IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 6, 2022, accepted February 3, 2022, date of publication February 7, 2022, date of current version February 22, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3149477

A Calibrated Ensemble Algorithm to Address Data Heterogeneity in Machine Learning: An Application to Identify Severe SLE Flares in Lupus Patients

YIJUN ZHAO[1], (Member, IEEE), MAN QIN[1], AND APRIL JORGE[2]
[1]Department of Computer and Information Sciences, Fordham University, New York, NY 10023, USA
[2]Division of Rheumatology, Allergy, and Immunology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115, USA

Corresponding author: Yijun Zhao (yzhao11@fordham.edu)

This work was supported by NIH under Grant K23-AR-079040.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Mass General Brigham Institutional Review Board, and informed consent was waived.

ABSTRACT Motivated to address the inconsistency between the essential i.i.d. assumption in machine learning theory and the data heterogeneity in real-world applications, we propose a novel calibrated ensemble (CE) algorithm to facilitate learning with diverse data subgroups. Unlike the traditional ensemble framework in which each learner is trained independently using the entire dataset, our method exploits the strengths of various machine learning models by training them simultaneously and forming model-ergonomic data subgroups as part of the training process. Consequently, each learner is calibrated to a unique subset of data based on their individualized predictive strength. Clinically, we can interpret each model as an expert specializing in treating patients with particular disease manifestations. We evaluate the CE model in our motivating domain of identifying lupus patients with severe SLE flares using 1541 clinical encounters in the Mass General Brigham (MGB) Lupus Cohort. Our experimental results demonstrate the efficacy of our CE model across seven performance evaluation metrics compared to five individual machine learning models and regular ensemble approaches. We further utilize ANOVA and Tukey HSD post-hoc statistical analysis to discover characteristic features of individual model clusters for clinical interpretations.

INDEX TERMS Data heterogeneity, ensemble learning, machine learning, lupus, SLE.

I. INTRODUCTION

Machine learning (ML) has attracted a significant amount of interest in recent years and has become a rapidly emerging field in artificial intelligence. Conceptually, ML can be viewed as discovering the underlying pattern in a large collection of data (i.e., training examples), guided by various learning algorithms. The effectiveness of this process relies on the assumption that the collected data are drawn independently from the same distribution. For example, logistic regression (LR) [1], neural networks [2], and many other standard algorithms implicitly make this assumption in their learning processes. However, this assumption is often violated in practice. Practitioners applying machine learning to real-world data often find themselves in a common predicament: data are collected from heterogeneous sources and, consequently, have different underlying distributions. This is particularly true in the medical domain where patients may belong to distinctive subgroups with altered disease characteristics, or different physicians can introduce biases due to subjective interpretations of the clinical test or lab results [3], [4].

The subgroups of comparable subjects which can be effectively modeled as coming from the same distribution are notoriously difficult to identify. Many efforts, such as multiple-task learning [5], multi-view learning (MVL) [6], and transfer learning [7], have been made to address the idiosyncrasies in subsets of training data. Another typical approach is to group the data using descriptive features guided by domain knowledge. However, this option could lead to reduced training data, and the domain knowledge may not always be available. We provide a brief survey of these related studies in Section II.

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

In our study, we observe that different machine learning algorithms tend to focus on a different set of features in their decision-making processes, suggesting that they are targeting patients with different disease manifestations. Based on our findings, we propose an iterative approach in which an ensemble of k ML models dynamically partitions the training data into k subgroups. Model parameters and subgroups of patients best suited for fitting by individual learners are adjusted in each iteration to maximize the models' performance. Unlike the traditional ensemble framework in which each learner is trained independently using the entire dataset, our method exploits the strengths of various machine learning models by training them simultaneously and forming model-ergonomic data subgroups as part of the training process. We term our model calibrated ensemble (CE) because it is an ensemble of multiple learners and each learner is calibrated to a unique subset of data.

In addition, instances in each algorithmically induced data cluster can be interpreted as patients exhibiting certain homogeneous traits that are apprehensible to a particular learned model (i.e., expert) but relatively opaque to others. To further understand these latent disease subgroups, we utilize the ANOVA [8], Bartlett's [9], and Tukey HSD post-hoc [10] statistical tests to identify individual model clusters' characteristic features and study their clinical interpretations.

The main contribution of this paper is a novel ensemble approach to explicitly address multiple underlying data distributions in building predictive models. In the clinical setting, the resulting models can be interpreted as experts and the data associated with each model is a patient subgroup in which the expert (i.e., the model) specializes. Our method can be extended to any dataset that is a mixture of multiple distributions, which is typical in real-world applications.

Another contribution of our work is a new method to discover latent disease subgroups in clinical data. In contrast to the majority of existing approaches where data clusters are typically identified via domain knowledge before the onset of model training, our algorithm iteratively forms model-ergonomic subsets in the model training process. Subsequently, the clinical interpretation of these patient groups can be inferred by performing statistical analysis on each feature across different model clusters to identify characteristic traits, and thus, leading to potential discoveries of patient subgroups that are not yet established in the clinical setting.

We demonstrate the efficacy of our CE algorithm in our motivating domain of identifying disease flares in patients with systemic lupus erythematosus (SLE), a heterogeneous disease characterized by a range of clinical manifestations and laboratory abnormalities. The heterogeneous nature of the disease lends difficulty to accurately predicting disease flares, as does the irregular nature of real-world clinical observational data with varying distributions.

II. RELATED WORK

Addressing data heterogeneity in machine learning is an active research area because data collected from a complex real-world environment hardly follows a single underlying distribution. As alluded in Section I, researchers have resorted to techniques including domain knowledge integration [4], [11], multi-view learning (MVL) [6], [12], multi-task learning (MTL) [5], [13], and transfer learning (TL) [3], [7].

Leveraging experts' domain knowledge, Zhao *et al.* introduced a domain induced Dirichlet mixture of Gaussian processes (DI-DPMGP) model to address the patient subgroups and physician subjectivity in predicting the disease course for multiple sclerosis patients [4]. In their approach, data subgroups generated by a k-means algorithms served as hierarchical constraints to a non-parametric model. Ross *et al.* [11] proposed a novel clustering with constraints method to identify new and clinically relevant categories of lung disease. In particular, they introduced a new way of looking at subtyping/clustering by recasting it in terms of discovering associations between individuals and disease trajectories.

