VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Machine Learning: Shaping the Future of Healthcare

FACULTY:
Dr. Ayesha Shaik
SCOPE

GROUP MEMBER'S:

Dogiparthi Aasrith-21BAI1702

Yarraguntla Kalyan Chakravarthy- 21BAI1759

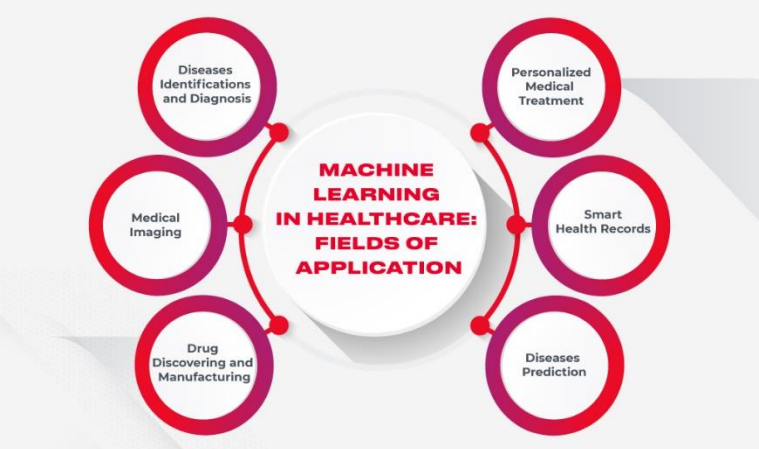
Slot:B2+TB2

Abstract

Machine learning is the automatic discovery of patterns and reasoning about data. Machine learning can enable personalized care called precision medicine. These techniques have been successful in healthcare. This article discuss about uses of machine learning in healthcare. Machine learning and AI will revolutionize healthcare in future.

Machine learning (ML) is a growing research area with many opportunities to explore. “Alth ough it has a negative impact on health, it’s the defining technology of this decade”, says James Collin of MIT. Many young start-up’s in the machine learning industry are trying to succeed in focus on treatment. Google has developed a machine learning algorithms to help identify cancerous tumors in mammograms. Stanford University uses deep techniques to identify skin cancers. Different researchers have suggested different behaviours, clinical researchers have proposed different risk factors to define chronic diseases. More information means more machine learning, but this large number of features requires more models to improve accuracy. Therefore, it will be better if the machine can eliminate the high medical risk. Classification of different types will enable the evaluation of different diseases in patients. In healthcare, the percentage of correct diagnosis (sensitivity) of patients is more correct diagnosis (sensitivity) of patients is more important than the percentage of diagnosis of healthy individuals. Patients always need care and relationships with caregivers. Machine learning will not eliminate, but it will be a tool clinicians will use to improve ongoing care.

Machine learning (ML) and its applications in healthcare have been very popular. While impr oving computing power with big data, there is an opportunity to use machine learning algori thms to improve health. Supervised learning is a type of machine learning that predicts reco rded data based on methods such as linear or logistic regression, support vector machines, d ecision trees, LASSO regression, K neighbors, and Naive Bayes classifiers. Unsupervised mac hine learning models can identify patterns in data in a database where there is no random d ata. Such models can be used for fraud or false positive detection.Examples of clinical applic ations of machine learning include the development of multiple clinical decision models. An important public health aspect of machine learning is identifying and predicting populations at high risk for certain adverse health conditions and targeting these populations with public health implications. More ideas about machine learning should be incorporated into the me dical literature so that doctors can guide and interpret research in this area.



Introduction

Health is one of the most important assets in society. However, due to rapid development, people's desire for medical treatment has surpassed low-cost equipment and healthcare services. As the need for medical treatment grows, providing people with adequate treatment is the most important thing to manage in the medical field. The shape of the clean area changes according to the population, cultural development, natural resources, political and economic conditions of the country. The emphasis on health care and quality of life has led to conflicts among health groups and has been an important factor for progress in the world.

Health problems affect people's lives. During the evaluation process, healthcare organizations ensure the accuracy of all relevant and relevant public information in the decision to manage the patient. Therefore, knowledge plays an important role in solving health problems and further analysis is essential to improve the care of affected people. One of the most important factors affecting treatment is undoubtedly age. Although the needs of people and human health are increasing rapidly, today's progress will be one of the most important health services in society in meeting these needs.

Fortunately, the complexity of our medical systems today helps us make informed decisions. This generation's ability in medical facilities is used to gather information on all relevant symptoms, detect certain diseases before effects occur, and prevent any disease through preventive measures. Thanks to this device, many patients have been saved from many deadly diseases. In fact, machine learning is making progress in many areas, including laptop vision, natural language processing, and automatic speech space to enable powerful models (e.g. driverless cars, non-civilian assistant speakers, automatic translation).

Consider silent people to identify the cause, risk factor, drug potency, and disease subtypes from the long-standing epidemic.

Disease prevention strategies, such as data management and regular control procedures, are the basis of evidence supporting drug therapy. However, these strategies are difficult, expensive, and not independent of the emotions they have to contend with, and their benefits may not be significant for sighted patients [1]. As systems and organizations improve their implementation strategies, the use of electronic health records (EHRs) is expanding around the world, influencing EHRs to answer epidemiologists' questions [2] and are now pervasive for setting the truth in the delivery of human services [3].

Data analysis techniques fall into the following categories: presentation, exploration, deduction, estimation, and reasoning [4]. Descriptive data analysis explains data that is not understood, while exploratory data analysis distinguishes between items in the measured data.

Deductive reasoning measures in which observed relationships in the population are outside the evidence base, while forward-thinking reporting attempts to measure the probability of a particular outcome. Finally, a causal analysis determines how a change in one variable affects another variable. It is important to identify the type of query that occurs in the analysis in order to determine the type of information search appropriate for the query. Predictive analytics is used to predict people's outcomes by constructing a model based on what people are looking at [4] and using that model to form expectations based on people's interest values. Predictive visualization is an algorithmic representation of how data can be visualized.

This evaluation method evaluates success through criteria such as accuracy, analysis, and revision, which evaluate multiple repeated ideas.

Artificial Intelligence is a method of using data analysis to derive models that are accurate enough to predict outcomes or classify predictions in future data. Specifically, management uses assessment insights to build AI processes, models in categories, or predictive models of already known (best level) collaboration outcomes. The next sample (usually a repetition of some samples with penalty) is often used for new samples to measure or predict results that were not clearly observed before, and the quality requirements are evaluated by comparing the quality of the various assessments. representation. In this way, AI "live" in the realm of algorithm representation and should be evaluated in that capacity.

Repetitive processes created using artificial intelligence cannot and should not be analyzed using methods in the data output field. Doing so may lead to mis-evaluation of the representation of the job for which the model is set, which may lead the customer to misunderstand the model's output.

