Reviewing Classification Methods on Health Care

Devvrath Malik and Geetika Munjal

1 Introduction to Supervised Learning

Machine learning is used to program a computer to make predictions or decisions about a certain scenario. The computer achieves this feat by using the experiences it gained while training on a set of data known as "training data." Machine learning can be of two sorts: supervised learning and unsupervised learning. Supervised learning is a technique which maps the input for an output based on input–output pair of examples. Unsupervised learning technique is the one which learns from test data that have not been labeled. The main aim of unsupervised machine learning is to model the underlying structure or distribution in the data to get more knowledge from the data [1].

Different Supervised Learning Methods

The various classification algorithms in machine learning are divided into two broad categories: lazy learners and eager learners. (i) Lazy learner algorithm simply sets aside the training data until the test set data comes up. It classifies the instances of test data by using the stored training set data that is most related to the test set data. Hence, it has higher predicting time compared to eager learners. Two most common lazy learner algorithms are case-based reasoning and K nearest neighbor. (ii) Eager learner algorithm creates a machine learning classifier using given training set data before taking test set data for predictions. Here, a single hypothesis works for the entire dataset and hence is the reason they take more time in training the model and

D. Malik · G. Munjal (✉)
Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India

S. Bhatia et al. (eds.), *Intelligent Healthcare*, EAI/Springer Innovations in Communication and Computing, https://doi.org/10.1007/978-3-030-67051-1_8

less time in making prediction, for example, Naive Bayes classifier, artificial neural networks, and decision trees [2].

Decision tree algorithm is used in both classification and regression (cart) tasks. It builds the classifier model in the form of a tree-type structure. It works by breaking the data into minor groups of data and in the same time frame an associated decision tree is stepwise constructed [3]. An extended version of decision tree is a random forest algorithm wherein large a number of decision trees come together to work as an ensemble classifier model. Each tree in random forest algorithm performs its own operation to predict a class for a set of input attributes and the class that is predicted by majority of decision trees is set as classifier's final output prediction [4]. The low correlation in individual trees is the key to the model's prediction. Another classifier is an artificial neural networks which can be adapted in deep learning model. It is influenced by the working of a human brain as it follows the concept of a neuron. The network comprises of an input layer, successive hidden layers, and an output layer, where each node in a layer represents a single neuron that is interlinked with every single node in the next layer. It is a computational classifier with neurons acting as processing units that receive inputs and deliver outputs based on their corresponding activation functions [5].

Support vector machine (SVM) algorithm is also a supervised learner that can be used for both classification and regression tasks but is preferably used for classifying the data. In SVM, we plot each instance of a data as a single point in an N-dimensional space (where N is the number of input attributes in the given dataset). The value of each feature for a given instance corresponds to the value of the coordinate for that point. The main motive of SVM classifier is to find a suitable hyperplane that can distinctly classify these data points [6]. Another classifier is the K nearest neighbor, which is a powerful yet simple classification algorithm mostly used in pattern recognition and recommendation systems. It is built upon the assumption that similar kinds of things exist close to each other. However, the algorithm becomes significantly slow for large size datasets [7].

Some classifiers are probabilistic such as Naive Bayes classifier which is based on Bayes theorem that provides a principled way to calculate conditional probability. The Naive Bayes algorithm works upon the assumption that all independent variable input predictors are independent of each other and no two different predictor's correlates with each other. In real-life scenario, the probability for this assumption to hold true is quite small. However, even for the data where this assumption does not hold, the Naive Bayes approach works surprisingly well for that data. Mathematically, the Bayes theorem formula is written as:

$$P\ (A|X) = \frac{P\ (X|A) \times P\ (A)}{P(X)} \tag{1}$$

where A and X are the events and $P(X) \neq 0$. The Naive Bayes classification algorithm uses this Bayes theorem equation to classify the input features into different classes [8]. In terms of machine learning ideology, the above equation can be rewritten as:

$$P\ (A_i\ |X)\ = \frac{P\ (X\ |A_i) \times P\ (A_i)}{P(X)} \tag{2}$$

where X represents all input features or independent variables and A_{i} is the *i*th category of output class. In accordance with the Naive Bayes assumption of independent relationship between the input features of data, the probability $P\ (X|A_i)$ can be calculated as the product of each feature's X_j's probability appearing in the category A_i (X_j being the *j*th feature of all the input feature in the dataset). The Naive Bayes algorithm calculates the $P\ (A_i|X)$ for all the i number of categories for a single instance of data and compare their values to select the category with the highest probability value, as the output class for that instance.