In the MVL domain, Liu *et al.* explored multi-view learning [6] in classifying mild cognitive impairment (MCI), an early stage Alzheimier's disease. They proposed an effective method to enhance the feature representation of multi-modal MRI data by combining multi-view information to improve the performance of MCI classification. Serra *et al.* [12] proposed a multi-view genomic data integration methodology, in which the information from different data layers (views) is integrated at the levels of the results of each single view clustering iteration.

In the MTL and TL domain, Hu *et al.* applied transfer learning to generate individualized patient models, grounded in the wealth of population data, while also detecting and adjusting for inter-patient variabilities based on each patient's own histologic data [13]. Zhao *et al.* [3] applied transfer learning techniques to address human subjectivity in predicting disease course for chronic progressive diseases.

It is worth noting that all of the aforementioned methodologies require pre-defined criteria for data groups or subtasks/views. Nevertheless, this information can often be unavailable. Our proposed method is motivated to address this limitation by identifying the latent data clusters algorithmically as part of an ensemble model's training process. Consequently, the resulting learners are calibrated to specialize in distinct subsets of data based on their individualized predictive strengths.

III. METHODS

This section illustrates our proposed calibrated ensemble model, which will iteratively partition the patients into subgroups during its model training process, leveraging individual learners' strengths.

A. CALIBRATED ENSEMBLE MODEL

1) MODEL TRAINING

We denote our training data as

$$D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)\}$$

where $\vec{x}_i = (x_i^1, x_i^2, \ldots, x_i^d)$ and $y_i \in \{0, 1\}$ $(i = 0, 1, 2., \ldots, n)$ are the attributes and the corresponding observed class labels for instance i, respectively.

Our CE algorithm starts with training k baseline models, denoted as $m_1, m_2, \ldots, m_k$, using the entire dataset D. The training process entails all necessary steps to obtain the best performing model, including 10-fold cross-validation, imbalanced data treatment, and nested 10-fold cross-validation for hyper-parameter selection as described in Section IV-C1.

Next, we will form k data clusters for the corresponding models. The cluster membership of each training instance will be set to the model group with the highest probability score for the class label. For example, assuming our CE algorithm employs three learners, m_1, m_2 and m_3 and $P(m(\vec{x}_i) = y_i)$ denotes the probability score for class y_i when model m is applied to instance $\vec{x}_i$. If $P(m(\vec{x}_i) = y_i)$ are 0.4, 0.7, 0.6 from m_1, m_2, and m_3 respectively, then $\vec{x}_i$ will be assigned to the m_2 group. In the case of a tie, the group membership is set randomly among the equal performing algorithms.

Once we have formed the initial clusters, we will retrain each model using its own group data, followed by reassigning the group membership of each instance according to the probability scores after applying the retrained models. This model training and membership assigning process is repeated until convergence, i.e., when there is no group membership change for all the data points. To prevent degenerative models, we ensure a minimum of 15 instances from each class in each cluster. As a result, a data point will not be moved to a higher performing group if its reassignment will violate the minimum instance requirement. The convergence proof of the CE algorithm is given in Section III-B. We provide the outline of the algorithm in Figure 1.

It is worth noting that our CE algorithm bears some resemblance to a k-means algorithm [14]. However, there are two fundamental differences between the two algorithms. First, k-means is an unsupervised clustering algorithm aiming to discover the underlying structure of the data, whereas CE is a supervised algorithm that exploits the strength of individual machine learning algorithms. Second, the k-means algorithm acts on a set of descriptive attributes of the dataset and the clusters are formed using similarity measures, whereas the clusters in CE are formed based on models' performance. Indeed, the descriptive characteristics of each cluster can be inferred afterward by performing statistical analysis on the obtained subgroups (see Section IV-E for details).

2) MODEL INFERENCE

We perform model inference by applying each learner to the new data point and select the prediction with the highest probability. If we consider each model as an expert specializing in treating a certain patient type, then the highest probability corresponds to the highest confidence. Assuming that each expert is reasonably skilled (i.e., better than random guessing), this approach is consistent with the design principle of the CE algorithm, i.e., for each instance, the best learner is the one with the highest predicted probability score in the corresponding label class. With this inference method, our CE model demonstrates significant performance gains over individual baseline models, as well as regular ensemble leaners. Detailed performance comparisons are presented in Section IV.

We also experimented with another intuitive inference approach, which is trying to place a new data point into the "right" algorithmic group and then apply the corresponding learned model. Nevertheless, assigning the correct group membership to a new instance can be challenging because the data clusters are formed algorithmically and there are no specific descriptions of each group. To this end, we experimented with both centroid-based and prediction-based methods. The centroid-based method calculates the cluster centroid for each algorithmic group by taking the average of all data instances in the group. A new instance is assigned to the most similar

Calibrated Ensemble Algorithm

INPUT:

- Dataset: $D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)\}$ where $\vec{x}_i = (x_i^1, x_i^2, \ldots, x_i^d)$ and $y_i \in \{0, 1\}$ $\forall i \in \{1, 2., \ldots, n\}$ are the attributes and the observed class labels of instance i, respectively.
- k ML algorithms: $M = \{m_1, m_2, \ldots, m_k\}$

DENOTE:

- $M^t = \{m_1^t, m_2^t, \ldots, m_k^t\}$: models at iteration t
- $G^t = \{g_1^t, g_2^t, \ldots, g_k^t\}$: data clusters associated with each model at iteration t
- $P(m(\vec{x}) = y)$: probability score of class y when model m is applied to instance $\vec{x}$

INITIALIZATION:

1) $t = 0$
2) Initialize M^0 by training each m_i^0 using D.
3) Initialize G^0 as follows:

$$g_i^0 = \{(\vec{x}, y) \in D | i = \arg \max_{1 \le j \le k} P(m_j^0(\vec{x}) = y)\}$$

REPEAT UNTIL CONVERGENCE

4) $t = t + 1$
5) Re-train m_i^t using g_i^{t-1}
   Set m_i^t = max(m_i^t, m_i^{t-1})
   where "max" returns the better model with higher $\sum_{\vec{x} \in g^{(t-1)}} P(m(\vec{x}) = y)$ for $m \in \{m_i^t, m_i^{t-1}\}$
6) Adjust each cluster in G^t as follows:

$$g_i^t = \{(\vec{x}, y) \in D | i = \arg \max_{1 \le j \le k} P(m_j^t(\vec{x}) = y)\}$$

OUTPUT: $M^t = \{m_1^t, m_2^t, \ldots, m_k^t\}$

FIGURE 1. Calibrated ensemble algorithm.

group measured by the Euclidean distances between the data point and the centroids. The prediction-based approach builds a 3-class classification model based on the class memberships of the training data. Both methods demonstrated worse outcomes than the above highest probability approach.

