Received August 1, 2016, accepted August 22, 2016, date of publication September 2, 2016, date of current version September 28, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2605641

AG-MIC: Azure-Based Generalized Flow for Medical Image Classification

SOHINI ROYCHOWDHURY, (Member, IEEE), AND MATTHEW BIHIS
Department of Electrical Engineering, University of Washington, Bothell-98011, WA, USA

Corresponding author: S. Roychowdhury (roych@uw.edu)

This work was supported in part by a Microsoft Azure Machine Learning Research Award and in part by the University of Washington Bothell Initiatives Group Research Grant.

ABSTRACT Medical image-based research requires heavy computational workload associated with image analysis and collaborative device independent platforms to incorporate expert opinions from multiple institutions. Cloud-based resources such as Microsoft Azure Machine Learning Studio (MAMLS) provide such a platform that is conducive to the medical-image-based data analysis. This paper fosters the advantages of the cloud-based computing frameworks (such as MAMLS) and presents a practical work-flow well-suited for the standard machine learning tasks seen in medical image research viz., binary classification, multi-class learning, regression and so on. The proposed automated generalized workflow allows medical researchers/practitioners to focus on data inferencing rather than dealing with the intricate details of predictive modeling, such as feature and model selection. The scalable architecture of the proposed flow utilizes the MAMLS framework to processes data sets that require partial core storage space in the virtual machine to one complete core storage space in a common flow. Also, the proposed flow invokes multiple feature ranking and predictive models in parallel for automated selection and parameterization of the optimal data model. The performance of the proposed flow is bench-marked on 14 public data sets and four local medical image data sets (∼0.12 MB–1.22 GB) using a single common flow, while ensuring better (∼8% improvement) or atleast similar generalization capability with respect to existing works.

INDEX TERMS Microsoft Azure, machine leaning, medical image, cloud-computing, hyper-parameter search, feature selection.

I. INTRODUCTION

Cloud computing for medical image-based research has attained significant attention in the recent years. While the number of medical image-based studies have grown at a steady rate of 3%-5% per year, data-storage requirements have significantly grown at 10%-25% per year [1]. The advent of major commercial cloud-service providers between 2006 and 2008 such as Amazon web service (AWS), Google App Engine (GAE) and Microsoft Azure has led to the development of platforms, software and infrastructure that promote collaborative research among multiple investigators at different institutions [2], [3], with the advantage of minimal overhead for maintaining the storage and computation systems. This is particularly useful for medical image-based studies using computed tomography (CT) or fundus images; that have been impacted by long wait times for storage and transfer across workstations [1]. Thus, there is an impending need for raw medical image data management, image processing and image-based evaluation systems that have cloud-based high-volume data storage, computation and sharable capabilities.

Over the past two decades machine learning algorithms have been used extensively for detecting underlying patterns in a variety of data streams, and in gaining insights for forecasting and prognostic purposes [4], [5]. However, with the advent of big-data, scalable machine learning solutions have become a necessity, where the basic idea is to distribute the computation in cloud to speed up the model building process [6]. The primary challenge posed by big-data is that the data does not fit in the memory of a single processing system [7]. In such situations, standalone physical systems with limited storage and processing capabilities suffer from long queuing delays and computational bottlenecks. This necessitates the need for harnessing the advantages of cloud computing infrastructures that use *scalable* machine learning modules for designing robust, reliable and reproducible predictive models. In this work, we present such a practical *cloud-based* work-flow for tackling typical machine learning tasks seen in medical image research viz., binary classification, multi-class learning, regression etc. The proposed *cloud-based* generalized work-flow allows medical researchers to focus more on data inferencing than the processing system and data modeling constraints.

Cloud platforms have been categorized into three categories based on the type of services: Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS). While IaaS allows several virtual systems to operate over a singular hardware infrastructure in an independent manner, SaaS provides installation, management and interoperability of software applications without the knowledge of the hardware infrastructure. The primary advantage of PaaS systems, such as the Microsoft Azure Machine Learning Studio (MAMLS), for medical research is that they allow multiple developers to share platform resources without having to install or maintain these resources [1]. Thus, platforms like MAMLS provide the collaborative medical research resources using user-end device independent shared work spaces. Additionally, the MAMLS platform provides secure data transmission, analytics, and remote visualizations that are key for maintaining patient data integrity for sensitive medical data processing [1], [8].

Existing literature surveys [1], [9] have indicated the usefulness of cloud-platforms for tomographic image reconstruction and monitoring applications. Besides, most new image processing algorithms need to be evaluated in comparison with existing algorithms to assess overall clinical improvements. This requires the development of benchmarks that allow image processing algorithms to be compared under common standards. In this paper, we present the comparative assessments of 4 medical image data sets that can facilitate for future bench-marking. This work motivates future collaborative research initiatives to further enhance the medical yield through cloud-based data storage, modeling and inferencing [1].

This paper makes three key contributions. First, a comprehensive cloud-based machine learning flow framework is presented. This is of utmost importance to medical researchers/practitioners agnostic of the underlying complex data modeling and feature selection steps. Further, the proposed flow leverages the system hardware-related independence of a scalable cloud-based platform. This allows the medical researchers/practitioners focus on data inferencing rather than handling the nitty-gritty details of storage or computation requirements for their analysis. Second, we automate the learning algorithm selection as well as hyper-parameter search for the underlying models for each algorithm. This automation is provided for several machine learning tasks typically seen in medical image research viz., binary/multi-class classification, regression etc. The efficacy of such an automation is validated through better (or at least similar) generalization capability of the overall flow compared to state-of-art methods on 14 public datasets, and 4 medical image datasets. Third, we provide comprehensive empirical results in support of the generalization capability of the proposed flow (AG-MIC), and the utility of the flow for deriving insights from medical image data. The proposed flow is bench-marked for classification and regression tasks on 14 public data sets with variable data sizes. Till date, such bench-marking experiments have not been conducted for the Microsoft Azure platform. Moreover, a very detailed study of the utility of the flow on 4 medical image datasets for deriving insights and obtaining state-of-art prediction performance is also provided.

In our prior work [10], we introduced an initial version of the proposed flow using built-in classifiers for optimal classification on three public datasets from the University of California Irvine, Machine Learning Repository [11] and one local medical image data set. In this work, we have further fine-tuned the data modeling process to include automated algorithm selection as well as hyper-parameter search for the underlying models for each algorithms for linear/non-linear parameterization. Now, the modified work-flow is bench-marked on 14 public datasets and 4 real-life large scale medical image datasets. The proposed AG-MIC has been tested on different sizes of datasets ($\sim$ 0.12 MB-1.22 GB) using a single common flow, while ensuring comparable (or sometimes better $\sim$ 8% improvement) classification performances when compared to existing state-of-the-art methods. This improvement in overall classification performances can be accredited to the variety of data models and feature selection methods that the proposed scalable work-flow invokes and the automation involved in the overall optimal model/algorithm selection. A schematic representation of the proposed flow is presented in the Fig. 1. As shown in the schematic flow, the stand alone user-end devices delegates the computational intensive tasks to the cloud computing framework. The comprehensive work-flow is completely run in the cloud. The final results are available as end-reports, which can be conveniently downloaded and analyzed in the user-end devices.

The organization of this paper is as follows. In Section II, prior work on medical imaging and machine learning tasks using cloud computing is summarized. In Section III, the data sets under analysis, proposed flow, and its modules are presented. In Section IV, the experimental results of the proposed flow are presented. Conclusions and discussions regarding the performance of the proposed cloud-based framework for classification and regression tasks are presented in Section V.

II. PRIOR WORK

Recent years have witnessed an exponential growth in data from Business Operations, Healthcare, Trading, Weather Patterns, Geographical phenomenon, Personalized Medicine and Internet of Things (IoT). This drastic increase in volume, velocity, variety and complexity of data has led to the development of data analytic systems that are capable of producing simple spreadsheet-like end reports for a variety of users [12]. Till date, several cloud-service based data analytic systems have been introduced such as:

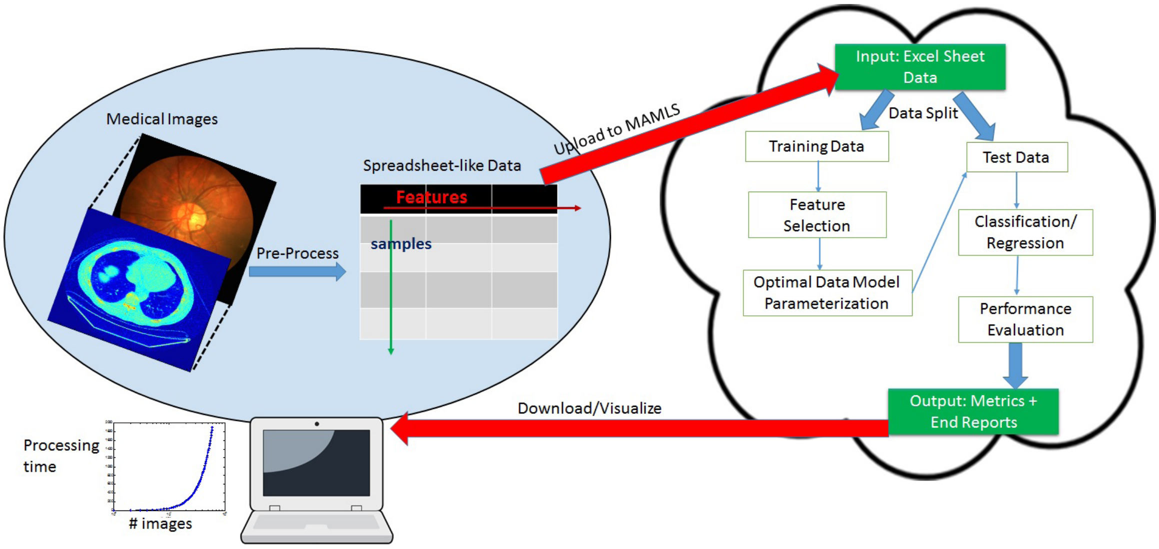


FIGURE 1. Summary of the steps in the proposed AG-MIC. For standalone physical computers (2.6 GHz and 2 GB RAM), the processing time per medical image is observed to increase exponentially and eventually run out of memory in certain cases. To counteract this system hardware dependency, first, medical image sets are pre-processed to produce spreadsheet-like input data for the cloud-based flow. The input data is split into training and test sets. The training data is used for optimal feature selection followed by data model parameterization. The optimally parameterized predictive data model is evaluated on the test data set for classification/regression tasks. The performance evaluation metrics are finally visualized as end-reports that can be downloaded to the user-end devices.

