developing CAAs compared to IVIG-responsive patients. So, the early identification of IVIG resistance is crucial to avail additional treatments, such as corticosteroids or a second IVIG dose, which can significantly improve outcomes.

However, regarding the rare disease research, data scarcity is a critical challenge as it results in small patient data sets. These datasets are also limited to certain hospitals or geographic regions, thus restricting the applicability and effectiveness of conventional ML techniques [1]. Currently deep learning (DL) models are widely used in disease diagnosis, but their application in rare disease research is still limited due to inadequate data and population diversity [2]. In such an environment the only solution is inter-institution collaborative research. But such collaboration is often restricted by strict privacy regulations and ethical concerns of sharing patient data, such as HIPAA and GDPR. In this scenario an emerging approach, privacy-preserving Federated learning (FL), learns to address this issue. Using FL, institutions can collaboratively train models without sharing their local raw data. This concept was first introduced by Google in 2016 [3]. In a typical FL scenario, each client trains a model with their data and sends model updates (gradients, weights, etc.) to a central server that aggregates them to generate a global model. This FL approach is useful in maintaining data privacy while exploiting distributed intelligence, which is critical in sensitive sectors such as healthcare. FL has also received a lot of interest in recent years and has been applied with success in several medical fields, including oncology, cardiology, and medical imaging. Its use in research into rare diseases, however, is still relatively underexplored.

Applying FL encounters a challenge stemming from the diversity of the clinical data from different hospitals. There is a collaboration paradox: collaboration is needed to overcome the scarcity of data; however, the global model is bound to underperform for local populations due to the mismatched data distribution. Such instances require an approach that welcomes collaboration and at the same time respects local variability. In this study, we design an Adaptive Intermediate Model Selection strategy in FL that enables the dynamic selection of best global or local best intermediate models based on their effectiveness. The design enables custom-tailored population models that respond to the needs of the population while safeguarding privacy.

We utilize the Flower framework [4] in this study, which is an open-source and flexible platform for simulating federated learning. It allows the distributed clients simulation with an independent data partition that actually mimics real-world multi-institutional healthcare scenarios. As the shared model, we use convolutional neural network (CNN) because of its proven effectiveness in learning hierarchical representations from tabular clinical data after appropriate feature encoding. Beyond diagnosis, CNN models have been applied across domains to enable secure, collaborative problem-solving, aligning with our use of FL for medical prediction across institutions [5]. Recent works further highlight the role of secure AI practices in healthcare, from reducing vulnerabilities such as phishing through targeted training [6] to leveraging CNN-based models for intrusion detection, reinforcing their relevance for secure FL systems [7]. We first train conventional ML models, then a

centralized DL model using CNN, and then convert the CNN model to a federated setting to implement our adaptive intermediate model selection mechanism. In the adaptive method, every client monitors the best-performing local model during training and compares it with the generated global model. This helps improve local models by leveraging global knowledge and allows clients to keep their best-performing local models. Clients can pick the global model or their saved local model, depending on which one performs better according to the used metrics. This guarantees that the advantages of FL are maintained while solving the problem of model appropriateness in heterogeneous, privacy-preserving settings.

2 RELATED WORKS

Even though rare diseases are individually uncommon, they collectively affect over 300 million people worldwide. To improve diagnostic accuracy, ML techniques have been increasingly employed in rare disease research to detect complex clinical patterns. For example, in research [8], they utilized random forests and support vector machines for biomarker identification in Sturge–Weber syndrome, while [9] employed CNNs with transfer learning for asbestosis diagnosis. Generative adversarial networks (GANs) with recurrent neural networks (RNNs) were applied by the researchers of [10] to address data scarcity in Brugada syndrome, and [11] used XGBoost for metabolic profiling in pulmonary tumor diagnosis. Low prevalence, heterogeneity of data, and fragmented data landscapes pose major challenges in the rare disease research. Further risks that limit the research are the privacy policies that prevent inter-institutional data sharing, along with the scarcity of biomarkers and small cohort sizes. Within this broader context, KD—a rare pediatric vasculitis—has received growing attention, particularly in predicting resistance to IVIG therapy, a key determinant of clinical outcome.

2.1 ML-Based IVIG resistance prediction in KD

Numerous studies have explored ML approaches to predict IVIG resistance in KD, aiming to support early clinical decision-making and reduce the risk of coronary artery complications. [12] developed and evaluated multiple ML models using 82 clinical features from 644 patients. The gradient boosting model outperformed existing clinical scoring systems, achieving the best performance with an AUC of 0.7423, accuracy of 0.8844, specificity of 0.9919, and sensitivity of 0.3043, offering a more accurate and robust decision-support tool. In another multi-center study conducted by [13], 1,398 KD patient records were used, and among the developed models (logistic regression, support vector machine, XGBoost, and LightGBM), the LightGBM model achieved the best performance with an AUC of 0.874, sensitivity of 0.702, and specificity of 0.903. Using a 10-year multi-center dataset, researchers of [14] developed a LightGBM model (AUC = 0.78, sensitivity = 0.50, specificity = 0.88) and a 3-variable scoring system (AUC = 0.72, sensitivity = 0.49, specificity = 0.82). A region-specific random forest model was developed by [15], achieving an AUC of 0.78, sensitivity of 0.52, and specificity of 0.92, using 10 key clinical features. The model was integrated into a web-based tool for real-time clinical use, enhancing early risk stratification. ML models trained on large Korean and U.S. cohorts by [16] achieved moderate predictive performance (AUC ~0.71), with minimal improvement from adding echocardiographic or clinical features.