B. PROOF OF CONVERGENCE

We show that our CE algorithm will converge within a finite number of steps. Following the same notion as in Figure 1, the cost (C) of the CE algorithm is defined as follows:

$$C = \sum_{j=1}^{k} \sum_{\vec{x} \in g^j} [1 - P(m_j(\vec{x}) = y)] \tag{1}$$

where

- k is the total number of ML algorithms employed by the CE ensemble.
- g^j is the data cluster associated with model j.
- y is the ground-truth label of instance $\vec{x}$.
- $P(m_j(\vec{x}) = y)$ is the probability score of class y when model m_j is applied to instance $\vec{x}$.

Thus, C is the total deficiency in predicted probability scores of all instances with respect to their ground-truth labels. We claim that the cost function C is strictly decreasing in each iteration of steps 4) to 6) in Figure 1. This can be shown as follows.

First, in step 6), we observe that the group membership for x changes from cluster g to g' only if $P(m'(\vec{x}) = y) > P(m(\vec{x}) = y)$, which means model m' makes a more accurate prediction for $\vec{x}$. This improved performance can either correct a wrong prediction or result in a higher predicted probability score towards the ground-truth class. In both cases, the membership reassignments will decrease the algorithm's total cost C.

In step 5), we observe that all models in iteration t (i.e., m_i^t) will be retrained using the adjusted corresponding data clusters (i.e., g_i^t) to obtain m_i^{t+1}, $\forall i \in \{1, 2, \ldots, k\}$. If the cost incurred by m_i^{t+1} is higher than that of m_i^t for instances in the cluster g_i^t, we will simply keep the old model by setting $m_i^{t+1} = m_i^t$. Thus, the cost C is non-increasing in step 5).

Since the C is strictly decreasing throughout the iterations and is lower-bounded by 0, the CE algorithm converges within finite steps.

IV. EXPERIMENTAL RESULTS

In this section, we first describe our motivating task of predicting lupus flares. We then demonstrate the efficacy of the CE algorithm by comparing its performance to that of individual baseline models and regular ensemble learners. Lastly, we illustrate the interpretation of the data clusters identified by the CE model's individual learners under the clinical setting.

A. PREDICTING LUPUS FLARES

Lupus is a chronic autoimmune disease with a prevalence of at least five million people worldwide [15]. Patients suffer from various symptoms, including pain, extreme fatigue, hair loss, cognitive issues, and physical impairments that affect every facet of their lives. Systemic lupus erythematosus (SLE) is the most common form of lupus, affecting approximately 70% of lupus patients [15]. The clinical course of SLE is heterogeneous and characterized by disease flares which can range from mild to life-threatening, affecting various organ systems [16]–[18]. Such flares can lead to irreversible organ damage and lower health-related quality of life, as well as considerable economic costs.

Identifying severe SLE disease flares in real-world data could provide unprecedented insight into nuanced patterns underlying disease activity. The stratification of patients by risk for SLE flares could lead to improved clinical monitoring and targeted treatment. However, accurate assessment of lupus flares is critical but problematic in clinical trials [19]. A gold standard measure is the SELENA-SLEDAI flare index (SFI) [20], a cumulative and weighted index used to assess disease activity across 24 different disease descriptors in patients with SLE. In practice, a revised SFI (i.e., rSFI) [21], [22] is preferred, which further incorporates additional information as an expert domain driven rule-based algorithm and classifies SLE patients into mild-flare, moderate-flare, severe-flare, and no-flare categories.

Our study applies machine learning techniques to identify patients in the most severe-flare category (class 1) against the remaining mild/medium/no flare patients (class 0) based on their rSFI scores, engendering a binary classification task. This undertaking is valuable because hospitalizations are needed for the most severe SLE flares, occurring in 7% of individuals with SLE per year and accounting for most of the direct costs of SLE care [23], [24].

B. DATA AND PREPROCESSING

Our data comes from the longitudinal EHR-based Mass General Brigham (MGB) Lupus Cohort, including patients with SLE from two large academic medical centers and multiple community hospitals. These subjects were identified by a previously validated SLE phenotype study and have been followed longitudinally between 2016-2020 [25]. Our dataset consists of 1,541 clinical encounters over this period. This study was approved by the Mass General Brigham Institutional Review Board, and informed consent was waived.

We extracted a total of 203 features from patients' encounter information, lab results, and medication records. Categorical features were further processed using one-hot encoding, a technique in which an integer encoded categorical variable is converted to a set of binary variables, each of which indicates a unique value in the category [26]. One-hot encoding eliminates the artificial ordering introduced by the integer values that a machine learning algorithm could exploit erroneously.

1) IMBALANCED DATA

Our dataset is highly imbalanced with a class 1 (severe-flare) to class 0 (mild/medium/no flare) ratio of 332 to

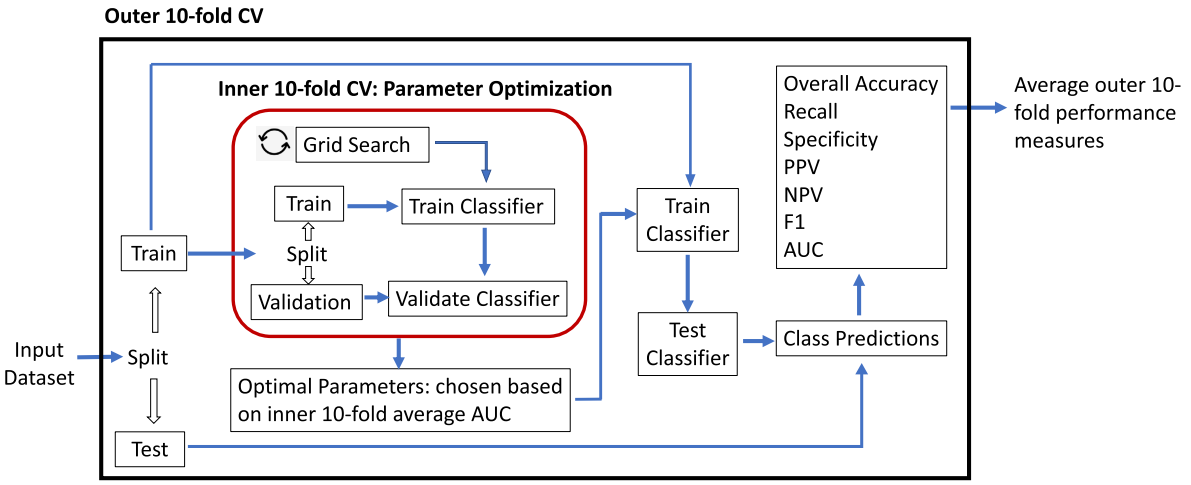


FIGURE 2. Individual model training and evaluation architecture.