Another class of supervised learners is "adaptive," such as AdaBoost or adaptive boosting, where boosting refers to an ensemble technique that combines many weak learner algorithms to create a strong classifier. AdaBoost classification algorithm can be used in conjunction with many different classification algorithms to boost their performances. However, the algorithm is best suited to boost the performance of decision tree classifiers as it is a weak learner for binary classification task. The AdaBoost algorithm makes use of the decision stumps instead of the complete decision tree as weak learners. The decision stumps refer to a decision tree with a depth of one that performs just better than the random classifier [9]. The final classification output of the AdaBoost classifier is the output predicted by that category of a stump whose net significance is higher [10].

Logistic regression is another classification tool taken by machine learning from the area of statistics. It is again a supervised linear classification algorithm that makes use of a logistic sigmoid function to transmute its output prediction into a probability value mapped between 0 and 1. A contradiction seems to occur with the term "regression" being used for classification, but that is what makes it special. The algorithm uses the linear regression equation to give discrete binary outputs. However, unlike the linear regression model that fits a straight line on the data, the logistic regression classifier fits an s-shape curve on the input data that correspond to its sigmoid function [11]. Its popularity has been increased progressively over the last two decades, particularly for binary classification tasks. A simple logistic regression model for binary classification is represented mathematically as shown below:

$$\ln\left(\frac{\tau}{1-\tau}\right) = \alpha + \beta X \tag{3}$$

where τ is the probability that the input X belongs to the default class $y = 1$. Formally, it is written as $\tau(x) = \tau(y = 1|x)$. The ratio on the left side of Eq. 3 is known as the odds ratio of the default class. Simple logistic regression formula can easily be extended to multiple input features $(x_1, x_2, \ldots x_n)$.

Comparative Summary of Supervised Methods

Based on our studies of above algorithms, we have summarized pros and cons of various classifiers in Table 1.

Table 1 Advantages and disadvantages of various classification algorithms

Classifiers	Pros	Cons
K nearest neighbor	No training period is required New training data can be added any moment Very easy to apply	Memory consuming Less efficient with large and high dimensional dataset
Logistic regression	Good accuracy Easy implementation Fast training	Not applicable to nonlinear problems No assumption about classes in feature space
Naive Bayes	Easy implementation and quick prediction [12] Requires less training dataset	Very sensitive about input data May predict wrong if the attributes are correlated
Support vector machine	Works well with clear margin of seperation Efficient with high dimensional datasets Memory efficient Effective even with unlabeled data	Does not work with large datasets Training time is more
Random forest	Very effective Provide a reliable feature importance estimate	Slow prediction Memory consuming High computational cost
Decision trees	Needs less labor in pre processing of data Normalization and scaling of data is not required Classifier model is intuitive	Small change in data leads to instability in the model Complex calculations Adding small amount of data causes big change in the structure of the tree
AdaBoost	Classification accuracy is better Versatile, can be combined with any machine learning algorithm Not much parameter tuning is required	Sensitive to uniform noise Weak learners need to perform better than the random chance If weak learners are too weak, overfitting occurs Learning time is longer
ANN	Ability to learn complex patterns and nonlinear relationships Parallel processing Has fault tolerance	Unknown duration of training the model Large number of parameters to be tuned Follows a black box approach Optimization time is longer

2 Applications of Supervised Learning in Healthcare

Classification methods have been applied in various sectors where health care has a lot of scope where supervised learning can be very beneficial in solving various problems. Some of them are to help in recognizing and tracing long-term diseases and patients with high risk, design appropriate medication, and reduce patients admitted to the hospitals, thus helping in the healthcare governance [13]. Adopting these supervised learning methods will reduce the pressure that the hospitals may face in times of epidemics and pandemics. One of the applications of supervised learning methods is in cardiovascular disease management, which occurs due to acute coronary syndrome (ACS), where a patient comes into the hospital with a chest pain that is mainly caused by benignant causes and thus enormous resources are needed for detection purposes [14]. The answer to this problem lies in using the current resources efficiently. To do this, we can take the help of supervised learning methods which involve the decision with regard to the magnitude of care, logical allotment of resources, and calculation of modifiable risk, which can make the patients better. We can tailor the treatment suitable for a particular patient based on his medical records, personal and family antecedents, electrocardiogram, biomarkers, noninvasive stratification tests, and coronary angiography. Various methods, including artificial neural networks and deep learning, decision trees, and support vector machines, are used for this problem. Machine learning is also helping the radiologists in various forms, such as (i) creating study protocols where machine learning can help radiologists create study protocols based on their priorities. This may involve assessing clinical information and commanding information stored in electronic devices; (ii) Refining image quality and reducing radiation dose in CT, in which there has always been a desire to minimize the dose of radiation during a CT scan. But if we reduce the radiation level in CT scan, it leads to noise in the image that is obtained, hence it is of poor quality. By using deep learning, we are essentially increasing the quality of images even though we are using low doses of radiation. (iii) Optimizing MR scanner utilization as an MRI scan takes a lot of time. Hence, by using machine learning, we analyze patients' clinical record and allot them a time slot accordingly to optimize MRI scan. (iv) Evaluating image quality as it is quite time-consuming. Hence, we use machine learning to automate it [15].

