8-1-2024

Application of Machine Learning Algorithms in Healthcare

Dwaipayan Mukhopadhyay

APPLICATION OF MACHINE LEARNING ALGORITHMS IN HEALTHCARE

By

Dwaipayan Mukhopadhyay

Bachelor of Science in Statistics
Calcutta University
2011

Master of Science in Mathematics
University of New Orleans
2016

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy – Mathematical Sciences

Department of Mathematical Sciences
College of Sciences
The Graduate College

University of Nevada, Las Vegas
August 2024

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

June 10, 2024

This dissertation prepared by

Dwaipayan Mukhopadhyay

entitled

Application of Machine Learning Algorithms in Healthcare

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Mathematical Sciences
Department of Mathematical Sciences

Dieudonne Phanord, Ph.D.
Examination Committee Co-Chair

Ashok Singh, Ph.D.
Examination Committee Co-Chair

Rohan Dalpatadu, Ph.D.
Examination Committee Member

Rachidi Salako, Ph.D.
Examination Committee Member

Laxmi Gewali, Ph.D.
Graduate College Faculty Representative

Alyssa Crittenden, Ph.D.
Vice Provost for Graduate Education &
Dean of the Graduate College

ABSTRACT

Machine Learning (ML) is a subset of artificial intelligence that has made substantial strides in predicting and identifying health emergencies, disease populations, and disease state and immune response, amongst a few fields of healthcare. Here we provide a brief overview of machine learning-based approaches and learning algorithms. Second, we discuss a general procedure of ML and review some studies presented in ML application for several healthcare fields. We also briefly discuss the risks and challenges of ML application to healthcare.

This dissertation also consists of four different cases in healthcare where we have applied ML techniques on real life data sets. In the first case study, Random Forest (RF) method has been used with high accuracy for classifying a rare skin disease Erythemato-squamous Dermatosis. In the second case study, Logistic regression analysis was utilized in finding the risk factors associated with Alcoholic hepatitis (AH). A sub-analysis was performed to determine variables associated with mortality in AH patients. In the third case study, Linear Discriminant Analysis (LDA) & RF were utilized for classifying five types of cancer (breast cancer, kidney cancer, colon cancer, lung cancer and prostate cancer) based on high dimensional microarray gene expression data. Principal component analysis (PCA) was used for dimensionality reduction, and principal component scores of the raw data for classification. In the fourth case study, we aim to discover the potential factors behind the initiation and then possibly sustain the desire to quit smoking using the LDA & RF method.

ACKNOWLEDGMENT

I would like to thank Dr. Dieudonne Phanord & Dr. Ashok Singh for agreeing to be my dissertation advisors, for their invaluable support and I am truly grateful for all their wisdom and guidance. I would also like to thank Dr. Rohan Dalpatadu & Dr. Rachidi Salako for their patience, this thesis would not have been possible without their constant and unwavering support. Finally, I am forever in debt to Dr Laxmi Gewali's support in my academic career and his advice on my dissertation.

PREFACE

This dissertation consists of a total of nine chapters. Chapters 1-4 provide a brief description of the commonly used machine learning tools, and a literature review of the application of machine learning in the healthcare field.

Chapter 5 describes the problems faced during the implementation of machine learning techniques. Chapters 6 – 9 are the four research articles, two of which have already been published in peer-reviewed research journals, and the other two have been submitted for publication.

Each of the Chapters 6 – 9 are standalone articles, with their own figures, tables, and references, so the figure and table numbers repeat in these sections and each section has its own set of references.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The application of machine learning dates to the 1950s when Alan Turing proposed the first machine that can learn and become artificially intelligent [1]. Machine learning (ML) is the engine which is helping to drive advances in the development of artificial intelligence. It is impressively employed in both academia and industry to drive the development of ‘intelligent products’ with the ability to make accurate predictions using diverse sources of data [2]. On of the many exciting features of ML-based models is that they learn automatically and experimentally and do not need to be explicitly programmed [3,4]. ML improves efficiency and reliability and reduces costs in computational processes. Moreover, it can accurately and rapidly generate models through data analysis. Machine learning presents tools that can process a large amount of data, the volume of which is far beyond human understanding. Since its advent, machine learning has been used in various applications, ranging from security services through face detection [5] to increasing efficiency and decreasing risk in public transportation [6, 7], and recently in various aspects of healthcare and biotechnology [8-13].

Healthcare is one of the most important fields which is overseeing a huge influx of applications of machine learning techniques [14]. Current machine learning advancements in healthcare have primarily served as a supportive role in a physician or analyst's ability to fulfill their roles, identify healthcare trends, and develop disease prediction models. Machine learning-based approaches have also been implemented to achieve increased efficiency in the organization of electronic health records [15], identification of irregularities in the blood samples [5], organs [6-8], and bones [16] using medical imaging and monitoring, as well as in robot-assisted surgeries [9, 17].

In the following, the dissertation is organized as follows: in **Chapter 2**, machine learning and its general framework used in healthcare are expressed. In **Chapter 3**, we present the different categories of a learning model used in the medical or healthcare field. In **Chapter 4**, a literature review of the application of ML techniques in healthcare is introduced. In **Chapter 5**, we describe some challenges on the use of machine learning in medicine briefly

Chapters 6 – 9 are the four research articles, two of which have already been published in peer-reviewed research journals, and the other two have been submitted for publication.