1209. Applying standard machine learning algorithms to an imbalanced dataset leads to insufficient performance on the minority class, which is often the more interesting and important class under investigation. Indeed, the primary interest of our classification task is to accurately predict patients with severe SLE flares. To address the class imbalance issue, we employed the cost-sensitive learning [27] technique in our model training process. Specifically, a higher cost (i.e., weight) is assigned to all minority instances to facilitate a larger penalty when they are misclassified. For each algorithm, the best weight was selected as a hyper-parameter using a nested 10-fold cross-validation detailed in Section IV-C1.

2) MISSING VALUE IMPUTATION

There are 30 features with missing values (MVs) in our dataset, with the missing percentage ranging from 0.13% to 97%. We removed 14 features missing in over 40% of patients. For the remaining features, we imputed the missing values using the mean or mode for the numeric and categorial features, respectively. Finally, we applied z-score normalization to standardize the data.

C. EXPERIMENTAL FRAMEWORK

1) BASELINE MODELS

We employ five baseline machine learning methods: *Decision Tree* (DT), *Random Forest* (RF), *Logistic Regression* (LR), *Naive Bayes* (NB), and *XGBoost*. Figure 2 illustrates our framework for training and evaluating each baseline model. Specifically, all experiments are conducted using an outer 10-fold (black box) cross-validation. Therein, we divide the training data into ten disjoint partitions (i.e., folds), and train/evaluate each classifier ten times with different training and test data. At each iteration t, ($t = 1, 2, \ldots 10$), fold i will be designated as the test data, and the remaining nine folds will be designated as the training data. We report the average performance of the ten test folds.

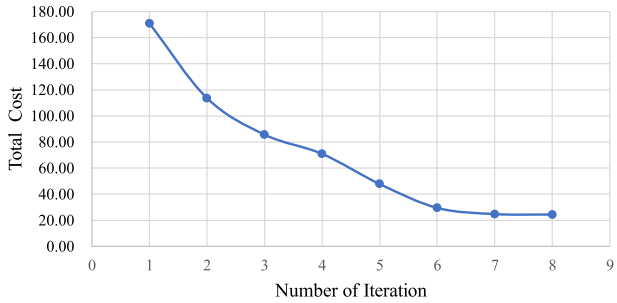


FIGURE 3. Total cost over iterations in CE model training.

Optimal hyper-parameters are selected using a grid search [28] and an inner 10-fold cross-validation (red box in Figure 2), aiming at the highest validation area under the receiver operator curve (AUC) [29]. The optimal parameter combination is then used to perform a final training on the complete 9-fold of data. The model performance measures are computed from the ground-truth and predicted class memberships based on the predictive probabilities. The performance is evaluated using seven metrics: overall accuracy, recall, specificity, PPV, NPV, F1, and AUC.

2) LEARNERS FOR THE CALIBRATED ENSEMBLE MODEL

We selected RF, LR, and NB to be the learners for our CE model based on a study of baseline models' principal predictors. Specifically, in our experiments, we observed that three (i.e., RF, LR, NB) out of these five algorithms exhibit notably different principal predictors, while DT and XGboost have a significant number of common predictors as the RF algorithm. Thus, we conjectured that RF, LR, and NB focus on different disease characteristics and target different patient subgroups. To capitalize on our findings, we chose RF, LR, and NB for our CE algorithm.

D. MODEL PERFORMANCE EVALUATION

Our CE model converged after eight iterations with 135, 634, and 622 instances in the RF, NB, and LR clusters, respectively. The total cost function, as defined in Equation (1),

TABLE 1. Model performance comparison.

Methods	Overall	Recall	Specificity	PPV	NPV	F1	AUC
Individual Models							
RF	0.62	0.78	0.58	0.34	0.91	0.47	0.68
LR	0.68	0.75	0.66	0.39	0.91	0.51	0.71
NB	0.69	0.82	0.65	0.41	0.93	0.54	0.74
DT	0.66	0.70	0.65	0.36	0.89	0.47	0.67
XGBoost	0.68	0.82	0.64	0.39	0.93	0.53	0.73
Average 1	**0.67**	**0.77**	**0.64**	**0.38**	**0.91**	**0.50**	**0.71**
Traditional Ensemble of RF, LR, NB, DT, and XGBoost							
Ensemble (soft)	0.71	0.82	0.68	0.39	0.94	0.53	0.75
Ensemble (hard)	0.66	0.84	0.61	0.35	0.94	0.49	0.73
Average 2	**0.69**	**0.83**	**0.65**	**0.37**	**0.94**	**0.51**	**0.74**
Gain (%) over Average 1	**3%**	**8%**	**2%**	**-3%**	**3%**	**2%**	**5%**
Calibrated Ensemble of RF, LR, and NB							
Calibrated Ensemble (CE)	**0.74**	**0.86**	**0.71**	**0.55**	**0.93**	**0.67**	**0.79**
Gain (%) over Average 1	**10%**	**12%**	**11%**	**45%**	**2%**	**35%**	**11%**
Gain (%) over Average 2	**7%**	**4%**	**9%**	**49%**	**-2%**	**32%**	**6%**

Performance comparison of five individual models, traditional ensembles, and the calibrated ensemble using seven evaluation metrics. Row "Average 1" presents the average scores of five individual models for each metric. The "hard" ensemble performs inference via a majority vote from the base learners' decisions, while the "soft" ensemble calculates the average predicted probability scores from the base learners and thresholds it at 0.5 to make a decision. Row "Average 2" presents the average performance of the two traditional ensemble approaches.

monotonically decreased from 171.02 to 24.38 across the iterations (Figure 3). Table 1 presents the main results of our study. Each model's performance is evaluated using seven metrics shown in columns 2-7.