A nomogram based on nine clinical features was developed by [17], achieving AUCs of 0.75 (internal), 0.66 (external), and 0.83 (prospective) sensitivity (0.74) and specificity (0.64). The model shows promise for early risk identification and clinical decision support in Eastern China. A GBDT model was developed by [18], achieving an AUC of 0.87, sensitivity of 72.6%, and specificity of 89%. Key predictive features included total bilirubin, albumin, and C-reactive protein, highlighting the model's regional suitability. A nomogram incorporating seven clinical predictors by [19] achieved AUROCs of 75.8% and 74.2% in training and validation cohorts, respectively, in Chinese KD patients. The model showed good calibration and offers practical utility for early treatment decisions. The retrospective study developed by [20], the Las Vegas Scoring System (LVSS), achieved higher specificity and comparable sensitivity to existing scoring systems, with a sensitivity of 76.2% and a specificity of 68.6%, with IVIG resistance observed in 30.4% of patients.

The study conducted by [21] compared ten IVIG resistance prediction scores in Turkish children and found that all models had very low sensitivity. But some of them showed high specificity, maybe due to the highly imbalanced nature of data. In the research by [22], they evaluated six ML models and found the best-performing model to be random forest. It achieved the best performance across internal and external validations. In the study conducted by [23], they developed an XGBoost model and achieved high sensitivity (0.889) with strong overall performance (AUC = 0.821), accuracy of 0.748, sensitivity of 0.889, and specificity of 0.683.

2.2 The need for FL in KD

Although recent ML models have improved IVIG resistance prediction, their performance is constrained by limited, single-center datasets and privacy regulations that hinder data sharing. KD, due to its rare nature, lacks sufficient data for robust model generalization. This nature eventually results in inconsistent performance metrics across populations. These limitations highlight the need for a collaborative but at the same time privacy-preserving learning approach. FL addresses this gap by enabling federated model training across institutions without exposing patients. FL is gaining increasing attention in healthcare research nowadays. Recently, in the research proposed by [24], they used an FL framework for liver disease prediction. The methodology adopted is the integration of ensemble models like Random Forests and boosting. In [25], they introduced a personalized tensor-based FL model for multi-site brain disease classification. [26] applied a **CNN-FL hybrid** for lung disease detection using decentralized chest X-rays. Additionally, [27] implemented FL for arrhythmia classification using 12-lead ECG signals, allowing accurate cardiac diagnosis without compromising patient data. Although FL has demonstrated value in other medical domains, it has not yet been explored in KD research. It is primarily because of its most advanced and technically demanding ML paradigms. In the KD domain, this FL approach stands as the only viable solution that enables secure, multi-institutional collaboration without compromising patient privacy. To the best of our knowledge, our study is the first to implement FL in the context of Kawasaki Disease, paving the way for its application in similarly underserved rare disease domains.

3 MATERIALS AND METHODS

In this study, we worked with a dataset of 644 medical records of KD patients collected from the research repository of [12] that contained 82 features that included

demographic details, clinical signs, and laboratory test results. These variables covered multiple clinical domains, such as basic patient information, clinical features, sonography measurements, comprehensive metabolic panel, complete blood count, and inflammatory markers, as summarized in Table 2 of [13]. Out of these, 124 patients were resistant to IVIG treatment, while the remaining 520 responded well. Since the dataset was highly imbalanced, with only about 19% of cases falling into the resistant category, we used the ADASYN (adaptive synthetic sampling) [28] technique to synthetically generate more samples from the minority class and help the model learn better from those cases. After applying ADASYN, approximately 376 synthetic IVIG-resistant samples were generated, resulting in a balanced dataset of about 1,040 records (≈520 per class) that was subsequently divided into training (85%) and testing (15%) subsets for model evaluation. This enhanced dataset was used to train and test our CNN-based FL model, allowing us to predict IVIG resistance while ensuring data privacy. To ensure robustness and maintain consistency in model evaluation, all comparative approaches in this study were trained and evaluated on the same dataset distribution, the ADASYN-augmented dataset. In addition to the initial dataset, we aimed to evaluate the model's generalization by leveraging a separate independent dataset from [13] that comprised 1,398 KD cases diagnosed between the years 2015 to 2020. This secondary dataset had 1,240 cases classified as IVIG responders and 158 as resistant cases, with 31 features spanning clinical, demographic, and imaging data, as well as laboratory data. The usage of both datasets enhanced the confidence in the model's performance by ascertaining its accuracy and reliability across diverse populations.