A comparative analysis of decision tree, random forest, multilayer perceptron, and Naive Bayes is presented [16], where decision tree along with correlation-based feature selection has performed well for detecting dementia. Classification algorithms are widely used in breast cancer categorization as well [17]. In medical datasets with large number of input features, preprocessing is required to identify relevant features, followed by classification task [18]. The deploying of classification algorithms in diagnosis has also helped in identifying the possibility of the return of the disease in patients who were cured earlier, in spotting a high-risk disease or illness [19]. It can help in identifying the transition of a patient from one disease state to another. Machine learning algorithms have been recently used in spotting the transition from prediabetes state to type-2 diabetes with the help of electronic health record data [20]. With the advancement of natural language processing (NLP) in machine learning and AI, the researchers and data enthusiast have been able to draw relevant information, insights from the unstructured data generated in the form of clinical reports, performance feedback of a doctor, and from other medical reports of patients after successful disease treatment. The use of classification algorithms in combination with NLP can help not only in drawing patterns from unstructured data into quality and performance, but also in early prediction and diagnosis of a disease. Recently, an automated speech analysis in combination with classification algorithms was performed on the free speech generated from the in-person interviews of individuals at clinical high risk for psychosis [21]. This study was able to predict transition to psychosis state with great accuracy for a group of individuals marked at high risk. With the evolution of health monitoring technology that is heavily dependent on machine learning and artificial intelligence, it is not only possible to keep track of one's health and to predict early symptoms of a disease but also to monitor slightly different aspects of health status like mental fatigue. Mental issues like depression, anxiety, addiction, and behavior disorders have become a serious health issue nowadays, and it comes at a very high public health cost. A recent study on predicting mental fatigue using eye-tracking data was successfully able to detect the problem in individuals with 91% accuracy [22]. There are enormous applications of machine learning algorithms in medicines; quite recently, Google developed a machine learning algorithm that is able to predict the cancerous tumor on mammograms. This new approach obtained an accuracy of 89% compared to 73% of a human pathologist [23].

3 Healthcare Datasets Used in the Study

Considering the importance of classification task on health care, we have tried to analyze performance of various classifiers on various medical datasets available in open platform. The results of all the classifiers are compared on various metrics based on confusion matrix. The data are preprocessed and classified using KNN, Naive Bayes, logistic regression, AdaBoost, decision tree, and artificial neural networks. We have applied all these supervised methods on various healthcare datasets, which are briefed in Table 2, followed by their detailed description.

Cleveland Heart Disease Dataset Heart disease refers to the broad number of health conditions that has a direct impact on a human heart. It is one of the leading reasons behind a large number of deaths across the world. The original source of the dataset is Cleveland Clinic Foundation, which includes about 303 observation samples, each with 13 input predictor attributes, which are gender, age, chest pain category, blood pressure, serum cholesterol, blood sugar level, ECG, maximum heart rate, induced angina, depression induced by exercise w.r.t rest, slope of peak exercise, number of vessel, and thal. All these are used to classify if the patient is suffering from a heart problem or not [24].

Table 2 Description of medical datasets used in the study

Medical datasets	No of samples	No of input attributes	Remark
PIMA Indian diabetes	768	8	Classifying if the patient has diabetes
Wisconsin breast cancer	699	30	Classifying if cancer is malign
Cleveland heart disease	303	13	Classifying if the patient has heart disease
Indian liver patient dataset (ILPD)	583	10	Classifying if the person is a liver patient
Cesarean section classification	80	5	Classifying if the patient needs c-section

PIMA Indian Diabetes Dataset This dataset is used to predict if a person with certain diagnostic measurements is at risk to develop diabetes in near future or not. The data were originally given by the National Institute of Diabetes and Digestive and Kidney Diseases [25]. It is a small dataset with 768 instances and 8 input variables including pregnancy, glucose level, blood pressure, skin thickness, insulin level, body mass index, diabetes pedigree, and age. One outcome variable is 0 if patient does not have diabetes and "1" if the patient has diabetes. All the instances in it are of females of PIMA Indian heritage with minimum age of 21 years [26].

Breast Cancer Wisconsin (Diagnostic) Dataset These data were originally generated by Dr. William H. Wolberg at the University of Wisconsin, USA and include 569 samples and 32 columns. Its input features were collected from a digital scanned image of a fine needle aspirate of a breast mass collected from patients. High-resolution graphical computer program called Xcyt was used to measure the features of a cell based on the digital scan performed. It uses the curve-fitting algorithm that computed different features for each cell in the sample including ID Number, diagnosis, cell radius, texture, cell perimeter, cell area, smoothness, compactness, concavity, concave point, symmetry, fractal dimension. It then calculates three different values for each feature of a cell image namely mean value, extreme value, and standard error, resulting in a 30 real value attribute [27]. All this information is used to check if cancer is malign or not.

