Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 11 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i11.56385

PAPER

Privacy-Preserving Federated Learning for Prognostic Modeling in Rare Diseases: A Scalable Case Study on Kawasaki Disease

Namitha T N[1] (✉), **Raghavendra S**[2], **Vinith R**[3]

[1]Department of CSE, Christ (Deemed to be University), Bangalore, Karnataka, India

[2]Department of AIML and Data Science, Christ (Deemed to be University), Bangalore, Karnataka, India

[3]Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India

namitha.tn@res.christuniversity.in

ABSTRACT

Predictive modeling in rare diseases faces major challenges, including data scarcity, class imbalance, and strict privacy regulations that limit cross-border collaboration. These challenges are particularly critical in Kawasaki disease (KD)—a rare vasculitis in children—where 10% to 20% of patients are resistant to intravenous immunoglobulin (IVIG), the standard first-line treatment. This significantly increases the risk of coronary artery abnormalities (CAA), making early and accurate prediction of resistance to IVIG essential for improving patient outcomes. Our work proposes a federated learning (FL) approach to address the constraints imposed by security and privacy concerns. We investigate convolutional neural networks (CNN) as the shared model, collaboratively trained across clients. Coupled with strategies to address class imbalance resulting from the rarity of the condition, the federated approach yielded promising results when evaluated against conventional machine learning (ML) models. The proposed approach demonstrated strong performance, achieving 94% accuracy, 93% precision, 89% recall, and 91% F1 score. To ensure robustness and generalizability, an independent dataset was also used, where the proposed model excelled similarly. These results highlight the potential of FL to overcome data privacy barriers and provide a scalable, secure solution for predictive modeling in rare diseases, supporting its integration into medical prediction workflows.

KEYWORDS

federated learning (FL), adaptive synthetic sampling, convolutional neural network (CNN), flower framework, rare disease, Kawasaki disease (KD), intravenous immunoglobulin resistance

1 INTRODUCTION

The deployment of artificial intelligence (AI) and machine learning (ML) technologies in healthcare gives the opportunity to enhance service delivery through

Namitha, T. N., Raghavendra, S., Vinith, R. (2025). Privacy-Preserving Federated Learning for Prognostic Modeling in Rare Diseases: A Scalable Case Study on Kawasaki Disease. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(11), pp. 66–80. https://doi.org/10.3991/ijoe.v21i11.56385

Article submitted 2025-05-02. Revision uploaded 2025-07-12. Final acceptance 2025-07-13.

advanced prediction and more sensitive classification of diseases. The new technologies outperform classical statistical techniques in discovering intricate relationships in complex datasets and hold great potential in diagnostics and treatment planning and in devising more personalized approaches to manage patients. On the other hand, the ML models are highly dependent on the existence of comprehensive, high-quality datasets. When it comes to rare diseases, there are often very few patients and not much data available. This lack of information makes it hard for models to learn properly, which leads to poor performance. Additionally, ethical and legal restrictions such as general data protection regulation (GDPR) also make it difficult to collate data from different institutions and countries, which makes it apparent that there is a need for new and innovative solutions developed to tackle the problem of data scarcity and privacy in rare disease research.

Federated learning (FL), introduced by Google for next-word prediction in the GBoard application [1], holds significant potential to address critical challenges by tackling both data scarcity and privacy concerns, making it especially valuable for advancing research in rare diseases. It trains models without centralizing data but through collaborative efforts, where clients can train their models without sending raw data. The decentralized framework respects strict privacy regulations while providing flexibility in cross-border collaborations. FL aggregates parameters of a model from many clients to develop a predictive global model that harnesses the power of available diverse data. Such an approach is highly useful in rare diseases, where every client might just have a small segment of the data required for the model. By protecting data privacy and improving collaboration, FL enables researchers to break the constraints that would otherwise disable the proper development of accurate and personalized models of prediction.

Kawasaki disease (KD) is an exceptional but serious condition in pediatrics, in most cases affecting children under the age of five. The disorder is characterized by prolonged fever, rash, and widespread inflammation of blood vessels. If left untreated, it may change to serious complications, such as coronary artery abnormalities (CAA), which occur in 25% of untreated cases, with the development of coronary aneurysms. Intravenous immunoglobulin (IVIG) can significantly mitigate the risk if given before the onset of CAA. However, in 10–20% of the patients, IVIG fails and results in a condition termed IVIG resistance. IVIG-resistant patients are nine times more susceptible to having coronary artery injury than IVIG-sensitive patients. This makes early prediction of IVIG resistance critical for tailoring aggressive treatment strategies, such as corticosteroid therapy or infliximab, which have shown better coronary outcomes than IVIG alone. The complexity of the condition and the limited availability of KD data underscore the need for innovative predictive models to support early and effective interventions.

Kawasaki disease presents a classical challenge of acutely scarce data combined with stringent confidentiality requirements. To address both issues, we developed a predictive model to estimate patients' responses to IVIG treatment using an FL framework. We built this setup with the Flower framework, chosen for its flexibility, easy integration with PyTorch or TensorFlow, and strong community support, which accelerated development in this healthcare context. To capture complex, non-linear patterns and ensure compatibility with FL, we selected a convolutional neural network (CNN). Unlike traditional ML models, CNNs support gradient-based optimization and parameter sharing, both essential for FL. After rigorous local training and output integrations, we developed a global model that can be used with a wide variety of data while retaining patient confidentiality. This model ensures better acknowledgement of privacy compared to conventional

centralized systems. The paper begins by reviewing recent advancements in IVIG resistance prediction and the role of FL in healthcare, emphasizing that this approach has not yet been applied in the rare disease context. It then details the dataset, methodologies, and comparative performance of conventional ML, deep learning, and our proposed FL approach. Our results show that the proposed model outperforms existing methods, highlighting the promise of privacy-preserving FL frameworks in rare disease prediction and their potential to facilitate timely, personalized treatment strategies.