MACHINE LEARNING

Everything in our lives is connected & digitally recorded [18,19]. The growth of Artificial intelligence (AI), particularly, machine learning (ML) in recent years in the context of data analysis allows the applications to function in an intelligent manner [20]. In the phrase “machine learning”, “learning” represents the search process in the possible representation space to create the best representation based on available data [21,22]. It enables computers to “self-learn” or obtain information from training data; recognize patterns in data and develop their own predictions, improving over time without being explicitly programmed [23].In a data driven system following techniques are substantial- Classification analysis, regression, data clustering, dimensionality reduction etc [24, 25]. Besides, deep learning originated from the artificial neural network that can be used to intelligently analyze data [26]. It is imperative in selecting a proper learning algorithm that is suitable to understand the basics of the aim of the analysis and their applicability to apply in several real-world application areas.

General Framework for Designing a Learning Model in Medicine:

Here we introduce designing a learning model in the healthcare field which consists of five main phases: problem definition, dataset, data preprocessing, ML model development, and evaluation [27,28]. These phases are shown in **Figure 1**. In the following, each of these phases is described in detail.

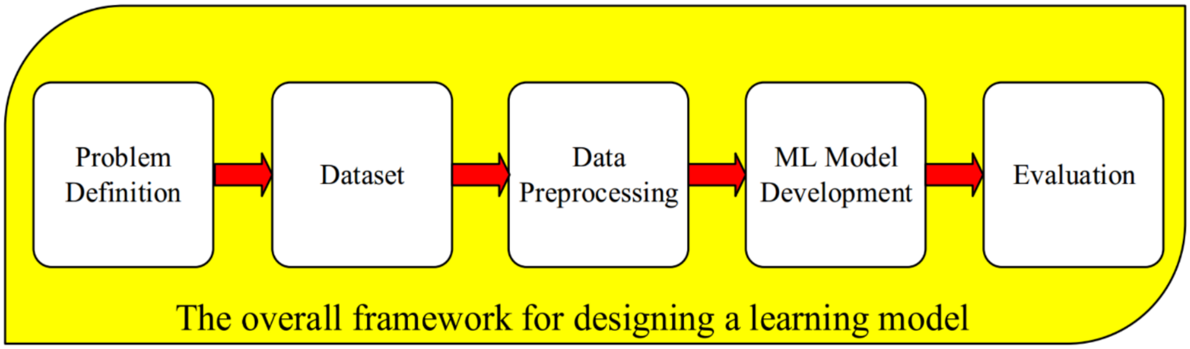


Figure 1 : Different Phases for Designing a Learning Model

Problem Definition: The first step for researchers is to identify problems and challenges in the healthcare field. They should examine the preexisting solutions and figure out how to improve those solutions using machine learning. In addition, [29].

Database: In the healthcare field, datasets are used for training, validating, and testing and may include demographic information, images, laboratory results, genomic data etc. [30,31]. Various platforms are used to produce or collect these data, for example network servers, e-health records, genome data, personal computers, smartphones, mobile applications, and wearable devices [32,33]. ML-based models are data centric. Therefore, they may be faced with a problem called overfitting or underfitting [34,35]. An efficient learning model has an appropriate bias and proper variance. **Figure 2** describes the overfitting and underfitting.

One of the basic methods in analysis is to divide the dataset into two parts: training set and testing set. The "training set" indicates a dataset used for training the learning model and adjusting its parameters. The "testing set" also indicates a dataset used for evaluating the performance of the learning model. Usually, the training set is larger than the testing set, for example, the ratio of 75 to 25 or 80 to 20.