EHRs have access to a wide range of features that allow state-of-the-art grouping and forecasting, while artificial intelligence provides techniques for processing the large volumes of data found in medical facilities. Then, the application of artificial intelligence in searching electronic medical records is at the forefront of current medical records [5], demonstrating its therapeutic and enforcement power. We explain the work and complex process of using these skills in practice and research.

Finally, our perspective on AI opens doors in healthcare, and applications have the most significant potential to impact health and safety

This chapter presents specific issues that need to be considered for medical companies, particularly within the AI framework to plan for success. framework and human expertise is narrow [6]. Failure to address these issues could undermine the legitimacy and effectiveness of AI in human services. We offer a chain of command for healthcare opportunities categorized into: computer engineering, medical assistance, and advanced healthcare. Finally, we explain the door to artificial intelligence research that matters in medicine: many documents and exciting tools, it's good to accept that the principle is interpretable and decide which one is different.

Data preparation

Data preparation is done to clean and process data. According to a Forbes study, data scientists spend 80% of their time preparing data. Data may have characteristics of different scales; therefore, parametric models of neighbor kNN (kNN), support vector machine (SVM) and neural network (NN) types need to be normalized [7]. Data preparation in machine learning (ML) pipelines takes a lot of time, so Pandas analytics in Python can be used to understand data. Almost all applications that are part of data preparation are discussed below.

Feature Cleanup

This step is important to identify the important and bad parts of the code. Histograms and bar graphs can be drawn to understand the same thing.

Missing values

Treat missing values by ignoring them or treating them as values of missing data, variables due to missing values are surprise and lack of significance in explanatory or predictive variables. The difference between the lack of importance should be different. The most obvious is to get real results by repeating the data collection process; but it is impossible.

Missing data can be divided into: completely random missing (MCAR), randomly missing (MAR), and randomly missing (MNAR) [9]. Let's look at strategies for resolving missing results:

I. Elimination: The best way to resolve missing results is elimination, which may not be effective if values are missing randomly.

a. Listwise: Delete rows with missing values.

b . Pairwise: Assumes missing data to be missing completely at random.

II. Plugging

The fancy impute package in Python provides several powerful machine learning models for loading missing values.

The Simple Imputer and Iterative Imputer classes in the Python sklearn.preprocessing library, commonly known as MICE (Multiple Imputation in Chains), can be used for assignment. Impute is also an incomplete library of evaluation algorithms. Here are a few missing item upload (MVI) methods:

A)**Popular Averaging Techniques:** Most Popular Most Popular Most Popular Report is suitable for small data that is widely distributed and not skewed; For numeric variables, median evaluation can cause bias, and mode assignment (commonly used) is better for categorical data.

Types can also be used with different numbers. Thus, we can quickly fill in missing values, but at the expense of changes in the data. Although this is easy and fast to use, it can lead to poor performance.

B)**Forecasting Techniques:** Estimation techniques are used to estimate MCAR-type data by selecting variables that have some relationship to missing observations.

Assignment; The data set is divided into a non-significant missing features set called the training set and another non-significant feature set called the test set where the features with missing values are considered as target variables. There are many statistical and/or machine learning methods for loading missing values.

C)**KNN (non-parametric) :** Evaluation is made using the most frequent value of the variable's neighbors and the mean/mod value of the constant variable. Calculate the similarity between the two samples using the distance.

If the features are similar, you should choose Euclidean distance like width and height, otherwise you can choose Manhattan distance like age, height and other things for different features.

The Hamming distance is the number of categorical features with different values. This method of assignment works for both qualitative and quantitative properties, although many of them have no value. It works with MCAR, MAR, MNAR loss data with numeric (continuous, discrete), patterns and categorical features. The KNN algorithm takes a lot of time when analyzing large datasets and is user-friendly as it searches all the data to find similar events. Also, the chosen value of k comes with a tradeoff.

Large values of k can have negative features, while low values of k indicate loss of important features and the effect of noise. Therefore, the results are not generalizable. In binary classification problems, single values of k can avoid correlation.

Genetic Algorithms (GA), Expectation Maximization (EM), and kMeans [11] can also be used for evaluation. EM works iteratively to predict a value called expectation, using other properties and verifying that that value is the most likely.

to. is called maximization. If it's not the best, it will re-estimate the value further until it reaches it. The following procedure stores the relationship between the predicted features for other features. The kMeans algorithm uses a distance measure to measure missing values.

The Missingpy library in Python has a missForest class that reuses Random Forest (RF) to identify missing values. It starts loading the column with the fewest missing values. Missing values in initially unmatched columns were filled in numerical ways and continuous type. Imputer fits the RF model with estimated missing values. The interpolator selects the next row with the second lowest zero value in the first round.

This process is repeated for each column with missing values, multiple iterations in each row, until the stop operation is complete.

Outliers in the data

"The difference between one observation and the other is so great that it suggests the hypothesis that it was produced by a different mechanism" - Hawkins (1980). Outliers are data points that are three or more standard deviations (MAD) from the mean.

We often ignore outliers when building models; this often corrupts data and reduces accuracy. A value of <-1 or >1 indicates high curvature. The boxplot() function [11] in the Seaborn library uses the median, lower and upper quartiles (defined as the 25th and 75th percentiles) to detect outliers. Interquartile range or median / IQR = Q3 − Q1, where Q1 is the lower quartile and Q3 is the upper quartile. [Q1 - (1.5 × IQR), Q3 + (1.5 × IQR)].

Kurtosis is a measure of outliers (values range from -2 to 2). Visualization is used to identify abnormalities. Methods include box plots, histograms, and scatterplots using plotly, bokeh, matplotlib, seaborn, and pandas plotting the Python libraries.

The following procedures can be used to resolve outliers:

A)**Remove observations:** If the procedure was specified by data entry or when processing data, or if the number of relevant observations is too few, observations can be deleted to remove outliers .

B)**Converted and binned values:** If values are converted, outliers will be removed. One way to transform variables is to use the natural logarithm, which minimizes variation due to extreme values.

C)**Imputation:** Imputation for artificial objects using statistical data.

One of the most efficient search methods is the standard deviation (Z-score) [2] $z = (x-\mu)\ \sigma$.

Calculates the number of standard deviations of data points from the sample, assuming your data is drawn from a Gaussian distribution. Therefore, the Z-score is a parametric method. A transformation, such as scaling, can be applied to make the data fit a Gaussian distribution.

The scipy.stats module can be used to calculate the same results.

Insufficient Data [3]

Data used in clinical practice generally contain fewer samples from one class than the other (<5%).

For this reason, in most machine learning models decision trees, logistic regression may be more accurate in most classes and thus give inaccurate performance measures. The accuracy of the model is very high in healthy patients and very low in sick people. The overall accuracy will be high not because the model is good but because the attributes of several classes are often broadcast as noise and the data is unequal.

