INVESTIGATING SUPER LEARNER ON HEALTHCARE DATA SETS

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Humaira Rahman

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

April 2021

Fargo, North Dakota

North Dakota State University
Graduate School

Title

INVESTIGATING SUPER LEARNER ON HEALTHCARE DATA SETS

By

Humaira Rahman

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Simone A. Ludwig
Chair

Dr. Pratap Kotala

Dr. María de los Ángeles Alfonseca-Cubero

Approved:

04/16/2021 — Date

Dr. Simone Ludwig — Department Chair

ABSTRACT

In the field of machine learning, classification is the essential task that predicts the target class or label for each sample in the data. Improving the performance of a classification model has been a challenging research problem. Researchers try to choose the proper techniques and combine several algorithms to be applied to the specific data set to get better predictions. Nowadays, researchers have used the method called super learner. The idea of super learning is that it combines multiple techniques as base learners and uses a meta-learner to get the final predictions and thus obtain more reliable results. In this paper, we investigated the super-learning techniques on various healthcare data sets. We displayed the results and compared the results with the single machine learning techniques that we choose as base learners. We observed that super learning provides more dependable performance than the individual machine learning methods in most cases.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Simone A. Ludwig, my research advisor, and my family members who supported me in the process of achieving my goals. I am also grateful to Dr. Pratap Kotala and Dr. María de los Ángeles Alfonseca-Cubero for their guidance and time to serve in the supervisory committee.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Based on the task's objective, Machine learning (ML) algorithms [1-3] can be classified into two groups, namely supervised or unsupervised learning. Supervised approaches are used when data has a label, and such a variable is referred to as a response, output, dependent, or class variable. The data is not labeled for an unsupervised method, and there is none to predict or classify. Classification is a supervised learning strategy in machine learning and statistics, in which a computer program learns from the input data and then applies that learning to classify new observations or data. There are binary classification or multi-class classification based on the target variable. An example of binary classification is identifying whether the person is male or female or if the email is spam or non-spam. In multi-class problems, we have more than two values to be predicted. Classification techniques have been used extensively in various fields, including healthcare, bioinformatics, network security, software engineering, and business. Researchers are striving to ascertain which algorithm will perform best given a specific research problem and available data. The primary purpose of machine learning methods is to create a model that can be used for classification, prediction, estimation, or any other task that requires classification [1-2].

The performance of the classification model is crucial [3-5]. Generally, the accuracy that defines the percentages of unknown instances that the model can correctly classify is widely practiced determining the performance. Moreover, the model’s sensitivity, specificity, F1-1 score, and area under the curve (AUC) are utilized to assess its performance. Researchers are using a single classifier to achieve better performance for the available data sets. Choosing the most suitable machine learning model for a given problem, on the other hand, is a difficult task, and there is no simple or straightforward solution for dealing with multiple problems at once.

Indeed, even if various models are well suited to a particular problem, finding one that performs optimally for different distributions may be challenging. To create a better model, the ensemble learning model allows us to combine various models or classifiers. Multiple learning algorithms are used to ensemble machine learning methods to obtain better performance than any single learning algorithms. The super learner is an ensemble machine learning algorithm that combines several algorithms to get better predictive performance. And generally, super learning provides a prediction as-good-as or better than any individual algorithms [6-7].

Researchers have used this technique to a group of base learners to improve predictive performance. The super-learning algorithm is a supervised learning technique that finds the best combination of several prediction algorithms. The Super learner algorithm applies stacked generalization, also known as stacking or blending, to k-fold cross-validation, in which all models use the same k-fold splits of the data and a meta-model is fitted to each model's out-of-fold predictions. We present two different types of super learners or stacked ensembles in this paper. The first technique uses two base learners, namely gradient boosting machine (GBM) and Distributed random forest (DRF), and the second one employs three base learners, namely GBM, DRF, and deep neural network (DNN). We use five well-known health care data sets to compare both super learners' performance with the individual base learners. Our evaluations confirm that the super learner provides better results than individual algorithms used as base learners on five different data sets that we have used in this paper.

2. RELATED WORK

Super Learner has tremendous potential to enhance prediction algorithms' quality in applied health sciences and reduce the dependency on parametric modeling assumptions in empirical findings [8]. In practical non-parametric and semi-parametric statistical models based on actual knowledge, the Super Learner algorithm allows researchers to use different algorithms to outperform a single algorithm. Any mapping from data to a predictor is referred to as an algorithm. This can include anything from simple logistic regression to more complex algorithms like neural nets [8]. The authors suggested using a k-fold CV to generate level one data and extended the previous stacking framework to regression problems. The authors also proposed nonnegativity constraints for the meta-learner. A general framework for stacking was proposed that combined regression and classification estimates and compared CV-generated level-one data to bootstrapped level-one data [9]. However, there is a concern that these methods may over-fit the data and may not be the best way to combine the candidate learners [10]. Ensemble or combining learners in different methods showed better performance than a single-candidate learner. Researchers propose a solution in the form of a new learner, which they call a super learner. In the context of prediction, a super learner is a prediction algorithm that applies a set of candidate learners to observed or training data and selects the best learner for a given prediction problem based on cross-validated risk [7][10]. Super learning is becoming increasingly popular for dynamic accuracy prediction in a variety of domains.

In [7], the authors investigated two forms of super learner techniques with two and three machine learning algorithms, respectively, on four different health care data. Their research showed that the super learner provides better performance than the single methods employed in their study. They also explained that super-learners with three base learners perform better than

the super learner with two machine learning algorithms. Furthermore, [11] researchers used a super learning model in anomaly detection. They showed that it provides a better decision than any single model, such as Naive Bayes (NB), decision tree (DT), neural network (NN), support vector machine (SVM), K-nearest neighbors (KNN), and Random Forest (RF) [11].