- Project Daytona that uses iterative MapReduce for optimal data analytics [13],
- The Google Prediction API that is developer-oriented and provides a selection of algorithms without significant user interfaces (UI) [14],
- Amazon Machine Learning that provides a single, opaque algorithm with which wizard-driven models can be built [15],
- MAMLS platform that is capable of running prediction modules by making a single web service call in a full-fledged flowchart-style data flow [16]. Additionally, MAMLS allows the use of R and Python codes and packages.

Cloud-based picture archiving and communication systems (Cloud PACS) have revolutionized medical image (PET/CT/MRI) transmission, storage and remote visualizations to the point where zero-footprint diagnostic image quality and browser-based applications are now feasible [1], [17]. Cloud-based PaaS services have further enabled device independence and streamlined patient referrals and consultations. The contributions of cloud computing services in medical research include image archival, co-operative trials between multiple institutions, quality assurance certifications and standardized analytical modules for test repeatability [1], [18]. Additionally, cloud-based resources support distributed computations on for data inferencing tasks. For instance, image-based bio-informatics research has been shown to significantly benefit from distributed computational analysis, parallelism and large data set storage capabilities of cloud-based platforms [1], [19]. Also, cloud based neuro-imaging genetic studies [20] have demonstrated the impact of functional signals in sub-cortical brain regions with genome-wide genotypes. The Azure PaaS has been used on large micro-arrays of gene expression datasets for importance analysis of bio-informatics [12]. Azure machine learning methodologies developed so far include the following: the work in [21] describes a method for real-time traffic viewing using data mined from 1100 social networks feeds from 4 cities; the work in [22] proposes machine learning experiments for grading sample student short answers; the method in [23] proposes predictive algorithms against credit card frauds; the method in [24] proposes classification of web proxy usage from captured packet data.

However, one of the primary concerns regarding the usage of cloud-based resources for research purposes involve data/network security, encryption and connectivity bandwidth issues. Protection of personal medical records face ethical and legal limitations. Thus, secure information transmission and reception, cloud-based backup policies and fast recovery of failed service calls are key to the use of cloud-based resources in medical image-based research [1]. While SaaS models such as Google Docs, Dropbox etc. require the clients to implement standard security and integrity controls, PaaS services primarily focus on protecting data [25]. Encrypted system logs, workspace and user authentication, user-end system failure logs and resource monitors have enabled the use of Azure-based PaaS services to maintain data privacy, integrity and security of web-bases service calls against data attacks and system failures [26].

Moreover, recent bench-marking of the Azure cloud platform storage capabilities in [26] demonstrate the storage, availability, scalability and fault tolerant statistics. Currently, the MAMLS platform supports blob storage ranging from 4 MB to 20 GB of data with cloud resource availability in the range of 99.9-99.95% service level agreements (SLA) for computation, search and storage requirements [27]. Based on the memory size and the number of virtual machines cores invoked for data storage, data sets under analysis in [26] were typically categorized as: extra small (up to 768 MB, shared cores), small (up to 1.75 GB, 1 core), medium (3.5 GB, 2 cores), large (7GB, 4 cores) and extra large (14 GB, 8 cores). However, from a practical standpoint, whenever the size of data exceeds the processing unit memory, there is a need for distributed computing resources with scalable architectures. The Azure PaaS supports horizontal and vertical scaling for load sharing [27], and presents it as a competitive alternative for such large-scale analysis. In view of the improved security and its scalable architecture, MAMLS is a realistic choice for medical image based research which requires heavy computation and a secure framework for collaborative research. In this paper we utilize these advantages of the MAMLS framework and propose a comprehensive machine learning work-flow on top of it. The proposed flow has a scalable architecture due to two reasons. First, the proposed flow utilizes the MAMLS framework to processes data sets that require partial core storage space in the virtual machine to one complete core storage space in a common flow without exponentially increasing the computational time complexities with growing data sizes. Execution of the proposed cloud based work-flow is independent of end-users system configuration. For example, running the current work-flow for CT Image dataset (∼1.22 GB) or Blood Vessel Image dataset (∼1 GB), using 'R' platform in a single machine system (2.6 GHz, 2 GB RAM) results in "out-of-memory" error. In contrast, the MAMLS based work-flow does not run into such memory issues. Second, the proposed flow invokes multiple feature ranking and predictive models in parallel which further alludes to the scalable architecture of the overall proposed flow.

Machine learning is an underutilized resource for the analysis of large medical research and clinical trial data sets [1]. For most medical data sets, the availability of labeled training data is rare. In this work, due to sufficient sample and reasonable balanced outcomes, we use 70/30 data split, where 70% data samples are used for training and model parameterization while the remaining 30% test samples are used for performance evaluation of the trained models [28]–[30], unless other data splits suggested by data authors.

III. DATA AND METHOD

The performance of the proposed generalized flow is bench-marked and analyzed in comparison with existing works for binary, multi-class, and hierarchical classification and regression tasks, respectively. The data sets under analysis, method notation, the proposed flow, and the processing modules are explained below.

A. DATA

The proposed flow is bench-marked using 14 publicly available machine learning data sets for classification and regression tasks and 4 locally generated medical image data sets. Among the 4 locally generated medical image data sets, 3 data sets of medical fundus images are pre-processed to extract several features per sample region/pixel in every image. These compositions of the 4 local data sets and the outputs from the modules of the AG-MIC are presented in the supplementary material. Thus, the features per sample are denoted as 'X' and the sample label is 'Y'. For these fundus image based data sets, 4 different categories of features are extracted: Structural (S), Gaussian coefficient-based (G), Intensity-based (I), Gradient Intensity-based (GI) and Gradient in image intensity-based (GII). For the fourth locally generated medical image data set, raw medical CT images are down-sampled and every image is converted to a sample row. Thus, each pixel of the CT images becomes a sample feature 'X' while the CT image quality serves as the class label 'Y'. The data transferred to the MAMLS platform in spreadsheet-like formats ('csv', 'txt', 'libsvm', etc.) containing the feature and label information.

For certain data sets the training and test data split is specified by the data set authors. For comparative analysis, the existing data split for such data sets are preserved. For example, the MNIST handwriting recognition data set [31] is defined with 60,000 samples for training and 10,000 samples for testing, and these data proportions are used in the proposed flow.

1) PUBLIC DATA FOR BENCH-MARKING

For bench-marking purposes, the proposed generalized flow is tested on 8 binary classification, 2 multi-class classification and 4 regression data sets obtained from the UCI Machine Learning database [11]. Although, the final goal of the proposed flow is medical image classification, the methods and modules used by the proposed flow include machine learning algorithms. Since the first step of all medical research is system/algorithm bench-marking [1], 14 standard publicly available machine learning data sets are chosen for bench-marking the proposed machine learning work-flow. Another reason for work-flow bench-marking using the public data sets is that most existing MAMLS experiments published in the Cortana Intelligence Gallery are analyzed using a subset of the public data sets presented in Table 1. Thus, the AG-MIC work-flow bench-marking will enable comparative assessment of incremental advances made in the MAMLS modules in the near future. The description of all the data sets including their class frequency distribution, storage size and the classification/regression tasks associated with these data sets are described in Table 1. Here, the Annealing data set uses 75/25 data split [32].

TABLE 1. List of data sets under analysis.

Binary Classification		Multi-class Classification	Regression
Telescope -19,020 samples. -Size: 1.46 MB. -Class Frequencies: 65% class h, 35% class g. -10 features: fLength, fWidth, fSize, etc. -Task: Classify gamma (class g) particle bursts from background noise (class h).	**Network Intrusion Detection** -148,517 samples. -Size: 17 MB. -Class Frequencies: 52% class normal, 48% class attack. -42 features: Duration, Login Status, System Information, etc. -Task: Predict (normal) for class by digitizing label column.	**MNIST: Handwritten Digit recognition** -70,000 samples (60,000/10,000 split). -Size: 125 MB. -Class Frequencies: 9.9% class 0, 11% class 1, 10% class 2, 10% class 3, 9.7% class 4, 9% class 5, 9.8% class 6, 10% class 7, 9.75% class 8, 9.94% class 9. -784 features: Pixel-based values. -Task: Predict class (Label) for hand-written digits (0-9).	**Bike Rental UCI** -17,379 samples. -Size: 1.32 MB. -16 features: Season, Working Day, Weather, etc. -Task: Predict the count of total rental bikes.
Wisconsin Breast Cancer -683 samples. -Size: 0.02 MB. -Class Frequencies: 65% class 0, 35% class 1. -9 features: Clump, Thickness Uniformity of Cell Size, Marginal Adhesion etc. -Task: Classify positives (l) from negatives (0).	**Prediction of Student Performance** -100,000 samples. -Size: 36.04 MB. -Class Frequencies: 20% class 0, 80% class 1. -22 features: Problem Information, Student Information Timestamps, etc. Task: Predict yes (1) or no (0) for Correct First Attempt.	**Annealing** -798 samples (75/25 split). -Size: 0.09 MB. -Class Frequencies: 76% class 3, 11% class 2, 7.5% class 5, 4.3% class U, 1% class 1. -38 features: Steel Characteristics, i.e., Hardness, Strength, Shape, etc. -Task: Predict steel grade (classes) for 1, 2, 3, 5, and U.	**Energy Efficiency** -768 samples. -Size: 0.04 MB. -8 features: Relative Compactness, Building Dimensions, Glazing Area, etc. -Task: Predict building efficiency (Heating Load and Cooling Load).
German Credit Card -1,000 samples. -Size: 0.08 MB. -Class Frequencies: 70% class 1, 30% class 2. -20 features: Status of Existing Checking Account, Credit History, Purpose, etc. -Task: Predict high (2) or low (1) risk for label.	**Direct Marketing** -64,000 samples. -Size: 4 MB. -Class Frequencies: 85% class 0, 15% class 1. -12 features: 9 features on prior behavior and demographics of users, 3 features on visit, conversion and spend. -Task: Predict customer response to email class 1 (for returning customer) vs. class 0 (for not returning customer).		**Forest Fires** -517 samples. -Size: 0.026 MB. -12 features: Spatial, Forest Fire Weather Index (FWI) System Variables, Weather, etc. Task: Predict forest fire area.
Blood Donation -748 samples. -Size: 0.012 MB. -Class Frequencies: 76% class 0, 24% class 1. -4 features: Recency, Frequency, Monetary, Time. -Task: Predict yes (1) or no (0) for Class.	**Adult Census Income** -32,561 samples. -Size: 3.473 MB. -Class Frequencies: 76% class $\leq$50k, 24% class>50k. -14 features: Age, Workclass, Education, etc. -Task: Predict above (>50K) or below ($\leq$50K) for income.		**Auto MPG** -392 samples. -Size: 0.018 MB. -8 features: Horsepower, Acceleration, Model, etc. -Task: Predict MPG of a variety of makes and models.

