International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Federated Learning Approach for Predicting the Growth Rate and Menace of COVID-19

Ram Prasad Reddy Sadi*[1], B. V. A. N. S. S. Prabhakar Rao[2], Rabindra Kumar Singh[3], Chadrika Dadhirao[4], Kupukotla Satish Kumar[5], Chengamma Chitteti[6]

Submitted: 17/11/2022 **Revised**: 16/01/2023 **Accepted**: 15/02/2023

Abstract: When attempting to use digital clinical data to predict the spread and threat of COVID-19, data available at a particular site is not sufficient for detecting COVID-19 detection. It also includes certain issues that include integrating data from multiple sources, and the concerns relevant to privacy while handling centralized database that comprises of sensitive data. Provides a framework which involves federated learning approach, that may use locally stored clinical data from several sites to develop a centralized COVID-19 prediction model. Suggest two unique approaches to local model aggregation to enhance the global model's predictive performance. This suggested method achieves performance on par with centralized learning and is better than localized learning models through extensive experimental assessment utilizing real-world health data from government sites. Additionally, aggregate approaches beat novel techniques in terms of Recall, Accuracy and Precision for a wide range of data distributions.

Keywords: *Federated Learning, EHR, COVID-19, SVM, Logistic Regression, Single-Layer Perceptron*

1. Introduction

Predicting the menace of COVID-19 is a significant concern for medical practitioners, the healthcare system, and the pharmaceutical industry. As patients can experience expected and sometimes unexpected symptoms from COVID-19, delayed detection of the disease can pose life-threatening risks to corona victims, posing unrest in society and the government. Clinical data, such as claims and electronic health records (EHR), has become common in providing rich insights into health services and supporting clinical investigation. Advancements in machine learning and artificial intelligence have produced several analytic methods that can be applied to high-dimensional data for impediment study[1]. However, making timely and accurate predictions remains a challenge. Due to the distributed nature of healthcare data, generating huge volumes of data for identifying rarely occurring events demands integrating of data across various sites. The analysis generated from various data sources can be conflicting or imprecise, necessitating methods to appropriately aggregate results.

Prior work to resolve these issues often has limitations in its approach. The medical research centers of the nation collect clinical data into a traditional, centralized database. A single database approach is the most straightforward way to explore the menace of the disease, but information owned by different entities is seldom shared due to significant privacy concerns. Moreover, creating and maintaining such a large data repository incurs resource and system-level constraints, including high latency and single points of vulnerability (failure, breach). To avoid such overhead and risks, the Indian Council of Medical Research created the sentinel system to monitor the safety of its regulated products using a distributed data network [2]. The network comprises multiple stakeholders, each maintaining a large claims database. Despite the distributed framework and large-scale data amassed from the active participation of data partners, it has limited analytic capabilities. Limitations of other progressive systems include access to potentially small-scale, sparse, and low-quality hospital records [3]. In addition, current claims-based frameworks experience a time lag between COVID-19 instances, claim submission, adjudication, and consolidation of the claim into a database. EHR data, collected in near real-time, is, therefore, a promising alternative but comes with the quality as mentioned earlier concerns. Hence, there is an unmet need for accurate, scalable, and efficient solutions for predicting COVID-19 using distributed health data that protects patients' privacy.

To address this challenge, we present a federated learning-based framework that permits health data to be distributed across multiple sites. Federated learning [4] has brought a paradigm shift in constructing machine learning models from distributed data sources maintained by various organizations. Under such a collaborative and decentralized setting, each site contributes to the computation of a global model while shielding its data from leakage to distrusted third parties. Our methodology enables us to train a global model using each site's local data without transferring the raw data from each site. To the preeminent of our ability, this is the first implementation of federated machine learning algorithms that leverage distributed digital clinical data for predicting the menace of COVID-19. COVID-19 prediction

[1] *Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam–530041, INDIA*
ORCID ID: 0000-0003-1815-246X
[2] *School of Computer Science and Engineering, Vellore Institute of Technology, Chennai – 600127, INDIA*
ORCID ID: 0000-0002-7695-8827
[3] *School of Computer Science and Engineering, Vellore Institute of Technology, Chennai – 600127, INDIA*
ORCID ID: 0000-0002-6587-365X
[4] *Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam-530041, INDIA*
ORID ID: 0000-0002-2205-9447
[5] *Department of Computer Science Engineering, Mahatma Gandhi Institute of Technology (A), Hyderabad –500075 ,INDIA*
ORCID ID: 0000-0002-4615-8300
[6] *Dept. of Data Science, School of Computing, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati– 517102, INDIA ORCID ID:0000-0002-8677-9064*
** Corresponding Author Email: reddysadi@gmail.com*

itself brings significant challenges for federated learning due to the huge imbalance between the majority class of individuals who do not have any of the presented symptoms of the disease and the minority class of individuals with severe symptoms. To address this issue, we introduce two methods with a novel approach for aggregating model updates from the sites and collating their performance with the progressive alternative.