From the row labeled "Average 1", we observe that the five individual models achieved an average overall performance of 67% test accuracy with 77% and 64% in Recall and Specificity, respectively. The average PPV and NPV scores are 0.38 and 0.91; their large discrepancy can be explained by the highly imbalanced data in our study. The average F1 score and AUC are 0.50 and 0.71, respectively.

Compared to the individual model, our CE methods demonstrated significant advantage across all seven evaluation metrics as shown in the row labeled "Gain (%) over Average 1" in the CE category. Specifically, the improvement in overall accuracy is 10% with 12% and 11% in class 1 and class 0, respectively. The gains in PPV, NPV, F1 scoare, and AUC are 45%, 2%, 35%, and 11%, respectively.

In addition to individual models, we further compared our CE model's performance to a traditional ensemble of the five baseline learners. To this end, we employed two types of inference for the regular ensemble model. The "hard" inference performs a majority vote from the base learners' final classification decisions (i.e., class 1 or 0), while the "soft" inference calculates the average predicted probability scores from five base learners and thresholds it at 0.5 to make the decision. We observe from Table 1, row "Average 2", that the regular ensemble learner made a noticeable improvement over the average of individual models in predicting class 1 (i.e., Recall, 8%), but the gain in class 0 is limited (Specificity, 2%). Other improvements are in NPV (3%), AUC (5%), and F1 score (2%). There is a 3% drop in NPV.

Last, we compare our CE model's performance to that of the ensemble approach. From row "Gain (%) over Average 2", We observe that the CE method outperformed the traditional ensemble approach with a 7% improvement in overall accuracy, 4% and 9% in class 1 and class 0, respectively. Most noticeably, the CE model offered a 49% gain in PPV with a marginal 2% trade-off in NPV, resulting in a 32% improvement in F1 score. The AUC improvement over the average ensemble approach is 6%.

E. CHARACTERISTIC FEATURE ANALYSIS

In this section, we present our study of the characteristics of patient subgroups formed by the CE model. We first aimed to identify features whose cluster means were statistically different among the model-specific groups. To this end, for each feature, we applied a one-way Analysis of Variance (ANOVA) test [8], which is an extension of the Student t-test for more than two groups. Specifically, ANOVA compares the means among the groups and determines whether any of those means are statistically significantly different from each other. Formally, for our application, it tests the null hypothesis:

$$H_0 : \mu^i_{RF} = \mu^i_{NB} = \mu^i_{LR}$$

where μ^i_{RF}, μ^i_{NB}, and μ^i_{LR} denote the average value of feature x_i over instances in the RF, NB, and LR subgroups, respectively. We further applied Bartlett's test [9] to ensure the homogeneity of variances assumption in the ANOVA analysis.

TABLE 2. Characteristics features of patient subgroups.

Subgroup	Characteristic Features	Cluster Means			p-values				
		RF	NB	LR	ANOVA	Bartlett's	Tukey HSD test		
							RF vs NB	RF vs LR	NB vs LR
RF	Initial - gastrointestinal	**0.07**	**0.00**	**0.01**	$< 10^{-4}$	$< 10^{-4}$	0.0010	0.0010	0.4341
	Initial - lupus headache	**0.08**	**0.02**	**0.03**	0.0035	$< 10^{-4}$	0.0023	0.0235	0.4406
	Initial - pleuritis	**0.27**	**0.12**	**0.12**	$< 10^{-4}$	$< 10^{-4}$	0.0010	0.0010	0.9000
	Historic - gastrointestinal	**0.08**	**0.02**	**0.01**	$< 10^{-4}$	$< 10^{-4}$	0.0010	0.0010	0.8354
	Historic - leukopenia	**0.04**	**0.17**	**0.19**	0.0001	$< 10^{-4}$	0.0010	0.0010	0.6462
	Current - anemia (hemolytic)	0.05	0.01	0.01	0.0162	$< 10^{-4}$	0.0158	0.0178	0.9000
	Current medication - sulfasalazine	**0.08**	**0.00**	**0.00**	$< 10^{-4}$	$< 10^{-4}$	0.0010	0.0010	0.9000
	Current medication - methotrexate	**0.18**	**0.07**	**0.09**	0.0005	$< 10^{-4}$	0.0010	0.0042	0.4385
	Arthritis details - polyarticular inflam. arthritis	**0.42**	**0.12**	**0.15**	$< 10^{-4}$	$< 10^{-4}$	0.0010	0.0010	0.2055
	lovenox	0.05	0.00	0.00	0.0000	$< 10^{-4}$	0.0010	0.0010	0.7860
NB	Initial - inflam. arthralgias/arthritis	0.76	0.59	0.80	$< 10^{-4}$	$< 10^{-4}$	0.0010	0.5114	0.0010
	Initial - nephritis	**0.10**	**0.18**	**0.09**	$< 10^{-4}$	$< 10^{-4}$	0.0200	0.9000	0.0010
	Initial - raynauds	**0.21**	**0.11**	**0.18**	0.0003	$< 10^{-4}$	0.0090	0.6454	0.0013
	Current - pericarditis	0.00	0.03	0.00	0.0010	$< 10^{-4}$	0.0408	0.9000	0.0019
	Initial lab - B2-glycoprotein 1 ab	**0.02**	**0.12**	**0.04**	$< 10^{-4}$	$< 10^{-4}$	0.0010	0.5701	0.0010
	phenotype - MCTD	**0.08**	**0.02**	**0.07**	0.0001	$< 10^{-4}$	0.0161	0.9000	0.0010
	warfarin	**0.03**	**0.08**	**0.02**	$< 10^{-4}$	$< 10^{-4}$	0.0242	0.8420	0.0010
LR	Initial - photosensitivity	**0.09**	**0.10**	**0.20**	$< 10^{-4}$	$< 10^{-4}$	0.9000	0.0031	0.0010
	Initial - other rash	**0.22**	**0.25**	**0.13**	$< 10^{-4}$	$< 10^{-4}$	0.7430	0.0315	0.0010
	Initial lab - SSA	**0.24**	**0.26**	**0.37**	0.0001	0.0419	0.8765	0.0124	0.0010
	Initial lab - SSB	**0.12**	**0.13**	**0.24**	$< 10^{-4}$	$< 10^{-4}$	0.9000	0.0029	0.0010
	Initial lab - RNP	0.22	0.28	0.45	$< 10^{-4}$	0.0064	0.3335	0.0010	0.0010
	Initial lab - dsDNA	**0.58**	**0.59**	**0.74**	$< 10^{-4}$	0.0077	0.9000	0.0010	0.0010
	Historic - photosensitivity	**0.08**	**0.10**	**0.18**	$< 10^{-4}$	$< 10^{-4}$	0.7091	0.0055	0.0010
	Historic - inflammatory arthralgias/arthritis	**0.43**	**0.44**	**0.72**	$< 10^{-4}$	0.0257	0.9000	0.0010	0.0010
	Historic - antiphospholipid syndrome	**0.17**	**0.14**	**0.04**	$< 10^{-4}$	$< 10^{-4}$	0.4226	0.0010	0.0010
	Historic - neurologic	0.01	0.02	0.05	0.0028	$< 10^{-4}$	0.7105	0.0327	0.0082
	Current - neurologic	0.00	0.01	0.03	0.0049	$< 10^{-4}$	0.8282	0.0608	0.0099
	aspirin 81 mg	0.10	0.07	0.03	0.0001	$< 10^{-4}$	0.4351	0.0022	0.0010
Unique to all subgroups	Arthritis details - arthralgias only	0.33	0.12	0.21	$< 10^{-4}$	$< 10^{-4}$	0.0010	0.0019	0.0010
	Current - inflam. arthralgias/arthritis	0.76	0.26	0.40	$< 10^{-4}$	0.0087	0.0010	0.0010	0.0010