2) NPDR LESION DATA SET

This local medical image data set for multi-class classification data set of size 8.04 MB is constructed from fundus images of the human retina from patients with varying severities of Non Proliferative Diabetic Retinopathy (NPDR) that can cause manifestations in the retina in the form of bright and red lesions [10]. Each fundus image is pre-processed by filtering operations described in [33]. Next, several lesion regions are isolated in each image. Thus, 15,495 different lesion regions with 66 region-based features per region are extracted as X from 89 images in the DIARETDB1 [29] image data set. Examples of the 6 classes of lesion regions that are extracted in this data set are shown in Fig. 2. The class frequencies for this highly imbalanced data set are: 4.9% class 0, 0.18% class 1, 2.6% class 2, 69% class 3, 13% class 4, 10% class 5, respectively.[1] In class imbalanced data sets for multi-class classification tasks, hierarchical classification strategies have been shown to be effective [33]. Based on the structural and intensity-based features of the bright and red lesions, the sample classification can be achieved by the following sequential or hierarchical levels of binary classification in 5 levels, representative of easy to tough sample partitioning, respectively.

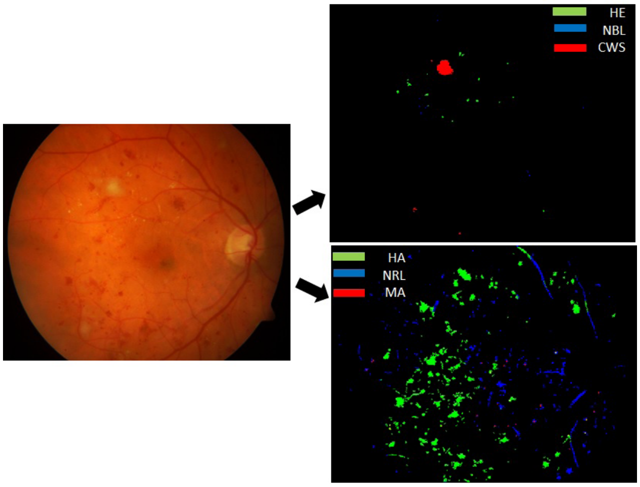


FIGURE 2. Pre-processing fundus images for NPDR lesion data creation. Green plane of each fundus image (left) is filtered to extract 3 classes of bright lesions (top right) and 3 types of red-lesion regions (bottom right).

[1] Available at https://sites.google.com/a/uw.edu/src/useful-links

- Level 1: Classification of bright lesions from red lesions (class 0,1,2 vs. class 3,4,5).
- Level 2: Separation of false positive bright lesions (class 0 vs. class 1,2).
- Level 3: Separation of false positive red lesions (class 3 vs. class 4,5).
- Level 4: Classification among bright lesions (class 1 vs. 2).
- Level 5: Classification among red lesions (class 4 vs. 5).

3) BLOOD VESSEL IMAGE DATA SET

This locally generated medical image data set of size 1 GB is developed for binary classification, where the task is to separate the blood vessel regions (class 1) from the false positive non-vessel regions (class 0) [34]. Accurate classification of blood vessels is crucial for detecting abnormal retinal vessel patterns that can be indicative of severe proliferative DR (PDR). The blood vessel image data represents fine vessel pixels that constitute fine vessel fragments, also called minor vessels [34] that are difficult to distinguish from small red lesions and non-vessels. This data set is created from hand annotated images of retinal vessels from the 20 STARE data set [35], 40 DRIVE data set [36] and 28 CHASE_DB1 [37] images, respectively. The pre-processing steps of extracting the fine blood vessel pixels is shown in Fig. 3. This data set contains 1,274,978 blood vessel pixel-based samples with 98 features per sample in X, with 229,386 samples from STARE,[1] 180,619 samples from DRIVE and 864,973 samples from the CHASE_DB1 data set, respectively.[2] The class label (Y) frequency distribution of this data set is: 67.05% class 0, 32.95% class 1.

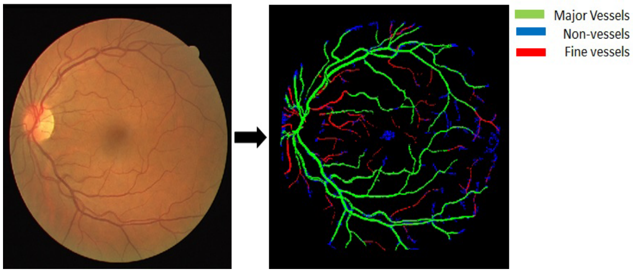


FIGURE 3. **Pre-processing fundus images for blood vessel image data creation. The major vessels are removed while the fine vessel pixels are subjected to binary classification.**

4) PDR IMAGE DATA SET

This local data set of size 9.13 MB is created using 57 fundus images where 30 images are normal and remaining 27 images have some degree of neovascularization, which in turn is a manifestation of PDR [38]. There are two kinds of neovascularizations that manifest in the human retina, namely neovascularization of the disc (NVD) and neovascularization elsewhere (NVE). While NVD manifests as a mesh of tortuous fine vessels in the optic disc (OD) region of the retina, NVE manifest as fine vessel-like region away from the OD. This data set is created by removing the major blood vessel regions and classifying the non-vessels and red lesion regions (class 0) from NVD vessels (class 1) and NVE vessel regions (class 2). This data set is more specialized than the blood vessel image data set to detect NVD and NVE as manifestations of PDR.

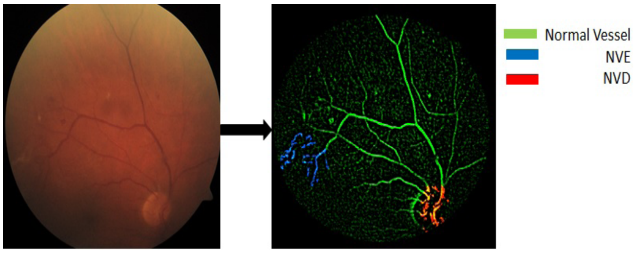


FIGURE 4. **Pre-processing fundus images for PDR image data creation. The vessel regions in the OD are classified to find NVD vessels and vessel regions away from the OD are classified to find NVE vessels.**

This data set contains 12,695 samples with 40 features per sample in X.[2] This imbalanced data set has the following class label (Y) frequencies: 88.61% class 0, 0.053% class 1, 0.061% class 2. The pre-processing steps for extracting the vessel regions with neovascularization are shown in Fig. 4. For this class imbalanced data set for multi-class classification task, 2 levels of hierarchical classification can be set up as follows:

- Level 1: Classification of non-vessels from neovascularization regions (class 0 vs. class 1,2).
- Level 2: Separation of neovascularization regions (class 1 vs. class 2).

Since NVD and NVE manifest in non-overlapping retinal regions, NVD and NVE must be perfectly classifiable from one another.

5) CT IMAGE DATA SET

This locally generated medical image data set of size 1.22 GB comprises of raw dicom images of size [512 × 512] each, corresponding to CT image stacks of the chest CT scans from a phantom. Six levels of CT image quality (CTIQ, class 1 through 6) are obtained as shown in Fig. 5. Prior work suggests that CT image quality can be detected from the variation of pixels in a relatively homogeneous image region [39]. Since the phantom images represent relatively uniform tissue regions, without loss in information, each image is pre-processed and downsized to [128 × 128] using bi-cubic interpolation to reduce data dimensionality. Thus, the resulting data contains 486 samples (81 images × 6 tube current settings), where each sample

[1] Available at https://sites.google.com/a/uw.edu/src/useful-links

[2] Available at https://sites.google.com/a/uw.edu/src/microsoft-azure-machine-learning-research

[2] Available at https://sites.google.com/a/uw.edu/src/microsoft-azure-machine-learning-research

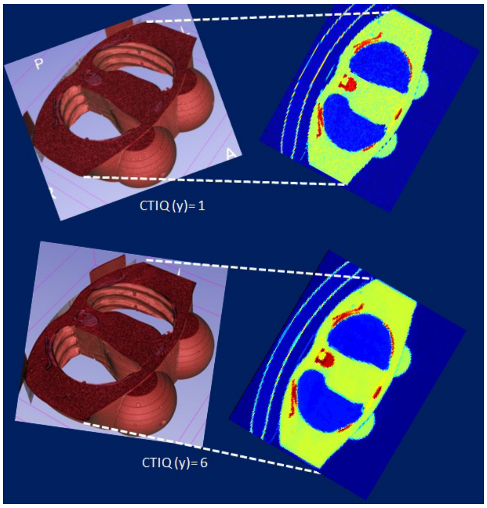


FIGURE 5. Raw CT images from phantom scans for the 6 classes of image qualities (CTIQ) are gathered for the CT image data set.

comprises of 16384 (128 × 128) columns in X. This data set is balanced with each class label having a sample frequency of 16.67% in Y. Since this data set has more features than samples and balanced class label frequencies, this data is subjected to 70/30 split to ensure more samples for training and learning.