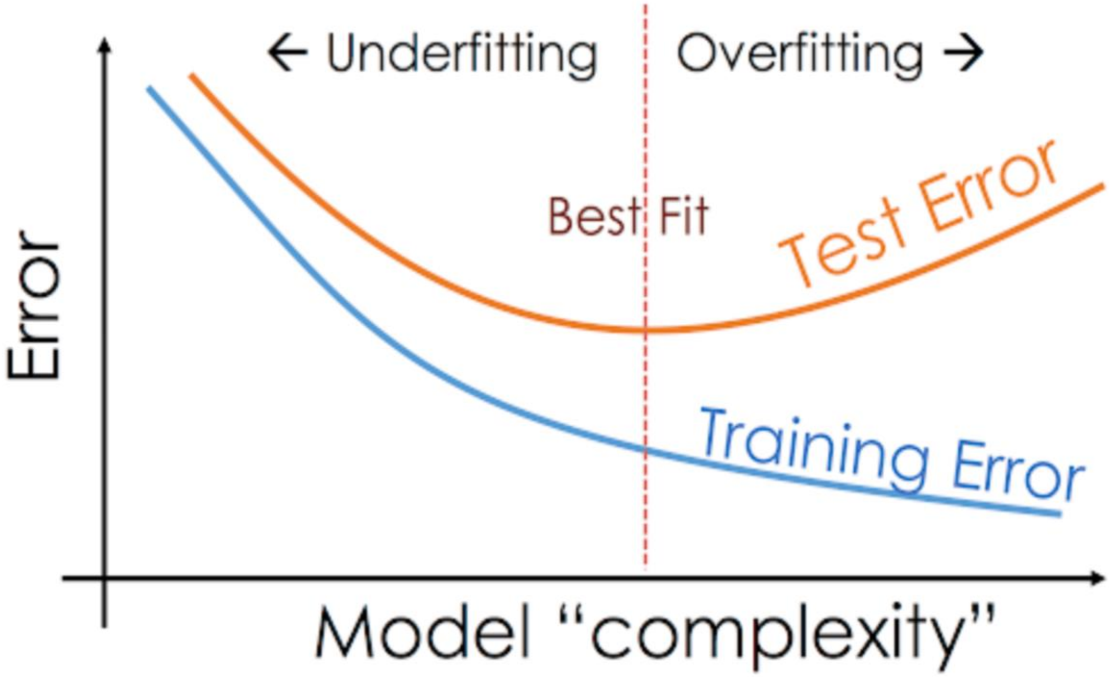
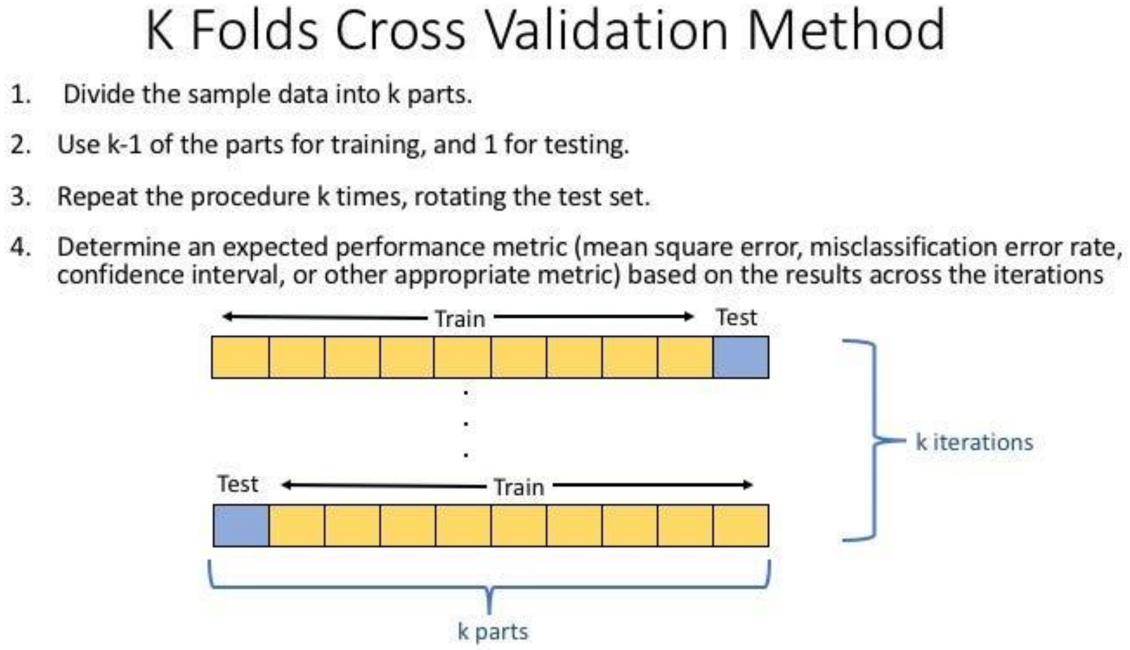


Figure 2: Overfitting and Underfitting Description

Whren the data set is small, a much preferable method is the K-Fold Cross-Validation technique is used [36,37]. In K-Fold Cross Validation, we split the dataset into k number of subsets (known as folds) then we perform training on all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purposes each time.

In **Figure 3** we describe a K-Fold Cross validation technique method.



K Folds Cross Validation Method

1. Divide the sample data into k parts.
2. Use k-1 of the parts for training, and 1 for testing.
3. Repeat the procedure k times, rotating the test set.
4. Determine an expected performance metric (mean square error, misclassification error rate, confidence interval, or other appropriate metric) based on the results across the iterations

Figure 3: K-Fold Cross Validation Technique

Data Pre-Processing: Machine learning models require high-quality data to achieve a higher quality in the training process and a more suitable performance, hence designing a learning model in the healthcare field is one of the most challenging issues. In data with high dimensions, some data reduction methods, such as feature selection [38,39] or feature extraction [40], can be used. Feature selection selects the best subset of features. On the other hand, feature extraction finds a new dataset with lower dimensions based on the initial data set.

ML Model Development: Learning models with more parameters can produce more accurate results but these models perform more computational operations and need a longer time for training. As a result, they cannot be used for real-time applications. Therefore, lightweight architectures are more appropriate for designing a leaning model. Considering the type of learning scheme is also very important when developing ML models [41,42]. In general, there are four main learning methods, including supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [43,44]. We describe these techniques more accurately in **Section 3.2.**

Evaluation: Study [45] shows that evaluating a machine learning-based system means assessing detecting differences between the current behavior of the system and the expected behavior. In the evaluation process, designers use various scales to examine the performance of the learning model which determines its strengths and weaknesses. In addition, after deploying the learning model in real environments, we must re-examine the performance of the learning model to evaluate its behavior when interacting with real users [46,47].

After constructing and training the final model, its performance must be evaluated based on the following factors:

Evaluating the performance of the final model: After constructing and training the final model, its performance must be evaluated based on the following factors:

- True positive (TP): The number of positive class members, which are properly predicted by the classifier and are labeled as positive class.
- True negative (TN): The number of negative class members, which are properly predicted by the classifier and are labeled as negative classes.
- False positive (FP): The number of negative class members, which are falsely predicted by the classifier and are labeled as positive class.
- False negative (FN): The number of positive class members, which are falsely predicted by the classifier and are labeled as negative class.