2 RELATED WORKS

Kawasaki disease, though its exact cause is unknown, can lead to severe complications such as CAA and long-term cardiovascular issues if untreated. Diagnosis, treatment, and long-term care details follow in reference [2]. These steps were first noted by Dr. Tomisaku Kawasaki in 1967. Identifying high-risk patients at an early stage allows the adoption of greater intensity therapies such as corticosteroids or infliximab alongside the standard treatment with IVIG. Such combined therapy results in greater coronary outcomes than when IVIG is used in isolation [3]. The first systematic study on predicting IVIG resistance in KD was administered by Kobayashi in 2006 [4]. The study highlighted the need for early intervention of IVIG non-responders to mitigate the need for advanced peripheral coronary artery blockage. Through examination of clinical profile and laboratory data, multivariable logistic regression (LR) was utilized to predict IVIG resistance, and ten laboratory criteria, alongside three demographic variables, were used to identify strong predictors. Additional research indicates use of corticosteroids alongside IVIG as initial treatment may also aid in the decline of detected coronary artery aneurysms.

A study conducted by [5] outlines the diagnosis and management of KD and highlights the role of adjunctive therapies to reduce the risk of CAA in high-risk patients. In [6], IVIG nonresponse was predicted using a gradient boosting decision tree, incorporating both demographic and laboratory variables. A more recent approach in [7] employed XGBoost, achieving an AUC of 77.4. Study [8] explored five ML models across two separate population datasets, noting consistently low sensitivity across all models, with a maximum of 0.216. In [9], a LightGBM model classified IVIG-resistant versus IVIG-responsive cases with a sensitivity of 0.50, compared to 0.58 for LR and 0.60 for support vector machine (SVM).

To assess the performance of existing prediction models, [10] compared ten different IVIG resistance scoring systems—including Kobayashi, Formosa, Egami, Sano, and Harada—on a Turkish paediatric population. The study found that all models were limited by low sensitivity. Due to the rarity of KD, datasets for IVIG resistance prediction are scarce. Study [11] developed and compared predictions using LR, SVM, XGBoost, and LightGBM, with LightGBM achieving the best performance (AUC = 0.874, sensitivity = 0.702). However, [12] presents a dataset with 753 observations and 82 features, applying seven linear and nonlinear ML techniques—gradient boosting machine (GBM), LightGBM, random forest (RF), decision tree (DT), LR (L1 and L2 regularized), and AdaBoost with GBM achieving the best performance (AUC = 0.7423, sensitivity = 0.3043, specificity = 0.9919, accuracy = 0.8844). In [13], a dataset comprising 5,277 subjects and 57 variables was used to predict IVIG resistance. The variables included one imaging variable, four demographic variables, and fifty-two laboratory variables. Various L algorithms—such as multivariate LR,

DT, RF, AdaBoost, GBM, and LightGBM—have been employed in the literature for this prediction task. IVIG resistance prediction has been explored using various ML and deep learning approaches. However, the rare nature of KD results in significant data scarcity, posing challenges to developing effective predictive models. Most of the studies described in this related work section from 2019 to 2024 utilized conventional models, as shown in Figure 1.

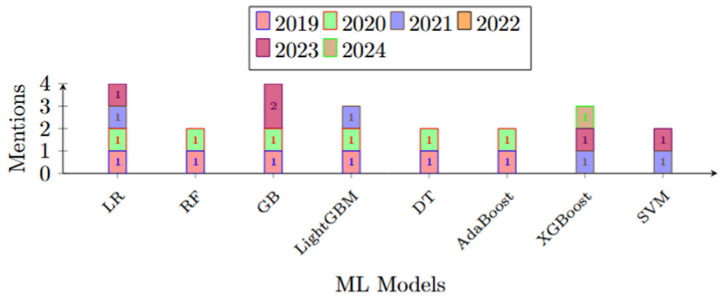

Fig. 1. Machine learning model usage in reviewed Kawasaki disease studies

Several recent studies have demonstrated the value of advanced AI techniques for medical prediction and secure collaborative systems. ML models have shown strong performance in tasks such as Alzheimer's [14] and lung cancer diagnosis [15]. CNN architectures have proven effective for disease prediction [16]. 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 [17]. Recent works further highlight the role of secure AI practices in healthcare, from reducing vulnerabilities such as phishing through targeted training [18] to leverage CNN-based models for intrusion detection, reinforcing their relevance for secure FL systems [19].

Furthermore, stringent privacy regulations across different countries, such as GDPR, restrict research collaborations and data sharing, further limiting the practical implementation of these models in rare disease scenario. To address these challenges, privacy-preserving FL can be employed to utilize data at endpoints without compromising patient privacy [20]. FL has been successfully applied in various medical domains as listed in [21], demonstrating its potential to overcome data-sharing barriers while maintaining compliance with stringent privacy regulations. Some example medical problems addressed using FL are listed in Table 1. To the best of our knowledge, the application of FL in KD, or even in the context of rare diseases, has not been explored yet.