3. DATA SETS

To measure the performance of our proposed models, we used five different healthcare data sets. Among them three data sets were collected from the UCI Machine Learning Repository [12-14]. It is a free data set domain which is available to the public. And other two data sets were obtained from IEEE data port [15-16]. IEEE data port is globally accessible data platform which is developed and offered by IEEE that provides significant benefits to researchers, data analysts, and the global technical community.

The first data set that we used is the Cervical Cancer Risk Factor data set. This data set focuses on the prediction of indicators/diagnosis of cervical cancer. There are 858 records, and each record has 36 attributes. This data set contains information of patients who attended gynecology service between 2012 and 2013 at Hospital Universitario de Caracas in Caracas, Venezuela [12][17].

Another data set is the EEG Eye State data set, which contributes EEG measurement with the Emotive EEG Neuroheadset. The duration of the measurement was 117 seconds. The eye state was detected via a camera during EEG measurement and added later manually to the file after analyzing the video frames where 1 indicates “the eye-closed” and 0 represents “the eye-open state”. This data set consists of 14,980 instances with 15 attributes and each record is representing the values of the electrodes and the eye state. Among the instances 8,255(55.12%) of the data set corresponds to the eye open and 6,722 (44.88%) instances to the eye closed state [13] [18].

The third data set we used is the Coronary Artery Disease data set. This data set consists of 1,190 instances with 11 features. The data was collected from the four following locations: Cleveland Clinic Foundation, Hungarian Institute of Cardiology, V.A. Medical Center, Long

Beach, CA and statlog project. The instance number for each data set is as follows: Cleveland:303, Hungarian:294, Long Beach VA:200 and Statlog project:270. These data sets were collected and combined to help research using machine learning algorithm on Coronary Artery Disease data set to detect advance clinical diagnosis and early treatments [15] [19].

The fourth data set is cardiovascular disease data set, which is collected from IEEE data port. The data consists of 70,000 patient records (34,979 presenting with cardiovascular disease and 35,021 not presenting with cardiovascular disease) and contains 11 features (4 demographic, 4 examination, and 3 social history). The data set has the following attributes:

- Age (demographic)
- Height (demographic)
- Weight (demographic)
- Gender (demographic)
- Systolic blood pressure (examination)
- Diastolic blood pressure (examination)
- Cholesterol (examination)
- Glucose (examination)
- Smoking (social history)
- Alcohol intake (social history)
- Physical activity (social history)

Some features are numerical, others are assigned categorical codes, and others are binary values. The classes are balanced, but there were more female patients observed than male patients. Further, the continuous-valued features are almost normally distributed [16] [20].

The final data set is a Hospital Readmission data set. The data was collected from the UCI repository from 130 hospitals in the U.S. for 10 years (1999-2008). The data set represents from the year 1999 to 2008 (10 years) of clinical care at 130 US hospitals and integrated delivery networks. It includes 17 features with 69,008 observations representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria:

1. It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

The data contains such attributes as patient gender, age, weight, number of lab test performed, HbA1c test result, and other health issues [14][21]. Table 1 shows the summary of the data sets [12-16] that we investigated in this paper.

For all of the data sets used in this report, we generated the training data and the test data. For training the model we used 75% of the data, while the remaining 25% was utilized for the test set. For the splitting technique, we used to scikit-learn (sklearn), a machine learning library for the Python programming language.

Table 1: Summary of the data sets.

Data sets	**Number of Attributes**	**Number of Instances**	**Class label**
Cervical Cancer	**3**6	858	Class 0: not cervical cancer Class 1: cervical cancer
EEG Eye State	15	14,980	Class 0: eye open Class 1: eye closed
Heart-Statlog	11	1,190	Class 0: no heart disease Class 1: heart disease
Cardio	11	70,001	Class 0: no cardiovascular disease Class 1: cardiovascular disease
Hospital Readmission	17	69,008	Class 0: not readmitted. Class 1: readmitted.

4. APPROACH

In this paper, we utilized the Super learning, also known as a stacked ensemble method. Super learning is a machine learning technique that uses two or more learning algorithms [8]. It is a supervised learning strategy that uses a loss-based method to find the best prediction algorithms. It is a cross-validation-based method for combining machine learning algorithms that generate predictions at least as good as the traditional single machine learning algorithm [6-7]. The detailed description and related techniques that we used in this paper are briefly outlined in this section.

4.1. Super Learning

Stacking is a broad class of algorithms that involves training a second level "meta learner" to ensemble several machine learning techniques. The type of ensemble learning implemented in H2O is called super learning, stacked regression, or stacking [22]. Super learning, also known as stacking, is an ensemble learning technique in which a meta-learner is educated on a group of base learners' performance. Cross-validation can be used to produce the results from the base learners, also known as level-one data. The original training data set is often referred to as the level-zero data. The K cross-validated predicted values from each of the M algorithms can be merged to form a new K x M matrix. This matrix, along with the original class vector, is called the level-one data [6-7]. Super learning consists of meta-learner and some traditional machine learning algorithms. The meta-learner [22 -23] and the machine learning algorithms that we utilized in this research are outlined as follows.