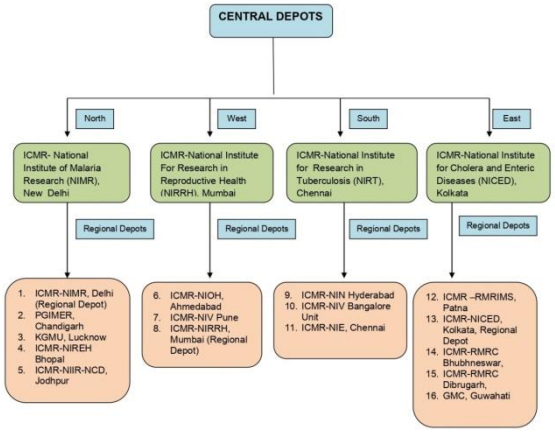

Fig. 1. The governance structure of the National Clinical Registry of COVID-19 (Source: https://www.icmr.gov.in/img/ncr/cncr_2.jpg)

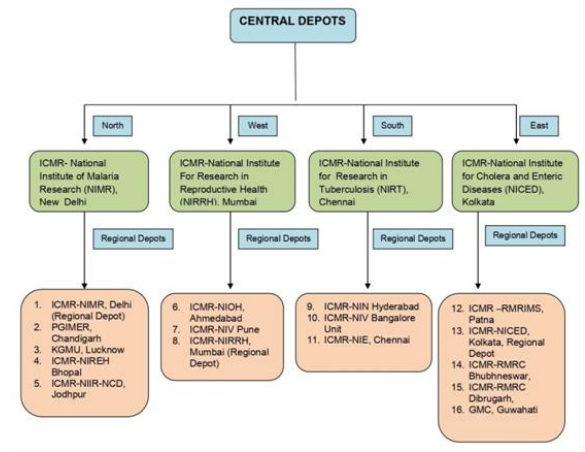

Fig. 2. Central Depots of COVID-19 (Source: https://www.icmr.gov.in/cdepot.html)

Figure 1 presents the organization in-charge of administering the COVID-19 national clinical registry. The figure indicates how the data is being managed at various levels. The monitoring, central implementation, and data management committees handle the data. The data is collected through clinical registry centers by capturing the data from the satellite centers. The data is centrally deposited at the main center after passing through the steering committee.

The data is collected from the respective zones, such as north, west, south, and east, where respective institutes monitor each zone. These zones are a collection of regional depots, as shown in figure 2. The over scenario indicates that the data is centrally being collected and stored.

To show the effectiveness of our proposed approach, we consider one use case i.e. prediction of growth rate and menace of COVID-19 across various centers. We conduct a comprehensive experimental evaluation using real-world patient data.

The key contributions of our work include in the following steps: Step 1: implementing federated models for prediction of the growth and menace of COVID-19 based on three supervised learning algorithms; Step 2: proposing and implementing two novel methods of aggregating local model updates in a federated setup; Step 3 demonstrating the effectiveness of our approach in analyzing sensitive, distributed, and highly imbalanced real-world digital clinical data; Step 4 conducting a comparative analysis to evaluate our approach against progressive alternatives; and Step 5 demonstrating scalability of our approach for varying number of sites, data size, and data distribution.

2. Background

In real-world circumstances, the sensitive and dispersed nature of electronic health information necessitates a method that can learn from data lying in silos while maintaining data privacy. This pushes us to investigate the capability and utility of federated learning for forecasting the development rate and threat of COVID-19. Federated learning allows the training of a global model from dispersed data without requiring the exchange of sensitive raw data across locations. The global model is dispersed to each location, where a local instance is trained. The changes from locally trained instances are then combined to enhance the globally developed model, which is subsequently notified to the participating sites for a second training session. Figure 3 depicts the conclusion of this iterative procedure when a performance condition is reached.

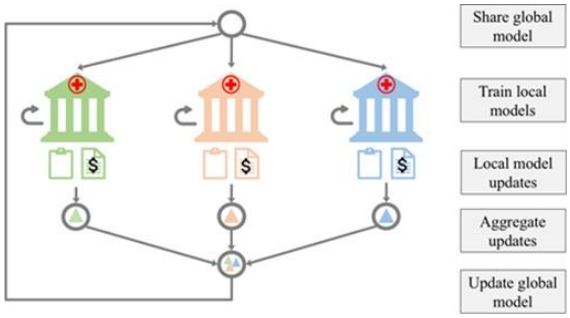

Fig. 3. System design of federated learning for prediction of growth rate and menace of COVID-19

Each Multiple patients' electronic health records are kept at each location. Once the parameters of the global model are provided to the participating sites, the model is trained using the local data of each site. The parameters of the local models are combined in order to enhance the global model. This is continued until a convergence requirement for the global model is met.