Table 1. Federated learning research in healthcare: selected prominent examples

Paper	Addressed Disease	Paper	Addressed Disease
[22]	Brain Tumor Segmentation	[23]	EEG Classification
[24]	Autism Spectrum Disorder	[25]	Skin Disease Detection
[26]	Brain Tumor Segmentation	[27]	Brain Disease Prediction from fMRI Brain Connection
[28]	Medical Images Privacy Preservation	[29]	MRI prostate segmentation
[30]	Female Pelvis Organ Segmentation	[31]	Cardiovascular Disease Diagnosis
[32]	Hospitalization Prediction from Electronic Health Records	[33]	Epidemic Diseases Screening

3 DATA AND METHODS

Due to the rare nature of KD, there are very few publicly available datasets. So, to build our prediction model, we have used the data available from various research works. To build the prediction model, we have used the datasets associated with [11], which consists of 1398 records, 31 clinical and laboratory features associated with KD, meant to predict IVIG resistance. These features include demographic data such as sex and age in months, along with the occurrence of conjunctival hyperemia, perianal changes, and cervical lymphadenopathy. More clinical indicators are available regarding the maximum temperature at initial treatment, rash, lip changes, and days of illness. Laboratory tests-erythrocyte sedimentation rate (ESR), lymphocyte count, hemoglobin (HB), platelet (PLT) count, neutrophil count, C-reactive protein (CRP), and the neutrophil-lymphocyte ratio (NLR)- reflect the manifestation of patient state, which are quantifiable through laboratory procedures. The target variable, IVIGRKD, revealed whether the patient is resistant to the treatment with IVIG (0 = not resistant, 1 = resistant). We used two separate datasets to see how well our model holds up. First, we trained the model and compared its performance with conventional ML models using the dataset [11]. Then, we ran the same steps on a second, completely independent dataset [12], which included 82 different variables such as patient demographics, clinical information, and lab results. This helped us see if the model could still perform well with new data, making it more trustworthy. This process is key to confirming the robustness and clinical applicability of the model in predicting IVIG resistance in real-world scenarios.

3.1 Data preprocessing

The datasets used in this study show a clear class imbalance because IVIG resistance is rare in KD. The first dataset, used for FL, contains 1,398 patient records with 31 features, but only 158 patients (about 11.3%) are IVIG-resistant. All 31 features were used as input, as no feature selection was applied; this is because CNNs can automatically learn useful feature representations from the raw input without manual feature engineering. The second dataset, used to test the robustness and generalizability of the model, has 644 records with 82 features, and just 124 cases (around 19.3%) are IVIG-resistant. This imbalance could lead to biased models that struggle to detect resistant cases. To tackle this, we chose ADASYN [34], Adaptive Synthetic Sampling, which generates synthetic samples especially for the harder-to-learn minority cases. This helps the model focus more on difficult decision boundaries and improves detection of IVIG resistance. The number of synthetic samples S is determined by $S = \gamma \cdot (N_{maj} - N_{min})$, where N_{maj} is the number of samples in the majority class, N_{min} is the number of samples in the minority class, and γ is a factor that controls how balanced the classes should be. After balancing the data, we filled missing values with the mean for each feature and standardized all features to have a mean of 0 and a standard deviation of 1. These steps helped create clean, balanced data for training our models.

3.2 Conventional ML models and deep learning using CNN

For IVIG resistance prediction in KD patients, three conventional ML models—LR, Naïve Bayes (NB), and gradient boosting (GB)—were employed. LR predicts the

probability of the positive class using the sigmoid function. NB applies Bayes' theorem, assuming feature independence, and calculates the posterior probability of a class given the input features.GB builds a strong predictive model by iteratively combining multiple weak learners. Each new learner corrects errors from the previous iteration by minimizing a specified loss function, with the final model aggregating their outputs. These ML models provide different approaches to classification, with LR offering probabilistic interpretation, NB handling feature independence assumptions efficiently, and GB leveraging ensemble learning for improved predictive power.

Convolutional neural networks are particularly useful in detecting intricate data patterns, which is important for forecasting IVIG resistance. They can detect local relationships and form complex hierarchies of features, which enables them to discern many meaningful associations in clinical and laboratory data that most models, even some deep learning models, would overlook. CNNs are also more economical than multilayer perceptrons or recurrent neural networks because of shared sparse connections and sparse weights, which lowers the number of parameters and helps control overfitting—something crucial with small or imbalanced datasets. In our model, 1D convolutional layers capture salient features, while ReLU activation, max pooling, and dropout layers help the signal extraction focus on the most prominent features, and noise reduction improves generalization. These attributes strengthen the reliability of CNNs in resolving the IVIG resistance prediction complexities in Kawasaki disease.

3.3 Federated learning using flower framework

Federated learning enables collaborative model training across institutions while preserving data privacy, addressing the challenges posed by the rarity of KD and the limited size of single-institution datasets. The Flower framework [35] provides flexibility, seamless integration with PyTorch, TensorFlow, and Scikit-learn, and a well-documented, modular architecture that supports efficient implementation. Its client-server design allows local clients to train models on private data while a central server aggregates updates to build a global model. Flower accommodates heterogeneous clients, managing variations in data distribution, hardware configurations, and network conditions. Its lightweight structure, support for custom strategies, and compatibility with edge computing and IoT environments offer significant advantages over alternatives such as TensorFlow Federated, PySyft, and FATE, making it well-suited for this study.