Therefore, the following procedure can help a class instructor find the least number of classes: Likewise, using the wrong gauge can be dangerous. If accuracy is a measure of how good a model is, a model that distributes all health metrics will be accurate, but it is clear that this model will not provide the same information. In such cases, other evaluation criteria can be used:

a. Precision/Specificity is the number of events selected.

B. Recall/Sensitivity how many times is selected.

c. TheF1 score is a compromise between precision and recall. It gives equal points.

d. MCC is the Mathew correlation coefficient between the observation and prediction binary classifications.

e. AUC_ROC (Area Under Curve_Receiver Operating Characteristics) is the relationship between a positive value and a false positive value. Like precision and recall, precision is divided into sensitivity and specificity, and models can be selected based on the balance of these values. For a 100% correct guess, the AUC should be 1

For computation performance, fusion_matrix and Precision_score can be used by the sklearn.metrics module.

II. Resample the training set: Other than using different evaluation criteria to deal with imbalanced data, one can also work on getting different dataset. Two approaches to make a balanced dataset out of an imbalanced one are under-sampling and over-sampling.

a. Under-sampling balances the dataset by reducing the size of the majority class by randomly selecting equal number of samples in the majority class as in the minority class. This method is used when no. of observations is large but it tends to discard potentially useful information which could be important for building rule classifiers. The sample chosen by random under sampling may be a biased sample not representative of the entire population. Thereby, resulting in inaccurate results with the actual test data set.

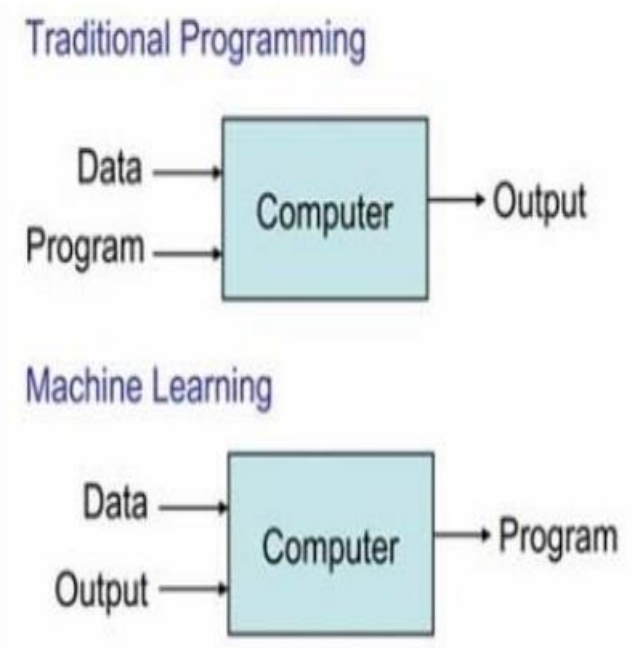
b. Over-sampling [4] is used when the quantity of data is insufficient. It tries to balance dataset by increasing the size of minority class samples by using e.g. repetition, bootstrapping or Synthetic Minority Over-Sampling Technique (SMOTE) [5] and can be accomplished utilising imblearn class in Python. Over-sampling increases the likelihood of overfitting since it replicates the minority class instances. Moreover, SMOTE is not effective for high dimensional data since it increases Recall at cost of Precision. ADaptive SYNthetic Sampling (ADSYN) is an improvement over SMOTE since it adds randomness. Both generate new samples by interpolation.

A combination of over- and under-sampling is often successful as well.

III. Use K-fold Cross-Validation (CV) appropriately: Over-sampling takes observed samples of minority class and applies Bootstrapping to generate new random data based on a distribution function. If CV is applied after over-sampling, the model would overfit to the artificially bootstrapped data. Hence, CV should always be done before over-sampling the data, just as how feature selection should be implemented. Only by resampling the data repeatedly, randomness can be introduced into the dataset to prevent overfitting.

IV. Ensemble different resampled datasets: To generate more data, so that the model generalizes well can be done by building ‘n’ models with all samples of the minority class and n-random samples of the majority class. Thus, Ensemble models tend to generalize better [3].

V. Design your own models: There is no need for resampling, if the model is suited for imbalanced data. e.g. XGBoost internally takes care that the bags it trains on are not imbalanced. The resampling happens secretly. The cost function used penalizes misclassification of the minority class more than misclassification of majority class. Thus, it is possible to design models that generalize naturally in favor of the minority class.



Feature Engineering

Feature Engineering is extraction of features from data and transforming them into formats that are suitable for ML algorithms. “The goal is to turn data into information, and information into insight.”—Carly Fiorina. There are 3 types of data—Numerical (Discrete and Continuous), Categorical (qualitative), and Ordinal (numbers with mathematical meaning). The predictor variable in a classification task may be binary or multi [3]. To reduce the complexity due to increase in number of classes, multiclass classifier is simplified into a series of binary classification such as One-Against-One and One-Against-All [5].

Feature Transformation

I. **Normalizing:** To eliminate the effect of outlier that may negatively affect the accuracy of conclusion, different types of variables need to be brought in same order of magnitude. Feature Scaling benefits Gradient Descent with faster convergence. Distance calculation algorithms are greatly influenced by difference in scale among variables whereas Tree based algorithms are not affected by difference in magnitude. Normalization and Standardization can be used for scaling. Normalization can be achieved through Min-Max scaling using equation xnormalized=(x−xmin) /(xmax−xmin)bringing all numeric values in the range [0, 1] with zero mean and one standard deviation. The formula for standardization is z=(x−μ)σand is commonly known as z-score, where μ is mean and σ is standard deviation. Scaling can be accomplished through StandardScaler and MinScaler utility classes of sklearn.preprocessing package. It is advisable to remove extreme outliers before applying normalization else it

would skew the values in your data to a small interval. RobustScaler class of Python can be used for skewed data.

II.**Feature encoding**: Major ML libraries work well with numerical variables. Nominal values can be misinterpreted by the learning algorithm. Missing values should be filled before encoding categorical features. Encoding allows algorithms which expect continuous features to use categorical features through scikit-learn's Label Encoder function. OneHotEncoder function is required for multilabelled features. OneHotEncoder converts n levels into n-1 new variables and can lead to dummy variable trap or curse of dimensionality (i.e. number of instances need to grow exponentially with number of features). It is not recommended to use OneHotEncoder with Tree based algorithms. Python provides sklearn.preprocessing package for the same. get_dummies() function of pandas package is a straightforward and easier way for the same.

III.**Discretisation/Binning**: Transforming continuous variables into discrete variables brings into non-linearity and thus improves the fitting power of model, minimizing the impact of extreme values and preventing overfitting possible with numerical variables.