Indian Liver Patient Dataset The liver is one of the primary internal organs that takes control of various critical functions happening inside the human body, such as protein production, detoxifying chemicals, filtering of blood coming from digestive tracks, and many more. Liver disease is a very broad category that includes any disturbance in the functioning of the liver causing illness. The data taken up from the UCI machine learning repository is originally collected from north eastern part of Andhra Pradesh, India. It consists of 583 instances of sample data, each having 11 columns, of which 10 input features including Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Sgpt Alamine Aminotransferase, Sgot Aspartate

Aminotransferase, Total Protiens, ALB Albumin and Albumin and Globulin Ratio and one selector field to split the data into two sets [28].

Cesarean Section Classification Dataset A cesarean section, also stated as c-section, is a surgical operation performed to deliver a baby through an incision made in the abdomen and the uterus of the mother [29]. The c-section is recommended only when the vaginal birth is too risky to perform. This is usually the case when baby is in a breech position inside the mother's womb or when the mother develops high blood pressure during her pregnancy (preeclampsia). The c-section dataset, also taken up from the UCI repository, consists of 80 observations of patient reports and 5 input attributes: age, delivery number, delivery time, blood pressure (BP), and heart problem, based on which it is predicted if the pregnant woman needs a cesarean section to give birth or not.

4 Classification Metrics

To measure the effectiveness of a classification model, we need some metrics that explain the performance of a classifier. The performance analysis of all the classifiers is done on the above-mentioned datasets based on the evaluation metrics that are derived from a confusion matrix [30]. A confusion matrix is shown in Table 3.

where true negative (TN) is represented as the amount of cases when machine classified the output class as 0 and actual output class is also 0. True positive (TP) is represented as the amount of cases when machine classified the output class as 1 and actual output class is also 1. False positive (FP) is represented as the amount of cases when machine classified the output class as 1, but actual output class was 0. False negative (FN) is represented as the amount of cases when machine classified the output class as 0, but actual output class was 1.

Accuracy It is the most common metric to measure the effectiveness of an algorithm. It is the ratio of the number of correct predictions made by the classifier model to the total number of predictions made by the model. Accuracy can also be calculated from the confusion matrix as:

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN}$$

Table 3 Confusion matrix for binary classification

Predicted True value	**Negative – 0**	**Positive – 1**
Negative – 0	True negative (TN)	False positive (FP)
Positive – 1	False negative (FN)	True positive (TP)

However, accuracy can sometimes be misleading while evaluating the model particularly for an imbalanced dataset. If we have 100 samples with 95 samples labeled as positive and 5 as negative, then a classifier which predicts the value for the most frequent class for all predictions will have an accuracy of 95%. This is called accuracy paradox.

Precision Precision is defined as the fraction of the number of true positive results to the total positive results predicted by the classifier. The significance of precision is that it measures the quality of the classifier's prediction on account of what the classifier claims to be positive. It can be calculated using the confusion matrix as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall It is defined as the fraction of the number of true positive results to the sum of true positive and false negative results predicted by the classifier. In simple words, it is the number of correct positive predictions made by the classifier divided by all the samples that should be positive. It represents the percentage of total relevant results correctly predicted by the classifier. It is calculated as shown below:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score F1 score is defined as the harmonic mean between the precision and recall. There is a trade-off that occurs between the recall and precision and F1 score represents the balance between them. It tells about the robustness of a classifier and how precise it is. The value of F1 score always lies between 0 and 1 and a higher value of F1 score represents a better classification model. It is calculated as follows:

$$f1 = 2 \times \frac{(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}$$

5 Methodology

The datasets mentioned in Sect 3 are analyzed using various classification algorithms mentioned in Sect. 2, where common sequential approach is followed as depicted in Fig. 1. Initially the dataset is collected and preprocessed so that all the noise in the data is removed. Data preprocessing is necessary to improve the quality of data that directly affects the performance of classifiers. It refers to the *trans* transformation of raw data into a format that is suitable for a supervised machine learning algorithm to perform its functions. Irrelevant information from the dataset is also removed as part of preprocessing. The preprocessing is different for all datasets, which varies from eliminating redundant features, categorical encoding, and standardization. As an example, the cesarean section and the PIMA Indian