B. METHOD NOTATION

The input data stream for each data set uploaded to the proposed flow is in the form: $[X_{[n \times d]}, Y_{[n \times 1]}]$, where '$X$' represents '$d$' features from '$n$' samples and '$Y$' represents a vector of sample class labels, respectively. $\forall x \in X$, the data is scaled in the range of [0,1]. Alphanumeric features are either converted to numeric forms or discarded for classification and regression tasks. The multi-class labels are denoted by i and the number of classes is n_κ, such that for binary problems $n_\kappa = 2$. A description of the mathematical notations used in the proposed flow is given in Table 2.

The performance of the proposed flow is evaluated based on the output metrics given in Table 3. Based on the task, i.e., binary classification, multi-class classification and regression, the performance metrics vary. For evaluation of these performance metrics, the number of test data samples is 'n_T'. In multi-class classification tasks, the number of samples in test data belonging to each class i is denoted as $n_i, i \in [1, ..n_\kappa]$, $\sum_{i=1}^{n_\kappa} n_i = n_T$.

C. THE PROPOSED FLOW

The MAMLS platform supports request response service (RRS) that is a low-latency, high-scale web service used to deploy the modules from the experimentation environment. The default endpoint is provisioned with 20 concurrent RRS requests per end point, while executing up to 4 modules in parallel in an experiment. Thus, the parallel processing capabilities of the MAMLS platform make it suitable for several decision making and parameterization steps to ensure high overall accuracies for classification and regression tasks. Fig. 6 shows the steps in the proposed generalized flow.

TABLE 2. Table of notations.

Notation	Meaning
n'	Number of samples in the training data set.
n_T	Number of samples in the test data set.
n	Total number of samples in the input data.
n_κ	Number of classes for multi-class classification.
d	Total number of features in the input data set.
ρ	Number of features in the reduced feature selected set.
$P^l(x_j)$	Regression data model output. Posterior probability of sample 'j' using $[1, ..l]$ input features.
$\nu^l(x_j)$	Classification data model output for sample 'j'. using $[1, ..l]$ input features.
$\omega^l(x_j)$	Trivial multi-class classifier output for sample 'j' using $[1, ..l]$ input features.
tp	Number of true positive samples. Number of class 1 samples that are classified as class 1.
tn	Number of true negative samples. Number of class 0 samples that are classified as class 0.
fp	Number of false positive samples. Number of class 0 samples that are miss-classified as class 1.
fn	Number of false negative samples. Number of class 1 samples that are miss-classified as class 0.
A^l_f	$= \frac{tp+tn}{tp+tn+fp+fn}$ Classification accuracy in validation fold f with $[1, ..l]$ features.
$\hat{A}^l$	Average classification accuracy across f folds with $[1, ...l]$ features.
$R(F)$	Feature ranks for a set of F features. Low rank implies a highly discriminating feature.
$R^l_S(F)$	Top l sorted features based on features ranks sorted in ascending order a set of F features.
F	Full feature set comprising of $[1, ...d]$ features.
F^1	Reduced feature set after Step 1 of cross validation containing $[1, ...m]$ features.
F^2	Reduced feature set after Step 2 of cross validation containing $[1, ...q]$ features.
$\mathbf{v}^l_f$	Vote for feature l in fold f following Step 2 of cross validation.
Φ	Final reduced feature set at the end of feature selection comprising of $[1, ...\rho]$ features.

TABLE 3. Performance evaluation metrics visualized in end-reports.

Metric	Definition
Binary classification	
Precision	$PR = \frac{tp}{tp+fp}$
Recall	$RE = \frac{tp}{tp+fn}$
Accuracy	$ACC = \frac{tp+tn}{tp+tn+fp+fn}$
Multi-class classification	
Micro Precision	$PR_\mu = \frac{1}{n_T} \sum_{i=1}^{n_\kappa} n_i \mathrm{PR}_i$
Micro Recall	$RE_\mu = \frac{1}{n_T} \sum_{i=1}^{n_\kappa} n_i \mathrm{RE}_i$
Micro Accuracy	$ACC_\mu = \frac{1}{n_T} \sum_{i=1}^{n_\kappa} n_i \mathrm{ACC}_i$
Macro Precision	$PR_M = \frac{1}{n_\kappa} \sum_{i=1}^{n_\kappa} \mathrm{PR}_i$
Macro Recall	$RE_M = \frac{1}{n_\kappa} \sum_{i=1}^{n_\kappa} \mathrm{RE}_i$
Macro Accuracy	$ACC_M = \frac{1}{n_\kappa} \sum_{i=1}^{n_\kappa} \mathrm{ACC}_i$
Regression	
Mean Absolute Error	$MAE = \frac{1}{n_T} \sum_{j=1}^{n_T} \lvert y_j - P^\rho(x_j) \rvert$
Root mean squared error	$RMSE = \sqrt{\frac{1}{n_T} \sum_{j=1}^{n_T} (y_j - P^\rho(x_j))^2}$
Relative absolute error	$RAE = \frac{\sum_{j=1}^{n_T} \lvert y_j - P^\rho(x_j) \rvert}{\sum_{j=1}^{n_T} \lvert y_j - \hat{y} \rvert}$
Relative squared error	$RSE = \sqrt{\frac{\sum_{j=1}^{n_T} (y_j - P^\rho(x_j))^2}{\sum_{j=1}^{n_T} (y_j - \hat{y})^2}}$
Coefficient of Determination	$CoD = 1 - \sum_{j=1}^{n} \frac{y_j - P^\rho(x_j)}{y_j - \hat{y}}$

First, each input data set is subjected to feature selection, where a subset of all the features that are highly

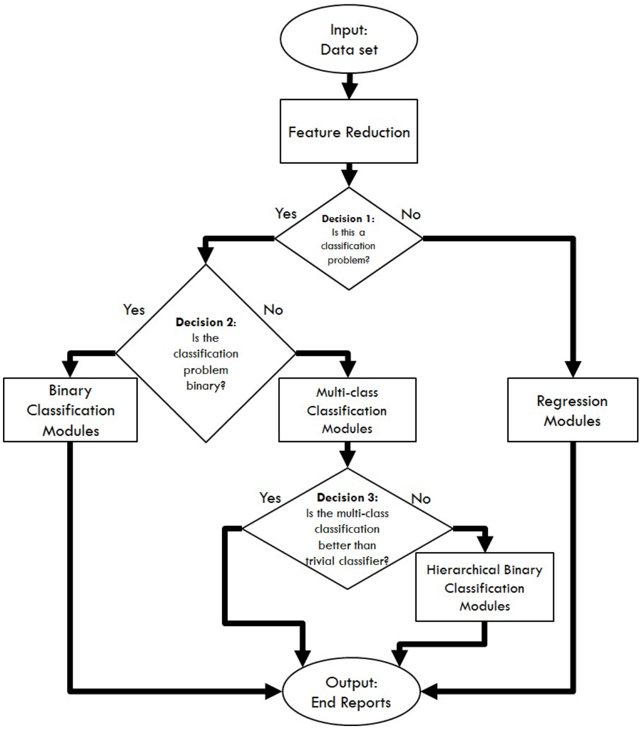

Input:
Data set
Feature Reduction
Decision 1:
Is this a classification problem?
Yes
No
Decision 2:
Is the classification problem binary?
Yes
No
Binary Classification Modules
Multi-class Classification Modules
Regression Modules
Decision 3:
Is the multi-class classification better than trivial classifier?
Yes
No
Hierarchical Binary Classification Modules
Output:
End Reports

FIGURE 6. Block diagram of the proposed flow (AG-MIC) in MAMLS. The 2 significant data processing modules include: feature selection module and data modeling module.

discriminating in nature for classification purposes are selected and retained. Thus, the input data set $[X_{[n\times d]}, Y_{[n\times 1]}]$ is reduced to $[X_{[n\times \rho]}, Y_{[n\times 1]}]$, where $\rho \leq d$. Next, the Decision 1 module directs the input data set to either classification or regression data models, based on the specified problem setting. Several predictive data models, are estimated by optimally tuning the model parameters using 5-fold cross validation over a grid of parameter values [40].

For regression tasks, a suite of predictive data models are parameterized and performance of the best parameterized model (i.e., one with lowest RMSE and highest CoD) on the test data set is visualized as user-end reports. For classification tasks, the feature reduced data set reaches Decision 2 checkpoint, where data sets for multi-class classification ($n_\kappa > 2$) are separated from binary classification data sets ($n_\kappa = 2$). For binary classification tasks, several binary classification data models are parameterized and performance of the best parameterized model (i.e., one with highest ACC) on the test data is visualized as user-end reports.

For the multi-class classification tasks, several multi-class data models are optimally parameterized followed by Decision 3. The Decision 3 checkpoint is used to detect if the best multi-class classification accuracy can be further improved. The underlying assumption here is that if a data set is heavily unbalanced, then a trivial classifier will classify all test samples as the class with the highest frequency. Hence, at Decision 3, shown in (1)-(3), a simple check is performed to determine if the estimated multi-class model performs worse than such a 'trivial' classifier. If invoked, the hierarchical binary classifiers perform sequential partitioning of samples, starting from the easiest partition to the toughest partition.

$$\textbf{Trivial Classification: } \omega^\rho \leftarrow \arg \max_{i\in[1,..n_\kappa]} n_i \tag{1}$$

$$PR^\omega_\mu \leftarrow \omega^\rho.$$

$$\textbf{Best Multi-class classification: } PR^c_M \leftarrow \nu^\rho. \tag{2}$$

$$\textbf{Decision 3 Module:} \tag{3}$$

If $(PR^\omega_\mu > PR^c_M)$

Hierarchical Classification.

Else

Multi-class classification retained.