IV.**Skewed data**: It is necessary for the data distribution to be in range [−0.5, 0.5]. However, it is common for health care data to be distributed unsymmetrically with a long tail of high values. To handle right skewed data, a particular power of the data or log transform is used to bring it to near normal distribution.

Feature Extraction

When the data to be processed through an algorithm is too large, it's generally considered redundant. Analysis with large number of variables is computationally expensive, therefore we should reduce the dimensionality of these types of variables.

Feature Selection

In healthcare, accumulating data is a costly aspect [13]. Even, there is chance of data overfitting the model when number of observations is less and need for significant computation time when number of features is more. Hence, if machines could extract most informative features, the cost overhead on patients would reduce tremendously. Feature Selection is essential for simpler, faster, more reliable and robust ML models. Aim is to maintain accuracy and stability, improve runtime and avoid overfit. A feature selection technique benefits with redundant or irrelevant data which can be removed without much loss of information. Feature Selection algorithms are Filter based, Wrapper Based [2], Embedded [14] and Hybrid [14]. Python provides feature_selection module for feature selection.

I. Filter based methods are further categorized as Basic, Multivariate and Statistical. Filter methods rank features independent of the relationship among them. There are various measures in Filter based methods as Correlation based and Information Theory. The first step is to remove constant information which provides no / minimal information since it has same value for all instances of the feature. This can be checked by checking the variance of feature values, if the variance is 0 then the feature values are redundant. There is possibility of feature values being quasi-constant i.e. the variance < 0.01. Similarly, post one hot encoding of large datasets or dataset with lots of categorical values, there is chance of duplicate rows. Hence it is required to transpose the dataset and perform same operations as for constant and quasi constant columns.

a. **Correlation [6**]: Next step should be to identify correlated features since effective results are appreciated if features correlate highly to the target and are uncorrelated to each other. Pearson Correlation Coefficient (PCC) [−1, 1] and Mathews Correlation Coefficient (MCC) perform the task. Coefficients closer to −1 designate strong negative relationship while coefficients closer to 1 designate strong positive relationship. Correlation is demonstrated through Correlation matrix. Correlated features do not necessarily affect the model performance (trees, etc.), but high dimensionality does and too many features hurt model interpretability. So, it's always better to reduce correlated features. Pearson Correlation is sensitive only to linear relationship and can be viewed through Heat Map. For Pearson correlation, it is recommended to drop features with values close to 0.

b. Statistical methods are fast but do not capture redundancy among features. These methods assign a score to each feature and thus rank them for inclusion or exclusion. Fisher score [7] and Chi square test can be used to measure the dependence among features. These methods often consider the feature independently and hence are univariate. Similarly, ANalysis of VAriance (ANOVA) parametric test identifies dependence among continuous variables.

c. Information theory is used extensively as it can measure nonlinear relationship among features. Mutual Information (MI) through Information Gain is used a lot in literature to identify how much knowing one variable reduces the uncertainty of other. MI measures similarity among features but is inconvenient to compute for continuous variables.

II. Wrapper methods are greedy, computationally intensive and exhaustive. They detect interaction among features by looking for subsets using Step Forward [7] or Step Backward [7]. Wrapper methods result in best feature subset for that particular type of model. Being exhaustive search, it builds a model and evaluates the subset for optimality through the score. This process is repeated until the performance of the model starts decreasing or increasing or till pre-determined number of features are extracted. Wrapper methods consider different combinations of selected features, and compare these to other combinations. Different

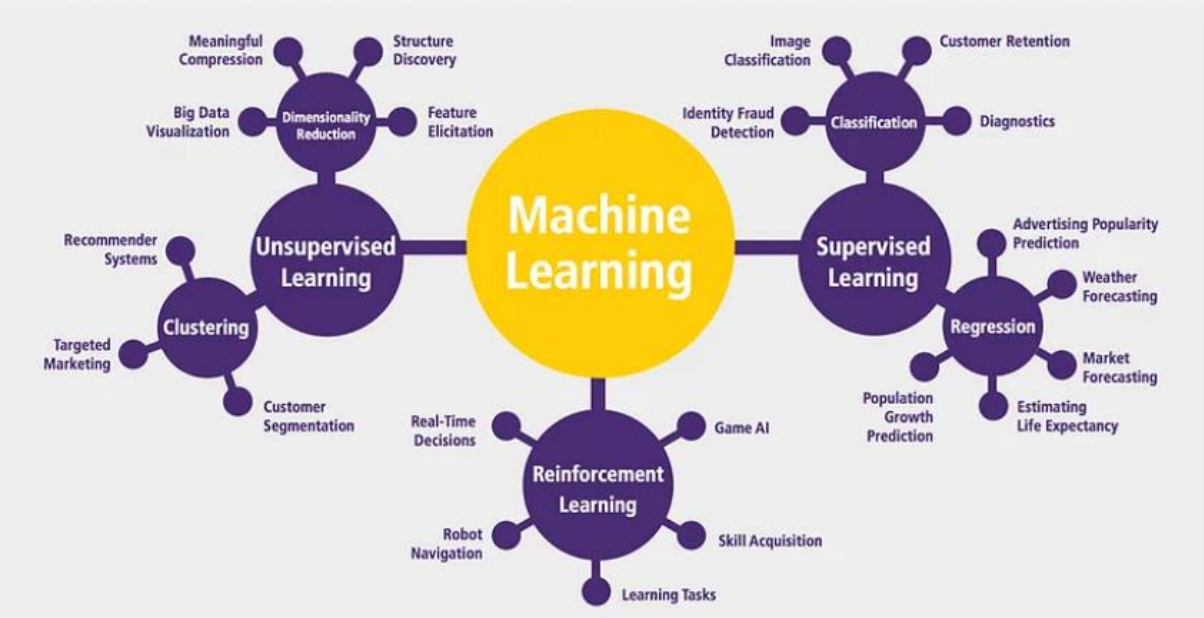
combinations can be tried using Relief, Gain Ratio and Entropy. They detect the possible interactions between variables. The search process may be methodical, stochastic or may use heuristic such as Best-First Search, Random Hill-Climbing algorithm, or Forward and Backward pass respectively to add and remove features. Boruta works as wrapper around RF by finding all informative features than finding a subset of features on which some classifier has a minimal error. GA selects optimal features to fit in the objective function for SVM classifier [2]. GA wrapper around NN has shown accuracy of 93.85% on Z-Alizadehsani dataset [8]. GA wrapper around Adaboost onto Parkinson disease dataset identified 7 important features and gave 98.28% accuracy whereas GA on Bagging returned 10 features with accuracy 96.55% [9]. kMeans and GA have also been used for dimensionality reduction [11].