In the following, we introduce some commonly used measures scales for evaluating a learning model. These scales are based on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN):

Sensitivity or Recall: This scale is defined as a probability so that a classifier truly predicts the result as positive when the corresponding ground truth is also positive. The other name of this scale is the true positive rate (TPR), and it is calculated as follows shown in **Figure 4**.

$$Sensitivity = \frac{TP}{TP + FN}$$

Figure 4: Sensitivity Formulae

Specificity: This scale is defined as the probability so that a classifier truly predicts the result as negative when the corresponding ground truth is also negative. The other name of the specificity is the true negative rate (TNR) and it is calculated as follows shown in **Figure 5**.

$$Specificity = \frac{TN}{TN + FP}$$

Figure 5: Specificity Formulae

Precision: This scale is defined as the probability so that a classifier truly predicts the result as positive, ratio of correct positive predictions to overall positive predictions. The other name of precision is Positive predicted value (PPV), and it is defined as shown in **Figure 6.**

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

Figure 6: Precision Formulae.

Accuracy: Accuracy simply measures how often the classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions as shown in **Figure 7.**

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

Figure 7: Accuracy Formulae

F1-Score or F-Measure: F1 score is the weighted average of precision and recall. The classifier will only get a high F-score if both precision and recall are high. This metric only favors classifiers that have similar precision and recall, and the formulae is shown in **Figure 7.**

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Figure 8: F1-Score Formulae

The higher the F1 score, the better the performance of our model. The range for F1-score is [0,1].

Receiver Operating Characteristic (ROC) Curve: The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR (True Positive Rate) against the FPR (False Positive Rate) at various threshold values**.**

The **Area Under the Curve (AUC)** is the measure of the ability of a classifier to distinguish between classes. From the graph, we simply say the area of the curve ABDE and the X and Y-axis.

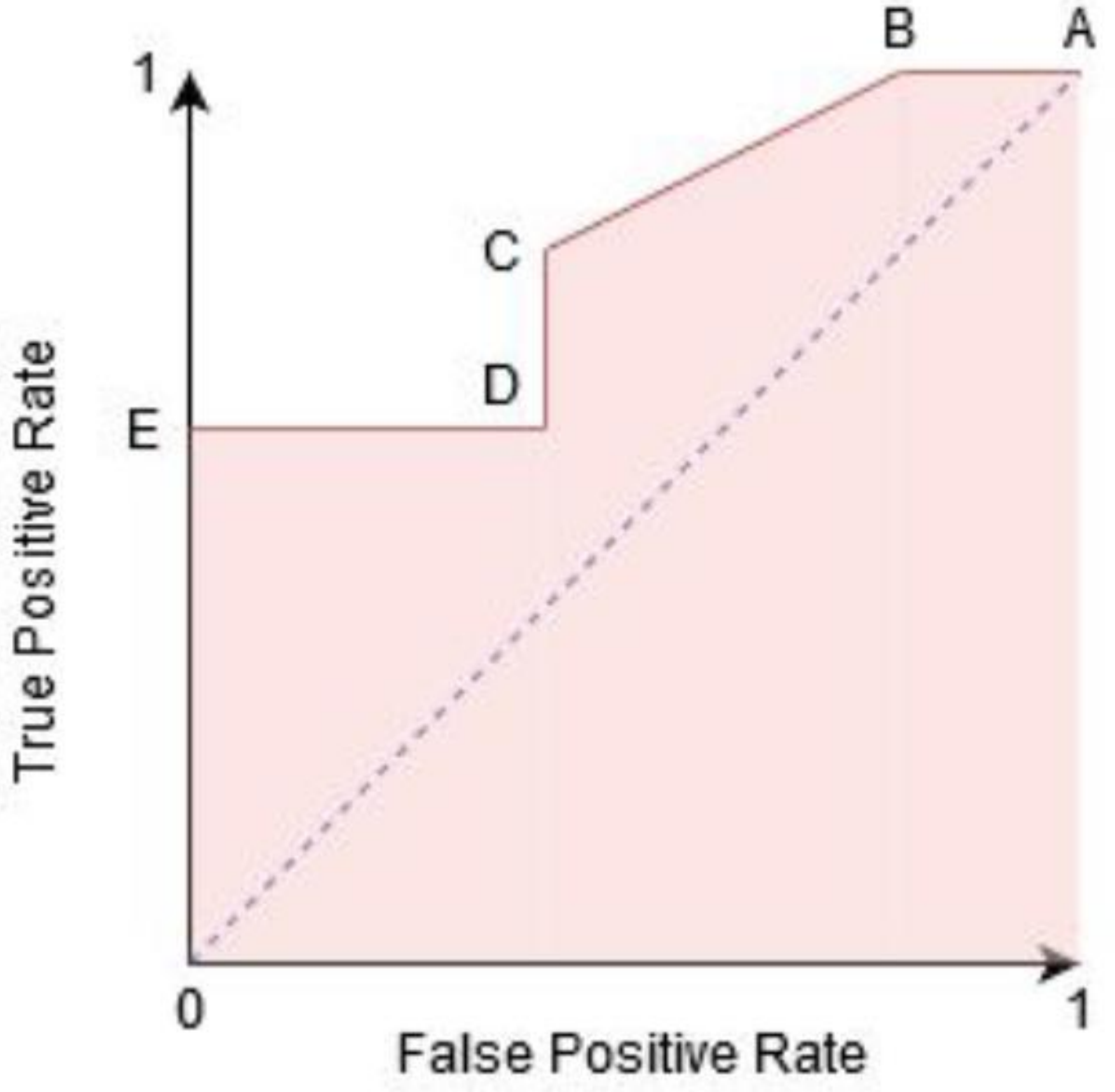


Figure 9: AUC-ROC Curve

From the graph shown above, the greater the AUC, the better is the performance of the model at different threshold points between positive and negative classes. This simply means that When AUC is equal to 1, the classifier is able to perfectly distinguish between all Positive and Negative class points. When AUC is equal to 0, the classifier would be predicting all Negatives as

Positives and vice versa. When AUC is 0.5, the classifier is not able to distinguish between the Positive and Negative classes.