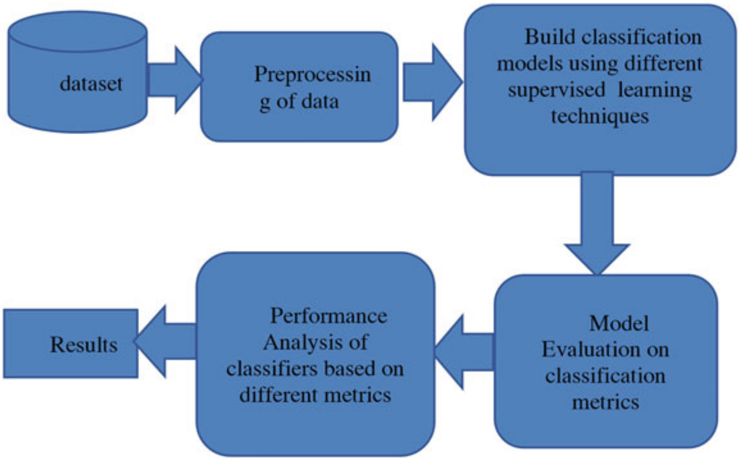

Fig. 1 Sequence of steps followed in classification task

diabetes datasets consisted of few observation samples with missing values for a primary attribute that were filled with mean imputation technique [31] to reduce the complexities in data.

Since each dataset consisted of attributes having numerical values in different range, feature scaling was also performed on all the datasets mentioned in Table 2 using standardization method [32]. Aside from PIMA Indian Diabetes dataset, each dataset consisted of one or more categorical attribute that were converted into a numerical value using label encoder and one hot encoder technique. The Wisconsin breast cancer dataset consisted of a couple of irrelevant input attributes that were removed before building the classifier models for prediction.

After the preprocessing step, the classifier models are built on each dataset using k-fold cross-validation technique. In this technique, the data sample is well shuffled and split into k number of groups. The classifier model is built/train on k-1 folds of data samples that accounts for training set and is evaluated on the *k*th fold of data sample representing a test set. The evaluation score is recorded and the process is repeated until all the k folds have been represented as test set. The mean of all the evaluation scores represents the overall performance of the classifier model [33]. After the evaluation, the parameters for the classifier model are tuned up to be fitted again on the data samples. When the performance of the classifier does not improve significantly, parameter tuning is stopped and the final performance metric values are calculated for the comparative analysis of the algorithms. Various models are compared based on their metrics values and a classifier with better performance is selected.

6 Results and Analysis

The performance analysis of all the eight classification algorithms (Table 1) is done on five different medical datasets collected from the UCI machine learning repository (Sect. 2.2). All distinguished classifier models are trained on the data samples using cross-validation technique to deal with the imbalance datasets. After the training, the models are evaluated on various performance metrics that are noted and tabulated. A perfect classifier is chosen as the one with best values for all the performance metrics.

The classification results of all the proposed supervised learning algorithms for the Cleveland heart disease dataset are reported in Table 4. Most of the models classified the data with an average accuracy of 80%. From the table it is observed that the artificial neural network (ANN) with an accuracy of 82.6% and support vector classifier with an accuracy of about 83% performed better compared to all other classification algorithms. The highest classification accuracy of 83% and F1 score of 0.856 is obtained by SVM classifier. Thus, SVM is chosen as the best algorithm to classify the Cleveland heart disease samples.

Results of PIMA Indian diabetes dataset are presented in Table 5. The diabetes dataset consisted of some missing values which were dealt with during the preprocessing stage. From the readings in Table 5, it is observed that most of the classifiers obtained an average accuracy of 75%, exception of SVM, KNN, and

Table 4 Performance metric results of algorithms for Cleveland heart disease dataset

Algorithms	Accuracy	Precision	Recall	F1 score
SVM	83.03	0.805	0.916	0.856
K nearest neighbors	81.8	0.802	0.893	0.843
Naive Bayes	81.38	0.783	0.924	0.845
Logistic regression	81.37	0.807	0.871	0.837
Random forest	80.57	0.806	0.856	0.827
AdaBoost	80.15	0.78	0.886	0.829
Decision tree	74.79	0.764	0.78	0.77
ANN	82.62	9.796	0.92	0.854

Table 5 Performance metric results of algorithms for PIMA Indian diabetes dataset

Algorithms	Accuracy	Precision	Recall	F1 score
SVM	77.45	0.758	0.916	0.615
K nearest neighbors	76.03	0.665	0.893	0.647
Naive Bayes	75.38	0.731	0.458	0.558
Logistic regression	76.24	0.70	0.562	0.618
Random forest	75.72	0.672	0.613	0.635
AdaBoost	75.7	0.712	0.527	0.599
Decision tree	68.74	0.554	0.582	0.565
ANN	73.62	0.772	0.412	0.505

Table 6 Performance metric results of algorithms for Wisconsin breast cancer dataset