III. Embedded methods use a wrapper [14] to consider interaction between feature and model but doesn't build a different model each time a different feature subset is picked. Embedded method is faster and cheaper than Wrapper and more accurate than Filter. It uses importance of features by identifying node impurity. The method is constrained to limitation of the associated algorithm. Types include Lasso Regularization, Decision Tree (DT), Random Forest and Gradient Boosted Trees derived importance. Least absolute shrinkage and selection operator (LASSO) regularization adds penalty on different parameters to reduce their freedom making the model fit noise in training data and thus generalize well on test data. We can ascertain that if penalty is too high, important features are dropped and the performance of model drops. For regression, the coefficients of predictors are proportional to how much it contributes to the target variable. These methods work under the assumption that there is a linear relationship between explanatory and predictor variable, and explanatory variables are independent, normally distributed and scaled. Thus, the features whose coefficients are more than mean of all coefficients are selected. The variants include L1 (Lasso), L2 (Ridge) and L1/L2 (Elastic Net). L1 might shrink some parameters to zero but in L2, parameters never shrink to zero but approach it. Bolasso, an improvement to Lasso bootstraps samples. For Tree derived importance algorithms, a feature is more important if it reduces the impurity more, example being Regularized Random Forest.

IV. Hybrid method can be employed to utilize advantages of both filter and wrapper methods. Recursive Feature Elimination (RFE) is commonly used with SVM or RF to repeatedly construct a model and remove features with low weights. RFE is a greedy optimization algorithm which repeatedly creates models and aims to find the best performing feature subset by keeping aside the best or the worst performing feature at each iteration. RFE ranks features according to the order of their elimination.

Machine Learning Models to Classify Healthcare Data

Machine Learning (ML) algorithms build a mathematical model based on sample data. ML tasks are classified as Supervised Learning and Unsupervised Learning. In supervised learning, the training data is labelled and the response variable may be discrete/qualitative (for classification task) or continuous/quantitative (for regression task). Machine Learning for Healthcare diagnostics is a classification task where the dependent variable may be split into binary or multiple classes. Below are discussed several ML algorithms which have proven to give good diagnostics for Healthcare.



1. Logistic Regression: Logistic regression is a statistical model used for binary classification problems, where the goal is to predict one of two possible outcomes. It models the relationship between the input features and the target variable by estimating the probabilities of the classes. Logistic regression assumes a linear relationship between the features and the log-odds of the target variable, making it simple and interpretable.

2. Decision Trees: Decision trees are hierarchical models that recursively split the data based on different feature thresholds to create a tree-like structure. Each internal node represents a decision based on a feature, and each leaf node represents a class prediction. Decision trees are easy to understand and visualize, as they can be represented graphically. They can handle both binary and multi-class classification problems and are capable of capturing non-linear relationships between features and the target variable.

3. Random Forest: Random forests are ensemble models that combine multiple decision trees to make predictions. Each decision tree in the random forest is trained on a different subset of the data and a random subset of the features. The final prediction is determined by aggregating the predictions of individual trees. Random forests improve generalization

and reduce overfitting compared to a single decision tree. They are robust, handle high-dimensional data well, and can capture complex interactions between features.

4. Support Vector Machines (SVM): SVMs are binary classifiers that aim to find a hyperplane in the feature space that separates the data points of different classes with the largest margin. SVMs can handle linear and non-linear classification problems through the use of different kernel functions. They work well with high-dimensional data and are effective when the number of features is larger than the number of samples. SVMs have a strong theoretical foundation and often perform well in practice.

5. Naive Bayes: Naive Bayes classifiers are probabilistic models based on Bayes' theorem and assume that the features are conditionally independent given the class. Despite the independence assumption, Naive Bayes classifiers can still perform well in practice, especially for text classification tasks in healthcare research. They are computationally efficient, require relatively small amounts of training data, and work well with high-dimensional data.

6. Neural Networks: Neural networks, particularly deep learning models, have gained popularity in healthcare research due to their ability to handle complex data types like medical images or time-series data. Convolutional Neural Networks (CNNs) are commonly used for image classification tasks, as they can automatically learn hierarchical representations of images. Recurrent Neural Networks (RNNs) are suitable for sequential data, such as time-series or textual data. Neural networks are highly flexible and can capture intricate patterns in the data, but they often require large amounts of training data and computational resources.

7. Gradient Boosting Models: Gradient boosting models, such as XGBoost or LightGBM, are ensemble models that sequentially train weak learners (usually decision trees) to correct the mistakes of the previous learners. Each subsequent model focuses on the samples that were misclassified by the previous models. Gradient boosting models achieve high predictive performance and handle heterogeneous data well. They are effective in scenarios where high accuracy is crucial.

When selecting a machine learning model for healthcare data classification, it is important to consider the characteristics of the dataset, the specific research question, the interpretability of the model, computational resources, and the available amount of labeled data. It is often recommended to experiment with multiple models, tune their hyperparameters, and evaluate their performance using appropriate metrics to choose the best model for the research article.

Diagnosing Model Performance

When we try to predict target variables using machine learning techniques, noise, variance and bias are the main reasons for the difference between actual and predicted values. Bias refers to how far the estimate is from the target, which may result from incorrect assumptions. Deviation = E[f(x)−f^[x]] Variance is how much a random variable deviates from its mean (expressed as a unit square). For highly unbiased models, pulling more information doesn't help because it leads to higher training and competitive costs for models with high variance; As the data progresses, the cross-validation rate continues to decrease.

Therefore, taking more training samples and testing smaller sets can improve variance and more features can improve bias. A learning curve showing training and cross-recognition scores as a function of frequency of training samples included. The difference between the two curves determines the interpretation of the sample in the bias-variance landscape. If the training and cross validation scores are lower than expected, we can define the model as biased and thus unsuitable. In this case, we will consider the introduction of new features and reduce the usual work.

If the two curves are symmetrical about the demand curve and separated, you have the deviation and the pressure difference; The pattern is clear but inconsistent. In this case, we may decide to add more samples to the data and increase the continuous run order. Underfit means high bias, i.e. high training and testing cost, while overfit means low training but high testing cost.

How do you know if your model predicts the future well? Not paying attention to the effectiveness of the training from the performance of the test can have serious consequences, because different subjects focus on different aspects of the data and do not care. Machine learning can be 99% accurate in memorizing training data, but poor in test data. One of the methods is called the "train test" approach. The model is trained on different types of data, and the expected future performance is derived from the mean and standard deviation of the validation results of the data used.

The final check was done on the test data, as the test data was not used to set the model. For small data a distinction would be 70%: 20%: 10%: train: validation: testing, evaluating the performance of the model on small samples. The sklearn.model_selection module can also be used for the same purpose. Machine learning's performance on the test set will generally be lower than its performance on the training set because the learning process cannot handle statistical outliers for training data and test data.