This study employs the Flower FL framework with a 1D CNN to predict IVIG resistance, where each client trains locally on partitioned data and shares only model weights with a central server. CNNs were selected as the shared model in the FL framework due to their compatibility with federated parameter sharing and their ability to support gradient-based optimization and backpropagation. CNNs provide native mechanisms, such as get_weights() and set_weights(), required for federated aggregation strategies such as FedAvg. In addition to CNNs, LR was incorporated into the FL setup by adapting it into a TensorFlow model compatible with the Flower framework. Although LR does not natively support federated weight sharing, restructuring it in this way enabled the use of functions such as get_weights() and set_weights(), allowing seamless integration with aggregation strategies such as FedAvg. In contrast, traditional ML models—including naive bayes (NB), SVM, GB, XGBoost, and LightGBM—do not natively expose model weights in a form compatible with

parameter exchange in FL, making them unsuitable for direct integration with such strategies. CNNs also enable automatic learning of complex, non-linear patterns and local feature interactions, even from tabular data, thereby improving representation learning and model generalization in the federated setting.

The conventional FL architecture is illustrated in Figure 2. The system consists of distributed clients (e.g., Client 1 to Client 4), each training a local model on its private, secured dataset while retaining all raw data within the local environment. Although we worked with a single dataset, we divided it into separate parts to simulate multiple clients (such as Client 1 to Client 4), with each client holding its own exclusive portion of the data. Each client used the majority of its local data for training and reserved a smaller portion for validation during each communication round to support local model evaluation. The raw data never left the local client, ensuring privacy. After local training, clients shared only their model updates, which were combined on the server using the federated averaging [20] algorithm to create a global model. This global model was then sent back to the clients for the next round of training. The entire process, including model distribution, aggregation, and evaluation, was managed by the Flower framework.

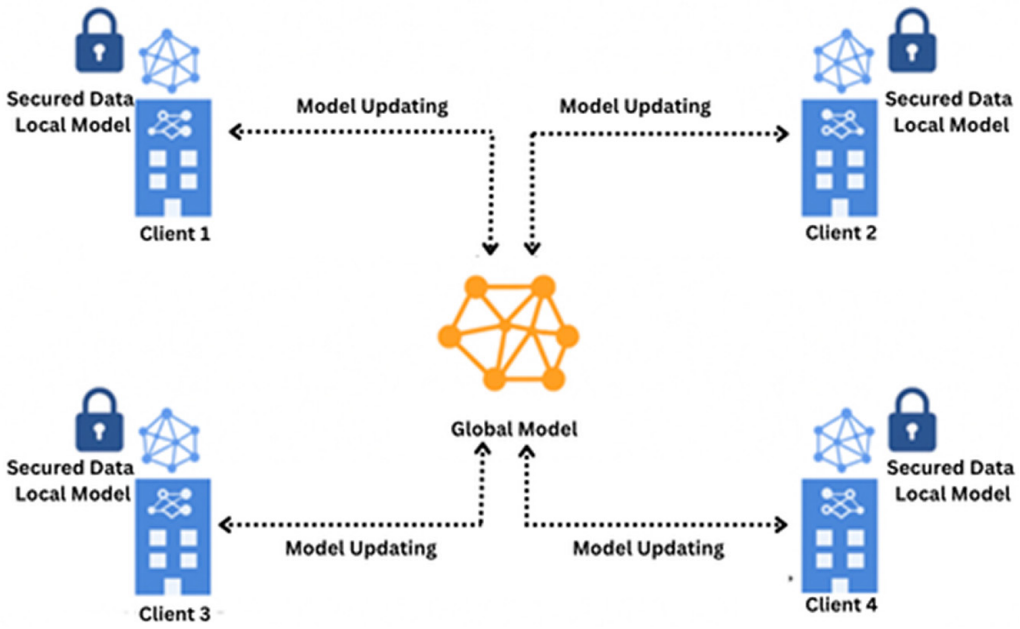

Fig. 2. Conventional federated learning architecture

Flower framework utilizes the getparameters method to fetch the current model parameter or weight from each client. This method is critical in allowing the server to obtain the most recent model state from all participating clients after every round of local training. This is done through a call to self.model.getweights(), which yields the model's weights and consequently permits the server to perform model aggregation and synchronization of the global model across clients. It enables the server to consolidate the updates that individual clients perform on their local data to refine the global model over time. Each model is trained on the dataset of each client, which is specific to them, through the fit method. Firstly, the method obtains the model weights from the server and executes self.model.setweights() to set them. Thereafter, the client trains the model for a given epochs with the local data. After the training, the client sends the updated model weights along with the number of processed

training examples to the server. This ensures that the appropriate weight is given to the updates because clients with more data will have a more profound effect on the improvement of the model.

Lastly, the evaluate method is called to evaluate the model's performance on the client's local test dataset. After getting the model parameters from the server, the client sets the local model weights using self.model.setweights(). The model is then evaluated using the evaluate() function, which computes the loss and accuracy on the local test set. The evaluation results—loss, accuracy, and the number of test samples—are returned to the server, enabling it to assess the global model's performance after each round of aggregation. By using this approach, FL allows the model to update and fine-tune the model locally on each client without revealing sensitive data to the server, which then aggregates updates to continually enhance the global model.