Initial implementations of federated learning were intended for image classification and language modeling on mobile devices [4, 5]. Existing literature aims to improve the performance of deep networks in a federated setting [6-9]. There is currently very limited research focusing on the application of federated learning in healthcare. Recent work noted the effectiveness of federated models in predicting hospital admissions using EHR data [10]. However, the potential of federated learning in healthcare applications that make use of claims or EHR data for prediction of growth rate and menace of COVID-19 is yet to be explored. Moreover, the existing method of aggregating updates from local models [4] relies on the size, rather than the inherent characteristics, of the data. This approach may not work well in healthcare applications, which often deal with skewed, sparse, and imbalanced datasets. Hence, exploring the underlying characteristics of distributed data to better the prediction accuracy of the global model is also an important research direction.

The usage of machine learning and deep learning methods for predicting various diseases [19-31] have been discussed by various authors. All the authors have considered centralized approach for analysis the results for various types of diseases.

Unlike the methods that focus on trend of COVID-19 [17-18] for given medications, specific prediction typically employs supervised learning algorithms. Decision Trees (DT), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Neural Networks(NN) are often used methods. Prior works on COVID-19 with machine learning methods are largely limited to centralized models, where all data are available to the researcher in a centralized data store. A majority of these works also lack evaluation on real-world datasets. For instance, distributed logistic regression based on multi-party computation, was studied using simulated data.

In this paper, we use three supervised learning algorithms viz., logistic regression, single-layer perceptron, and support vector machines to realized federated learning model, using stochastic gradient descent (SGD)-based optimization, which provides a generic approach for the algorithms to learn local models and aggregate their parameters to improve the global model and is the method currently supported by federated learning.

3. Methods

3.1 Data and Cohort Selection

To evaluate our approach, we collected data from clinical centers and used the data from
https://www.kaggle.com/imdevskp/corona-virus-report/data.

3.2 COVID-19 Prediction Model

The model explores the existing binary classification problem. The features are selected from the feature space X and are represented as x_k (for the k^{th} feature). The labels necessary for forming y_k are picked up from the label space Y := $\{-1, 1\}$. The features representing positive labels are denoted by X_+ and the feature relevant to negative labels are denoted by X_-, that is

$X_+ = \{x_k \in X: y_k = +1\}\ and\ X_- = \{x_k \in X: y_k = -1\}$

For any $x_k^+ \in X_+ and\ x_k^- \in X_-$, the primary goal of the binary classification is to build a function $f: X \rightarrow Y$ such that

$$f(x_k^+) = +1\ and\ f(x_k^-) = -1$$

In this paper, we denote y_k = +1to indicate the positive COVID-19 cases, and $y_k = -1$ to represent negative non-COVID-19 cases.

3.3 Cost Sensitive Learning

Class imbalance is intrinsic to COVID-19 prediction. Since most classification algorithms assume balanced class distributions or equal misclassification costs, they fail to represent the characteristics of imbalanced data and are more likely to classify new observations to the majority class [11]. For COVID-19 prediction, the cost of a false negative classification should be much higher than that of a false positive classification. Recent work on imbalanced learning can be categorized into sampling methods [12], cost-sensitive methods [13], and active learning methods [14]. As discussed in [15], sampling methods, such as under sampling the majority class or oversampling the minority class, either discard potentially useful data or can lead to overfitting. Since our dataset does not comprise unlabeled samples, active learning is not applicable. Hence, to mitigate the challenge of skewed data distribution, we incorporate cost-sensitive learning, wherein we in-crease the cost associated with misclassifying a minority class sample. Specifically, if C_{FN} and C_{FP} denote the cost of false negative and false positive in a cost matrix, respectively, then we set $C_{FN} > C_{FP}$. The magnitude of cost depends on the problem at hand and we determine their values using grid search.

3.4 Centralized Model

For the purpose of binary classification of samples into COVID-19 and non-COVID-19 cases, we consider three supervised classification methods: SVM, single-layer perceptron, and logistic regression. We implemented these algorithms using scikit-learn version 0.20.2. In order to produce standard results, the performance of the classifiers in a centralized learning approach is evaluated initially. This represents the scenario of gathering data from multiple sites for training a machine learning model. For each cohort, the dataset is partitioned into training and testing datasets with 70% and 30% respectively. The training features are labeled by X_{train} and Y_{train} and the testing features are labeled as X_{test} and Y_{test}. As the splits are stratified, the proportion of positive and negative cases in each split is the same as the entire dataset. After standardizing the features, we use 5-fold cross-validation to train the models on X_{train} and Y_{train}, and test them on X_{test}. To incorporate cost-sensitive learning, we update the *class_weight* parameter in scikit-learn based on class frequencies.