Statistically unique features (Column 2) of each model-specific subgroup (Column 1) with a 95% confidence interval for the statistical tests. The p-values from the ANOVA and Barlett's tests indicate if the three group means are statistically different. We infer the features unique to each algorithmic cluster from the p-values of the Tuck HSD test (last three columns). For example, the feature "Initial - pleuritis", is statistically *insignificant* between the NB and LR clusters (p-value = 0.9) but *significant* for RF vs. NB (p-value = 0.001) and RF vs. LR (p-value = 0.001), suggesting it is a characteristic feature for the RF subgroup. Bold features are plotted in Figure 4.

A statistically significant result (i.e., rejecting hypothesis H_0) from an ANOVA analysis indicates at least one group differs from the other groups. However, the omnibus test does not inform where the significance lies. To further analyze the pattern of difference between means, we performed the Tukey HSD ("Honestly Significant Difference") *post-hoc* test [10] for those statistically significant features. The Tukey HSD test is similar to a pairwise t-test, but more reliable for data with more than two independent groups.

Table 2 presents the statistically significant features (Column 2) of each model-specific subgroup (Column 1) with a 95% confidence interval for all three tests. The cluster means and p-values for each feature are displayed in Columns 3-5 and Columns 6-10, respectively. In particular, the last three columns in Table 2 present the pairwise p-values of the three clusters, from which we inferred the features unique to each algorithmic cluster. For example, we observe that the pairwise p-values for the feature "Initial - pleuritis" are 0.001, 0.001, and 0.9 for RF vs. NB, RF vs. LR, and NB vs. LR, respectively. Thus, this feature is statistically *insignificant* between the NB and LR clusters (p-value $= 0.9$) but *significant* for RF vs. NB (p-value $= 0.001$) and RF vs. LR (p-value $= 0.001$), suggesting it is a characteristic feature for the RF subgroup. As illustrated in Table 2, there is a total of 10, 7, and 12 features unique to the RF, NB, and LR subgroups, respectively. Lastly, two features (i.e., "arthralgias only" in arthritis details and current symptom of inflammatory arthralgias/arthritis, are used in all three clusters.

F. CLINICAL INTERPRETATIONS OF DATA SUBGROUPS

We observe in Table 2 that the patients in the NB cluster have nearly twice higher initial manifestation of nephritis (RF:0.10, NB:0.18, LR:0.09) than the other clusters