Each client k trains its local model θ_k using its local data $D_k = (X_k, y_k)$. The training process follows this update rule as in Eq. (1):

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

Where, $\theta_k^{(t)}$ represents the model parameters at round t for client k, η is the learning rate, and $\nabla_\theta \mathcal{L}(\theta_k^{(t)}; D_k)$ is the gradient of the loss function for the model parameters for client k. The client trains for E epochs on its local data, then sends the updated model parameters $\theta_k^{(t+1)}$ to the server for aggregation. After each round of local updates, the server aggregates the model updates using the FedAvg algorithm. The global model parameters $\theta_{global}^{(t+1)}$ are updated by computing the weighted average of the local model updates as in Eq. (2):

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

where S_t is the set of clients participating in round t, n_k is the number of training samples on client k, $\theta_k^{(t+1)}$ are the updated model parameters for client k, and $\theta_{global}^{(t+1)}$ are the aggregated global model parameters for round $t + 1$. This weighted average ensures that the updates from clients with larger datasets have a greater influence on the global model. After model aggregation, the server evaluates the global model on a test set. The evaluation metrics from each client are aggregated using the weighted average method. For accuracy, the global accuracy $Accuracy_{global}$ is computed as in Eq. (3):

$$Accuracy_{global} = \frac{\sum_{k \epsilon St} n_k Accuracy_k}{\sum_{k \epsilon St} n_k} \tag{3}$$

Similarly, for other metrics such as Precision in Eq. (4), Recall in Eq. (5), and F1-score in Eq. (6):

$$Precision_{global} = \frac{\sum_{k \epsilon St} n_k Precision_k}{\sum_{k \epsilon St} n_k} \tag{4}$$

$$Recall_{global} = \frac{\sum_{k \epsilon St} n_k Recall_k}{\sum_{k \epsilon St} n_k} \tag{5}$$

$$F1-Score_{global}=\frac{\sum_{k\epsilon St} n_k\, F1-Score_k}{\sum_{k\epsilon St} n_k} \tag{6}$$

Where $accuracy_k$, $precision_k$, $recall_k$, and $F1\text{-}score_k$ are the metrics from client k, and n_k is the number of test samples on client k. The global evaluation metric is calculated by averaging the metrics from all participating clients, weighted by their test set sizes. Evaluation of the global model on the test set D_{test} is performed after each round. The evaluation loss $L(\theta)$ is computed as in Eq. (7):

$$\mathcal{L}(\theta_{global}^{(t)})=\frac{1}{|D_{test}|}\sum_{i\epsilon D_{test}}\mathcal{L}\left(y_i,y_i'\right) \tag{7}$$

where $\theta_{global}^{(t)}$ is the global model at round t, D_{test} is the global test dataset, y_i is the true label for test sample i, and y_i' is the predicted label from the model.

4 RESULTS AND DISCUSSIONS

Results of IVIG resistance prediction using three distinct approaches: conventional ML models, a deep learning approach employing a CNN, and an FL approach using the Flower framework are presented here. Each of the methods was rigorously tested against the dataset in [11] and the dataset in [12] for model generalization and robustness. Performance of the approaches was tested and compared in terms of various evaluation metrics. Traditional ML models provide a baseline for evaluating IVIG resistance prediction. CNNs enhance accuracy through hierarchical feature extraction, while FL enables privacy-preserving, collaborative training across distributed clients, improving robustness and scalability. The comparative evaluation of these methods is elaborated in the following sections. The findings emphasize the improvements brought about by deep learning and FL over traditional models in predicting IVIG resistance outcomes accurately. The following abbreviations are used throughout the results section for the models compared in this study: LR, NB, GB, DL (Deep Learning with CNN), FL, federated learning with logistic regression (FL-LR) as shared model, and federated learning with CNN (FL-CNN) as shared model.

4.1 Conventional machine learning vs CNN (centralized implementation)

Logistic regression, NB, and GB were used as baseline ML models in correlation to their linear, probabilistic, and ensemble nature. All models were assessed based on four important metrics of measurement: accuracy, precision, recall, and F1 score. All these metrics together provided a robust evaluative framework of predictive performance, given the dataset imbalance between the majority of responsive cases and the minority of IVIG-resistant cases. Accuracy describes correct predictions in total; precision details the portion of predicted resistant cases that were actually resistant; recall, or sensitivity, describes the capability of detecting true resistant cases; and the F1 score tells us the average of the precision and recall, measuring the model's performance from both metrics. Table 2 demonstrates that all predictive traditional models show systematic underperformance across all evaluation metrics, indicative of such models' inability to realistically estimate the comprehensive and intricate relationships underpinning IVIG resistance.

The 1D CNN outperformed earlier traditional models owing to its capture of complex non-linear relationships in the given tabular data. The 1D convolutional layers provided local feature interactions and subtle pattern extraction linked to IVIG resistance. Max-pooling operations abstracted these features by emphasizing the most informative signals while reducing dimensionality. Dense layers integrated these representations for classification, with dropout regularization enhancing generalization by reducing overfitting. These architectural components contributed to performing better recall and F1 score, which are critical in this imbalanced setting that focuses on minimizing false negatives. The 1D CNN proved to be highly effective in learning intricate patterns as well as generalizing beyond the training dataset, thereby serving as a useful method for IVIG resistance prediction. With multiple epochs and these advanced learning mechanisms, the CNN captures complex patterns and generalizes better than conventional ML models, resulting in superior performance.