3.5 Localized Model

Since healthcare and biomedical data is rife with sensitive information, sharing such data across sites or transferring it to a centralized database is often restricted. In such a scenario, the model uses its own predictive analytics measures on its own data. We consider this scenario while designing localized models for COVID-19 prediction. We train each classifier on a site's data with no dependence on the data existing at other sites. Let us suppose there are *N* sites, representing hospitals or data owners. We use horizontal partitioning to split the training data into N disjoint subsets. We partition X_{train}into $\{X_{train}^i\}_{i=1}^N$, where $\cup_{i=1}^N X_{train}^i = X_{train}\ and\ X_{train}^i \cap X_{train}^j = \emptyset, \forall_{i,j} \in \{1, \cdots, N\}, for\ i \neq j$. We follow the same logic to partition the corresponding label set Y_{train} into $\{Y_{train}^i\}_{i=1}^N$. In the case of localized learning, the classifiers are trained on a single site's data $\{X_{train}^i\}_{i=1}^N\ and\ \{Y_{train}^i\}_{i=1}^N$, and tested on X_{test}. The limited availability of data may fail to account for detection of rare events [16]. We consider the results obtained from the localized models for benchmark analysis with federated and centralized learning models.

3.6 Federated Model

In this paper, we focus on classification models that can be trained using gradient descent optimization, as currently supported by federated learning. Similarly to the scenario of localized model, for *N* sites, we randomly partition the training data into *N* disjoint subsets of feature set $\{X_{train}^i\}_{i=1}^N$ and corresponding label set $\{Y_{train}^i\}_{i=1}^N$. Let *T* denote the rounds of aggregating local round updates. For stochastic gradient descent, let η, *E*, and *Batch* denote the learning rate, number of epochs, batch based on a given batch size B, respectively. Let $F_i(W)$ be the local loss function of the i^{th} site with respect to its model parameter *w*. As described in Section 2, the universal model developed is sent to all the participating sites, where in turn the received global model in trained on the local data. During local model training, based on given η, *E*, and Batch, at each site, the average gradient $(\nabla F_i(w))$ is computed by considering the current model parameter w. The parameter updates form the local participating sites are collected and their weighted average is calculated for aggregation. The process iterates as long as the

convergence criterion is not achieved. The converge function could be something like loss function minimization. The overall training of the global model depends on the continuous updates received from the models existing at the local sites [35]. Algorithm 1 presents the core algorithm of federated learning, where the weight w_D^i is equal to $\frac{D_i}{D}$, where $|D_i|$ and $|D|$ denote the size of data at the i[th] site and the entire dataset, respectively. Such an approach may fail to consider the inherent characteristics of data distribution at the sites. For the use case of COVID-19 growth rate prediction, federated averaging would not account for imbalanced data and the varying distribution of COVID-19 cases across sites. Since such scenarios are common when dealing with real-world health data, particularly in predicting rare events, it is important to explore other aggregation approaches.

3.7 Aggregation of Local Model Updates

In this paper, we suggest two novel approaches for aggregating updates of the model trained locally. The first method is particularly designed for training data with imbalanced classes. For each site, we estimate the class ratio of its training data to assign a corresponding weight, as denoted by w_D^i. This would imply that sites with cases of rare events would have higher impact when improving the global model. For the second approach, we consider loss per sample, the change in the loss function during local model training. Since a gradient descent-based method attempts to minimize the loss function, we determine its rate of convergence [36]. This is measured by the metric *epoch*, which is the maximum number of passes over the training data until convergence. Based on each site's epoch and training data size, we assign a weight, corresponding to w_D^i, for future aggregation. Using this approach, sites that require less training samples to reach convergence faster will be assigned a higher weight during aggregation.

To evaluate these methods, we create a separate partition of the training data, based on the opioid cohort, to represent unequal distribution of class labels, as shown in Table 1. We do not conduct the same experiment with the antipsychotic cohort due to the limited number of minority class labels (COVID-19).

report the runtime incurred in training the models for each of the setups. The system with Intel(R) Xeon(R) E5-2683 v4 2.10 GHz CPU equipped with 16 cores and 64 GB of RAM configuration is used to run the experiments.

Algorithm 1 Federated Learning Model for COVID-19 growth rate Prediction

function UPDATEGLOBALMODEL
initialize w_0
for $t = 1$ **to** T **do**
for $i = 1$ **to** N **do**
    $w_{t+1}^i = UPDATELOCALMODEL(i, w_t)$
$w_{t+1} = \sum_{i=1}^{N} w_D^i * w_{i+1}^i$

function UPDATELOCALMODEL(i, w)
for $e = 1$ **to** E **do**
for $b \in Batch$ **do**
    $w = w - \eta \nabla F_i(w)$
    return w

For the set of experiments comparing centralized, localized, and federated learning models, we examine the observed differences in performance metrics in two ways: (a) by calculating the % relative error of federated learning and localized learning with respect to centralized learning when using federated averaging [5], and (b) by testing the statistical significance of the difference using the Wilcoxon signed-rank test at 0.05 significance level.