3.1 Conventional ML models and DL using CNN

As a foundational implementation, we started our study with traditional ML models. We used three different models here, logistic regression, Naive Bayes, and decision tree, to establish a baseline performance. Logistic regression was selected due to its effectiveness and straightforward handling of basic classification tasks. Naive Bayes was selected because of its speed and its effectiveness in providing useful results, even when the data does not conform to its assumptions. The decision tree is used because of its simplicity and effectiveness in capturing non-linear composite relationships. As the deep learning approach, we used a one-dimensional CNN to examine the data for deeper, more complex patterns. Even though our dataset consists of tabular clinical values without explicit temporal order, we applied 1D CNNs to explore the CNN's ability to identify potential local interactions between features. This approach allowed us to capture dependencies between adjacent features while maintaining parameter efficiency compared to fully connected deep networks.

3.2 FL using Flower framework

To implement FL, we used the Flower framework [4], which is an open-source, modular framework for simulating federated systems with popular ML frameworks like TensorFlow and PyTorch. In our study, we created a federated environment by partitioning the data and assigning a subset of data to each client. For local model evaluation, clients individually created a training set and a testing set. Each client trained the shared model with local training data and returned the modified model parameters to a central server. The server aggregated these updates using the FedAvg

(federated averaging) algorithm [3] and returned the updated global model to the clients. This was repeated for multiple rounds, enabling collaborative learning within a framework that preserved the raw data and decentralized learning.

We implemented a 1D CNN (1D CNN) as the shared model in a federated setup because it effectively learns local connections and hierarchical features in the data through its convolutional layers, and it has a fairly low trainable parameter count. This is a positive aspect in a federated environment, which is bandwidth and computation constrained. Older models, such as decision trees or Naive Bayes, are not appropriate for federated training cycles, as they do not allow for trainable-gradient optimizations—which are crucial for parameter aggregation in FL. Logistic regression, while federatable, is simply not powerful enough to capture and model the complex, non-linear correlations within the data. Therefore, in light of the balance between expressiveness, training efficiency, and modelling alignment with the FL structure, the choice of 1D CNN makes sense. At the start of training, the initial global model parameters $\theta_{global}^{(0)}$ are obtained from the first instantiated client and broadcast to all participating clients. Each client receives the current global weights $\theta_{global}^{(t)}$ from the server and initializes its local CNN with these parameters. Training then proceeds on the client's private dataset $D_k = (X_k, y_k)$ where the CNN updates its parameters through backpropagation using convolutional and fully connected layers. The update rule for client k at round t is given as in Eq. (1):

$$\theta_k^{(t+1)} = \theta_k^{(t)} - \eta \nabla_\theta L(\theta_k^{(t)}; D_k) \tag{1}$$

where η is the learning rate, and L (·) is the CNN's loss function computed on the local dataset. After training for E epochs, the client returns the updated parameters $\theta_k^{(t+1)}$ along with the number of training samples n_k. This ensures that clients with larger datasets exert more influence on the global update, a critical step when combining CNN models across heterogeneous data sources. The server aggregates these CNN updates using the federated averaging (FedAvg) [3] algorithm, which computes a weighted mean of the client models as in Eq. (2):

$$\theta_{global}^{(t+1)} = \frac{\sum_{k \in S_t} n_k \theta_k^{(t+1)}}{\sum_{k \in S_t} n_k} \tag{2}$$

Where S_t is the set of participating clients in round t. After aggregation, the server evaluates the 1D CNN on a validation/test dataset to monitor performance. Global metrics are derived by weighting client-specific results, e.g., global accuracy as in Eq. (3): and similar weighted formulas apply for precision, recall, and F1-score

$$Accuracy_{global} = \frac{\sum_{k \in S_t} n_k Accuracy_k}{\sum_{k \in S_t} n_k} \tag{3}$$

The overall loss is also measured as in Eq. (4):

$$L(\theta_{global}^{(t)}) = \frac{1}{|D_{test}|} \sum_{i \in D_{test}} L(y_i, y_i') \tag{4}$$

Through this iterative process, the 1D CNN learns hierarchical representations from each client's data locally, while the server aggregates updates to refine a single, privacy-preserving global model that improves over successive rounds.

3.3 Proposed adaptive intermediate model selection in FL

In KD research, the limited availability of data at individual clinical sites and stringent privacy regulations render centralized model training impractical. FL offers a promising solution by enabling collaborative model development without exchanging raw data. However, a single globally trained model may not perform optimally across heterogeneous client populations due to site-specific variations in data distributions and clinical features. To address this challenge, we propose an adaptive intermediate model selection strategy in FL. In this approach, each client tracks the performance of the global model received at every training round and retains the version—either the raw global model or a locally fine-tuned variant—that achieves the best performance on its local validation dataset, as shown in Figure 1. This strategy balances global knowledge transfer with local personalization. Importantly, clients are given the flexibility to define their model selection criteria based on local requirements. In our implementation, we initially used the F1-score as the selection metric and later extended it to include recall, given its effectiveness in handling class imbalance by accounting for both false positives and false negatives—an essential consideration in rare disease scenarios. Nevertheless, the framework is flexible, allowing clients to adopt alternative metrics (e.g., accuracy, specificity) based on clinical priorities or the characteristics of their data.