Algorithms	Accuracy	Precision	Recall	F1 score
SVM	98.04	0.983	0.986	0.984
K nearest neighbors	96.94	0.966	0.986	0.975
Naive Bayes	93.42	0.947	0.95	0.947
Logistic regression	97.59	0.979	0.982	0.98
Random forest	96.92	0.969	0.982	0.975
AdaBoost	97.14	0.975	0.978	0.977
Decision tree	94.08	0.958	0.947	0.952
ANN	96.49	0.969	0.975	0.972

Table 7 Performance metric results of algorithms for Indian liver patient dataset

Algorithms	Accuracy	Precision	Recall	F1 score
SVM	71.36	0.71.35	1	0.832
K nearest neighbors	64.84	0.755	0.752	0.752
Naive Bayes	93.42	0.947	0.95	0.947
Logistic regression	71.71	0.740	0.930	0.824
Random forest	70.83	0.764	0.855	0.804
AdaBoost	71.18	0.716	0.985	0.829
Decision tree	64.29	0.749	0.749	0.745
ANN	71.35	0.713	1	0.831

logistic regression that classified the data with a little higher accuracy. Here again, it is the SVM classifier that obtained the highest accuracy of 77.45% with an F1 score of 6.1. However, it is the K nearest neighbors (KNN) classifier with accuracy of 76% that obtained the best F1 score of 0.64.

Table 6 depicts the results of breast cancer dataset. It can be observed that all algorithms classified the data with high accuracy. The best values for all the classification metrics (accuracy, precision, recall, and F1 score) were obtained by the SVM classifier that classified the data with an impressive accuracy of 98% and having F1 score of 0.98. Aside from the SVM classifier, the logistic regression and AdaBoost algorithms classified the data with an accuracy of 97.59% and 97.14%, respectively.

Performance metrics results of all the classification algorithms experimented on liver patient dataset are shown in Table 7. From the table, it is observed that five algorithms (namely logistic regression, SVM, AdaBoost, random forest, and ANN) obtained a classification accuracy between 70% and 72%. The logistic regression classifier classified the data with highest accuracy of 71.71% having F1 score of 8.2. However, the best value of F1 score (8.3) were obtained by two other classifiers, namely, SVM and ANN, having model accuracy of 71.36% and 71.35%, respectively.

Table 8 Performance metric results of algorithms for cesarean section dataset

Algorithms	Accuracy	Precision	Recall	F1 score
SVM	70.97	0.78	0.669	0.714
K nearest neighbors	63.64	0.708	0.629	0.665
Naive Bayes	73.63	0.781	0.758	0.767
Logistic regression	72.53	0.767	0.761	0.762
Random forest	57.23	0.636	0.606	0.621
AdaBoost	63.74	0.653	0.783	0.712
Decision tree	54.85	0.608	0.607	0.607
ANN	57.6	0.580	0.980	0.724

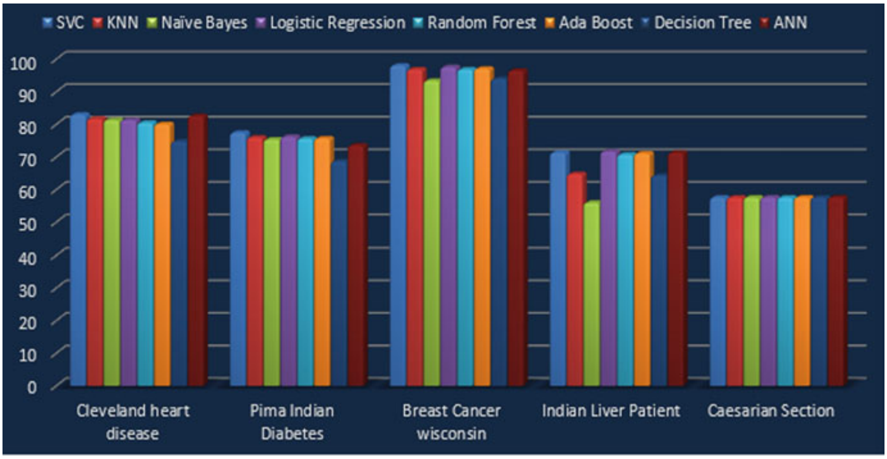

Fig. 2 Performance report of all classification algorithms in terms of F1 score metric

Performance metric results of proposed supervised learning algorithms for cesarean section dataset can be seen in Table 8. It is observed that only Gaussian Naive Bayes, logistic regression, and SVM classifiers classified the data with an accuracy of 70% or more.

The best values of classification accuracy (73.63%), precision (78.12), and F1 score (7.67) were obtained by the Gaussian Naive Bayes classifier. However, the rest of the classification algorithms failed to classify the cesarean section dataset. This failure is mainly due to the low volume of the data as it consists of only 80 observation samples. Thus, Gaussian Naive Bayes algorithm is chosen as the best algorithm to classify the cesarean section dataset.

The complete performance report of the algorithms in terms of accuracy and F1 score is much easily analyzed graphically as shown in Figs. 2 and 3, respectively.