4.2 Proposed federated learning with convolutional neural network

Federated learning makes it possible to train models across different clients without sharing raw data, helping protect privacy. In our study, we simulated this by partitioning the dataset across multiple clients to emulate a multi-institutional setting. Each client trained its model locally, and the updates were combined to create a single global model that performs well while keeping data secure. In our study, we selected 15 communication rounds based on empirical evaluation. We experimented with various round counts (10, 15, and 20) and found that 15 rounds offered an optimal balance between performance and training efficiency, as model improvements plateaued or slightly declined beyond this point.

Federated learning-logistic regression, which used LR as the shared model, produced suboptimal results, with lower precision and F1 scores, highlighting its limited ability to capture the non-linear and complex patterns. In contrast, FL-CNN achieved superior performance by effectively learning these intricate relationships and local feature interactions within the tabular data. The CNN architecture's compatibility with FL allowed efficient parameter sharing and aggregation, leading to marked improvements in recall and F1 score. These findings, as shown in Table 2, demonstrate the clear advantage of using CNNs in federated setups for IVIG resistance prediction.

Table 2. Performance comparison of various ML, DL, and FL models on dataset 1 (primary evaluation dataset) from [11]

Metric	LR	NB	GB	DL	FL-LR	FL-CNN
Accuracy	78	75	92	89	78	94
Precision	71	60	90	83	46	93
Recall	48	56	77	83	72	89
F1-Score	57	58	82	83	56	91

The metrics improvement in FL-CNN, as shown in Table 2, such as increased accuracy of 94% and significant gains in recall and F1-score, denotes its ability to incorporate the knowledge from distributed datasets while adhering to privacy-preserving principles. Unlike standalone models, FL can unlock the potential of diverse but fragmented datasets, a crucial capability for rare diseases where individual institutions might hold only a small fraction of cases. Furthermore, privacy laws such as GDPR often restrict the sharing of sensitive medical data, limiting centralized

dataset creation. FL overcomes this barrier, allowing institutions to collaborate without exposing raw data, making it an ideal framework for generating predictive models in healthcare, particularly for rare diseases such as Kawasaki disease.

To ensure consistent performance, multiple train-test splits (80:20, 70:30, 75:25, and 85:15) were applied. In all settings, FL-CNN consistently outperformed conventional approaches, as shown in Table 3. This performance improvement is especially critical for rare diseases, where the availability of data is often limited. FL's ability to utilize decentralized datasets, each with distinct patient populations, enabled the model to generalize better and capture the variability inherent in rare disease data. Centralized DL models, on the other hand, may struggle with overfitting when trained on small, potentially biased datasets. FL also guarantees data privacy by ensuring sensitive patient data remains local to each client, which is especially crucial in healthcare use cases. Not only does this collaborative framework improve model performance, but it also resolves the privacy issues involved with health care research, thus making FL an attractive solution for use cases such as IVIG resistance prediction in rare disease scenarios.

Table 3. FL-CNN performance on dataset 1 with varying train-test splits (to assess model stability across partitioning strategies) from [11]

Train Test Split	Accuracy	Precision	Recall	F1 Score
80:20	94	88	90	90
70:30	92	89	85	87
75:25	92	87	86	86
85:15	94	93	89	91

The effectiveness of the FL approach was further evaluated using an independent dataset consisting of 82 features, as shown in Tables 4 and 5. All procedures—including conventional ML models (LR, NB, GB), CNN, FL-LR, and the proposed FL-CNN—were applied. In this scenario as well, FL-CNN consistently outperformed all other models. In Figure 3, we present a graphical representation of the performance of all models on both datasets. Given the imbalanced nature of the datasets and the critical clinical goal of identifying IVIG-resistant patients—who are most likely to benefit from additional therapy—recall (sensitivity) was prioritized as a key metric. Sensitivity reflects the model's ability to correctly identify IVIG-resistant cases, and its improvement has been highlighted as a major unmet need in prior studies, where low sensitivity has limited the clinical utility of predictive models. In this work, the reported sensitivity values from the original studies (from which the datasets were obtained) were directly referenced and compared with our proposed system, as shown in Figure 4. In both cases, the proposed FL-CNN consistently outperformed all models reported in those studies.

Table 4. Performance comparison of various ML, DL, and FL models on an independent dataset from [12] to evaluate the proposed model's consistency and generalizability

Metric	LR	NB	GB	DL	FL-LR	FL-CNN
Accuracy	75	64	83	83	75	94
Precision	67	54	78	93	82	99
Recall	82	82	83	74	66	88
F1-Score	73	66	81	82	74	93