4.1.1. Base Learner

An ensemble is comprised of a group of learners known as base learners. An ensemble's generalization capacity is typically much higher than that of base learners [24-26]. Ensemble

learning is desirable because it can elevate weak learners who are just marginally better than a random guess to strong learners who can make incredibly accurate predictions. As a result, “poor learners” are often referred to as “base learners.” However, that while most theoretical studies focus on weak learners, the base learners used in practice are not always weak, as using less-weak base learners also leads to better results. Typically, base learners are derived from training data. A base learning algorithm, such as gradient boosting machine, Random forest, deep neural network, or other form of machine learning algorithm, is used to generate base learners from training data. Most ensemble methods produce homogeneous base learners using a single base learning algorithm, but some methods produce heterogeneous learners using multiple learning algorithms.

For this paper, to evaluate the better performance, we choose three base learners from H2O, namely gradient boosting machine (GBM), distributed random forest (DRF), and deep neural network (DNN). The approach that we applied here is a super learner. The super learner is also called a stacked ensemble. H2O’s stacked Ensemble method is a supervised ensemble machine learning algorithm that finds the optimal combination of a prediction algorithm collection using a process called stacking. The base learner that we used in this paper are described below [22-24].

4.1.2. Gradient Boosting Algorithm (GBM)

Gradient boosting is one of the most powerful techniques for building predictive models. The Gradient Boosting Machine, or GBM, generates final predictions by merging predictions from various decision trees. Keep in mind that in a gradient boosting machine, all of the weak learners are decision trees. It builds the model in stages and extends it by allowing the use of every differentiable loss function. GBM is a regression and classification forward learning

ensemble system. The leading heuristic is that good predictive results can be obtained with increasingly more refined approximations. GBM is part of H2O, a distributed, open-source, Java-based machine learning platform for big data. H2O's GBM constructs regression trees on all of the data's features in a sequential manner. Each tree is built in parallel in an entirely distributed way. The program has been updated to include new functionality. We utilized the modernized features available in H2O, such as per-row observation weights-fold cross-validation, per-row offsets, and support for a more significant number of distribution functions (Gamma Poisson and Tweedie) [6].

4.1.3. Distribute Random Forest (DRF)

Distributed Random Forest (DRF) is a robust classification and regression algorithm. The DRF method is a nonlinear predictive technique based on a forest of decision trees [22-23]. A variety of decision trees are trained using randomly selected training data subsets. Training subsets are chosen at random to reduce the variance of the model. DRF generates a forest of classification or regression trees instead of a single one when given a set of data. Each of these trees is based on a subset of rows and columns and is thus a weak learner. There will be a minor variation if there are more trees. In order to make a final prediction, both classification and regression use the average forecast across all of their trees.

4.1.4. Deep Neural Network (DNN)

Deep Neural Networks (DNN) is typically referred to as Feed Forward Networks (FFNNs), in which data flows from the input layer to the output layer [27], and the links between the layers are only one way, forward, and never touch a node again. H2O's Deep Learning is based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation [22] [27]. The network can contain many hidden layers

consisting of neurons with tanh, rectifier, and max-out activation functions. Advanced features such as adaptive learning rate, rate annealing, momentum training, dropout, L1 or L2 regularization, checkpointing, and grid search enable high predictive accuracy [7]. Each compute node trains a copy of the global model parameters on its local data with multi-threading (asynchronously) and contributes periodically to the global model via model averaging across the network.

4.1.5. Meta-learner

H2O's Stacked Ensemble method is a supervised ensemble machine learning algorithm that learns the optimal combination of a collection of prediction algorithms using a process called stacking. The algorithm that learns the base learners' optimal combination is called the meta-learning algorithm or meta learner. By default, the Stacked Ensemble meta learner is GLM with non-negative weights.

4.2. Experiment Set-up

For this experiment, we followed the steps below to set up the super learner.

4.2.1. Classification Model Data and Sample Data for Classification

We create the classification model data and sample data for classification, with the training data set's class information known and the testing data set's class information as unknown. The data sets are known as level-0 data, where X is the training data set with n rows and m columns, and the class value column is separated from the training data set, which is referred to as Y.

4.2.2. Classifiers and Model Selection

We determined the base learners and a meta-learner algorithm to set up the stacked ensemble or super learner. We began by choosing two base learners, GBM and DRF, for this

study. After that with the previous two base learners, we also selected DNN as a third base learner. We used the Cartesian grid search and defined a set of values for specific parameters to search over each base learner for the model selection process. Base learners and meta-learner selection process and training are described as follows.

4.2.2.1. Base Learner set-up

We used the grid search to train GBM, RF, and DNN on the training data set with the same parameters obtained from the grid search. Each of these algorithms, ten-fold cross-validation were used, and for that, the cross-validation prediction parameter was set to true. Since the response column is categorical with two groups, the Bernoulli distribution was selected for all three base learners. Besides, the fold assignment modulo was chosen for the base learners, which is a simple deterministic way to divide the data set into folds equally. It's essential to recognize that in our tests, we used two base learners (GBM and DRF) at first and then three base learners (GBM, DRF, and DNN).

The best model was selected for each of the base learners based on the mean squared error (MSE), which is the average squared difference between the computed values and the actual values. This was done once the grid search on the training data was complete, and then we queried the grid object and sorted the models by the performance metric MSE. Finally, for each base learner, the model with the minimum MSE was selected.

4.2.2.2. Meta-learner set-up

For the meta-learner in the stacked ensemble, we used GLM as default meta learner [22-23]. GLM was used to train the level-1 data. GLM uses the default parameters for training data. Here, it is kept in mind that all base models must have been cross validated, and they must all use the same cross validation folds for the stacked ensemble to work. In addition, the value of a

parameter called “keep cross-validation prediction” was set to true. In our case, we used ten-fold cross-validation and set the keep cross-validation prediction parameter to true for all of the base learners to consider this.