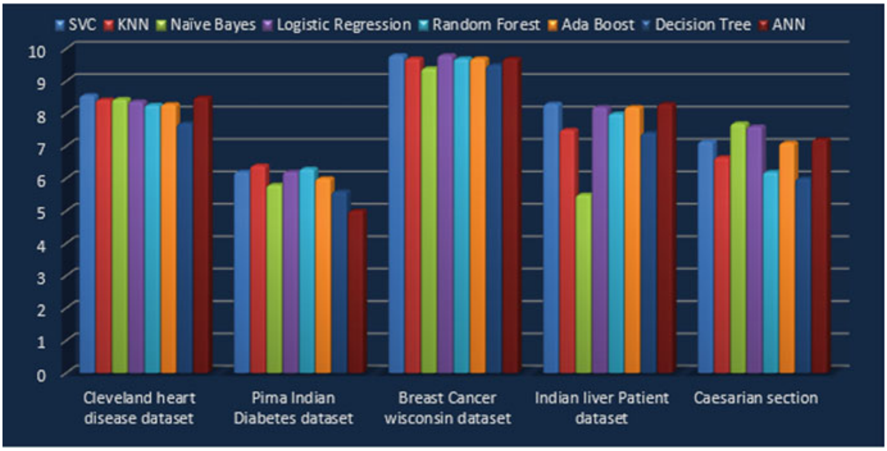

Fig. 3 Accuracy analysis of all the classifiers on five healthcare datasets

7 Conclusion and Future Work

Classifiers play a critical role in giving new insights into healthcare field from predicting prognosis, deciding treatment plan, or may be just for research purposes so that more precise and fruitful studies can be carried out.

In our study, we analyzed the performances of eight different supervised machine learning classification algorithms on five different healthcare datasets. Based on examining the classification metric results for all the algorithms on each dataset, it is observed that each classification algorithm performs differently on different kind of data, however, the SVM and logistic regression algorithm gave the best and the most consistent results for all the concerned datasets. For imbalance dataset like diabetes, F1 score is an important metric to consider. For small dataset, the Naive Bayes algorithm performs better than all other classifiers and can be preferred over SVM and logistic regression algorithms for all such cases.

The scope of supervised techniques is not limited to any extent and applying it on data can lead us to solutions to most of the problems we face today. The medical science itself is a very huge field which generates a lot of data. In the current study, we discussed techniques to limited datasets which can be extended to other different medical diagnosis. Also, with the introduction of deep learning architectures in supervised learning it can give us opportunity to explore improvement in existing results.

References

1. R. Deo, Machine learning in medicine. Am. Heart Assoc. J. **132**(20), 1920–1930 (2015)
2. D. Bhavani, A. Vasasvi, P. Keshava, Machine learning: a critical review of classification techniques. Intl. J. Adv. Res. Comp. Commun. Eng. ISSN: 2319-5940
3. O. Obaid, M. Mohammed, A. Ghani, S. Mostafa, F. Taha AL-Dhief, Evaluating the performance of machine learning techniques in the classification of Wisconsin breast cancer. Intl. J. Eng. Technol. (IJET) **7**, 160–166 (2018)
4. C. Latha, S. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlocked **16**, 100203 (2019)
5. M. Saritas, A. Yasar, Performance analysis of ANN and Naive Bayes classification algorithm for data classification. Intl. J. Intelligent Systems Appl. Eng **7**, 88–91 (2019)
6. G. Rumbe, H. Youth, Comparative study of classification techniques on breast cancer FNA biopsy data. Intl. J. Artif. Intellig. Interact. Multim. **1**, 5–12 (2010). https://doi.org/10.9781/ijimai.2010.131
7. A. Christobel, P. Sivaprakasam, Improving the performance of k-nearest neighbor algorithm for the classification of diabetes dataset with missing values. Intl. J. Comp. Eng. Technol. (IJCET) **7**(3), 155–167 (2012)
8. I. Rish, An empirical study of the Naïve Bayes classifier. IJCAI 2001 Work Empir. Methods Artif. Intell. **3**, 41–46 (2001)
9. R. Wang, AdaBoost for feature selection, classification and its relation with SVM. Intl. Conf. Solid State Devices Mat. Sci. **25**, 800–807 (2012)
10. A. Wyner, M. Maolson, J. Bleich, Explaining the success of AdaBoost and random forests as interpolating classifiers. J. Mach. Learn. Res. **18** (2017)
11. J. Peng, C.-Y.J. Peng, K.L. Lee, G.M. Ingersoll, An introduction to logistic regression analysis and reporting. J. Educ. Res. **96**, 3–14 (2002)
12. M. Gladence, M. Karthi, A. Maria, A statistical comparison of logistic regression and different Bayes classification methods for machine learning. ARPN J. Eng. Appl. Sci. **10**, 5947–5953 (2015)
13. M. Amin, A. Ali, *Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions* (Department of Computer Science & Engineering University of Engineering & Technology, Lahore, 2019)
14. E. Zriqat, A. Altamimi, M. Azzeh, A comparative study for predicting heart diseases using data mining classification methods. Intl. J. Comp. Sci. Inf. Security (IJCSIS) **14**(12), 869–879 (2017)
15. W. Shijun, R. Summers, Machine learning and radiology. Med. Image Anal. **16**, 933–951 (2012). https://doi.org/10.1016/j.media.2012.02.005
16. D. Bansal, R. Chhikara, K. Khanna, P. Gupta, Comparative analysis of various machine learning algorithms for detecting dementia. Procedia Comput. Sci. **132**, 1497–1502 (2018). https://doi.org/10.1016/j.procs.2018.05.102
17. G. Munjal, M. Hanmandlu, S. Srivastava, Novel gene selection method for breast cancer classification. J. Biochem. Technol. **8**(4), 1116–1120
18. M. Lamba, G. Munjal, Y. Gigras, Feature selection of micro-array expression data (FSM) – a review. Proc. Comput. Sci. **132**, 1619–1625 (2018)
19. C. Sidey-Gibbons, J. Sidey-Gibbons, Machine learning in medicine: a practical introduction. BMC Med. Res. Methodol. **19** (2019). https://doi.org/10.1186/s12874-019-0681-4
20. J. Anderson, J. Parikh, D. Shenfeld, Reverse engineering and evaluation of prediction models for progression to Type 2 diabetes: application of machine learning using electronic health records. J. Diabetes Sci. Technol. **10** (2016). https://doi.org/10.1177/1932296815620200
21. G. Bedi, F. Carrillo, G.A. Cecchi, D.F. Slezak, M. Sigman, N.B. Mota, S. Ribeiro, D.C. Javitt, M. Copelli, C.M. Corcoran, Automated analysis of free speech predicts psychosis onset in high-risk youths. Nat. Partner J. (NPJ) Schizophrenia **1**(1), 15030 (2015)