4.2 Comparative Analysis

The classification metrics such as accuracy, recall, and precision of the proposed federated learning model (FL) are compared the standard approaches centralized learning model (CL) and localized learning model (LL). The experiment is run for a finite number of times, for example 10 times, for every setup to consider aggregate classification accuracy of the models. Figures 4, 5, and 6 report these metrics for the COVID-19 dataset. As seen in Figure 4, SVM and perceptron yield similar scores and perform better than logistic regression. For all classifiers and datasets, the federated learning model performs better than the centralized learning model. At the same time, due to the lack of sufficient training data, localized models do not perform well.

Table 1 Partitioning of the COVID-19 cohort training data with varying class ratio

Site#	1	2	3	4	5	6	7	8	9	10
# COVID-19	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000
#Non-COVID-19	5,000	5,000	5,000	100,000	150,000	200,000	250,000	16,258	16,258	16,258
Class Ration	1:1	1:1	1:1	1:20	1:30	1:40	1:50	1:3.2	1:3.2	1:4.1

4. Empirical Results

This section highlights the results of the proposed approach. We evaluate the results to state the efficiency of the proposed model. We consider the classification metrics used for the study, and then compare the results to the existing models the illustrate the accuracy of the proposed system.

4.1 Evaluation Metrics

To measure the capability in terms of prediction accuracy and analysis, of the centralized, localized, and federated learning models, we compute *precision*, *recall*, and *accuracy* scores. As noted in prior work [12,32], precision and recall are better indicators for models dealing with imbalanced data. We also

It must be noted that the precision score for the antipsychotic data is higher than that for the opioid data. This is due to having a lower number of false positives, possibly because the former dataset is severely imbalanced with a class ratio of 1:65.

Figure 5 presents the recall scores of the models for the two datasets. Perceptron generated the highest recall score, followed by SVM and logistic regression. Federated learning performed as good as centralized learning and outperformed the localized learning models. Since the implementation of cost-sensitive learning reduced the cases of false negative, even with such imbalanced data, centralized and federated learning models for SVM and perceptron achieved high recall.

The relevance of accuracy metric for logistic regression, single-

layer perceptron, and the support vector machines is depicted in Figure 6 for the given dataset. Similarly, to previous observations, Perceptron and SVM perform better than logistic regression. The proposed learning algorithm yields acceptable results as compared to centralized and localized approaches, respectively.

We observe that the % relative error values from federated learning are smaller than those from localized learning (Table 3). To put the numbers into a context, a difference of 5% in recall can translate to missing 5 out of 100 COVID-19 cases compared to using centralized learning. Higher recall is desirable given the potential cost of missing severe COVID-19 cases, and therefore federated learning with low % relative error is preferred. Based on statistical testing, we observe acceptable results from federated and centralized approaches for all the considered metrics, in both opioid and antipsychotic data. On the other hand, localized learning shows no encouraging results compared to either centralized or federated learning for the three evaluated metrics (all p values < 0.05).

As previously discussed, we partition the COVID-19 cohort such that the sites have a varying distribution of COVID-19 and non-COVID-19 cases (see class ratios in Table 1). We compare the effectiveness of our two proposed aggregation methods, in terms of precision, recall, and accuracy, with respect to default averaging (without weights) and federated averaging (based on data size). As seen in Table 5, for all evaluation metrics, our methods, particularly aggregation based on loss per sample, outperforms the progressive method of aggregation. This result implies that for skewed datasets, it is very important to consider the underlying characteristics of the data when aggregating local models.

4.3 Discussion

The accessibility of electronic health data brings countless scope to investigate and predict the growth rate of COVID-19, provided that the hurdles in gathering and using such data are overcome.

Table 2. Comparison of relative error (%) for federated learning (FL) and localized learning (LL) with respect to centralized learning (CL). The values denote average (standard deviation) over 10 iterations

		Precision		**Recall**		**Accuracy**	
Dataset	Classifier	FL vs CL	LL vs CL	FL vs CL	LL vs CL	FL vs CL	LL vs CL
COVID-19	SVM	2.84(1.64)	6.44(1.59)	3.10(2.58)	8.49(3.88)	3.96(2.33)	13.34(2.50)
	Perceptron	1.11(.73)	7.16(2.49)	7.32(5.45)	9.06(6.06)	5.31(3.99)	12.88(4.58)
	LogReg	1.86(1.11)	11.28(2.56)	3.21(2.81)	11.55(3.37)	2.66(2.54)	12.55(4.49)

Table 3. Time (in seconds) incurred in training the centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression. The times denote average (standard deviation) over 10 iterations.