4.2.2.3. Output generation

The final part was to use the super learner or ensemble-model to find predictions on the test or unknown data.

4.3. Model Evaluation Criteria

To measure the performance of our model, several evaluation measures were used such as sensitivity, specificity, accuracy, and AUC. These were derived from the confusion matrix and applied to the classifier evaluation [3] [4][5][7][28] and are shown in Equations (1) through (4).

$$\text{Accuracy} = (TP + TN)/ (TP + FP + TN + FN) \quad (1)$$

$$\text{Sensitivity} = TP/ (TP + FN) \quad (2)$$

$$\text{Specificity} = TN/ (TN + FP) \quad (3)$$

$$\text{F-1 Score} = 2TP / (2TP + FP + FN) \quad (4)$$

Here:

TP = number of positive examples correctly classified

TN = number of negative samples correctly classified

FN = number of positive observations incorrectly classified

AUC: Area Under the ROC Curve

AUC stands for Area under the ROC Curve [28] that measures the entire two-dimensional area underneath the entire ROC curve. It is a metric for measuring the performance of learning algorithms. An outstanding model has an AUC close to 1, indicating that it has a high level of separability. AUC near 0 indicates a weak model, which means it has the lowest measure

of separability. In fact, that means that the result is being reciprocated. It predicts 0s to be 1s and 1s to be 0s. When AUC is 0.5, the model cannot distinguish between classes.

5. RESULTS

This section presents the experimental results and performance evaluation of our model. For our experiment, we used H2O. We favored Python as the programming language for the implementation using H2O. It is worth mentioning that we used 75 % as training data, and the rest of the 25 % were applied for test data.

This paper compares our approach to the individual base learners used in this investigation. Test data set is considered to determine the performance of the model. Table 2 shows the super learner's performance consisting of two base learners, GBM and DRF, on the test data for various data sets. Table 3 shows the super learner's performance composed of three base learners, specifically GBM, DRF, and DNN, on test data for all the data sets used in this research.

Table 2: Performance of the super learner consisting of two base learners - GBM and DRF.

Data set	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**F1 Score (%)**	**AUC**
Cervical Cancer	98.95	77.78	97.13	98.44	0.9698
EEG Eye State	94.87	97.44	95.93	96.36	0.9878
Heart Statlog	91.04	93.17	92.20	98.15	0.9448
Cardio	77.84	66.75	71.08	67.76	0.8007
Hospital Readmission	77.77	67.02	71.26	68.12	0.8004

The performance in terms of accuracy for all base learners and super learners consisting of two algorithms, namely GBM and DRF on test data set, is shown in Table 4. The performance in terms of AUC for all base learners and super learners consisting of three base learners, specifically GBM, DRF, and DNN on the test data set, is shown in Table 5. For Table 4 and Table 5, bold values indicate the best results.

Table 3: Performance of the super learner consisting of three base learners - GM, DRF and DNN.

Data set	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**F1 Score (%)**	**AUC**
Cervical Cancer	99.47	71.43	96.65	98.16	0.9724
EEG Eye State	95.42	96.85	96.04	96.45	0.9883
Heart-Statlog	92.13	90.48	91.19	90.00	0.9435
Cardio	76.69	68.76	72.14	70.17	0.8011
Hospital Readmission	76.74	68.91	72.27	70.34	0.9271

Table 4: Accuracy comparison using single-base learners, and super learner with two base and three base learners (**Bold** indicates the best value)

Data set	**Accuracy (%) GBM**	**Accuracy (%) DRF**	**Accuracy (%) DNN**	**Accuracy (%) SL with GBM and DRF**	**Accuracy (%) SL with GBM, DRF, and DNN**
Cervical Cancer	96.65	95.69	96.65	**97.13**	96.65
EEG Eye State	**96.14**	91.93	**96.14**	96.03	96.04
Heart-Statlog	91.19	**92.20**	91.19	**92.20**	91.19
Cardio	71.94	71.47	71.94	71.08	**72.14**
Hospital Readmission	71.39	71.19	71.39	71.26	**72.27**

For the Cervical Cancer Risk Factor data set, the super learner with three base learners has the best accuracy (97.13%), followed by the individual learner DNN (96.65%). For the EEG Eye state data set, the best accuracy (96.14%) was obtained with both Super Learner methods followed by individual algorithms GBM and DNN.

Table 5: AUC comparison using single-base learners, and super learner with two base and three base learners (**Bold** indicates the best value)

Data set	**AUC GBM**	**AUC DRF**	**AUC DNN**	**AUC SL with GBM and DRF**	**AUC SL with GBM, DRF, and DNN**
Cervical Cancer	0.9666	0.9711	0.9666	0.9698	**0.9724**
EEG Eye State	**0.9903**	0.9689	**0. 9903**	0.9878	0.9883
Heart-Statlog	0.9394	**0.9448**	0.9394	**0.9448**	0.9435
Cardio	**0.8019**	0.7994	**0.8019**	0.8007	0.8011
Hospital Readmission	0.8016	0.7992	0.8016	0.8004	**0.8019**

Similar trends are also observed in Table 5; the best AUC value was obtained using the super learner having three base learners for all the data sets used in this research. For the cervical cancer data, the best AUC was obtained when SL consisting of two base learners and three base learners.

For the EEG Eye state data set, the best AUC value (0.9903) was reported with GBM and DNN followed by SL consisting of three base learners (0.9883). For SL with two base learners, the AUC value was observed (0.9878).