End

D. FEATURE SELECTION

One significant contribution of this work is the selection of a discriminating set of features that aid classification/regression tasks. In our previous work [10], we demonstrated that feature reduction can often lead to increased accuracy in classification tasks. In this work, we have developed a robust 5-fold double cross validation (CV) module that utilizes a suite of feature ranking strategies for optimal feature set selection in 2 steps followed by feature voting. This feature selection module incorporates scalable feature ranking/voting and selection mechanisms motivated by existing works in [34] and [38]. This module can be modified according to the user-end needs to provide $f-$fold double CV, where f can be varied as [5, 10, 20]. Here, we use a stratified 5-fold double CV [40] where the training data, with n' samples, is partitioned into $f = 5$-folds, such that each fold comprises of 80% training and 20% validation samples with similar sample class frequencies maintained in each fold. The folding operation ensures that each sample is used for validation at least once. In step 1 of CV (4), for every fold, the full feature set of the training data (F) with dimension: $Dim(F) = d$, is ranked and the ranked features are sorted in ascending order $(R^d_S(F))$. As the number of top ranked features are varied, $l = [1, ..d]$, the set of the top l ranked features are selected for classification of the validation samples in each fold (5). The accuracy of validation sample classifications using top l ranked features in each fold (A^l_f) is averaged across all f folds in (6) and the number of features for which this averaged classification accuracy is maximized is l^1 in (7). Now, top l^1 features can be different across the f folds, hence top l^1 features from all folds are gathered and a unique combination of all the selected features is F^1 in (8).

In step 2 of CV, for every $f-$fold, the reduced set of F^1 features are re-ranked and rank-order sorted in (9) followed by classification of the validation samples using top l ranked feature combinations in (10). Next, the average classification error across all folds is computed in (11) and the number of features that maximize this average validation sample accuracy is computed as l^2 in (12). A set of unique top ranked l^2 features across all f folds is selected as F^2 in (13). Finally, the vote of each feature in F^2 per fold is computed in (14).

All features in F^2 that appear in the top l^2 ranked features per fold repeatedly are considered to be discriminating and they are assembled in the final feature set Φ (15) with ρ number of reduced features.

$$\textbf{Step 1: For every f-fold,} \tag{4}$$

$$\textbf{Full Feature Ranking and Sorting:}$$

$$[R(F), R_S^d(F)] \leftarrow (X_{[1:n',1:d]}, Y_{[1:n',1]}).$$

$$\textbf{Classification:} \forall l = [1, ..d], j \in [1:n'], \tag{5}$$

$$v^l(x_j) \leftarrow X_{[1:n', R_S^l(F)]}$$

$$\textbf{Evaluation:} \text{ Mean accuracy across folds} = \hat{A}^l \leftarrow (\mathbf{v^l}, Y) \tag{6}$$

$$l^1 = \arg \max_{l=[1,...d]} \hat{A}^l \tag{7}$$

$$\text{Feature set of unique top } l^1 \text{ features from } f\text{-folds:} \tag{8}$$

$$F^1 = unique(R_S^{l^1}(F), f), Dim(F^1) = m, l^1 \leq m \leq d.$$

$$\textbf{Step 2: For every f-fold:} \tag{9}$$

$$\textbf{Feature Ranking and Sorting:}$$

$$[R(F^1), R_S^m(F^1)] \leftarrow (X_{[1:n',1:m]}, Y_{[1:n',1]}).$$

$$\textbf{Classification:} \forall l = [1, ..m], j \in [1:n'], \tag{10}$$

$$v^l(x_j) \leftarrow X_{[1:n', R_S^l(F^1)]}$$

$$\textbf{Evaluation:} \text{ Mean error across folds} = \hat{A}^l \leftarrow (\mathbf{v^l}, Y) \tag{11}$$

$$l^2 = \arg \max_{l=[1,...m]} \hat{A}^l \tag{12}$$

$$\text{Feature set of unique top } l^2 \text{ features from } f\text{-folds:} \tag{13}$$

$$F^2 = unique(R_S^{l^2}(F^1), f), Dim(F^2) = q, l^2 \leq q \leq m.$$

$$\textbf{Feature Voting for every f-fold:} \forall l = [1, ...q], \tag{14}$$

$$\begin{aligned} \mathbf{v}_f^l &= [1, \text{ if feature } l \text{ belongs to } R_S^{l^2}(F^2). \\ &= [0, \text{ otherwise.} \end{aligned}$$

$$\textbf{Top voted features across f-folds:} \forall l = [1, ...q] \tag{15}$$

$$\Phi \leftarrow \text{If } (sum(v_f^l, f) \geq f/2), Dim(\Phi) = \rho.$$

An example of the 2-step CV approach for feature reduction on the NPDR lesion image data set is shown in Fig. 7. In the proposed flow, the following 3 types of built-in feature ranking strategies are used: F-score, chi-squared and mutual information [10]. Also, the multi-class k-nearest neighbor (kNN) classifier is used for classification of the validation samples. This module can be expanded to support additional feature ranking strategies and other classifiers based on user needs. The kNN classifier was selected for its computational simplicity and speed. For the feature selection operation in regression data sets, the class label values are re-scaled in the positive range and rounded off using the floor function $\lfloor\rfloor$ as $y_j' = \lfloor 10log(1 + y_j) \rfloor$. This operation is motivated by the prior work [41] that uses the logarithmic scaling operation on the public data of Forest Fires [42].

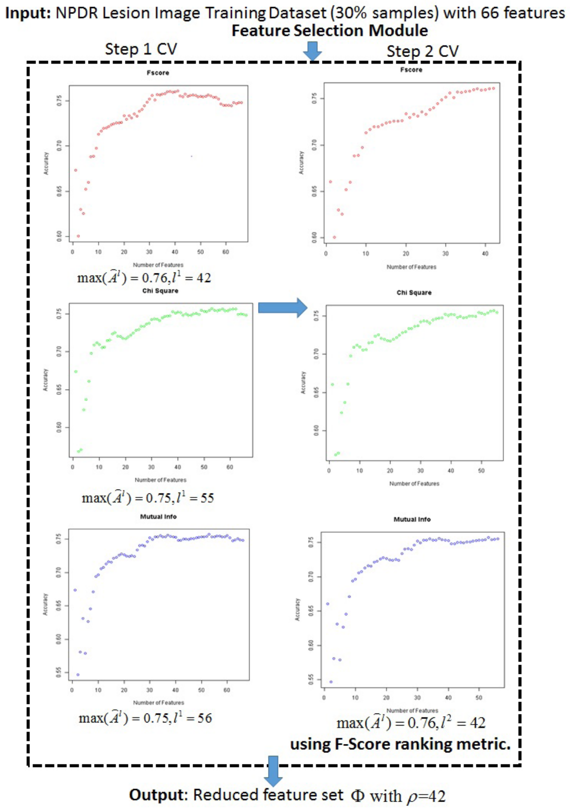


FIGURE 7. The 2 steps in 5-fold double CV for feature selection on the NPDR Lesion Image data set. At the end of step 1, out of 66 input features, the number of unique identified features in feature set F^1 are 42, 55 and 56 using the F-score, chi-squared and mutual information ranking methods, respectively. At the end of step 2, F^2 with 42 unique features result in highest average validation accuracy using the F-score ranking method. All these features are the top ranked 42 features in atleast 3 out of the 5 folds. Hence Φ contains $\rho = 42$ features.

E. DATA MODELS

Once the discriminating feature set is selected, a suite of classification and regression predictive data models are invoked. A description of the data models used in the proposed flow and their respective parameters that are optimally trained are presented in Table 4. Each data model is optimally parameterized using the 'sweep parameters' module in MAMLS that performs a grid search for parameters in 5-fold CV mode. The significance of each classification parameter is discussed in [43]. The performance of each data model is evaluated in terms of the output metrics defined in Table 3. For binary and hierarchical classifications, the Receiver Operating Characteristic (ROC) curves [33] are generated and the area under the ROC curves (AUC) are evaluated for classification robustness. Higher AUC implies robust data model to classification thresholds.

IV. EXPERIMENTS AND RESULTS

Three categories of experiments are performed to analyze the contributions of the proposed flow. First, the performance of the AG-MIC on 14 publicly available machine learning data sets is comparatively analyzed with existing state-of-the-art methods. Second, the selected set of highly discriminating features for the medical image data sets are analyzed for data inferencing. Third, the performance of the proposed flow is

TABLE 4. Data models under analysis.

Model, Tasks	Parameters	Parameter Range
Support Vector Machines (SVM)	c: Regularizer associated with soft thresholding	$[10^{-6}, ..10]$
-Binary classification	K: Non-linear kernel	Radial Basis Function
-One-vs-all multi-class classification	γ: Kernel Parameter	$[2^{-15}, 2^{-13}, ...2^{5}]$.
Logistic Regression (LR)	L1 regularizer	[0,..1]
-Binary classification	L2 regularizer	[0,..1]
-Multi-class classification	tol: Threshold for iterative improvement	[0,...1]
-Regression	Memory size limitation	[5,..50]
Boosted Decision Trees (BDT)	Maximum leaves per tree	≤ 20
-Binary classification	Minimum samples per leaf node	≤ 10
-One-vs-all multi-class classification	Learning Rate	[0,...1]
-Regression	Number of trees	[10,.. 200].
Decision Forest (DF)	Maximum tree depth	[1,..64]
-Binary classification	Splits per node	[1,..1024]
-Multi-class classification	Learning Rate	[0,...1]
-Regression	Samples per leaf	[1,...16].
	Number of trees	[2,...32].
Decision Jungle (DJ)	Number of directed acyclic graphs (DAG)	[1,..024]
-Multi-class classification	Depth and Width of DAG	[1,..1024]
	Steps of DAG optimization	[0,...1]
Neural Networks (NN)	Loss Function	Cross Entropy
-Binary classification	Number of hidden layer neurons	≤ 100
-Multi-class classification	Learning Rate	[0,...1]
-Regression	Number of iterations	[1,.. 200].
Poisson Regression	L1 regularizer	[0,..1]
-Regression	L2 regularizer	[0,..1]
	tol: Threshold for iterative improvement	[0,...1]
	Memory size limitation	[5,..50]
k-Nearest Neighbor (kNN)	k: Neighborhood Parameter	[3,5,7....150]
-Binary classification		
-Multi-class classification		

analyzed for the local medical image data sets in comparison with existing state-of the-art methods.