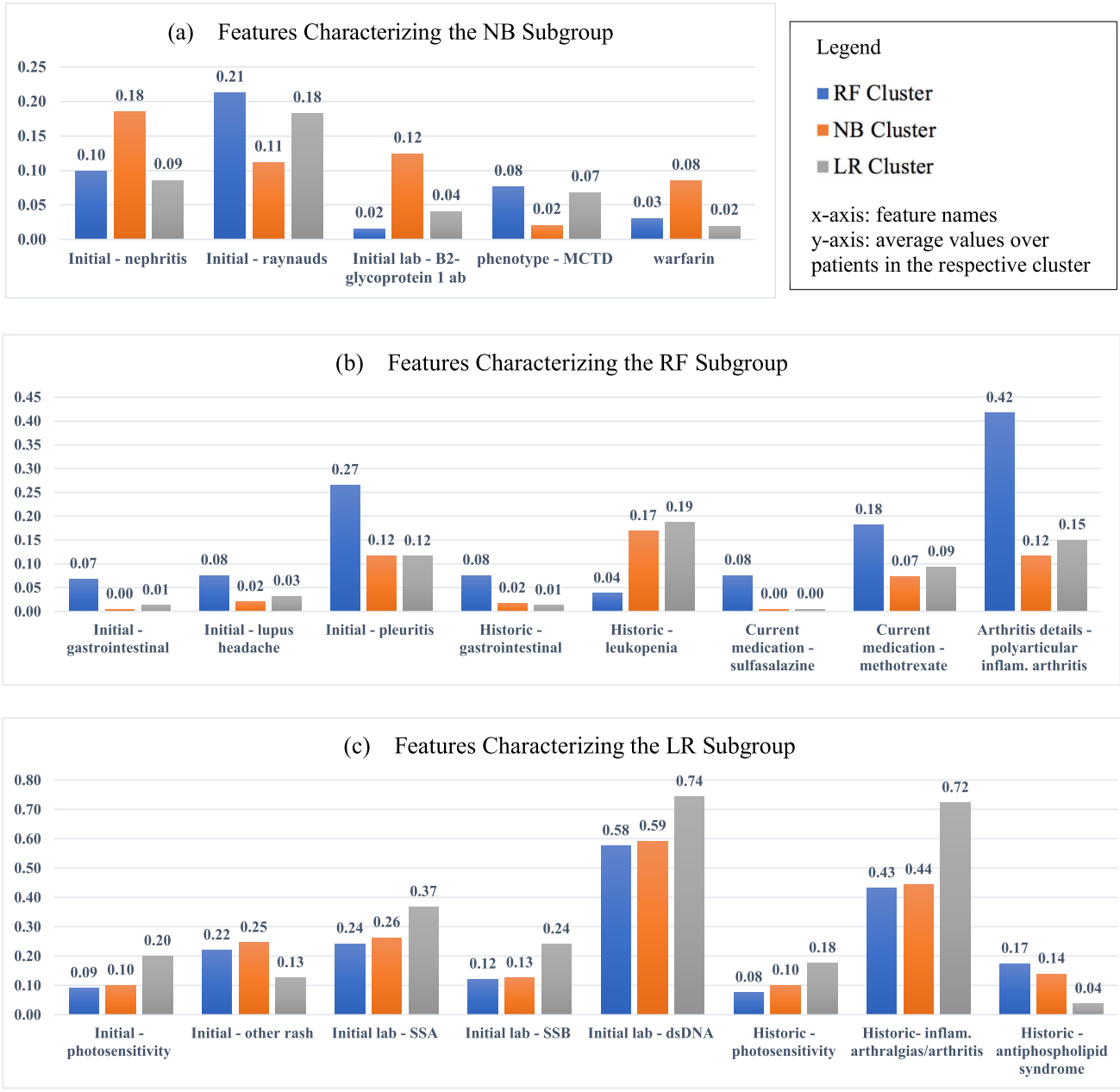


FIGURE 4. Characteristic features of individual patient subgroups identified by CE learners.

and a significant higher level of B2-glycoproten 1 ab in the initial lab test (RF:0.02, NB:0.12, LR:0.04). They have lower manifestations of raynauds (RF:0.21, NB:0.11, LR:0.18) and a low chance of having MCTD (i.e., mixed connective tissue disease); RF: 0.08; NB: 0.02, LR:0.07). This group also has a considerably higher usage of blood thinning medication warfrain (RF:0.03, NB:0.08, LR:0.02). A comparison of these features is presented in Figure 4(a).

The patients in the RF cluster are characterized by their unique initial and historical symptoms, and current medications. They have a high manifestation of gastrointestinal symptoms, both initially (RF:0.07, NB:<0.005, LR:0.01) and historically (RF:0.08, NB:0.02, LR:0.01). Compared to other clusters, a significantly higher percentage of patients in the RF cluster exhibit initial lupus headache (RF:0.08, NB:0.02, LR:0.03) and pleuritis (RF:0.27, NB:0.12, LR:0.12). Patients in the RF cluster demonstrate 4-5 times lower chance of historic manifestation of leukopenia. It is also worth noting that these patients have a significant current usage of medications, including sulfasalazine (RF:0.18, NB:<0.005, LR:<0.005) and methotrexate (RF:0.18, NB:0.07, LR:0.09). Lastly, these patients have noticeably higher polyarticular inflammatory arthritis (RF:0.42, NB:0.12, LR:0.15). A visual comparison of these features is presented in Figure 4(b).

For the LR subgroup, patients demonstrate approximately twice higher level of photosensitivity both initially (RF:0.09, NB:0.10, RF:0.20) and historically (RF:0.08, NB:0.10, RF:0.18) when compared to other clusters. These patients have a high level of SSA (RF:0.24, NB:0.26, LR:0.37), SSB (RF:0.12, NB:0.13, LR:0.24) and dsDNA (RF:0.58, NB:0.59, LR:0.74) in their initial blood lab tests. Additionally, they have a low historic manifestation of antiphospholipid syndrome (RF:0.17, NB:0.14, LR:0.04) and a high chance of inflammatory arthralgias/arthritis.

V. CONCLUSION

In this work, we proposed a new calibrated ensemble (CE) approach to address the heterogeneity in real-world data, which often violates the fundamental i.i.d. assumption in machine learning theory. Unlike the traditional ensemble framework in which each learner is trained independently with the entire dataset, our method exploits various ML models' strengths by training them simultaneously and instituting model-specific data subgroups as part of the training process. As a result, each learner is calibrated to a unique subset of data based on their individualized proficiencies. Clinically, we can interpret each model as a specialist for patients with a particular set of disease manifestations.

We evaluated the CE model in our motivating domain of identifying lupus patients with severe SLE flares in 1,541 clinical encounters. Our experimental results demonstrated consistent efficacy of our CE model across seven evaluation metrics when compared to five individual ML models and regular ensemble methods. We further conducted statistical analysis to identify characteristic features of each model-specific patient subgroup and examined their clinical interpretations.

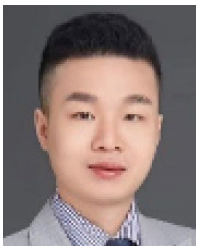
Two factors could have contributed to the success of our CE algorithm. The first is the enforcement of the i.i.d. assumption for each ML algorithm. In particular, by restricting data to a high-performing subset for each learner, the i.i.d. assumption is enhanced for each algorithm, thereby allowing greater potential for success. The second factor is that different ML algorithms can be most effective for different disease manifestations due to their intrinsic designs. For example, the discriminative (e.g., RF, LR) and generative approaches (e.g., NB) are fundamentally different in model structure and learning principle. The CE algorithm exploits each learner's strength and dynamically selecting an ergonomic subset for each model as part of its training process. Our method can be extended to other applications where the collected data may come from multiple underlying distributions.