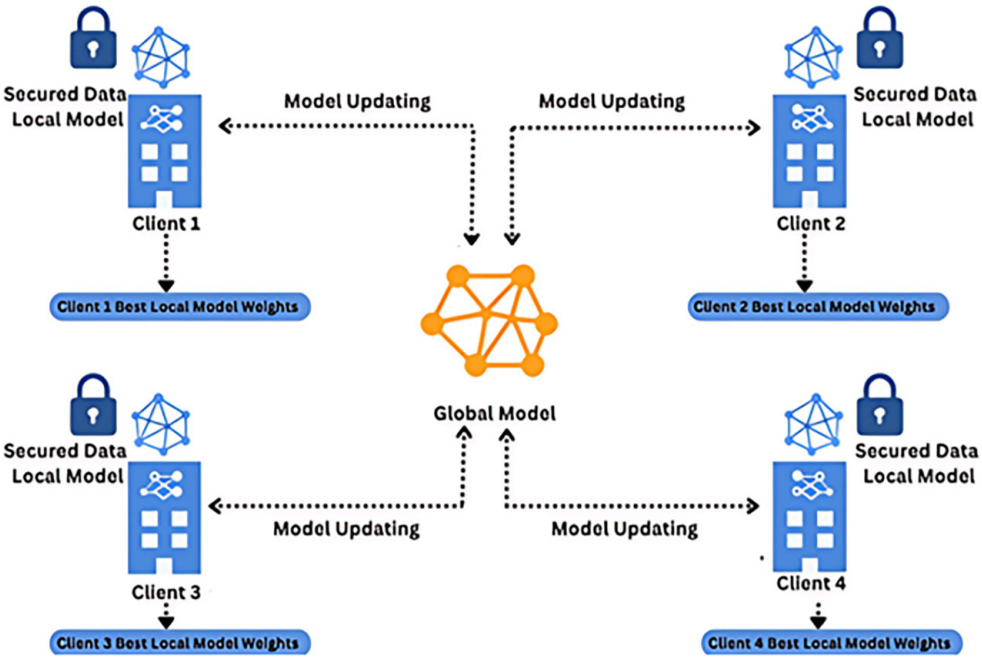

Fig. 1. Architecture of adaptive intermediate model selection in federated learning

Let K be the total number of participating clients in the federated system. During each communication round t, the central server broadcasts the current global model parameters $\theta^{(t)}$ to all clients. Upon receiving the global model, each client $k \in \{1, 2, ..., K\}$ trains it on its private dataset D_k^{train}, yielding updated local model parameters $\theta_k^{(t)}$. To enable personalized model selection, each client evaluates both the received global model and its locally updated model on a local validation set D_k^{val}.

The evaluation is performed using a performance metric M_k chosen independently by the client, such as F1-score, recall, or accuracy. The scores for the local and global models are computed as in Eq. (5):

$$M_k^{(t)} = M_k(f(.;\theta_k^{(t)}), D_k^{val}), M_k^{(g)} = M_k(f(.;\theta^{(t)}), D_k^{val}) \tag{5}$$

Based on this evaluation, the client selects the model that performs better according to its chosen metric, as in Eq. (6):

$$\theta_k^* = \begin{cases} \theta_k^{(t)}, if\ M_k^{(t)} > M_k^{(g)} \\ \theta^{(t)}, otherwise \end{cases} \tag{6}$$

Throughout the training process, each client maintains the best-performing model on its validation set by tracking the maximum observed performance over time, as in Eq. (7):

$$\theta_k^{best} = \arg \max_{\tau \in \{1,2,\ldots,t\}} M_k(f(.;\theta_k^*(\tau)), D_k^{val}) \tag{7}$$

Employing the FedAvg algorithm, the server side updates the global model by aggregating contributions from local updates, each weighted by the number of samples n_k at the corresponding client. This modelling technique integrates the fusion of global model learning with client personal model adaptation. This technique is ideal for situations involving rare diseases since the data present scarcity, imbalance, and considerable heterogeneity across clinical sites. We used the Flower FL framework to implement this. In this instance, we simulated three clients, each training a shared 1D CNN on its private dataset. After each training round, clients assessed the global model using the F1-score to determine which version to keep. A modified federated averaging (FedAvg) algorithm on the server-side captured client updates, while centralized monitoring of global metrics assessed performance during evaluation sprints. The training sessions concluded with an assessment of the global model, which incorporated metrics of the accuracy, precision, recall, and F1-score of the best local model each client kept. Results showed that offering clients personal model adaptation through intermediate model selection is beneficial to performance, especially in data-limited and heterogeneous clinical environments.

4 RESULTS AND DISCUSSIONS

Initially, we assessed standard ML techniques, namely logistic regression, Naive Bayes, and decision tree, and subsequently a CNN from a deep learning perspective. The initial training and testing of these models occurred on the primary dataset from research [12] containing 644 instances of KD cases, for which we used ADASYN for class imbalance. The same dataset was also used for later experiments under an FL setting, with the CNN as the global model on simulated clients. To determine the approach's generalizability and robustness, we applied the same method on a different, independent KD dataset from research [13], which contained 1,398 records collected over five years. The independent dataset served to demonstrate model performance—particularly models trained under the FL setting—on different populations and varying distributions of data.