A. FLOW BENCH-MARKING WITH PUBLIC DATA SETS

In Table 5, the public classification and regression data sets from [11] are used to bench-mark the proposed flow. The classification data set of Direct Marketing [44] has been previously bench-marked on the MAMLS platform. For this data set, the experiment designed in [45], built uplift and response models for population of customers who were sent women's email, or those who were not sent any email. This MAMLS experiment was shown to better uplift (7.3%) when compared to the existing response model (6%) in [44]. On the Direct Marketing data set, the proposed AG-MIC results in $[ACC, PR, RE, AUC] = [0.8623, 1, 0.062, 0.6]$, respectively. From Table 5, we observe that the proposed flow has similar to superior performance of classification and regression on all the public data sets. Additionally, the scalable architecture of the proposed flow allows automated model selection across a wide range of optimally tuned models, thereby ensuring high classification accuracies and low regression errors using a common flow.

B. DISCRIMINATING FEATURE SETS

In this experiment, we analyze the selected set of features in comparison with existing works to identify the new features that contribute towards increasing classification accuracies for the local medical image data sets. For the 3 medical image data sets generated by pre-processing fundus images, the categories of features in the full feature set (F) and the reduced feature set (Φ) in comparison with the features identified in existing works is shown in Table 6. In this experiment, we assess the importance of certain feature categories for medical image classification tasks.

For the NPDR lesion data set, $\rho = 42$ unique features, or 63.6% of the original features are identified as highly discriminating using the proposed flow. In the existing work [33], 30 features are extracted per lesion region using AdaBoost [40] for feature ranking. From Table 6, we observe that when compared to [33], the proposed flow extracts 9 lesser S-category features and additional 12 G-category, 10 GI-category features. Since the proposed flow relies on several ranking strategies and selects only the top voted features, it has superior classification performance for the significantly tougher sample partitioning problems such as hierarchical classification level 3,4,5 as shown in Table 7.

For the blood vessel image data set, $\rho = 85$ unique features, or 86.7% of the original features are selected by the AG-MIC. In the existing work [34], 8 pixel-based features are used for vessel/non-vessel pixel classification. From Table 6, we observe that the 77 additional region-based features belonging to S, G, I, GI, GII-categories extracted by the AG-MIC are significantly important across vessel image data sets for very fine vessel/non-vessel classification tasks as shown in Table 7.

In the PDR image data set, $\rho = 33$ unique features, or 82.5% of the original features are selected by the proposed flow. When compared to the existing work in [38],

TABLE 5. Bench-marking the proposed flow for classification and regression tasks on public data sets [11].

Data Set	Existing Method	Existing Performance	AG-MIC Performance
Classification			

Telescope	Gang et. al. [46] implemented 10-fold CV using Trees, rules, Bayesian classifiers, lazy classifiers etc. in WEKA 3.7.	Adaboost M1 algorithm changed weights of training samples iteratively to increase emphasis on previous misclassified samples. ACC=0.8694	BDT classifier, $\rho = 6$:
			ACC=0.883, AUC=0.915
			PR=0.864, RE=0.739

Wisconsin Beast Cancer	Wang et. al [47] performed 10-fold CV using Naive Bayes, NN, SVM, DT models.	Naive Bayes had worst performance while SVM and NN had high classification accuracy. ACC=0.956-0.985	NN classifier, $\rho = 9$:
			ACC=0.975, AUC=0.994
			PR=0.938, RE=0.994

Blood Donation	Sundaram et. al. [48] implemented CART classifier using WEKA to predict a blood donor's contribution (class 1) in 2007.	PR=0.53
		RE=0.31, AUC=0.69
			LR classifier $\rho = 4$:
			ACC=0.784, AUC=0.76
			PR=0.658, RE=0.2

German Credit Card	Luo et. al. [49] performed 10-fold CV using SVM and Clustering-Launched Classification (CLC) with a polynomial degree of 4 .	CLC method had superior performance when compared to SVM and a hybrid genetic algorithm (GA-SVM) classifier. The ACC are: -SVM=0.737, -CLC=0.833, -GA+SVM=0.779	LR classifier, $\rho = 14$:
			ACC=0.756, AUC=0.773
			PR=0.61, RE=0.514

Adult Census Income	Chen et. al. [50] performed quality embedding for nearest neighbor classification (NNMap), that reduced the classification problem to classical boosting.	Classification performances were evaluated using brute force, 3-NN, 5-NN, FastMAP, boostNN and NNMAP. ACC evaluated by 5-runs of 10-fold CV: -NNMap=0.805, -BoostNN=0.814	BDT classifier, $\rho = 5$:
			ACC=0.867, AUC=0.922
			PR=0.74, RE=0.691

Network Intrusion Detection	Panda et. al. [51] performed 10-fold CV to distinguish between 'Normal'/'Attack'. Performances of DT, principal component analysis, gradient solvers in SVM, and Random Forest were analyzed.	The ACC are:
		-Balanced Nested Dichotomy+Random Forest Normal= 0.999, Attack: 0.995
		-Decision Tree Classifier with RBF kernel Normal= 0.944, Attack= 0.907
			BDT classifier, $\rho = 41$:
			ACC=0.999, AUC=0.9999
			PR=0.999, RE=0.999

Prediction of Student Performance	Drumond et. al. [52] implemented matrix factorization models and reported AUC and Hinge Loss.	AUC range: 0.72-0.73	DF classifier, $\rho = 13$
			ACC=0.84, AUC=0.8014
			PR=0.85, RE=0.976

MNIST	McDonnel et. al. [53] trained on 60,000 samples, a single hidden layer feed forward neural network with 1600-15000 hidden neurons.	On test set of 10,000 samples, ACC=0.98-0.9855	NN, $\rho = 190$ ACC^{mi}=0.9839, PR^{mi}=0.983, RE^{mi}=0.984
Annealing	Nguyen et. al. [32] compared performances of DT, RF with Naive Bayes Classifiers.	Train on 75% and test on 25% samples, ACC=0.78-0.99	One-vs-all BDT, $\rho = 15$: ACC^{mi}=0.987, PR^{mi}=0.964.
Regression			

Bike Rental (UCI)	Yin et. al. [54] performed feature engineering by digitization of discrete values, converting periodic and conditional expectation value mapping. Feature selection was performed by forward/backward searches. Classification with 30/70 split and 10-fold CV.	Classifiers: Ridge Linear Regression (RLR), Support Vector Regressions (SVR), and RF. Metric: Root Mean Square Log Error (RMSLE). -RLR (all features)=0.80, -SVR (top 10 features)=0.33. -RF (with and without feature selection)=0.30-0.33	DF, $\rho = 14$
			RMSLE=0.38,
			RMSE=0.05, MAE=0.02,
			RAE=0.01,
			RSE=0.001
			CoD=0.99

Energy Efficiency	Tsanas et. al. [55] performed statistical analysis to create a correlation matrix using Spearman's coefficient followed by Regression using iteratively reweighted least squares (IRLS) and RF. Results reported on 10-fold CV from 100 averaged runs.	Metrics: [MAE (MSE)](%)
		IRLS Heating: [2.14 ± 0.24 (9.87 ± 2.41)]
		IRLS Cooling: [9.87 ± 2.41 (2.21 ± 0.28)]
		RF Heating: [0.51 ± 0.11 (1.03 ± 0.54)]
		RF Cooling: [1.42 ± 0.25 (6.59 ± 1.56)]
			BDT, $\rho = 6$,[MAE (MSE)](%):
			Heating: [3.7 (2.7)]
			RAE=0.041, RSE=0.002
			CoD=0.997
			Cooling: [2.04 (7.3)], RAE=0.2,
			RSE=0.015, CoD=0.95

Forest Fires	Cortez et. al. [41] predicted the log area affected by forest fire using spatial, temporal and fire weather index features (STFWI), spatial, temporal and weather features (STM),fire weather index features (FWI) and M weather features (M). 10-fold search for hyperpameterization of NN and SVM.	Metrics: [MAE (MSE)](%) STFWI:-RF=[13.31 (64.3)], -NN=[13.09 (64.5)] -SVM=[13.07 (64.7)], -DT=[13.46 (64.4)] Using STM: -Naive= [18.61 (63.7)] -DT= [13.43 (64.6)], -RF= [13.04 (64.5)] -NN=[13.92 (68.9)],-SVM= [13.13 (64.7)]	NN, $\rho = 4$,[MAE (MSE)](%): [14.28 (43.65)] BDT, $\rho = 4$ [16.85 (45.67)]

Auto MPG	Nobrega et. al. [56] performed Kalman filter-based Extreme Machine learning on regression data. Optimal parameters for sequential learning were learned using 25 hidden neurons, 75 training samples.	RMSE on test data is in the range [0.067-0.07].	BDT, $\rho = 6$
			MAE=0.02, RMSE=0.04
			RAE=0.33, RAE=0.16
			CoD=0.82

21 additional features belonging to the I-category are recognized as significant. This leads to $ACC_\mu = 0.917$ and $ACC_M = 0.944$ when compared to the existing $ACC_M = 0.912$ in [38] using for multi-class classification by the LR model. This classification accuracy is further enhanced by hierarchical classification on this data set as shown in Table 7.

For CT image data set, 395 pixel features are retained by the feature selection module for 2.41% feature retention. These features primarily constitute the surface of the chair over which the phantom is placed and the outer edge of the phantom as shown in Fig. 8. Thus, based on the definition of the chair over which the patient is imaged, the CT image quality can be predicted with 12.6% higher classification accuracy when compared to standard spatial image segmentation and quantification method in [39].

C. FLOW PERFORMANCE ON MEDICAL IMAGE DATA SETS

Once the AG-MIC is bench-marked, we analyze its performance on all the local medical image data sets in Table 7.

TABLE 6. Performance of feature selection on medical image data sets in comparison with existing works.

Feature Category	#Full Features	# AG-MIC Features	#[33] Features	#Full Features	# AG-MIC Features	#[34] Features	#Full Features	# AG-MIC Features	# [38] Features
Data:	NPDR	Lesion		Blood	Vessel	Image	PDR	Image	
S	14	5	14	14	9	-	7	6	6
G	12	12	0	12	11	-	-	-	-
I	16	13	14	16	13	-	28	25	4
GI	24	12	2	24	23	-	4	1	1
GII	-	-	-	24	22	-	-	-	-
Others	-	-	-	8 pixel window-based	7 pixel window based	8 pixel window-based	1 watershed transform-based	1 watershed transform-based	1 watershed transform-based
Total	66	42	30	98	85	8	40	33	12

TABLE 7. Classification performance of the proposed flow in comparison with existing works.