In a ROC curve, the X-axis value shows False Positive Rate (FPR) which is defined as 1-Specificity, and Y-axis shows True Positive Rate (TPR). The higher the value of X means higher the number of False Positives (FP) than True Negatives (TN), while a higher Y-axis value indicates a higher number of TP than FN. So, the choice of the threshold depends on the ability to balance between FP and FN.

CATEGORIES OF ML-BASED TECHNIQUES

In this section, we provide a step-by-step procedure of ML-based techniques used in medicine and it has four categories-

- Different data pre-processing methods (data cleaning methods, data reduction methods).
- Various learning methods (unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning);
- Evaluation methods (simulation-based evaluation and practical implementation-based evaluation in real environment);
- Application (diagnosis, treatment).

Data Processing Methods

Data processing methods are subdivided into Data cleaning and reduction methods which are described below.

Data Cleaning Methods: ML-based methods utilized in healthcare implement data cleaning methods to eliminate missing data or noisy data, because such problems are common in the health datasets. These problems have several reasons: (1) Data collection devices are not accurate in the healthcare field. (2) Some data samples may incorrectly be recorded due to human errors; (3) Some patients do not disclose proper information about their illness inadvertently or deliberately. In general, there are several data cleaning methods, including missing value management, noisy data management, and data normalization [48,49].

Data Reduction Methods: High dimensions in healthcare data reduces the quality of the training process and the accuracy of the learning model. Dimensionality reduction means that health data are presented in a compressed form. An appropriate dimensionality reduction scheme in the healthcare field should maintain useful features. Data reduction methods are divided into two main categories: feature selection and feature extraction. The primary distinction between the selection and extraction of features is that the “feature selection” keeps a subset of the original features [50], while “feature extraction” creates brand new ones [51]. In the following, we briefly discuss these techniques.

Feature Selection: This is the process of choosing a subset of unique features (variables, predictors) to use in building machine learning and data science models. An optimal subset of the selected features can minimize the overfitting problem through simplifying and generalizing the model as well as increases the model’s accuracy [50]. Thus, “feature selection” [52, 53] is considered as one of the primary concepts that greatly affects the effectiveness and efficiency of the target machine learning model. Chi-squared test, Analysis of variance (ANOVA) test, Pearson’s correlation coefficient, recursive feature elimination, are some popular techniques that can be used for feature selection.

Feature Extraction: This technique usually provides a better understanding of the data by keeping the main features of the database and removes its noise and correlations, alongside improving prediction accuracy, and to. These methods are used for compressing health data that have high dimensions, which will help to accelerate the learning process and reduce computational cost or training time [54].

Many algorithms have been proposed to reduce data dimensions in the machine learning and data science literature [24, 55]. In the following, we summarize the popular methods that are used widely in various application areas.

Linear Discriminant Analysis (LDA): Linear Discriminant Analysis (LDA) is a linear decision boundary classifier created by fitting class conditional densities to data and applying Bayes' rule [56, 57,58]. This method is also known as a generalization of Fisher's linear discriminant, which projects a given dataset into a lower-dimensional space, i.e., a reduction of dimensionality that minimizes the complexity of the model or reduces the resulting model's computational costs. LDA seeks directions (eigenvectors) that maximize the ratio of the determinant of the between-class scatter matrix to the within-class scatter matrix. These eigenvectors are also referred to as discriminant vectors, and they define the directions along which the data should be projected to achieve maximum class separation. One of the foundational assumptions of underlying LDA is the Gaussian distribution of features within each class, alongside the assumption of equal covariance matrices across different classes [57]. In **Figure 10**, a typical LDA approach for dimensionality reduction is shown. LDA is closely related to ANOVA (analysis of variance) and regression analysis, which seeks to express one dependent variable as a linear combination of other features or measurements.