For the Heart-Statlog-Cleaveland data set, the best AUC (0.9448) was obtained with SL consisting of two base learners and DRF followed by SL with three base learners (0.9435). For the cardio data set, the highest AUC values (0.8019) were attained for both GBM and DNN, followed by SL with three (0.8011) and two base learners (0.8007), respectively. Here, it is to be mentioned that for this data set, the individual algorithm provides better results than the super learners, but for accuracy, the super learner provides better performance.

Table 6: Sensitivity comparison using single-base learners, and super learner with two base and three base learners (**Bold** indicates the best value)

Data Sets	**Sensitivity GBM**	**Sensitivity DRF**	**Sensitivity DNN**	**Sensitivity SL with GBM and DRF**	**Sensitivity SL with GBM, DRF, and DNN**
Cervical Cancer	99.47	98.94	99.47	98.95	**99.47**
EEG Eye State	95.32	91.2	95.32	94.97	**95.42**
Heart-Statlog	90.84	91.04	90.84	91.04	**92.13**
Cardio	76.89	77.12	76.89	**77.84**	76.69
Hospital Readmission	77.63	77.68	77.63	**77.77**	76.74

For the hospital readmission data set, the best values for AUC obtain when we applied the super learner with three base learners (0.8019) followed by individual algorithms GBM and DNN (0.8016).

For the Cervical Cancer data set, the best Sensitivity (99.47%) was obtained with SL consisting of three base learners and GBM and DNN followed by SL with three base learners (99.47%). For the EEG Eye State data set, the highest Sensitivity value (95.42%) was attained for SL with three base learners. From Table 6, we can find that the Sensitivity values came marginally better for super learners than the single base learner for all the data sets.

From Table 7, the best Specificity values have been found in the super learner methods for both two base and three base learners. The best Specificity was achieved (97.44%) with Super Learner using two base learners following DNN (97.21%). In the Heart-Statlog data set, the Specificity value was found with SL using two base learners (93.17%) followed by DRF (93.17%).

For the Cervical Cancer data set, we achieved the highest F-1 score (98.44%) with SL consisting of two base learners, GBM and DNN, followed by SL with three base learners (98.16%).

Table 7: Specificity comparison using single-base learners, and super learner with two base and three base learners (**Bold** indicates the best value)

Data set	**Specificity GBM**	**Specificity DRF**	**Specificity DNN**	**Specificity SL with GBM and DRF**	**Specificity SL with GBM, DRF, and DNN**
Cervical Cancer	71.43	66.67	71.43	**77.78**	71.43
EEG Eye State	97.21	92.9	97.21	**97.44**	96.85
Heart-Statlog	91.46	93.17	91.46	**93.17**	90.48
Cardio	68.35	67.57	68.35	66.75	**68.76**
Hospital Readmission	67.25	66.97	67.25	67.02	**68.91**

For the EEG Eye State data set, we attained the highest F-1 Score value (96.54%) for SL with three base learners. From Table 8, we can observe that the F-1 Score values came marginally better for all the data sets than the single base learner. We found the best F-1 score for Heart-Statlog data set for SL consisting of two base learners (91.39%) followed by DRF (91.39%). Most reliable results were attained using the super learner methods (either SL consisting of two base learners or SL composed of three base learners) in most of the cases. However, in few cases, we observed that individual machine learning algorithms provide slightly better or equal performance.

Table 8: F-1 score comparison using single-base learners, and super learner with two base and three base learners (**Bold** indicates the best value)

Data set	**F-1 Score GBM**	**F-1 Score DRF**	**F-1 Score DNN**	**F-1 Score SL with GBM and DRF**	**F-1 Score SL with GBM, DRF, and DNN**
Cervical Cancer	98.16	97.64	98.16	**98.44**	98.16
EEG Eye State	96.54	92.8	**96.54**	96.46	96.45
Heart-Statlog	90.15	91.39	90.15	**91.39**	90
Cardio	69.71	68.8	69.71	67.76	**70.17**
Hospital Readmission	68.41	68.04	68.41	68.12	**70.34**