COVID-19			
	CL	FL	LL
SVM	612.8 (8.5)	122.2 (3.4)	4.1 (.3)
Perceptron	842.8 (9.0)	117.6 (2.9)	6.3 (.7)
LogReg	513.7 (6.4)	102.7 (3.4)	4.8 (.6)

In Table 4, we report the running time (in seconds) incurred in training the models for different setups. As expected, centralized learning requires a lot of time as it involves training the models on the entire training dataset. Federated learning requires significantly less time to train the models. Localized learning models train on a subset of the data on a single round, due to which they incur the lowest running time. For both datasets, perceptron required higher running time, compared to SVM and logistic regression. Due to the considerably large scale of opioid data, it consistently required more time to train the models.

To demonstrate the scalability of federated learning models, we further measure their predictive capability, in terms of precision and recall, for a varying number of sites and data sizes. As the number of sites increases, the size of training data residing at each site proportionally decreases. Due to the imbalanced nature of the data, this has a pronounced impact on the recall score, as evident in Figure 7. This scenario also accounts for the ability of the system to handle varying sizes of training data.

In this work, federated learning approach has been proposed and the model is evaluated to address the COVID-19 growth rate prediction frameworks based on centralized learning. We demonstrated that SVM and perceptron perform better than logistic regression with respect to precision, recall, and accuracy. Perceptron has higher recall values, making it the preferable classifier for COVID-19 growth predication, where false negatives generally have more significant consequences than false positives. We also demonstrated that the performance of federated learning models is comparable to that of centralized learning, implying that a federated learning framework can be used to predict growth rate of COVID-19 without affecting the performance of the proposed model, and thereby the addressing the limitations of the centralized model. An important finding of our evaluation regards the quality of our proposed aggregation approach with loss to sample ratio weighting, which achieves superior performance compared to progressive federated averaging. This approach is advantageous in federated learning applications with real-world health data, where severe class polarity is a norm, rather than an anomaly.

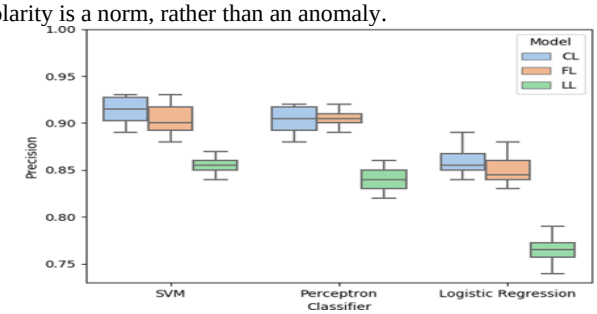

Fig. 4 Comparison of precision score for centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression with COVID-19 data

Table 4 Comparison of our proposed aggregation methods (based on class ratio and loss/sample) with respect to averaging and federated averaging methods. For SVM, perceptron (Perc), and logistic regression (LR), we report the average (standard deviation) values of precision, recall, and accuracy scores

Aggregate	Precision			Recall			Accuracy		
	SVM	**Perc**	**LR**	**SVM**	**Perc**	**LR**	**SVM**	**Perc**	**LR**
Average	0.93 (0.01)	0.91 (0.02)	0.91 (0.01)	0.63 (0.02)	0.68 (0.02)	0.59 (0.01)	0.64 (0.02)	0.68 (0.03)	0.59 (0.01)
Fed Avg	0.93 (0.02)	0.91 (0.01)	0.90 (0.02)	0.58 (0.03)	0.64 (0.01)	0.54 (0.02)	0.58 (0.01)	0.64 (0.02)	0.54 (0.02)
Class ratio	0.94 (0.01)	0.92 (0.01)	0.90 (0.01)	0.72 (0.02)	0.68 (0.01)	0.61 (0.02)	0.71 (0.01)	0.67 (0.01)	0.62 (0.01)
Loss/sample	0.94 (0.01)	0.92 (0.01)	0.91 (0.01)	0.75 (0.02)	0.69 (0.01)	0.63 (0.01)	0.74 (0.01)	0.69 (0.01)	0.63 (0.02)

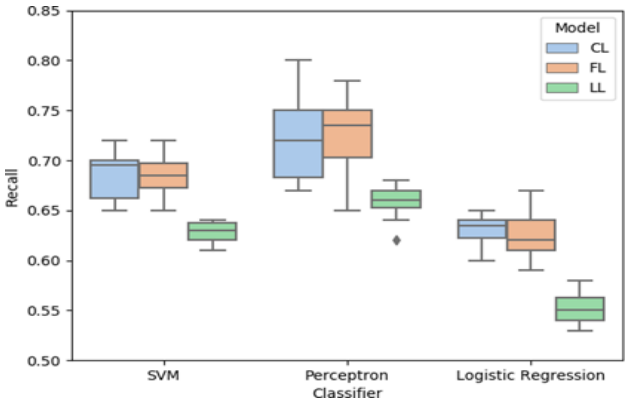

Fig. 5 Comparison of recall score for centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression with COVID-19 data