Proposed	Flow							Existing work		
Data	ρ	Task	Model	PR	RE	ACC	AUC	Method	ρ	ACC
NPDR Lesion	41	Hierarchical Classification, Level 1:	BDT	0.9995	0.9999	0.9995	0.993	[33]	30	1
		Hierarchical Classification, Level 2:	BDT	0.9557	0.9183	0.9548	0.976			0.89
		Hierarchical Classification, Level 3:	BDT	0.7438	0.6362	0.8545	0.89			0.80
		Hierarchical Classification, Level 4:	BDT	0.9631	1	0.9640	0.962			0.97
		Hierarchical Classification, Level 5:	BDT	0.7133	0.5884	0.7168	0.775			0.65
Blood Vessel Image:	85	Binary Classification	DF	0.791	0.73	0.848	0.91	[34]	8	0.795
DRIVE Vessel				0.81	0.885	0.83	0.907			0.83
STARE Vessel				0.8	0.81	0.83	0.908			0.751
CHASE_DB1 Vessel				0.751	0.58	0.854	0.884			0.8062
PDR Image	33	Hierarchical Classification, Level 1:	BDT	0.573	0.766	0.9314	0.9	[38]	6-10	0.912
		Hierarchical Classification, Level 2:		1	1	1	1			1
CT Image	395	Multi-class Classification	LR	0.876_{μ} 0.8923_{M}	0.876_{μ} 0.8855_{M}	0.876_{μ} 0.959_{M}	- -	[39]	16384	0.75_{μ} 0.75_{M}

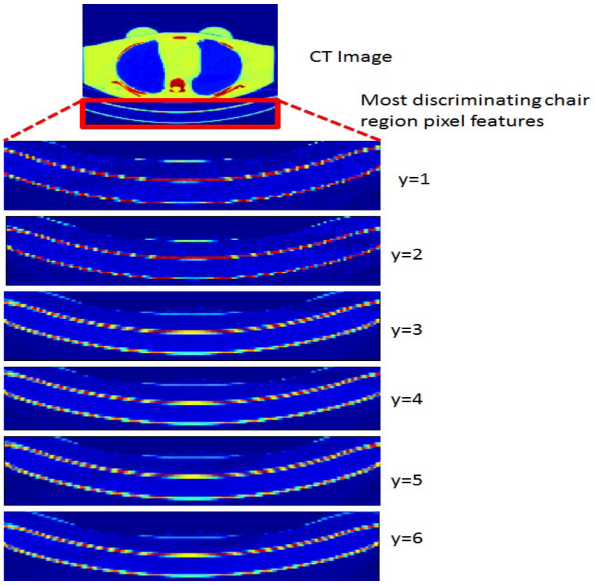


FIGURE 8. The highly discriminating feature pixels for the CT image data set correspond to the chair and outer edge of the phantom. The qualitative analysis of the chair region shows significant variations that are indicative of CT image quality (*y*).

Here, we observe that for all the local data sets, the selected set of reduced features have superior classification performance on the medical image data sets when compared to existing works.

V. CONCLUSIONS AND DISCUSSION

This paper proposes a comprehensive machine learning work-flow (AG-MIC) that is developed on a sharable cloud-computing platform. The proposed flow invokes multiple feature ranking and predictive models in parallel followed by the selection of the optimal data model for classification and regression tasks on a variety of machine learning and medical image data sets. The AG-MIC performs two primary tasks: data dimensionality reduction by a scalable feature ranking and feature selection module, and optimal tuning of binary, multi-class and hierarchical classification models to ensure high overall classification accuracy. It is noteworthy that the end-to-end run-times for the experiments in the proposed flow ranges between 3-35 minutes to 2-6 hours of cloud-processing time based on the storage size of the data sets. The optimal feature set selection step reduces the classification time complexities from 1-354 minutes to 0.9-48 minutes as shown in the supplementary material. The variations in end-to-end processing times is due to the RRS turnaround times. Each experiment in the cloud is processed in a queue, and the wait times can vary significantly depending on the service traffic load. Since cloud-based platforms such as the MAMLS are capable of parallel processing of experiment modules, computation time complexities do not pose as bottlenecks.

Several experiments are performed using the proposed flow on publicly available data sets and on medical image data sets for classification and regression tasks. Firstly, on 14 publicly available machine learning data sets, the performance of the AG-MIC is comparatively analyzed with existing state-of-the-art methods. This helps to validate the generalizability of the proposed flow for a gamut of real-life applications. Secondly, on 4 locally created medical image data sets, a selected set of features using the AG-MIC flow are found to provide better insights into pathology/image quality classification tasks when compared to the full feature set. Such an analysis validates the practical utility of such an automated feature selection framework, typically well suited for medical research domains. Finally, the performance of the proposed flow is analyzed for local medical image data set classifications in comparison with existing state-of the-art methods. This analysis demonstrates the adaptability and robustness of the proposed flow to sample class imbalances and data storage size variabilities in real-life medical data sets.

One key contribution of this work is the selection of highly discriminating set of features that present unforeseen dependencies in the data. For the pre-processed fundus image based medical data sets, 63-87% of the pixel-based and region based features are considered useful for classification tasks by the proposed flow. While the features extracted per pre-processed fundus image data set are motivated by domain knowledge, the final reduced feature set aids standardized data inferencing. We observe that for the NPDR lesion classification, Gaussian coefficient based and Gradient image based features are more important than structural features. For fine blood vessel classification, region based features are significantly important along with pixel-based features, while for classification of neovascularization, regional intensity-based features are significantly important. However, for the CT image data set, the proposed flow retains only 2.4% of the raw pixel features, which significantly reduces the computational complexity for image quality classification tasks. It is noteworthy that the CT image down-sampling operation significantly improves the image quality-based classification accuracy. The original CT image data set of size [512 × 512] pixels results in $ACC_\mu = 0.79$ since the data set has too many features when compared to the number of samples. Sub-sampling the images significantly reduces dimensionality per sample, that leads to stable classification. Thus, the proposed flow is limited by the input data dimensionality. In the supplementary material we observe that for training data sets with low ratio between the number of input features and the number of samples $\frac{d}{n'} \leq 100$, the proposed feature selection module significantly increases the overall classification accuracy when compared to the full feature set. Future work will be directed towards combining deep learning strategies for analyzing high-dimensional medical image data sets with the AG-MIC. Additionally, the proposed flow and data sets can be used to bench-mark future scalable methods involving medical data based classification tasks.

Another significant contribution of the proposed flow is that for unbalanced data sets, hierarchical classifiers are invoked to improve the overall classification accuracy when compared to multi-class classifiers. For the 5 multi-class classification data sets under analysis, the classification performances of trivial classifiers (PR^{ω}_{μ}) and best multi-class classifiers (PR^{c}_{M}), respectively, are shown in Table 8. We observe that hierarchical classifiers are invoked only for the NPDR lesion and PDR image data sets, where the trivial classification is better than multi-class classification. The hierarchical classifiers further improve the overall classification accuracy for these 2 data sets.

TABLE 8. Decision making for hierarchical classification.

Data	PR^{ω}_{μ}	PR^{c}_{M} (Model)	Classification
MNIST	0.11	**0.9086** (NN)	Multi-class
Annealing	0.76	**0.97** (BDT)	Multi-lass
NPDR Lesion	**0.69**	0.6181 (BDT)	Hierarchical
PDR image	**0.89**	0.746 (BDT)	Hierarchical
CT image	0.17	**0.892** (LR)	Multi-class

The proposed multi-disciplinary and application oriented flow connects the efficacy of cloud-computing frameworks with machine learning algorithms for medical image analysis. The proposed method utilizes the system hardware independence of a cloud-based platform and builds a systematic work-flow that further reduces the dependencies of feature selection and data modeling from medical classification tasks. The resulting flow ensures low system and data modeling dependencies, which can lead to vital medical research contributions, such as identification of new feature sets for personalized medicine and classification tasks.

ACKNOWLEDGMENT

The authors would like to thank Devin Nakahara and Vinh Le for their contributions on 'R' module development. Also, the authors are grateful to the anonymous reviewers for their numerous constructive comments.

REFERENCES

[1] G. C. Kagadis *et al.*, "Cloud computing in medical imaging," *Med. Phys.*, vol. 40, no. 7, p. 070901, 2013.
[2] L. Wang *et al.*, "Cloud computing: A perspective study," *New Generat. Comput.*, vol. 28, no. 2, pp. 137–146, 2010.
[3] Y. Shen, Y. Li, L. Wu, S. Liu, and Q. Wen, "Cloud computing overview," *Enabling the New Era of Cloud Computing: Data Security, Transfer, and Management*. Pennsylvania, PA, USA: IGI Global, 2013, p. 1.
[4] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Mach. Learn.*, vol. 30, nos. 2–3, pp. 195–215, 1998.
[5] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA, USA: MIT Press, 1996.
[6] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[7] G. Press. (Sep. 2014). *12 Big Data Definitions: What's yours?* [Online]. Available: http://www.forbes.com/sites/gilpress/2014/09/03/12-big-datadefinitions-whats-yours/#3314c45921a9
[8] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, "Interactions with big data analytics," *Interactions*, vol. 19, no. 3, pp. 50–59, May/Jun. 2012.
[9] T. Shepherd *et al.*, "Comparative study with new accuracy metrics for target volume contouring in PET image guided radiation therapy," *IEEE Trans. Med. Imag.*, vol. 31, no. 11, pp. 2006–2024, Nov. 2012.
[10] M. Bihis and S. Roychowdhury, "A generalized flow for multi-class and binary classification tasks: An Azure ML approach," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct./Nov. 2015, pp. 1728–1737.

[11] University of California, Irvine, CA, USA. *Machine Learning Repository*, accessed on Sep. 30, 2014. [Online]. Available: http://archive.ics.uci.edu/ml/

[12] H. P. Shanahan, A. M. Owen, and A. P. Harrison, "Bioinformatics on the cloud computing platform Azure," *PloS ONE*, vol. 9, no. 7, p. e102642, 2014.