4.1 Results

The results provided in Tables 1 and 2 demonstrate the effectiveness of the FL model on this dataset. The FL outperforms the conventional approaches in all performance metrics, as shown in Figure 2. Statistical validation using 5-fold cross-validation and a paired t-test confirmed that the performance improvement of the FL-CNN (Accuracy = 89%, F1 = 88%) over the centralized CNN (Accuracy = 83%, F1 = 82%) was statistically significant ($p < 0.05$).

Table 1. Performance comparison of various models on **Dataset 1** (primary evaluation dataset) from [12]

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	75	82	67	73
Naïve Bayes	64	82	54	66
Decision Tree	83	82	80	81
Centralized CNN	83	93	74	82
FL-CNN	89	98	80	88

Table 2. Performance comparison of various models on **Dataset 2** from [13] to evaluate the proposed model's consistency and generalizability

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78	71	48	57
Naïve Bayes	75	60	56	58
Decision Tree	83	81	82	81
Centralized CNN	89	83	83	83
FL-CNN	90	98	82	89

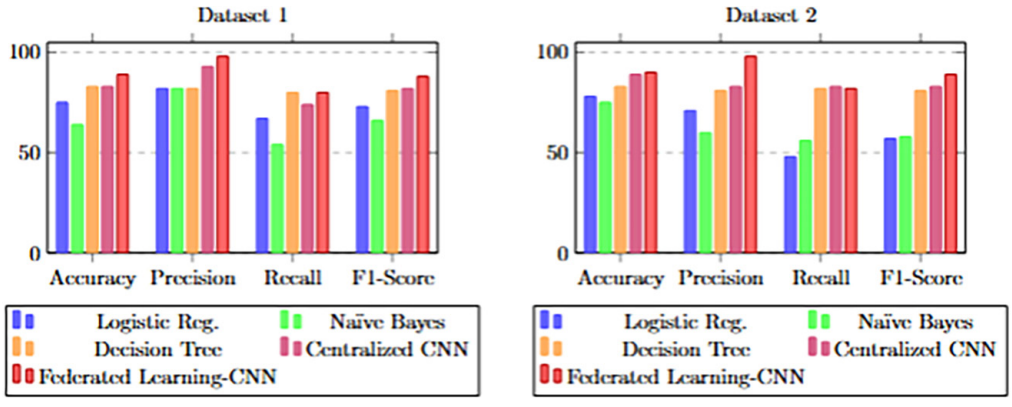

Fig. 2. Performance comparison of various models on Dataset 1 and Dataset 2

Note: FL outperforms all conventional models across all metrics.

To evaluate the effectiveness of our proposed adaptive intermediate model selection strategy, we implemented an FL system using the same Flower framework across three simulated clients and two clinical datasets. Every client evaluated the

results of the global models they received at each training round and kept the version that had the highest local validation performance, primarily based on F1-score, as shown in Table 3. In Dataset 1, although the FL global model performed well overall (F1-score: 88, Recall: 80), Client 0's best local model even bettered it, achieving an F1 score of 90 and recall of 86, which emphasizes the value of personalization. Client 1's result was on par with the global model, while Client 2's best local model underachieved, which could imply data sparsity or noise. This shows that client-specific model selection can still provide benefits when local data distribution shifts away from the global distribution.

On the other hand, the results from Dataset 2, displayed in Table 4, indicated that in the final FL global model evaluation, the global model version consistently surpassed the performance of all client-specific best models, based on the F1-score, Accuracy, and Recall metrics (F1-score: 89, Accuracy: 90%, Recall: 82%). Even so, Client 0's best model highlighted recall (87%), which is arguably the best performance metric in clinical settings, for false negatives are the most critical errors. Comparison of performance for the global model and the client-side best models can also be seen in Figure 3. This illustrates that while the FL methodology and collaborative training, effectively generalize to different use cases in the population, the use of adaptive selection provides clients the ability to fine-tune within specific frontiers based on their local validation results. This is particularly important in healthcare situations such as Kawasaki disease and research on rare diseases, where the clinical data diversity, clinical focus, and data priorities differ across healthcare centers.

Table 3. Performance of FL global and client-specific best models on dataset 1 using adaptive intermediate model selection (F1 as customization metric) [12]

Model	Accuracy	Precision	Recall	F1-Score
FL global model	89	98	80	88
FL client 0 Best model	**90**	**94**	**86**	**90**
FL client 1 Best model	89	97	81	88
FL client 2 Best model	76	98	65	78

Table 4. Performance of FL global and client-specific best models on dataset 2 from [13] using adaptive intermediate model selection (F1 as customization metric)

Model	Accuracy	Precision	Recall	F1-Score
FL global Model	**90**	**98**	**82**	**89**
FL client 0 Best model	85	79	87	83
FL client 1 Best model	85	98	76	85
FL client 2 Best model	85	97	76	85

To continue testing the robustness and consistency of the adaptive intermediate model selection method, we opted for recall as the metric for client-side customization, as described in Tables 5 and 6 for different datasets. Choosing recall makes sense in the context of imbalanced classes, as in rare disease prediction, where minimizing the false negatives is of utmost importance. In such scenarios, recall is critical

as it measures performance in identifying cases in the minority class, directly affecting the clinical value of the model.