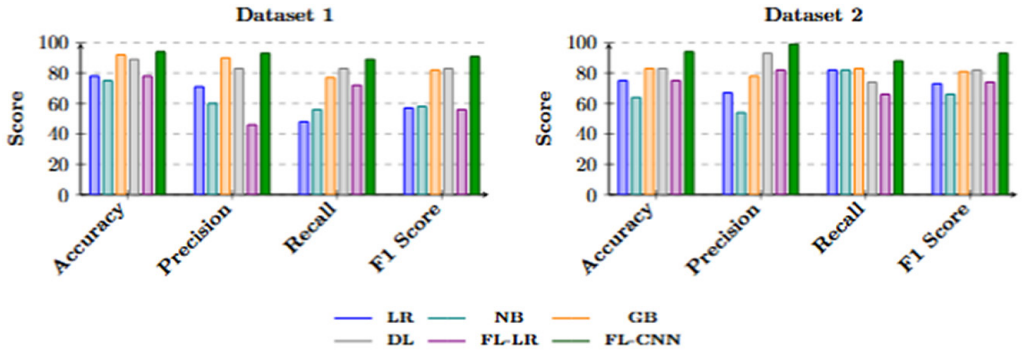

Fig. 3. Comparison of performance metrics across models using dataset 1 [11] and dataset 2 [12]

Table 5. FL-CNN performance under different train-test split strategies on an independent dataset from [12] to evaluate the proposed model's consistency and generalizability

Train Test Split	Accuracy	Precision	Recall	F1 Score
80:20	91	98	84	90
70:30	86	98	85	91
75:25	91	99	83	90
85:15	94	99	88	93

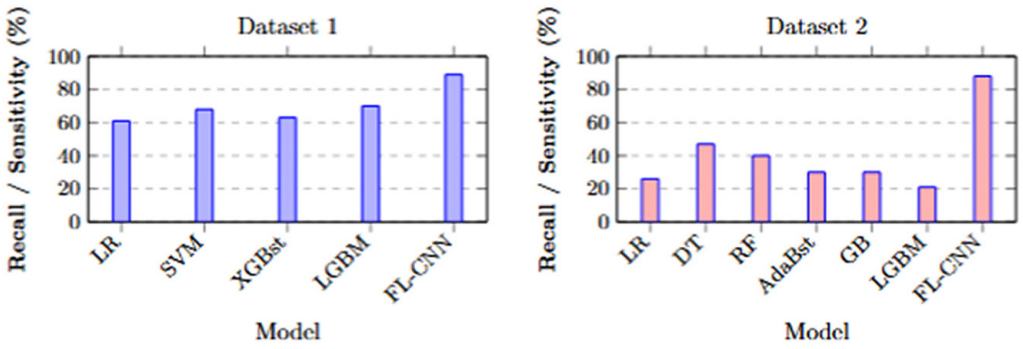

Fig. 4. Comparison of the Recall/Sensitivity results of the proposed FL-CNN with results reported in [11] and [12]

5 CONCLUSION

This study has addressed the challenges of class imbalance, data sparsity, and strict privacy regulations in predictive modeling for rare diseases, using KD, a rare and complex condition that mainly affects children under five, as a case study. Predicting resistance to IVIG—the standard treatment—is crucial because resistant patients face a much higher risk of developing serious heart complications. Our approach uses FL with the Flower framework, where a CNN serves as the shared model. This setup ensures scalability, privacy, and secure collaboration across clients. To address the class imbalance in the data, we applied adaptive synthetic oversampling. We compared the conventional ML models with the proposed FL model with CNN as the shared model (FL-CNN). Proposed FL-CNN outperformed all other methods, enabling secure and efficient model aggregation across distributed nodes. It achieved 94% accuracy, 93% precision, 89% recall, and a 91% F1 score, and showed strong

performance on an independent dataset, with 94% accuracy, 99% precision, 88% recall, and 93% F1 score. Our results show that FL is a powerful approach for predictive modeling in rare diseases, enabling collaboration without compromising data privacy. To our knowledge, this is the first application of FL in this context. Building on our use of CNN, future work can explore integrating advanced models such as transformers or graph neural networks within the FL framework to further improve performance and capture complex patterns in the data.

6 REFERENCES

[1] A. Hard *et al.*, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018. https://doi.org/10.48550/arXiv.1811.03604

[2] B. W. McCrindle *et al.*, "Diagnosis, treatment, and long-term management of Kawasaki disease: A scientific statement for health professionals from the American heart association," *Circulation*, vol. 135, no. 17, pp. e927–e999, 2017. https://doi.org/10.1161/CIR.0000000000000484

[3] N. Chantasiriwan, S. Silvilairat, K. Makonkawkeyoon, Y. Pongprot, and R. Sittiwangkul, "Predictors of intravenous immunoglobulin resistance and coronary artery aneurysm in patients with Kawasaki disease," *Paediatr. Int. Child Health*, vol. 38, no. 3, pp. 209–212, 2018. https://doi.org/10.1080/20469047.2018.1471381

[4] T. Kobayashi *et al.*, "Prediction of intravenous immunoglobulin unresponsiveness in patients with Kawasaki disease," *Circulation*, vol. 113, no. 22, pp. 2606–2612, 2006. https://doi.org/10.1161/CIRCULATIONAHA.105.592865

[5] T. Sosa, L. Brower, and A. Divanovic, "Diagnosis and management of Kawasaki disease," *JAMA Pediatr.*, vol. 173, no. 3, pp. 278–279, 2019. https://doi.org/10.1001/jamapediatrics.2018.3307

[6] 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

[7] 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, pp. 405–414, 2024. https://doi.org/10.3345/cep.2024.00549