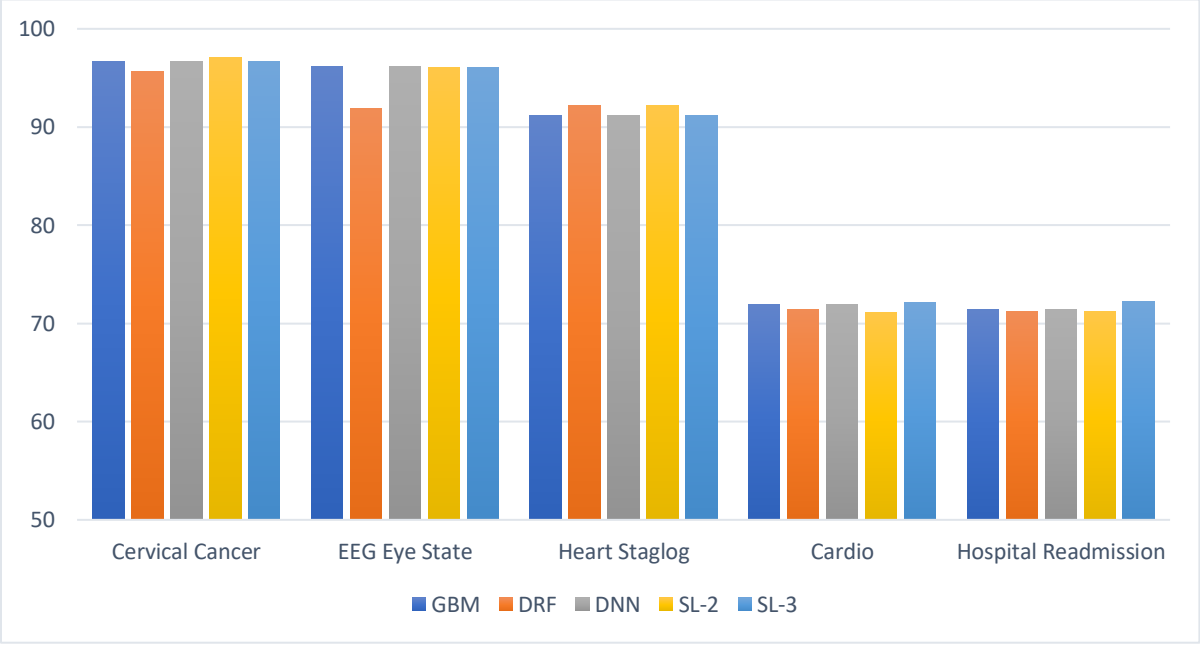

Figure 1: Accuracy comparison in super learners with individual base learner

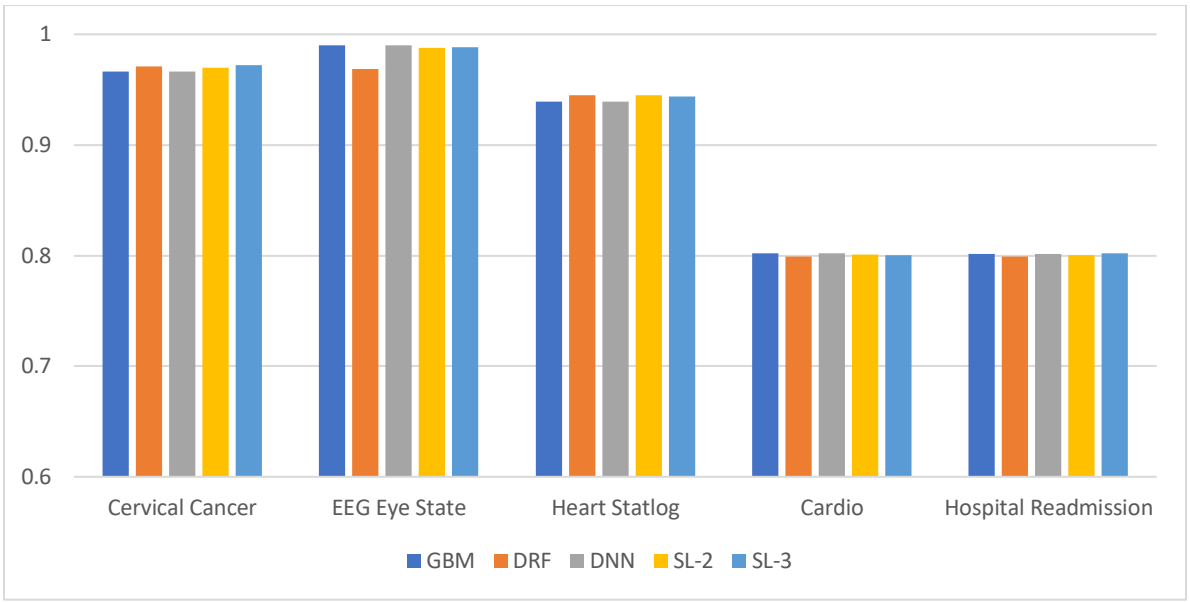

Figure 2: AUC comparison in super learners with individual base learner

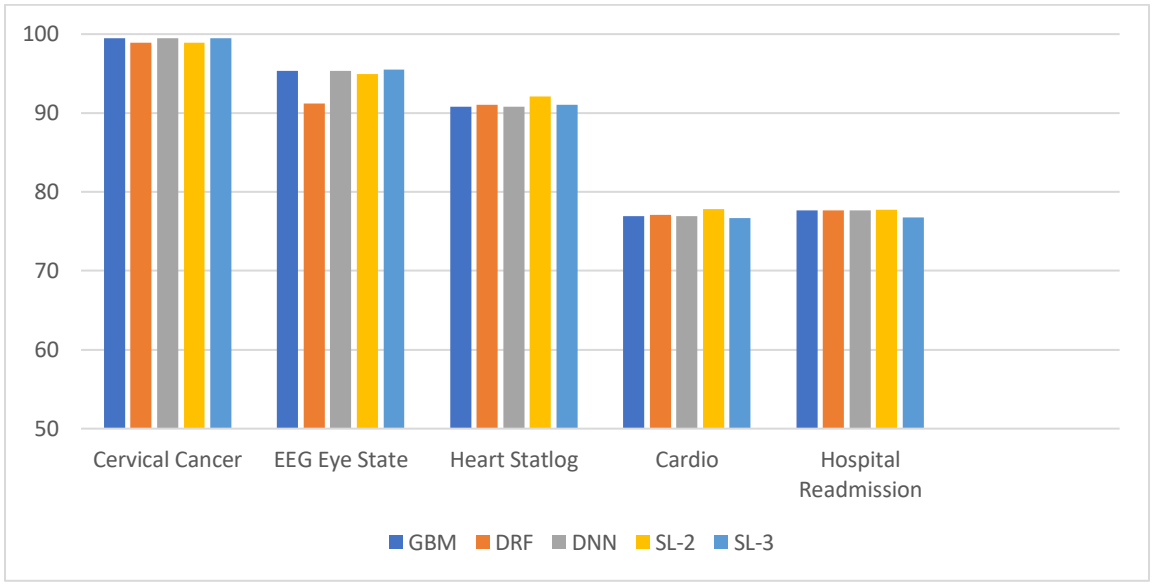

Figure 3: Sensitivity comparison in super learners with individual learner

For the Cervical Cancer Risk Factor data set, the super learner with three base learners has the best Sensitivity (99.47%) followed by the individual learner DNN (99.47%). For the EEG Eye state data set, the best Sensitivity (95.42%) was obtained with the super Learner method followed by Individual algorithms GBM and DNN (95.32%).