This problem can be solved using k-fold CV [7]. To minimize variance, performance is averaged over multiple CV variables so memory and overall performance can be predicted. A variant of the Kfold CV is the one-out CV (LOOCV) / Jackknife. It is a technique used to obtain unbiased estimates and reduce the risk of overtime [16]. Additionally, hierarchical K-fold reduces bias and variance.

The strength of machine learning algorithms is in the good position of hyperparameter values. Do not give negative results or try all possible combinations by trial and error; Genetic

algorithms are a form of evolutionary computation that is well suited for hyperparameter optimization. The power function for the genetic algorithm is chosen to evaluate the quality of the drug in terms of dividing the population of the algorithm by different parameters. If we can find the best hyperparameter values, the consequences can be many. GA can be used in conjunction with roulette selection and SVM to divide by a match and then mutate to preserve genetic diversity and beliefs.

Accuracy increased from 83.70% to 88.34% [2]. The researchers worked to increase the efficiency of the neural network from 84.62% to 93%.

85% by increasing the starting weight using GA with 10 fold CV [8]. The selected feature vector contains the exposure and number of selected features. Compared to grid search, GA-NN shows greater accuracy. Other hyperparameters None. The number of layers, the number of nodes per layer [12], the choice of optimizer, the learning speed and power can also be optimized [8].

Similarly, genetic algorithms can optimize the number of base classifiers in a group [17]. Optimization for accuracy and variety always leads to the selection of a small number of classifiers. TPOT is a Python automated machine learning tool that optimizes machine learning pipelines using genetic programming. Grid search and random search achieve their goals in terms of computation time. Not even tested on the combination.

Usually when using the algorithm, we make several choices when choosing the hyperparameters. As in the case of k-NN, this can be set with the k value; In the case of RF, it can be speed and depth of the DT group. The validation curves show how well the model fits as a function of the hyperparameter values.

Challenges in implementing ML models in health care

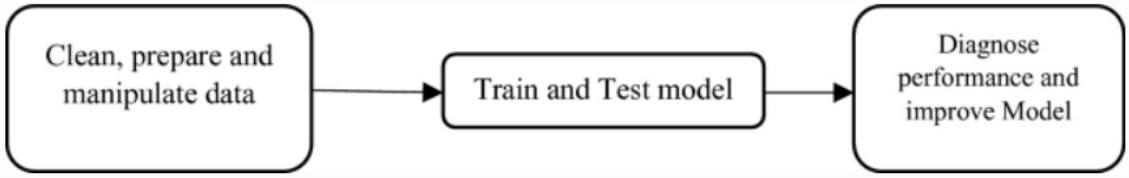
Numerous challenges that arise in the application of ML algorithms in healthcare need to be explored and discussed by experts . Any machine learning model depends on high-quality data that are representative of the population to which the model's results are to be generalized. Thus, if we intend to integrate ML models into health care, formulating effective data management at all levels becomes an essential requirement. In addition, pipelines for data processing and ML with user-friendly front ends for the products must be formulated. These pipelines can transform the raw data into datasets that can be used to train various ML models. The relevant stakeholders need to formulate an effective data governance strategy to leverage the generated data. Another critical challenge is that prediction based on ML usually does not provide reasons for the prediction unless models such as decision trees are used that allow intuitive interpretation . In situations where the ML model is used to predict a health outcome, the legal procedures are not optimized in case of a potential error. This point can be quite challenging in practice, given the complexity of legal procedures in different countries.

Experimental Dataset

To test above knowledge critical care disease datasets available online on UCI ML repository and Kaggle have been used. Several datasets have been chosen w.r.t. different types and number of independent variable, missing values, binary/multiclass prediction, etc.

Pima Indian Diabetes dataset [11, 15, 17] from National Institute of Diabetes, Digestive and Kidney Diseases has 768 female patient record at least 21 years old with 9 features each. All features are numeric. It has been shown that 50% of patients with diabetes are not properly diagnosed.

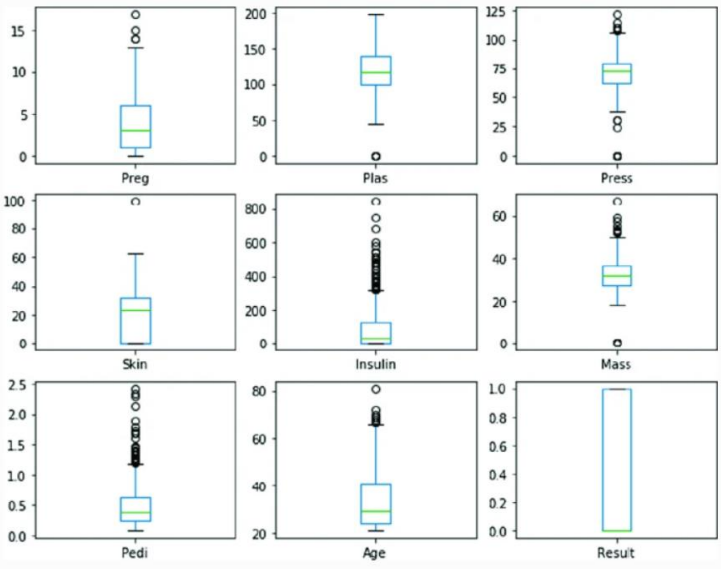
Results and Discussion



Machine learning block diagram

Feature Cleaning

On analysis of Pima Indian Diabetes dataset, missing values were found in skin thickness and insulin in 30% and 49% of rows. But domain knowledge tells that these two features can't have 0 value. Hence, these should be treated as missing values. Similarly, outliers can be detected through Box Plot/Whisker's Plot in almost all features (no. of pregnancy, glucose level, B.P., insulin level, BMI, diabetes pedigree function and age)