[8] 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

[9] Y. Sunaga *et al.*, "A simple scoring model based on machine learning predicts intravenous immunoglobulin resistance in Kawasaki disease," *Clin. Rheumatol.*, vol. 42, no. 5, pp. 1351–1361, 2023. https://doi.org/10.1007/s10067-023-06502-1

[10] 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

[11] 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, 2021. https://doi.org/10.3389/fped.2021.756095

[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. 16, no. 6, 2021. https://doi.org/10.1371/journal.pone.0237321

[13] X. H. Tan *et al.*, "A new model for predicting intravenous immunoglobin-resistant Kawasaki disease in Chongqing: A retrospective study on 5277 patients," *Sci. Rep.*, vol. 9, no. 1, 2019. https://doi.org/10.1038/s41598-019-39330-y

[14] M. Cabanillas-Carbonell and J. Zapata-Paulini, "Evaluation of machine learning models for the prediction of Alzheimer's: In search of the best performance," *Brain, Behav., & Immun. Health*, vol. 44, p. 100957, 2025. https://doi.org/10.1016/j.bbih.2025.100957

[15] J. Zapata-Paulini and M. Cabanillas-Carbonell, "Performance analysis of 10 machine learning models in lung cancer prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 2, pp. 1352–1364, 2025. https://doi.org/10.11591/ijeecs.v37.i2.pp1352-1364

[16] O. Iparraguirre-Villanueva and M. Cabanillas-Carbonell, "Application of convolutional neural networks in skin disease prediction: Accuracy and efficiency in dermatological image analysis," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 21, no. 2, pp. 18–37, 2025. https://doi.org/10.3991/ijoe.v21i02.52871

[17] 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

[18] 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

[19] 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

[20] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, in Proceedings of Machine Learning Research (PMLR), vol. 54 , 2017, pp. 1273–1282. Accessed: Jul. 09, 2025. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[21] T. N. Namitha, S. Raghavendra, and R. Vinith, "Privacy-preserving federated learning in healthcare: Fundamentals, state of the art and prospective research directions," in *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, 2024, pp. 1438–1443. https://doi.org/10.1109/AIC61668.2024.10730929

[22] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, BrainLes 2018*, in Lecture Notes in Computer Science, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds., vol. 11383, Springer, Cham, 2019, pp. 92–104. https://doi.org/10.1007/978-3-030-11723-8_9

[23] C. Ju, D. Gao, R. Mane, B. Tan, Y. Liu, and C. Guan, "Federated transfer learning for EEG signal classification," in 2020 *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2020, pp. 3040–3045. https://doi.org/10.1109/EMBC44109.2020.9175344

[24] D. Yang *et al.*, "Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan," *Med. Image. Anal.*, vol. 70, p. 101992, 2021. https://doi.org/10.1016/j.media.2021.101992

[25] H. Elayan, M. Aloqaily, and M. Guizani, "Sustainability of healthcare data analysis IoT-based systems using deep federated learning," *IEEE Internet Things J.*, vol. 9, no. 10, pp. 7338–7346, 2021. https://doi.org/10.1109/JIOT.2021.3103635

[26] W. Li *et al.*, "Privacy-preserving federated brain tumour segmentation," in *Machine Learning in Medical Imaging*, *MLMI 2019*, in Lecture Notes in Computer Science, HI. Suk, M. Liu, P. Yan, and C. Lian, Eds., vol. 11861, 2019, pp. 133–141. https://doi.org/10.1007/978-3-030-32692-0_16

[27] X. Tan, T. Ma, and T. Su, "Fast and privacy-preserving federated joint estimator of multi-sUGMs," *IEEE Access*, vol. 9, pp. 104079–104092, 2021. https://doi.org/10.1109/ACCESS.2021.3099400

[28] J. Cui, H. Zhu, H. Deng, Z. Chen, and D. Liu, "FeARH: Federated machine learning with anonymous random hybridization on electronic medical records," *J. Biomed. Inform.*, vol. 117, p. 103735, 2021. https://doi.org/10.1016/j.jbi.2021.103735

[29] F. Zhu *et al.*, "Model-level attention and batch-instance style normalization for federated learning on medical image segmentation," *Information Fusion*, vol. 107, p. 102348, 2024. https://doi.org/10.1016/j.inffus.2024.102348

[30] E. Czeizler *et al.*, "Using federated data sources and varian learning portal framework to train a neural network model for automatic organ segmentation," *Physica Medica*, vol. 72, pp. 39–45, 2020. https://doi.org/10.1016/j.ejmp.2020.03.011

[31] A. Linardos, K. Kushibar, S. Walsh, P. Gkontra, and K. Lekadir, "Federated learning for multi-center imaging diagnostics: A study in cardiovascular disease," *Research Square*, 2021. https://doi.org/10.21203/rs.3.rs-688924/v1

[32] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *Int. J. Med. Inform.*, vol. 112, pp. 59–67, 2018. https://doi.org/10.1016/j.ijmedinf.2018.01.007

[33] S. Otoum, I. Al Ridhawi, and H. T. Mouftah, "Preventing and controlling epidemics through blockchain-assisted AI-enabled networks," *IEEE Netw.*, vol. 35, no. 3, pp. 34–41, 2021. https://doi.org/10.1109/MNET.011.2000628

[34] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[35] D. J. Beutel *et al.*, "Flower: A friendly federated learning framework," *arXiv preprint arXiv:2007.14390*, 2020. https://arxiv.org/pdf/2007.14390

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).