REFERENCES

[1] S. Menard, *Applied Logistic Regression Analysis*, vol. 106. Newbury Park, CA, USA: Sage, 2002.
[2] H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan, *Neural Network Design*. Atlanta, GA, USA: Martin Hagan, 2014.
[3] Y. Zhao, C. E. Brodley, T. Chitnis, and B. C. Healy, "Addressing human subjectivity via transfer learning: An application to predicting disease outcome in multiple sclerosis patients," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 965–973.
[4] Y. Zhao, T. Chitnis, B. C. Healy, J. G. Dy, and C. E. Brodley, "Domain induced Dirichlet mixture of Gaussian processes: An application to predicting disease progression in multiple sclerosis patients," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 1129–1134.
[5] Y. Zhang and Q. Yang, "A survey on multi-task learning," 2017, *arXiv:1707.08114*.
[6] J. Liu, Y. Pan, F.-X. Wu, and J. Wang, "Enhancing the feature representation of multi-modal MRI data by combining multi-view information for MCI classification," *Neurocomputing*, vol. 400, pp. 322–332, Aug. 2020.
[7] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey, PA, USA: IGI Global, 2010, pp. 242–264.
[8] P. Vik, *Regression, ANOVA, and the General Linear Model: A Statistics Primer*. Newbury Park, CA, USA: Sage, 2013.
[9] H. Arsham and M. Lovric, "Bartlett's test," *Int. Encyclopedia Stat. Sci.*, vol. 1, pp. 87–88, 2011.
[10] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (HSD) test," *Encyclopedia Res. Design*, vol. 3, no. 1, pp. 1–5, 2010.
[11] J. Ross and J. Dy, "Nonparametric mixture of Gaussian processes with constraints," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1346–1354.
[12] A. Serra, M. Fratello, V. Fortino, G. Raiconi, R. Tagliaferri, and D. Greco, "MVDA: A multi-view genomic data integration methodology," *BMC Bioinf.*, vol. 16, no. 1, pp. 1–13, Dec. 2015.
[13] L. S. Hu, H. Yoon, J. M. Eschbacher, L. C. Baxter, A. C. Dueck, A. Nespodzany, K. A. Smith, P. Nakaji, Y. Xu, L. Wang, and J. P. Karis, "Accurate patient-specific machine learning models of glioblastoma invasion using transfer learning," *Amer. J. Neuroradiol.*, vol. 40, pp. 418–425, Feb. 2019.
[14] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4, no. 4. New York, NY, USA: Springer, 2006.
[15] *Lupus Facts and Statistics*, Lupus Found. Amer., Washington, DC, USA, 2012.
[16] M. A. Petri, R. F. van Vollenhoven, J. Buyon, R. A. Levy, S. V. Navarra, R. Cervera, Z. J. Zhong, and W. W. Freimuth, "Baseline predictors of systemic lupus erythematosus flares: Data from the combined placebo groups in the phase III belimumab trials," *Arthritis Rheumatism*, vol. 65, no. 8, pp. 2143–2153, Aug. 2013.
[17] E. Y. Yen and R. R. Singh, "Brief report: Lupus—An unrecognized leading cause of death in young females: A population-based study using nationwide death certificates, 2000–2015," *Arthritis Rheumatol.*, vol. 70, no. 8, pp. 1251–1255, Aug. 2018.
[18] A. M. Jorge, N. Lu, Y. Zhang, S. K. Rai, and H. K. Choi, "Unchanging premature mortality trends in systemic lupus erythematosus: A general population-based study (1999–2014)," *Rheumatology*, vol. 57, no. 2, pp. 337–344, Feb. 2018.
[19] A. Thanou, E. Chakravarty, J. A. James, and J. T. Merrill, "How should lupus flares be measured? Deconstruction of the safety of estrogen in lupus erythematosus national assessment–systemic lupus erythematosus disease activity index flare index," *Rheumatology*, vol. 53, no. 12, pp. 2175–2181, Dec. 2014.
[20] J. Mikdashi and O. Nived, "Measuring disease activity in adults with systemic lupus erythematosus: The challenges of administrative burden and responsiveness to patient concerns in clinical research," *Arthritis Res. Therapy*, vol. 17, no. 1, pp. 1–10, Dec. 2015.
[21] M. A. Petri, J. T. Merrill, R. Maciuca, J. C. Davis, and W. Kennedy, "FRI0293 validation of the revised selena flare index in systemic lupus erythematosus," *Ann. Rheumatic Diseases*, vol. 72, no. 3, pp. A473–A474, Jun. 2013.
[22] D. Isenberg, J Sturgess, E. Allen, C. Aranow, A. Askanase, B. Sang-Cheol, S. Bernatsky, I. Bruce, J. Buyon, R. Cervera, and A. Clarke, "Study of flare assessment in systemic lupus erythematosus based on paper patients," *Arthritis Care Res.*, vol. 70, no. 1, pp. 98–103, 2018.
[23] K. Gu, D. D. Gladman, J. Su, and M. B. Urowitz, "Hospitalizations in patients with systemic lupus erythematosus in an academic health science center," *J. Rheumatol.*, vol. 44, no. 8, pp. 1173–1178, Aug. 2017.
[24] J. Lee, C. Peschken, C. Muangchan, E. Silverman, C. Pineau, C. Smith, H. Arbillaga, M. Zummer, A. Clarke, S. Bernatsky, M. Hudson, C. Hitchon, P. Fortin, and J. Pope, "The frequency of and associations with hospitalization secondary to lupus flares from the 1000 faces of lupus Canadian cohort," *Lupus*, vol. 22, no. 13, pp. 1341–1348, Nov. 2013.
[25] A. Jorge, V. M. Castro, A. Barnado, V. Gainer, C. Hong, T. Cai, T. Cai, R. Carroll, J. C. Denny, L. Crofford, K. H. Costenbader, K. P. Liao, E. W. Karlson, and C. H. Feldman, "Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms," *Seminars Arthritis Rheumatism*, vol. 49, no. 1, pp. 84–90, Aug. 2019.

[26] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012, p. 35.
[27] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," 2013, *arXiv:1305.1707*.
[28] M. Claesen and B. De Moor, "Hyperparameter search in machine learning," 2015, *arXiv:1502.02127*.
[29] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Dec. 2006.

YIJUN ZHAO (Member, IEEE) received the Ph.D. degree in computer science from Tufts University, in 2017, for her research in addressing bias and subjectivity in machine learning. She is currently an Assistant Professor with the Department of Computer and Information Sciences, Fordham University, where she also serves as the Director of the Master of Data Science (MSDS) Program. Her research focuses on applying state-of-the-art machine/deep learning models to help provide greater diagnostic and treatment capabilities in the medical domain.



MAN QIN received the M.S. degree in data science from the Department of Computer and Information Sciences, Fordham University, in 2021. He conducted this research during the MSDS studies. He currently works at China Industrial Bank, as a Data Analyst. His research interests include deep learning, machine learning, and data mining.

APRIL JORGE graduated from the Georgetown University School of Medicine and completed her training in internal medicine at Northwestern University and in rheumatology at Massachusetts General Hospital. She also received the M.D. degree. She is currently an Instructor in medicine at the Harvard Medical School and an Assistant in medicine at the Division of Rheumatology, Allergy, and Immunology, Massachusetts General Hospital. Her research interests include systemic lupus erythematosus (lupus), other related autoimmune disorders, and women's health issues for patients with rheumatic diseases.