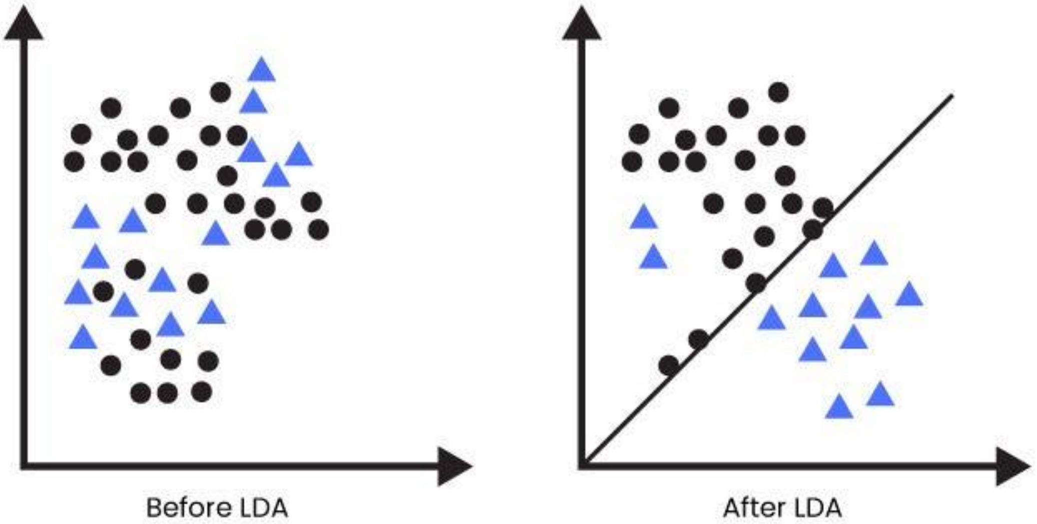


Figure 10: LDA for Dimensionality Reduction

Principal Component Analysis (PCA): Principal component analysis (PCA) is a well-known unsupervised learning approach in the field of machine learning that transforms a set of correlated variables into a set of uncorrelated variables known as principal components [59,60]. Thus, it can be used as a feature extraction technique that reduces the dimensionality of the datasets, and to build an effective machine learning model [51]. Technically, PCA identifies the completely transformed with the highest eigenvalues of a covariance matrix and then uses those to project the data into a new subspace of equal or fewer dimensions [58]. **Figure 11** demonstrates a PCA.

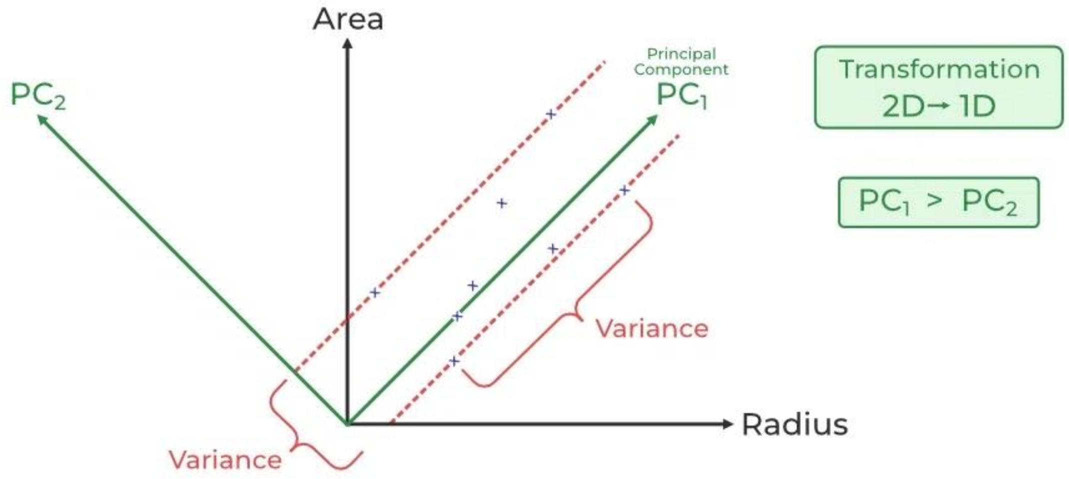


Figure 11: Principal Component Analysis

LASSO and Ridge Regression: LASSO and Ridge regression are well known as powerful techniques which are typically used for building learning models in the presence of a large number of features, due to their capability to preventing over-fitting and reducing the complexity of the model. The LASSO (least absolute shrinkage and selection operator) regression model uses L1 regularization technique [58] that uses shrinkage, which penalizes "absolute value of magnitude of coefficients" (L1 penalty). As a result, LASSO appears to render coefficients to absolute zero and aims to find the subset of predictors that minimizes the prediction error for a quantitative response variable. On the other hand, ridge regression uses L2 regularization [58], which is the "squared magnitude of coefficients" (L2 penalty). Thus, ridge regression forces the weights to be small but never sets the coefficient value to zero and does a non-sparse solution. Overall, LASSO regression is useful to obtain a subset of predictors by eliminating less important features, and ridge regression is useful when a data set has “multicollinearity” which refers to the predictors that are correlated with other predictors.

Types of Learning Methods

Machine Learning algorithms are mainly divided into four categories: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning [61], as shown in **Figure 12.**