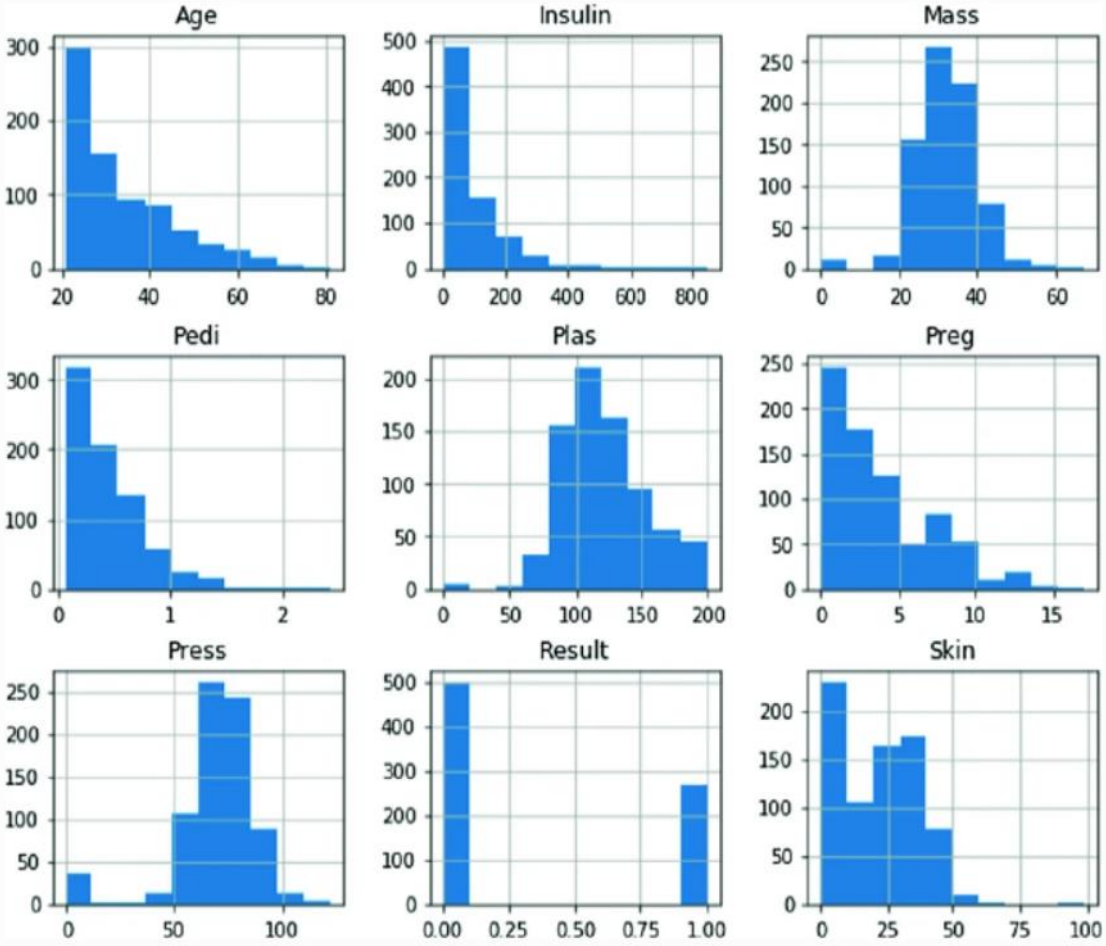


Outlier detection for Pima Indian Diabetes dataset

Missing Data

The experimentation involved choosing Statlog heart disease dataset with no missing values, randomly deleting values and imputing through SimpleImputer, kNN and MICE. SimpleImputer utilises mean for numerical and most frequent for categorical and reduces accuracy considerably, no change in accuracy is observed with kNN and there is enhancement in accuracy with MICE.

Skewness



Skewed features in Pima Indian Diabetes dataset

Feature Selection

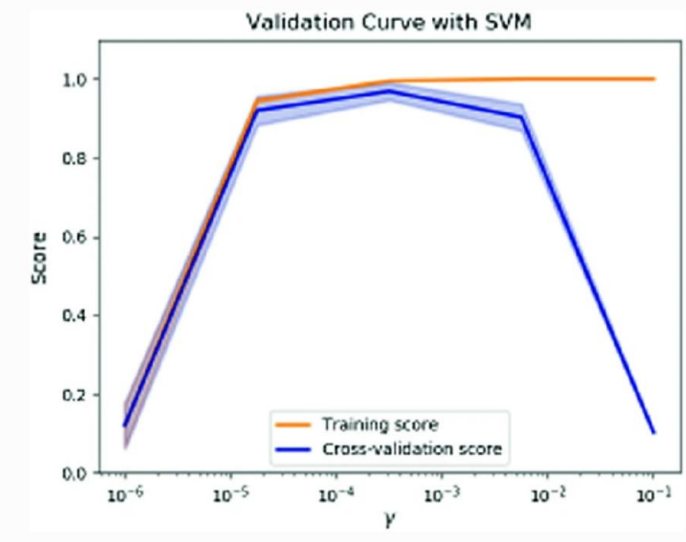
Applying Feature selection (Filter followed by Embedded and followed by SFS Wrapper) onto Arrythmia dataset with 280 features reduced it to 7 with more than 15% enhancement in accuracy. Even for Z-AlizadehSani dataset with 55 features reduction came to 34 features with 3% enhancement in Area under the Curve (AUC) utilizing filter-based feature selection [13]. Feature selection through GA when applied to Framingham Heart Disease dataset provided enhancement in accuracy by 5% through SVM classifier and 16% through NN Classifier.

Building Model

Experimentation is done on Pima Indian Diabetes dataset with Bagged DT (CART), RF, ExtraTree, AdaBoost, GradientBoost and VotingEnsemble (of LoR, CART and SVM). All achieve almost similar CV score. MLP Classifier and SVM show high sensitivity and F1 score onto Framingham heart disease dataset post oversampling with SMOTE, removing missing values through kNN imputation, scaling and splitting.

Evaluating Model Performance

Cross validation enhances performance. Overfitting occurs due to selection of model onto same data for training and testing. Thus, it is advisable to split the data into training and testing set. But a static split is not using your data efficiently. By varying what you learn on and what you're tested on, you generalize better. Experimentation was performed on Framingham dataset, balanced through under sampling and group imputation. Accuracy enhanced by 3%–6%. Figure 6 demonstrates Validation curve where the training and CV score with varying values of gamma hyperparameter of SVM classifier are shown onto Pima Indian Diabetes dataset. Results demonstrate an optimal value of the hyperparameter.



Validation curve onto Pima Indian Diabetes dataset

Hyperparameter tuning of ensembles provide good results onto Breast Cancer dataset with Stratified KFold when applied with NN (KerasClassifier), SVM, XGBoost, CatBoost, LGBM, RGF and AdaBoost ensembles. Ensembles show much better performance than NN and corresponding hyperparameters are identified. Utilising GA for hyperparameter tuning SVM enhanced accuracy by 1% for RBF kernel parameters C and ɣ. Similarly, the accuracy enhanced through GA hyperparameter tuning by 2% utilising adam optimiser. No. of hidden layers, no. of nodes in hidden layers, learning rate and momentum were optimised.

Summary:

The research article delves into the dynamic and evolving landscape of Machine Learning (ML) within the healthcare domain. Despite being considered the defining technology of the current decade, ML's impact on healthcare has been limited. However, many young start-ups in the ML industry are directing substantial efforts towards healthcare applications. Notable examples include Google's machine learning algorithm aiding in the identification of cancerous tumors on mammograms and Stanford's use of Deep Learning for skin cancer identification.