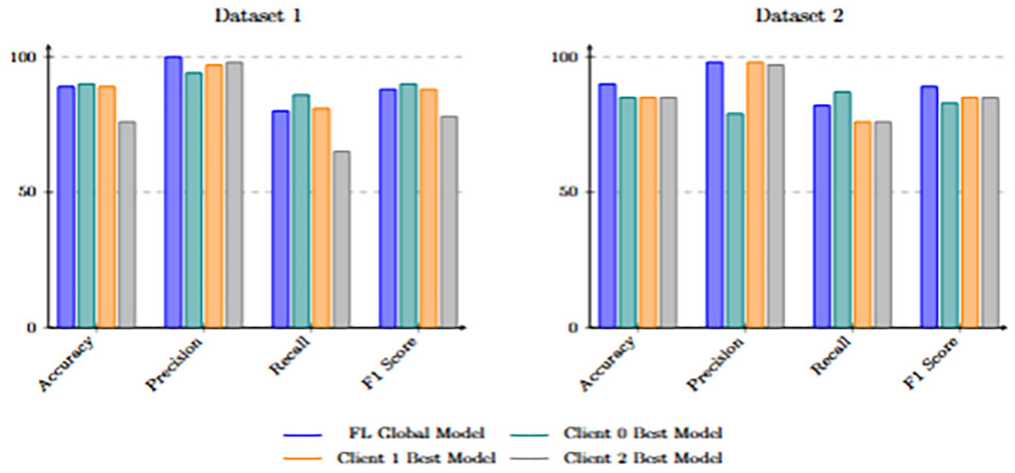

Fig. 3. Performance of FL global and client-specific best models using adaptive intermediate model selection (F1-score as customization metric)

4.2 Discussions

In Dataset 1, for example, Client 2's local best model was able to reach a higher recall of 84% in contrast to the global model, which was 83%, while still obtaining reasonably high values of precision and F1-score. This suggests that for some populations, the positive class distribution may be more easily captured and represented by client-specific models. In contrast, in Dataset 2, the FL global model performance was superior to all client models for all evaluated criteria, including recall, which suggests more generalized performance when datasets have less skewed distributions. For the performance of the global model and the best client-side models with respect to recall as the client-side customization metric, refer to Figure 4. This confirms that model flexibility provided by adaptive model selection is meaningful. It permits clients to keep and utilize models that are better suited to their unique needs and clinical objectives related to their population, especially in situations that require minimizing false negatives. The limitation of the study is that this study was conducted in a simulated federated setup with limited datasets, without considering factors like client heterogeneity and communication constraints. While certain client-specific models exhibited higher performance than the global model, this cannot be conclusively attributed to true population-level differences. The observed variations may also arise from differences in training and validation splits or inherent dataset bias. Further investigation with detailed demographic and clinical metadata across clients is required to confirm whether such local improvements reflect genuine population-specific diagnostic characteristics of IVIG resistance. Future work will involve evaluating the approach on larger, real-world medical datasets with diverse clients and integrating privacy-preserving mechanisms. In addition, comparative studies using different oversampling techniques will be conducted to analyze their impact on model performance. Additionally, extending the framework with transformer-based and multimodal models could further enhance scalability and generalization.

Table 5. Performance of FL global and client-specific best models on dataset 1 from [12] using adaptive intermediate model selection (Recall as customization metric)

Model	Accuracy	Precision	Recall	F1-Score
FL global model	90	97	83	89
FL client 0 Best model	82	98	72	83
FL client 1 Best model	85	85	81	83
FL client 2 Best model	88	89	**84**	87

Table 6. Performance of FL global and client-specific best models on dataset 2 from [13] using adaptive intermediate model selection (Recall as customization metric)

Model	Accuracy	Precision	Recall	F1-Score
FL global Model	90	97	84	90
FL client 0 Best model	83	69	**90**	78
FL client 1 Best model	88	98	79	88
FL client 2 Best model	80	97	70	81