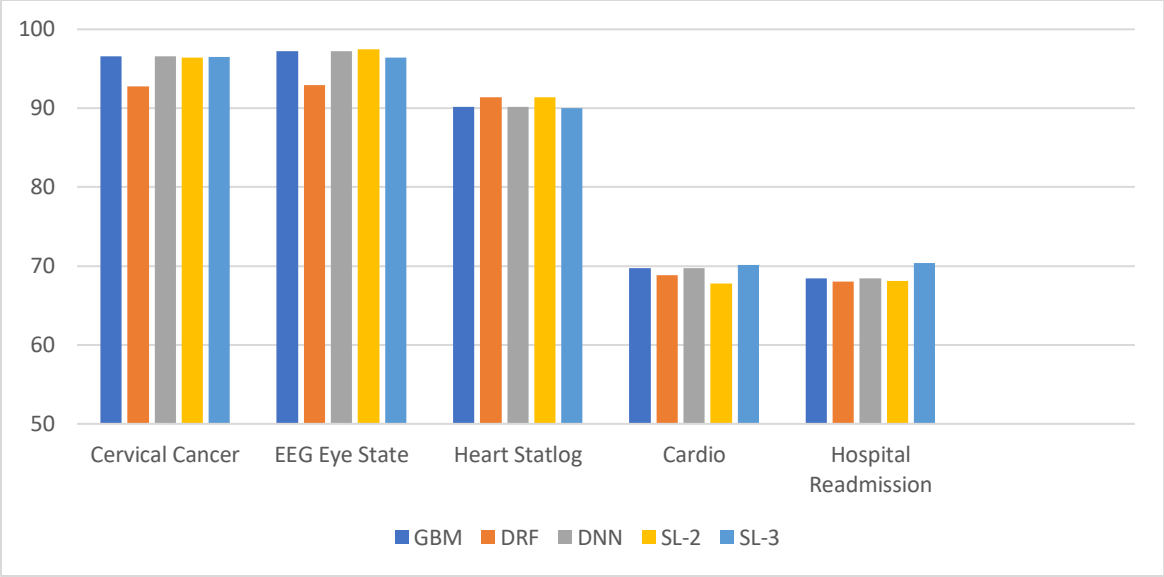

Figure 4: Specificity comparison in super learner with individual base learner

For the EEG Eye State data set, the super learner with two base learners has the best Specificity (97.44%) followed by the individual learner DNN (97.21%). For Heart-Statlog data set, the best Specificity (93.17%) was obtained with the Super Learner method followed by Individual algorithms DRF.

For the Cervical Cancer Risk Factor data set, the super learner with two base learners has the best F-1 Score (99.47%), followed by the Individual algorithms GBM and DNN. For Heart-Statlog data set, the best F-1 Score (92.13%) was reported with the Super Learner method followed by Individual algorithm DRF (91.04%).

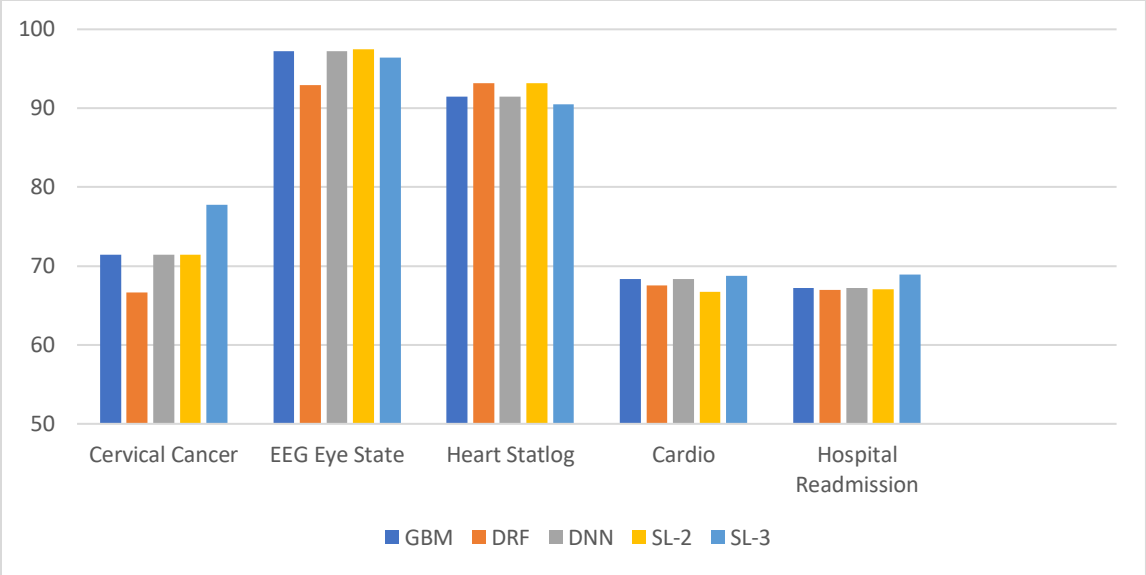

Figure 5: F-1 Score Comparison in super learners with individual base learner

In Figure 1 to Figure 5, we illustrated the comparisons of super learners with the single base learning algorithms for all the five data sets that have been investigated in this paper. Figure 1 and Figure 2 compares the accuracy and AUC respectively. Figure 3, Figure 4, and Figure 5 demonstrates the sensitivity, specificity, and F-1 score respectively for all the data sets.