The abundance of healthcare data generated in the US, reaching approximately one trillion GB annually, offers great potential for ML applications. Nevertheless, challenges arise from the vast number of features proposed by academic researchers and risk factors suggested by clinical researchers to identify chronic diseases. Handling a large number of features necessitates a corresponding increase in sample size for enhanced accuracy. Thus, the extraction of medically high-risk factors by machines is proposed to address this issue

The article emphasizes the importance of data pre-processing through Exploratory Data Analysis and Feature Engineering to improve accuracy in healthcare ML applications. Additionally, the use of multiclass classification enables the assessment of different risk levels for patients' diseases.

In healthcare, correctly identifying sick individuals (Sensitivity) takes precedence over correctly identifying healthy ones (Specificity). As a result, the article advocates for further research to increase the sensitivity of ML algorithms in healthcare applications.

While ML is a powerful tool for improving ongoing care, the article emphasizes that it will not replace the caring and compassionate relationship between patients and healthcare providers. Instead, ML is envisioned as a tool clinicians can use to enhance the quality of care they deliver.

Overall, the research article provides an insightful introduction to the challenging and emerging field of applying ML in healthcare. It highlights the need for continued exploration, research, and development in this domain to unlock the full potential of ML for improving patient outcomes and healthcare practices.

Conclusion

The data doctors use to diagnose patients in intensive care can be used to build machine learning models that help doctors understand a patient's health and thus make better decisions. Thus, the disease prediction or diagnosis will be "learned" from the recorded data to estimate the probability that the patient will have that disease.

For patients, accuracy is the percentage of times the patient was sick and the machine detected the virus; precision/recall is the percentage of time the machine correctly predicts disease for each patient. In other words, perception in medicine is very important compared to reality; not this. Most of the time the machine is right.

KNN interpolation and SMOTE sampling for improved precision. Therefore, a type II error reflecting a false positive is not required for treatment.

Therefore, well-known hospitals must collect and analyze electronic data to help patients and doctors make timely and accurate diagnoses. Hence, research prospects for improving the capabilities of machine learning algorithms, especially for non-uniform classes and multi-class data.

References

1. S. Gupta, R.R. Sedamkar, Apply Machine Learning for Healthcare to enhance performance and identify informative features, in *IEEE INDIACom; 6th International Conference on "Computing for Sustainable Global Development"*, BVICAM, New Delhi, India, 13–15 Mar 2019 - https://ieeexplore.ieee.org/abstract/document/8991386

2. C.B. Gokulnath, S.P. Shantharajah, *An Optimized Feature Selection Based on Genetic Approach and Support Vector Machine for Heart Disease* (Springer Nature, Iran, 2018)

- https://link.springer.com/article/10.1007/s10586-018-2416-4

3. E.R.Q. Fernanded, A.C.P.L.F. de Carvalho, X. Yao, Ensemble of classifiers based on multiobjective genetic sampling for imbalanced data. IEEE Trans. Knowl. Data Eng. **14**(8) (2015)- https://ieeexplore.ieee.org/abstract/document/8640265

4. F. Babič, J. Olejár, Z. Vantová, J. Paralič, Predictive and descriptive analysis for heart disease diagnosis, in *FedCSIS*, vol. 11 pp. 155–163, IEEE Catalog Number: CFP1785N-ART c 2017, Slovakia, https://doi.org/10.15439/2017f219. ISSN 2300-5963

5. R. Pari, M. Sandhya, S. Sankar, *A Multitier Stacked Ensemble Algorithm for Improving Classification Accuracy* (IEEE, 2018)- https://ieeexplore.ieee.org/abstract/document/8509171

6. S. Mahendru, S. Agarwal, *Feature Selection Using Metaheuristic Algorithms on Medical Datasets* (Springer Nature, Singapore, 2019)- https://link.springer.com/chapter/10.1007/978-981-13-0761-4_87

7. S.M. Saqlain, M. Sher, F.A. Shah, I. Khan, M.U. Ashraf, M. Awais, A. Ghani, *Fisher Score and Matthews Correlation Coefficient-Based Feature Subset Selection for Heart Disease Diagnosis Using Support Vector Machines* (Springer, London, 2018)

https://link.springer.com/article/10.1007/s10115-018-1185-y

8. Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A.A. Yarifard, Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Comput. Methods Programs Biomed. **141**, 19–26 (2017). (Elsevier ScienceDirect)

https://www.sciencedirect.com/science/article/abs/pii/S0169260716309695?via%3Dihub

9. N. Fayyazifar, M. Samadiani, *Parkinson's Disease Detection Using Ensemble Techniques and Genetic Algorithm* (IEEE, Pakistan, 2017)

https://ieeexplore.ieee.org/abstract/document/8324074

10. I. Chlioui, A. Idri, I. Abnane, J.M.C. de Gea, J.L.F. Alemán, *Breast Cancer Classification with Missing Data Imputation* (Springer Nature, Switzerland, 2019)

https://link.springer.com/chapter/10.1007/978-3-030-16187-3_2

11. T. Santhanam, M.S. Padmavathi, Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Comput. Sci. **47**, 76–83 (2015). (Elsevier ScienceDirect)

https://www.sciencedirect.com/science/article/pii/S1877050915004536

12. Y. Khan, U. Qamar, N. Yousaf, A. Khan, Machine learning techniques for heart disease dataset: a survey, in *ICMLC*, ACM, China, 22–24 Feb 2019

https://dl.acm.org/doi/abs/10.1145/3318299.3318343

13. S. Gupta, R.R. Sedamkar, Feature Selection to reduce dimensionality of heart disease dataset without compromising accuracy. Int. J. Comput. Trends Technol. (IJCTT) **67**(6) (2019)

14. X.Y. Liu, Y. Liang, S. Wang, Z.Y. Yang, H.S. Ye, Hybrid genetic algorithm with wrapper embedded approaches for feature selection. IEEE Access **6**, 22863–22874 (2018)
https://ieeexplore.ieee.org/abstract/document/8326701

15. Z. Yang, Y. Zhou, C. Gong, Diagnosis of diabetes based on improved Support Vector Machine and Ensemble Learning, in *ICIAI*, ACM, China, 15–18 Mar 2019
https://dl.acm.org/doi/abs/10.1145/3319921.3319954

16. A. Ogunleye, Q.G. Wang, XGBoost model for Chronic Kidney Disease diagnosis. IEEE/ACM Trans. Comput. Biol. Bioinform. (2019)
https://ieeexplore.ieee.org/abstract/document/8693581

17. S. Fletcher, B. Verma, Z.M. Jan, M. Zhang, The optimized selection of base-classifiers for ensemble classification using a multi-objective genetic algorithm, in *2018 IEEE International Joint Conference on Neural Networks (IJCNN)*, Australia

https://ieeexplore.ieee.org/abstract/document/8489467

18. H.A.G. Elsayed, L. Syed, An Automatic early risk classification of hard coronary heart diseases using framingham scoring model, in *ICC* (ACM, Cambridge, UK, 2017) https://dl.acm.org/doi/abs/10.1145/3018896.3036384