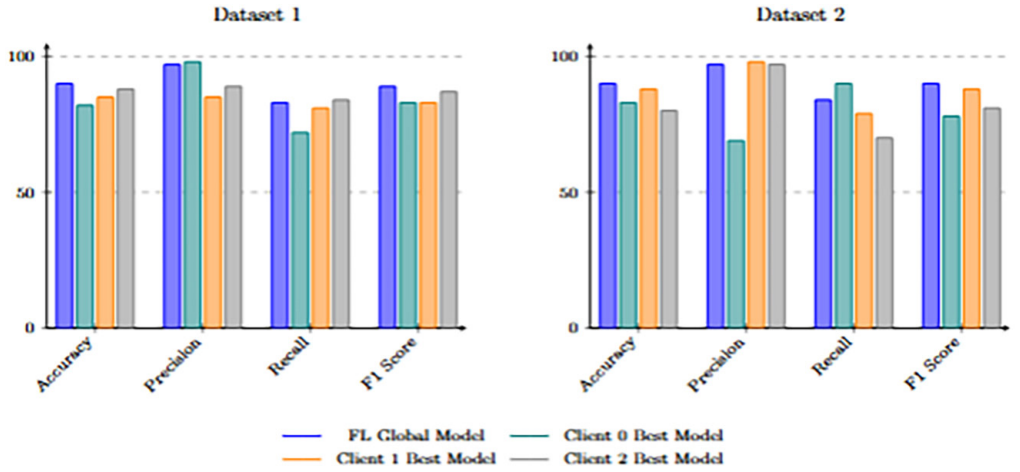

Fig. 4. Performance of FL global and client-specific best models using adaptive intermediate model selection (Recall as customization metric)

5 CONCLUSION

This study proposed an adaptive intermediate model selection strategy within a FL framework to enhance the prediction of IVIG resistance in KD. The approach addressed the study's objectives of improving personalization and robustness in heterogeneous, privacy-sensitive clinical environments by allowing each client to evaluate intermediate global models during training and retain the one that best fits its local validation data based on metrics such as F1-score or recall. Experimental results demonstrated that on Dataset 1, Client 0's local model achieved an F1-score of 90%, outperforming the global model (88%), while on Dataset 2, the global model achieved 90% accuracy, 97% precision, 84% recall, and a 90% F1-score, outperforming all clients. These results confirm that the adaptive mechanism enhances client-level flexibility while maintaining global model reliability. The findings

contribute to advancing personalization in FL by showing that adaptive model retention can balance global and local performance without compromising privacy. Future research should extend this approach to larger, real-world medical datasets, integrate privacy-preserving techniques, and explore transformer-based and multimodal architectures to further assess scalability and generalization.

6 REFERENCES

[1] J. Schaefer, M. Lehne, J. Schepers, F. Prasser, and S. Thun, "The use of machine learning in rare diseases: A scoping review," *Orphanet J. Rare Dis.*, vol. 15, no. 1, pp. 1–10, 2020. https://doi.org/10.1186/s13023-020-01424-6

[2] J. Lee *et al.*, "Deep learning for rare disease: A scoping review," *J. Biomed. Inform.*, vol. 135, p. 104227, 2022. https://doi.org/10.1016/j.jbi.2022.104227

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics, PMLR*, 2017, pp. 1273–1282. Accessed: Jul. 1, 2024. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[4] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, and N. Lane, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[5] I. U. Haq, M. Pifarré, and E. Fraca, "Natural language processing approach to evaluate real-time flexibility of ideas to support collaborative creative process," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 19, no. 5, pp. 93–107, 2024. https://doi.org/10.3991/ijet.v19i05.47465

[6] D. J. Challacombe and E. N. McElhiney, "Phishing susceptibility among healthcare workers: The impact of awareness, email type, and location," *International Journal of Advanced Corporate Learning (iJAC)*, vol. 18, no. 1, pp. 4–15, 2025. https://doi.org/10.3991/ijac.v18i1.51671

[7] S. Alshattnawi and H. R. Alshboul, "Combined deep learning approaches for intrusion detection systems," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 18, no. 19, pp. 144–155, 2024. https://doi.org/10.3991/ijim.v18i19.49907

[8] S. S. Gupta, K. E. Joslyn, K. D. McKenney, and A. M. Comi, "Biomarker development in Sturge-Weber syndrome," *J. Neurodev. Disord.*, vol. 17, no. 1, pp. 1–8, 2025. https://doi.org/10.1186/s11689-025-09640-6

[9] I. Smesseim *et al.*, "Prospective validation of an artificial intelligence assessment in a cohort of applicants seeking financial compensation for asbestosis (PROSBEST)," *Eur. Radiol. Exp.*, vol. 9, no. 1, pp. 1–8, 2025. https://doi.org/10.1186/s41747-025-00619-5

[10] K. Saleh *et al.*, "4-025 facilitating AI-ECG models for rare cardiac diseases: Transfer learning and synthetic data generation for brugada ECG classification," *Heart*, vol. 111, no. Suppl 3, pp. A139–A140, 2025. https://doi.org/10.1136/heartjnl-2025-BCS.137

[11] F. Amin, H. Khalid, M. Khalid, M. Talha, and A. Waafira, "Integrative AI-metabolomics: A new frontier in diagnosing pulmonary tumor thrombotic microangiopathy," *Annals of Medicine & Surgery*, vol. 87, no. 10, pp. 6870–6871, 2025. https://doi.org/10.1097/MS9.0000000000003707

[12] T. Wang, G. Liu, and H. Lin, "A machine learning approach to predict intravenous immunoglobulin resistance in Kawasaki disease patients: A study based on a Southeast China population," *PLoS One*, vol. 15, no. 8, p. e0237321, 2020. https://doi.org/10.1371/journal.pone.0237321

[13] J. Liu *et al.*, "A machine learning model to predict intravenous immunoglobulin-resistant kawasaki disease patients: A retrospective study based on the Chongqing population," *Front. Pediatr.*, vol. 9, p. 756095, 2021. https://doi.org/10.3389/fped.2021.756095