[13] R. S. Barga, J. Ekanayake, and W. Lu, "Project daytona: Data analytics as a cloud service," in *Proc. IEEE 28th Int. Conf. Data Eng. (ICDE)*, Apr. 2012, pp. 1317–1320.

[14] E. R. Sparks *et al.*, "MLI: An API for distributed machine learning," in *Proc. IEEE 13th Int. Conf. Data Mining (ICDM)*, Dec. 2013, pp. 1187–1192.

[15] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein, "Distributed GraphLab: A framework for machine learning and data mining in the cloud," *Proc. VLDB Endowment*, vol. 5, no. 8, pp. 716–727, 2012.

[16] S. Mund, *Microsoft Azure Machine Learning*. Birmingham, U.K.: Packt Publ., 2015.

[17] J. Philbin, F. Prior, and P. Nagy, "Will the next generation of pacs be sitting on a cloud?" *J. Digit. Imag.*, vol. 24, no. 2, pp. 179–183, 2011.

[18] K.-W. Chang *et al.*, "iSMART: An integrated cloud computing Web server for traditional chinese medicine for online virtual screening, de novo evolution and drug design," *J. Biomolecular Struct. Dyn.*, vol. 29, no. 1, pp. 243–250, 2011.

[19] X. Chen and F. Luo, "Cloud computing in bioinformatics," *Sci. Comput.*, vol. 29, no. 5, pp. 21–24, 2012.

[20] B. Da Mota *et al.*, "Machine learning patterns for neuroimaging-genetic studies in the cloud," in *Recent Advances and the Future Generation of Neuroinformatics Infrastructure*. 2015, p. 188.

[21] A. Pathak, B. K. Patra, A. Chakraborty, and A. Agarwal, "A city traffic dashboard using social network data," in *Proc. 2nd IKDD Conf. Data Sci.*, 2015, Art. no. 8.

[22] R. Krithika and J. Narayanan, "Learning to grade short answers using machine learning techniques," in *Proc. 3rd Int. Symp. Women Comput. Inform.*, 2015, pp. 262–271.

[23] A. Tselykh and D. Petukhov, "Web service for detecting credit card fraud in near real-time," in *Proc. 8th Int. Conf. Secur. Inf. Netw.*, 2015, pp. 114–117.

[24] S. Miller, K. Curran, and T. Lunney, "Cloud-based machine learning for the detection of anonymous Web proxies," in *Proc. IEEE Int. Symp. Softw. Crowdsour.*, Jun. 2016, pp. 1–6.

[25] V. M. O'Neill. (Jan. 2011). *SaaS, PaaS, and IaaS: A Security Checklist for Cloud Models*. [Online]. Available: http://www.csoonline.com/article/2126885/cloud-security/saas–paas–and-iaas-a-security-checklist-for-cloud-models.html

[26] D. Agarwal and S. K. Prasad, "AzureBench: Benchmarking the storage services of the Azure cloud platform," in *Proc. IEEE 26th Int. Parallel Distrib. Process. Symp. Workshops PhD Forum (IPDPSW)*, May 2012, pp. 1048–1057.

[27] A. Glick. (May 2016). *Disaster Recovery and High Availability for Applications Built on Microsoft Azure*. [Online]. Available: https://azure.microsoft.com/en-us/documentation/articles/resiliencydisaster-recovery-high-availability-azure-applications/

[28] D. Zikic *et al. Machine Learning for the Automatic Analysis of Medical Images*, accessed on Jun. 5, 2015. [Online]. Available: http://research.microsoft.com/en-us/events/2012summerschool/acriminis iss2012.pdf

[29] T. Kauppi *et al.*, "The DIARETDB1 diabetic retinopathy database and evaluation protocol," in *Proc. BMVC*, 2007, pp. 1–10.

[30] N. Situ, X. Yuan, J. Chen, and G. Zouridakis, "Malignant melanoma detection by bag-of-features classification," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2008, pp. 3110–3113.

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[32] N. Nguyen, A. Subramanian, and B. King. Benchmarking random forests against naive bayes. University of California, Berkeley, CA, USA, accessed on Jan. 12, 2016. [Online]. Available: http://bid.berkeley.edu/cs294-1-spring13/images/0/01/CS294_Project_Report_NAB.pdf

[33] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "DREAM: Diabetic retinopathy analysis using machine learning," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 5, pp. 1717–1728, Sep. 2014.

[34] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "Blood vessel segmentation of fundus images by major vessel extraction and subimage classification," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 3, pp. 1118–1128, May 2015.

[35] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 203–210, Mar. 2000.

[36] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abramoff, "Comparative study of retinal vessel segmentation methods on a new publicly available database," *Proc. SPIE*, vol. 5370, pp. 648–656, May 2004.

[37] M. Fraz *et al.*, "CHASE_DB1," KIngston Univ. Res., Tech. Rep., 2011.

[38] S. Roychowdhury, D. Koozekanani, and K. Parhi, "Automated detection of neovascularization for proliferative diabetic retinopathy screening," in *Proc. IEEE 38th Int. Conf. Eng. Med. Biol.*, Aug. 2016.

[39] S. Roychowdhury, N. Hollraft, and A. Alessio. (2016). "Blind analysis of CT image noise using residual denoised images." [Online]. Available: https://arxiv.org/abs/1605.07650

[40] V. S. Cherkassky and F. Mulier, *Learning From Data*. New York, NY, USA: Wiley, 1998.

[41] P. Cortez and A. Morais, "A data mining approach to predict forest fires using meteorological data," in *Proc. 13th Portuguese Conf. Artif. Intell. (EPIA)*, 2007, pp. 512–523.

[42] University of California, Irvine, CA, USA. *Forest Fires Data Set*, accessed on Sep. 30, 2014. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Forest+Fires

[43] M. Azure. (Mar. 2016). *Machine Learning Initialize Model Classification*. [Online]. Available: https://msdn.microsoft.com/enus/library/azure/dn905808.aspx

[44] K. Hillstrom. *The MineThatData e-Mail Analytics and Data Mining Challenge*. [Online]. Available: http://blog.minethatdata.com/2008/03/minethatdata-e-mailanalytics-and-data.html

[45] D. Pechyony. Uplift modeling for direct marketing. Cortana Intelligence Gallery. [Online]. Available: https://gallery.cortanaintelligence.com/Experiment/Uplift-Modeling-for-Direct-Marketing-2?r=legacy&share=1

[46] G. Kou, Y. Lu, Y. Peng, and Y. Shi, "Evaluation of classification algorithms using MCDM and rank correlation," *Int. J. Inf. Technol. Decision Making*, vol. 11, no. 1, pp. 197–225, 2012.

[47] J.-Y. Wang. (2003). Data mining analysis (breast-cancer data). Nanyang Technology University. [Online]. Available: http://www.csie.ntu.edu.tw/p88012/AI-final.pdf

[48] S. Sundaram and T. Santhanam, "A comparison of blood donor classification data mining models," *J. Theor. Appl. Inf. Technol.*, vol. 30, no. 2, pp. 98–101, 2011.

[49] S.-T. Luo, B.-W. Cheng, and C.-H. Hsieh, "Prediction model building with clustering-launched classification and support vector machines in credit scoring," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7562–7566, 2009.

[50] J. Chen, Y. Y. Tang, C. L. P. Chen, B. Fang, Z. Shang, and Y. Lin, "NNMap: A method to construct a good embedding for nearest neighbor classification," *Neurocomputing*, vol. 152, pp. 97–108, Mar. 2015.

[51] M. Panda, A. Abraham, and M. R. Patra, "A hybrid intelligent approach for network intrusion detection," *Procedia Eng.*, vol. 30, pp. 1–9, 2012.

[52] L. Drumond, N. Thai-Nghe, T. Horváth, and L. Schmidt-Thieme, "Factorization techniques for student performance classification and ranking," in *Proc. UMAP Workshops*, 2012, pp. 1–6.

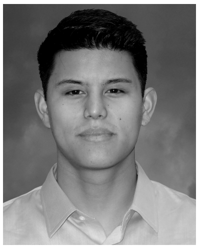
[53] M. D. McDonnell, M. D. Tissera, T. Vladusich, A. van Schaik, and J. Tapson, "Fast, simple and accurate handwritten digit classification by training shallow neural network classifiers with the 'extreme learning machine' algorithm," *PloS ONE*, vol. 10, no. 8, p. e0134254, 2015.

[54] Y.-C. Yin, C.-S. Lee, and Y.-P. Wong. (2012). Demand prediction of bicycle sharing systems. Stanford University. [Online]. Available: http://cs229.stanford.edu/proj2014/Yuchun%20Yin,%20Chi-Shuen%20Lee,%20Yu-Po%20Wong,%20Demand%20Prediction%20of%20Bicycle%20Sharing%20Systems.pdf

[55] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy Buildings*, vol. 49, pp. 560–567, Jun. 2012.

[56] J. P. Nobrega and A. L. I. Oliveira, "Kalman filter-based method for online sequential extreme learning machine for regression problems," *Eng. Appl. Artif. Intell.*, vol. 44, pp. 101–110, Sep. 2015.

SOHINI ROYCHOWDHURY received the Ph.D. degree in electrical and computer engineering from the University of Minnesota in 2014 and the M.S. degree from Kansas State University in 2010. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Washington, Bothell. Her research interests include image processing, signal processing, pattern recognition, machine learning, artificial intelligence, low power system design, and cloud computing. She is a recipient of two best paper awards, one best poster award and one best paper finalist at the Osmosis Student Paper Contest (2006), the IEEE Student Paper Contest Alborg University (2007), the IEEE Asilomar Signals, Systems, and Computers Conference (2012), and the Institute of Engineering and Medicine Conference (2013). She also received the Graduate School Fellowship for the year 2010 and numerous travel grants at the University of Minnesota.



MATTHEW BIHIS received the B.S. (*cum laude*) degree in electrical and computer engineering from the University of Washington, Bothell, in 2016. He has been an Active Member of the IEEE UWB Chapter and the Tau Sigma National Honor Society. He participated in the UW NASA Summer Undergraduate Research Program 2015. He is a recipient of the Mary Gates Scholarship from the University of Washington in 2015 and 2016.