For all the data sets used in this research, the best results were attained using the super learner methods (either SL consisting of two base learners or SL composed of three base learners) in most of the cases. However, in few cases, we observed that individual machine learning algorithms provide slightly better or equal performance. The result can be observed from Table 2 through Table 8 and Figure 1 to Figure 5.

6. CONCLUSION

One of the essential tasks of machine learning is classification, which predicts the target class for each example in the data. Researchers are usually using single classifiers to achieve better performance on the available data sets. Choosing the best data mining or machine learning algorithm for a particular problem is challenging. Since these researchers use a variety of models to solve a problem, they can achieve good results. We focused on investigating the classification performance in terms of sensitivity, precision, accuracy, and AUC for five healthcare data sets in this paper. For that, we used the super learning or stacked-ensemble method that finds the optimal weighted average of diverse learning models. We used GBM and RF for the base learners, followed by another base learner DNN with the previous two—GBM and RF.

We found that super learning outperforms individual base learners and other machine learning techniques based on our experimental findings for these five healthcare data sets. We obtained better or equivalent performance using the stacked ensemble or super learner methods (using two or three base learners) compared to individual base learners for all the measurement metrics considered in this study.

As for future work, we will examine research problems by including more diverse base learners and other meta-learners. The super learning technique could also be applied to different real-world problem domains such as cybersecurity, software engineering, bioinformatics.

REFERENCES

[1] Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124.

[2] Rahman, S. M., Kabir, M. F., & Rahman, M. M. (2014). Integrated Data Mining and Business Intelligence. In Encyclopedia of Business Analytics and Optimization (pp. 1234-1253). IGI Global.

[3] Kabir, M. F., & Ludwig, S. A. (2018, December). Classification of breast cancer risk factors using several resampling approaches. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1243-1248). IEEE.

[4] Kabir, M. F., & Ludwig, S. A. (2019, December). Classification Models and Survival Analysis for Prostate Cancer Using RNA Sequencing and Clinical Data. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 2736-2745). IEEE.

[5] Kabir, M. F., Ludwig, S. A., & Abdullah, A. S. (2018, December). Rule discovery from breast cancer risk factors using association rule mining. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 2433-2441). IEEE.

[6] LeDell, E. (2016). Scalable Super Learning. Handbook of big data, 339.

[7] Kabir, M. F., & Ludwig, S. A. (2019). Enhancing the performance of classification using super learning. Data-Enabled Discovery and Applications, 3(1), 5.

[8] Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: an introduction to super learning. European journal of epidemiology, 33(5), 459-464.

[9] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[10] Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6(1).

[11] Van der Laan, M. J., & Rose, S. (2011). Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media.

[12] UCI Machine Learning Repository: Cervical Cancer Risk Factor Data Set, https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29, last retrieved: March 2021.

[13] UCI Machine Learning Repository: EEG Eye State Data Set, https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State, last retrieved: March 2021.

[14] UCI Machine Learning Repository: Hospital Readmission with Diabetes Dataset, https://www.hindawi.com/journals/bmri/2014/781670, last retrieved: March 2021.

[15] IEEE DATAPORT: Heart Disease, https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive, last retrieved: March 2021.

[16] IEEE DATAPORT: Cardiovascular Disease Dataset, https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive, last retrieved: March 2021.

[17] Lu, J., Song, E., Ghoneim, A., & Alrashoud, M. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. Future Generation Computer Systems, 106, 199-205.

[18] Rösler, O., & Suendermann, D. (2013). A first step towards eye state prediction using eeg. Proc. of the AIHLS.

[19] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., ... & Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 111(1), 52-61.

[20] Padmanabhan, M., Yuan, P., Chada, G., & Nguyen, H. V. (2019). Physician-friendly machine learning: A case study with cardiovascular disease risk prediction. Journal of clinical medicine, 8(7), 1050.

[21 Min, X., Yu, B., & Wang, F. (2019). Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD. Scientific reports, 9(1), 1-10.

[22] Aiello, S., Click, C., Roark, H., Rehak, L., & Stetsenko, P. (2016). Machine learning with python and H2O. H2O. ai Inc.

[23] Vanerio, J., & Casas, P. (2017, August). Ensemble-learning approaches for network security and anomaly detection. In Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks (pp. 1-6).

[24] Nykodym, T., Kraljevic, T., Wang, A., & Wong, W. (2016). Generalized linear modeling with H2O. Published by H2O. ai Inc.

[25] Zhou, Z. H. (2009). Ensemble learning. Encyclopedia of biometrics, 1, 270-273.

[26] LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. Journal of the American Statistical Association, 91(436), 1641-1650.

[27] Monzurur Rahman, S. M., Kabir, M. F., & Siddiky, F. A. (2012). Rules mining from multi-layered neural networks. International Journal of Computational Systems Engineering, 1(1), 13-24.

[28] Ling, C. X., Huang, J., & Zhang, H. (2003, August). AUC: a statistically consistent and more discriminating measure than accuracy. In Ijcai (Vol. 3, pp. 519-524).