[14] Y. Sunaga and A. Watanabe, "A simple scoring model based on machine learning predicts intravenous immunoglobulin resistance in kawasaki disease," *Research Square*, 2022. https://doi.org/10.21203/rs.3.rs-1215051/v1

[15] Y. He *et al.*, "Interpretable web-based machine learning model for predicting intravenous immunoglobulin resistance in Kawasaki disease," *Ital. J. Pediatr.*, vol. 51, no. 1, pp. 1–17, 2025. https://doi.org/10.1186/s13052-025-02036-1

[16] J. Y. Lam *et al.*, "Intravenous immunoglobulin resistance in Kawasaki disease patients: Prediction using clinical data," *Pediatric Research*, vol. 95, no. 3, pp. 692–697, 2023. https://doi.org/10.1038/s41390-023-02519-z

[17] H. Huang *et al.*, "Nomogram to predict risk of resistance to intravenous immunoglobulin in children hospitalized with Kawasaki disease in Eastern China," *Ann. Med.*, vol. 54, no. 1, pp. 442–453, 2022. https://doi.org/10.1080/07853890.2022.2031273

[18] Y. Yang *et al.*, "Research on early identification model of intravenous immunoglobulin resistant kawasaki disease based on gradient boosting decision tree," *Pediatric Infectious Disease Journal*, vol. 42, no. 7, pp. 537–542, 2023. https://doi.org/10.1097/INF.0000000000003919

[19] J. Wang, X. Huang, and D. Guo, "Predictors and a novel predictive model for intravascular immunoglobulin resistance in Kawasaki disease," *Ital. J. Pediatr.*, vol. 49, no. 1, pp. 1–8, 2023. https://doi.org/10.1186/s13052-023-01531-7

[20] R. K. Natarajan, S. V. Bhoopalan, C. Cross, R. Shah, and A. Rothman, "Novel score to predict immunoglobulin resistance in Kawasaki disease," *Pediatr. Cardiol.*, vol. 44, no. 7, pp. 1546–1551, 2023. https://doi.org/10.1007/s00246-023-03175-0

[21] U. Kaya Akca *et al.*, "Comparison of IVIG resistance predictive models in Kawasaki disease," *Pediatr. Res.*, vol. 91, no. 3, pp. 621–626, 2022. https://doi.org/10.1038/s41390-021-01459-w

[22] Y. Xia *et al.*, "A machine learning-based model to predict intravenous immunoglobulin resistance in Kawasaki disease," *iScience*, vol. 28, no. 3, p. 112004, 2025. https://doi.org/10.1016/j.isci.2025.112004

[23] L. Deng *et al.*, "Construction and validation of predictive models for intravenous immunoglobulin–resistant Kawasaki disease using an interpretable machine learning approach," *Clin. Exp. Pediatr.*, vol. 67, no. 8, p. 405, 2024. https://doi.org/10.3345/cep.2024.00549

[24] D. Kumar, C. Verma, and Z. Illés, "Federated learning with explainable AI for liver disease prediction: A privacy-preserving approach," *Intell. Based Med.*, vol. 12, p. 100285, 2025. https://doi.org/10.1016/j.ibmed.2025.100285

[25] Y. Gao, G. Zhang, C. Zhang, J. Wang, L. T. Yang, and Y. Zhao, "Federated tensor decomposition-based feature extraction approach for industrial IoT," *IEEE Trans. Industr. Inform.*, vol. 17, no. 12, pp. 8541–8549, 2021. https://doi.org/10.1109/TII.2021.3074152

[26] M. Karmakar, A. Hota, and A. Nag, "Convolutional neural network (CNN) and federated learning-based approach for lung disease detection," *Iran Journal of Computer Science*, vol. 8, pp. 2387–2408, 2025. https://doi.org/10.1007/s42044-025-00320-1

[27] D. M. Jimenez Gutierrez, H. M. Hassan, L. Landi, A. Vitaletti, and I. Chatzigiannakis, "Application of federated learning techniques for arrhythmia classification using 12-Lead ECG signals," in *Algorithmic Aspects of Cloud Computing (ALGOCLOUD 2023)*, in Lecture Notes in Computer Science, vol. 14053, LNCS, pp. 38–65, 2024. https://doi.org/10.1007/978-3-031-49361-4_3

[28] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the International Joint Conference on Neural Networks*, 2008, pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

7 AUTHORS

Namitha T N is a Research Scholar in the Department of Computer Science and Engineering at Christ (Deemed to be University), in Bangalore, India (E-mail: namitha.tn@res.christuniversity.in).

Raghavendra S is an Associate Professor in the Department of Artificial Intelligence and Machine Learning and Data Science, School of Engineering and Technology, Christ (Deemed to be University), in Bangalore, India (E-mail: raghav.trg@gmail.com).

Vinith R is an Assistant Professor (Senior Grade) in the Department of Artificial Intelligence at Amrita Vishwa Vidyapeetham, in Coimbatore, India (E-mail: r_vinith@cb.amrita.edu).