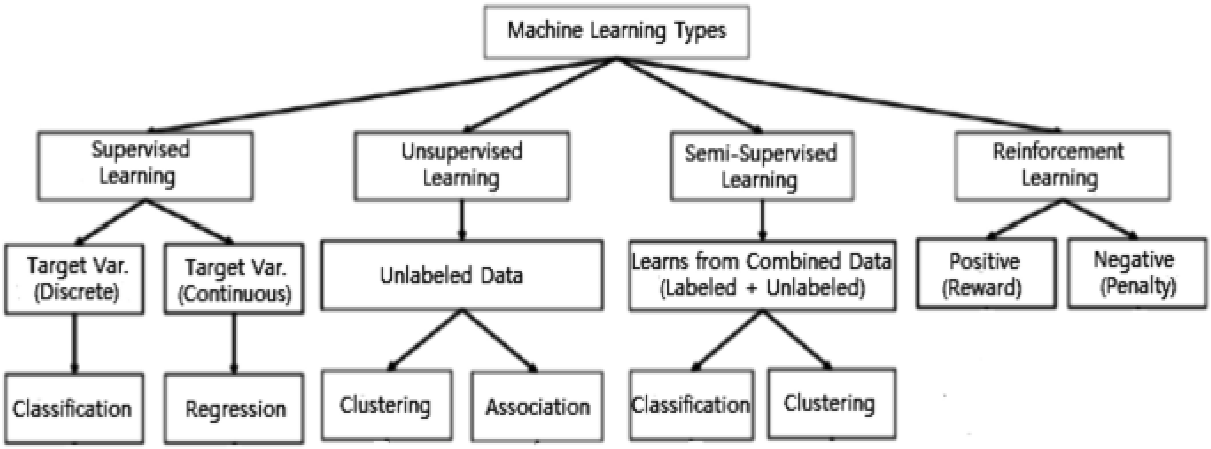


Figure 12: Types of Machine Learning Methods

Supervised: Supervised learning is the task to learn a function that maps an input to an output based on sample input-output pairs [62]. It uses labeled training data and a collection of training examples to infer a function. Supervised learning is carried out when certain goals are identified to be accomplished from a certain set of inputs [63], i.e., a task-driven approach. Supervised tasks are "classification" that separates the data, and "regression" that fits the data. For instance, predicting the class label or sentiment of a piece of text, like a tweet or a product review, i.e., text classification, is an example of supervised learning. In the following, we introduce some of the most important supervised learning schemes-

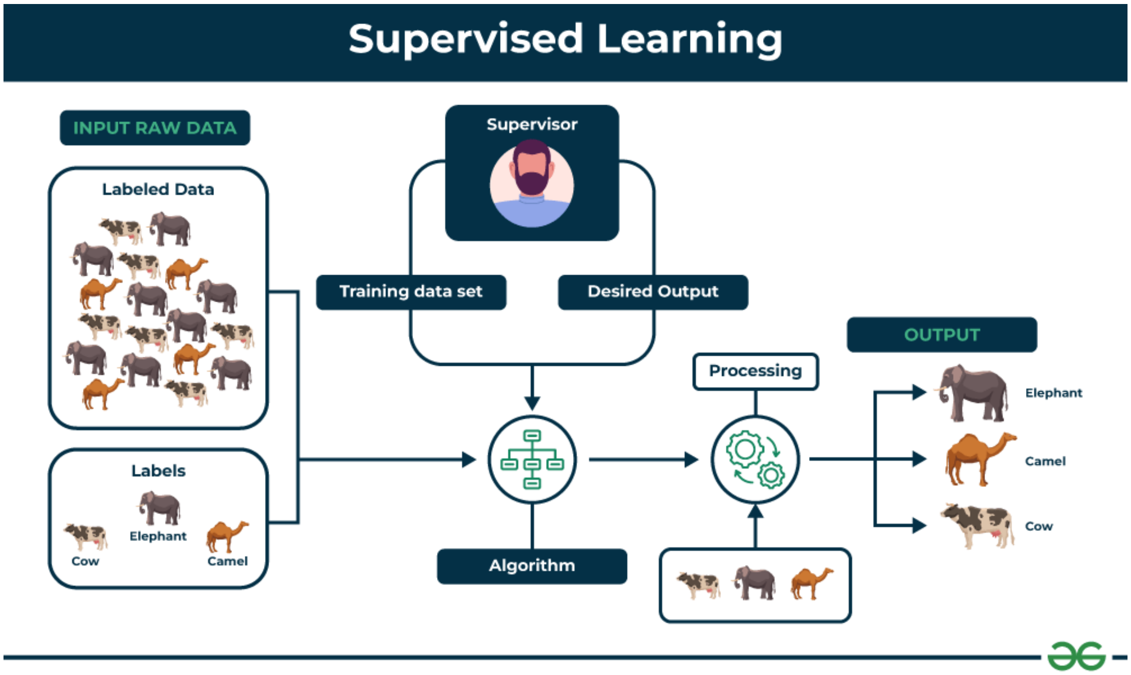


Figure 13: Supervised Learning Method

Naive Bayes (NB): The naive Bayes algorithm which works well for both binary and multi-class categories, is based on the Bayes' theorem with the assumption of independence between each pair of features [57]. To effectively classify the noisy instances in the data and to construct a robust prediction model, the NB classifier can be used [64]. The key benefit is that, compared to more sophisticated approaches, it needs a small amount of training data to estimate the necessary parameters and quickly [58]. However, its performance may affect due to its strong assumptions of independence of feautres. Gaussian, Multinomial, Complement, Bernoulli, and Categorical are the common variants of NB classifier [58].

Logistic Regression (LR): Logistic Regression (LR) [65] is a common probabilistic based statistical model used to solve classification issues. The logistic function is utilized here to estimate the probabilities, which is also referred to as the mathematically defined sigmoid function in Eq. 1. It can overfit high-dimensional datasets and works well when the dataset can be separated linearly. The regularization (L1 and L2) techniques [58] can be used to avoid over-fitting in such scenarios. The assumption of linearity between the dependent and independent variables is considered as a major drawback of Logistic Regression. It can be used for both classification and regression problems, but it is more commonly used for classification.

$$g(z) = \frac{1}{1+e^{-z}} \qquad (1)$$

Decision Tree (DT): Decision tree (DT) [66] is a well-known non-parametric supervised learning method. DT learning methods are used for both the classification and regression tasks [58]. ID3 [67], C4.5 [66], and CART [68] are well known for DT algorithms. By sorting down the tree from the root to some leaf nodes, as shown in **Fig. 10**, DT classifies the instances. Instances are classified by checking the attribute defined by that node, starting at the root node of the tree, and then moving down the tree branch corresponding to the attribute value. The hierarchical tree is created based on features in the dataset. In the decision tree, there are three types of nodes: root node (the highest node in the decision tree), the internal node (it indicates an experiment (or comparison) on each feature), leaf node/terminal node (class label or final result).