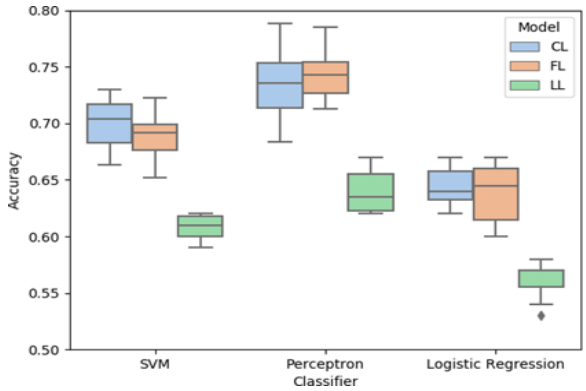

Fig. 6. Comparison of accuracy score for centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression with COVID-19 data

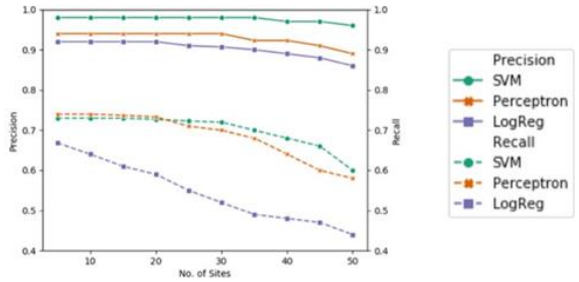

Fig. 7 Effect of varying number of sites on precision and recall scores of federated learning models (SVM, perceptron, logistic regression) with COVID-19 data

5. Conclusions and Future Work

In this paper, we focused on taxonomy of algorithms that are adaptable to distributed solution using Gradient Descent(GD), as currently supported by the federated learning paradigm. In the future, we aim to encompass other supervised learning algorithms into the proposed federated learning framework, as well as applications where large-scale distributed datasets are common and deep learning models are applicable. Will leverage other characteristics of data, such as quality, relevance, and rate of generation, to determine the impact of sites when aggregating their local model updates. Will also explore potential approaches for tuning hyper-parameters of the global model in a federated setup. Intend to work on approaches for privacy-preserving federated learning, which protect patients' privacy against adversarial attacks, in addition to not exchanging raw data while training the models.

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor and anonymous reviewers for their comments that help improve the quality of this work.

References

[1] Peter B Jensen, Lars J Jensen, &SoÃÿrenBrunak. (2012). Mining electronic health records: Towards better research applications and clinical care. Nature Reviews Genetics, 13, 395–405.

[2] Qoua L. Her, Jessica M. Malenfant, Sarah Malek, YuryVilk, Jessica Young, Lingling Li, Jeffery Brown, &SengweeToh. (2018). A Query Workflow Design to Perform Automatable Distributed Regression Analysis in Large Distributed Data Networks. eGEMs.

[3] Bruce K Bayley, Tom Belnap, Lucy Savitz, Andrew L Masica, Nilay Shah, & Neil S Fleming. (2013). Challenges in using electronic health record data forcer: Experience of 4 learning organizations and solutions applied. Medical Care, 51, S80–S86.

[4] JakubKonecˇny`, H Brendan McMahan, Felix X Yu, Peter Richtárik, AnandaTheertha Suresh, & Dave Bacon. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

[5] H Brendan McMahan, Eider Moore, Daniel Ramage, &

Seth Hampson. (2016). Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629.

[6] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, &Karn Seth. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, (pp. 1175– 1191), ACM.

[7] H Brendan McMahan, Daniel Ramage, KunalTalwar, & Li Zhang. (2017). Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963.

[8] Shiqiang Wang, Tiffany Tuor, TheodorosSalonidis, Kin K Leung, Christian Makaya, Ting He, & Kevin Chan. (2018). When edge meets learning: Adaptive control for resource-constrained distributed machine learning. In IEEE INFOCOM 2018-IEEE Conference on Computer Communications, (pp. 63–71), IEEE.

[9] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, &Rui Zhang. (2018). A hybrid approach to privacy-preserving federated learning. arXiv preprint arXiv:1812.03224.

[10] Theodora S Brisimi, Ruidi Chen, TheofanieMela, Alex Olshevsky, IoannisChPaschalidis, & Wei Shi. (2018). Federated learning of predictive models from federated Electronic Health Records. International journal of medical informatics, 112, 59–67.

[11] Haibo He and Edwardo A Garcia. (2008). Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, 9, 1263–1284.

[12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, & W Philip Kegelmeyer. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321–357.

[13] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, & Roberto Togneri. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. IEEE transactions on neural networks and learning systems, 29(8), 3573–3587.

[14] MaciejZie˛ba&Jakub M Tomczak. (2015). Boosted SVM with active learning strategy for imbalanced data. Soft Computing, 19(12), 3357–3368.