22. Y. Yamada, M. Kobayashi, Detecting mental fatigue from eye-tracking data gathered while watching video: evaluation in younger and older adults. Artif. Intell. Med. **91** (2018). https://doi.org/10.1016/j.artmed.2018.06.005
23. Y. Liu, K. Gadepalli, M. Norouzi, G. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Nelson, G. Corrado, J. Hipp, L. Peng, M. Stumpe. *Detecting Cancer Metastases on Gigapixel Pathology Images* (2017). arxiv:1703.02442.
24. N. Lutimath, C. Chethan, B.S. Pol, Prediction of heart disease using machine learning. Intl. J. Recent Technol. Eng. (IJRTE). ISSN: 2277-3878 **8**(2S10) (2019)
25. S. Vanaja, K. Rameshkumar, Performance analysis of classification algorithms on medical diagnoses-a survey. J. Comput. Sci. **11**, 30–52 (2015)
26. G. Munjal, S. Kaur. Comparative study of ANN for pattern classification WSEAS. *International Conference on Mathematical Methods and Computational Techniques in Electrical Engineering, Bucharest, Romania* (2006)
27. R.L. Borges, Analysis of the Wisconsin Breast cancer dataset and machine learning for breast cancer detection. *Proceedings of XI Workshop de Visao Computacional* (October 5–7th, 2015)
28. M. Singaravelu, S. Rajapraksh, S. Krishnan, K. Karthik, Classification of liver patient dataset using machine learning algorithms. Intl. J. Eng. Technol. **7**, 323–326 (2018). https://doi.org/10.14419/ijet.v7i3.34.19217
29. S. Abbas, R. Riaz, S. Kazmi, S. Rizvi, S.J. Kwon, Cause analysis of caesarian sections and application of machine learning methods for classification of birth data. IEEE **6**, 67555–67561 (2018)
30. F. Sardouk, A. Duru, O. Bayat, Classification of breast cancer using data mining. Am. Sci. Res. J. Eng. Technol. Sci. (ASRJETS) **51**(1), 38–46 (2019)
31. M. Rahman, N. Davis Darryl, Machine learning based missing value imputation method for clinical datasets. IAENG Trans. Eng. Technol. **229** (2012). https://doi.org/10.1007/978-94-007-6190-2_19
32. S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Data preprocessing for supervised learning. Int. J. Comput. Sci. **1**, 111–117 (2006)
33. F. Ahmed, Y. Ali, S. Shamsuddin, Using K-fold cross validation proposed models for Spikeprop learning enhancements. Intl. J. Eng. Technol. (UAE) **7**, 145–151 (2018). https://doi.org/10.14419/ijet.v7i4.11.20790