[15] Gary M Weiss, Kate McCarthy, &BibiZabar. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? (7, pp. 35–41), DMIN.

[16] Jenna Wiens, John Guttag, & Eric Horvitz. (2014). A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. Journal of the American Medical Informatics Association, 21(4), 699–706.

[17] ShreshthTuli, ShikharTuli, RakeshTuli, Sukhpal& Singh Gill. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing , Internet of Things, 11, 1-16.

[18] Punn, Narinder&Sonbhadra, Sanjay &Agarwal, Sonali.(2020). COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms. 10.1101/2020.04.08.20057679.

[19] PhaniMadhuri, N., Meghana, A., PrasadaRao, P. V. R. D., &Prem Kumar, P. (2019). Ailment prognosis and propose antidote for skin using deep learning. International Journal of Innovative Technology and Exploring Engineering, 8(4), 70-74.

[20] NarasingaRao, M. R., Sajana, T., Bhavana, N., Sai Ram, M., & Nikhil Krishna, C. (2018). Prediction of chronic kidney disease using machine learning technique. Journal of Advanced Research in Dynamical and Control Systems, 10, 328-332.

[21] Razia, S., SwathiPrathyusha, P., Vamsi Krishna, N., &SathyaSumana, N. (2018). A comparative study of machine learning algorithms on thyroid disease prediction. International Journal of Engineering and Technology(UAE), 7(2.8), 315-319.

[22] Razia, S., &Narasingarao, M. R. (2017). A neuro computing frame work for thyroid disease diagnosis using machine learning techniques. Journal of Theoretical and Applied Information Technology, 95(9), 1996-2005.

[23] Shinde, S. A., &Rajeswari, P. R. (2018). Intelligent health risk prediction systems using machine learning: A review. International Journal of Engineering and Technology(UAE), 7(3), 1019-1023.

[24] Bommadevara, H. S. A., Sowmya, Y., &Pradeepini, G. (2019). Heart disease prediction using machine learning algorithms. International Journal of Innovative Technology and Exploring Engineering, 8(5), 270-272.

[25] Srinivas, V., Aditya, K., Prasanth, G., Babukarthik, R. G., Satheeshkumar, S., &Sambasivam, G. (2018). A novel approach for prediction of heart disease: Machine learning techniques. International Journal of Engineering and Technology(UAE), 7(2.32), 108-110.

[26] Rajesh, N., Maneesha, T., Hafeez, S., & Krishna, H. (2018). Prediction of heart disease using machine learning algorithms. International Journal of Engineering and Technology(UAE), 7(2.32), 363-366.

[27] Sajana, T., &Narasingarao, M. R. (2017). Machine learning techniques for malaria disease diagnosis - A review. Journal of Advanced Research in Dynamical and Control Systems, 9(6), 349-369.

[28] Sajana, T., &Narasingarao, M. R. (2018). A comparative study on imbalanced malaria disease diagnosis using machine learning techniques. Journal of Advanced Research in Dynamical and Control Systems, 10, 552-561.

[29] Deutschmann, C; Sowa, M; Murugaiyan, J; Roesler, U; Rober, N; Conrad, K; Laass, MW; Bogdanos, D; Sipeki, N; Papp, M; Rodiger, S; Roggenbuck, D; &Schierack, P. (2019). Identification of Chitinase-3-Like Protein 1 as a Novel Neutrophil Antigenic Target in Crohn's Disease, Journal Of Crohn's& Colitis, 13(7), 894-904.

[30] Sivakumar, S.; Nayak, SoumyaRanjan; Vidyanandini, S.; Kumar, J. Ashok; &Palai, G. (2018). An empirical study of supervised learning methods for breast cancer diseases, Optik, 175, 105-114.

[31] Raghav, R. S. &Dhavachelvan, P. (2019).Bigdata fog based cyber physical system for classifying, identifying and prevention of SARS disease JournalOf Intelligent & Fuzzy Systems, 36(5), 4361-4373.

[32] Nitesh V Chawla, Nathalie Japkowicz, &AleksanderKotcz. (2004). Special issue on learning from imbalanced datasets. ACM Sigkdd Explorations Newsletter, 6(1):1–6.

[33] Kandati, D.R.; Gadekallu, T.R. Genetic Clustered Federated Learning for COVID-19 Detection. Electronics 2022, 11, 2714. https:// doi.org/10.3390/electronics11172714.

[34] Madhura Joshi, Ankit Pal, MalaikannanSankarasubbu.

"Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges" , ACM Transactions on Computing for Healthcare, 2022.

[35] Choong Seon Hong, Latif U. Khan, Mingzhe Chen, Dawei Chen, Walid Saad, Zhu Han. "Federated Learning for Wireless Networks", Springer Science and Business Media LLC, 2021.

[36] Tanzir Ul Islam, Noman Mohammed, Dima Alhadidi. "Private Federated Framework for Health Data", 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2022.