For splitting, the most popular criteria are "gini" for the Gini impurity and "entropy" for the information gain as shown in **Equation 2 & 3** that can be expressed mathematically as [58].

Entropy:

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \qquad (2)$$

Gini(E):

$$E = 1 - \sum_{i=1}^{c} p_i^2 \qquad (3)$$

In the following **Figure 14** we demonstrate Decision tree framework.

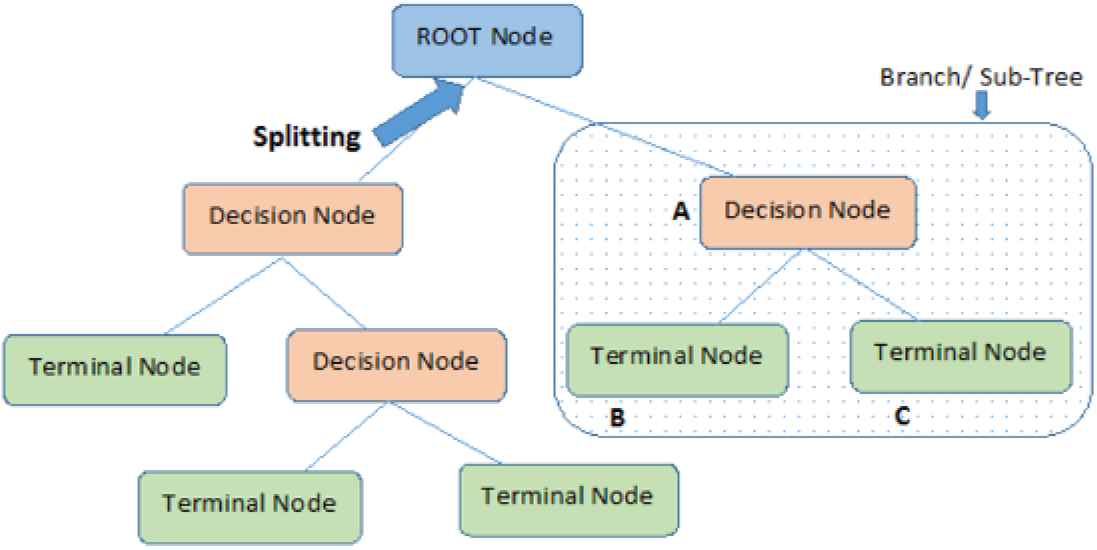


Figure 14: Decision Tree

Random Forest (RF): A random forest classifier [69] is well known as an ensemble classification technique that uses “parallel ensembling” which fits several decision tree classifiers in parallel, as shown in **Fig. 5,** on different samples and uses majority voting or averages for the outcome or final result. It thus minimizes the over-fitting problem and increases the prediction accuracy and control [58]. Therefore, the RF learning model with multiple decision trees is typically more accurate than a single decision tree-based model [70]. To build a series of decision trees with controlled variation, it combines bootstrap aggregation (bagging) [71] and random feature selection [72]. It is adaptable to both classification and regression problems and fits well for both categorical and continuous values.

In the following **Figure 15** we demonstrate a typical Random Forest framework.

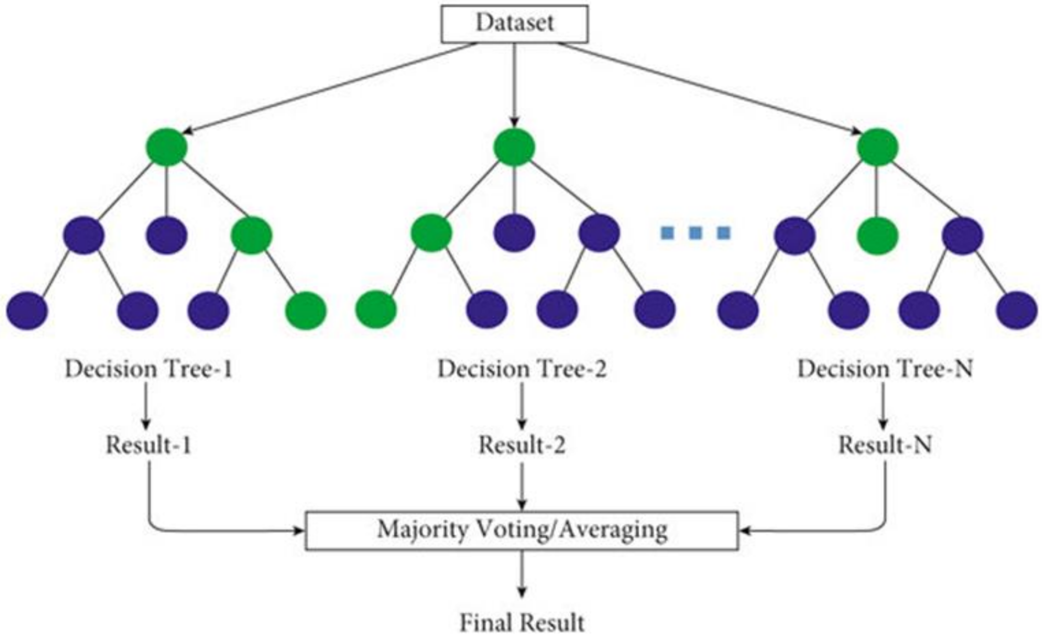
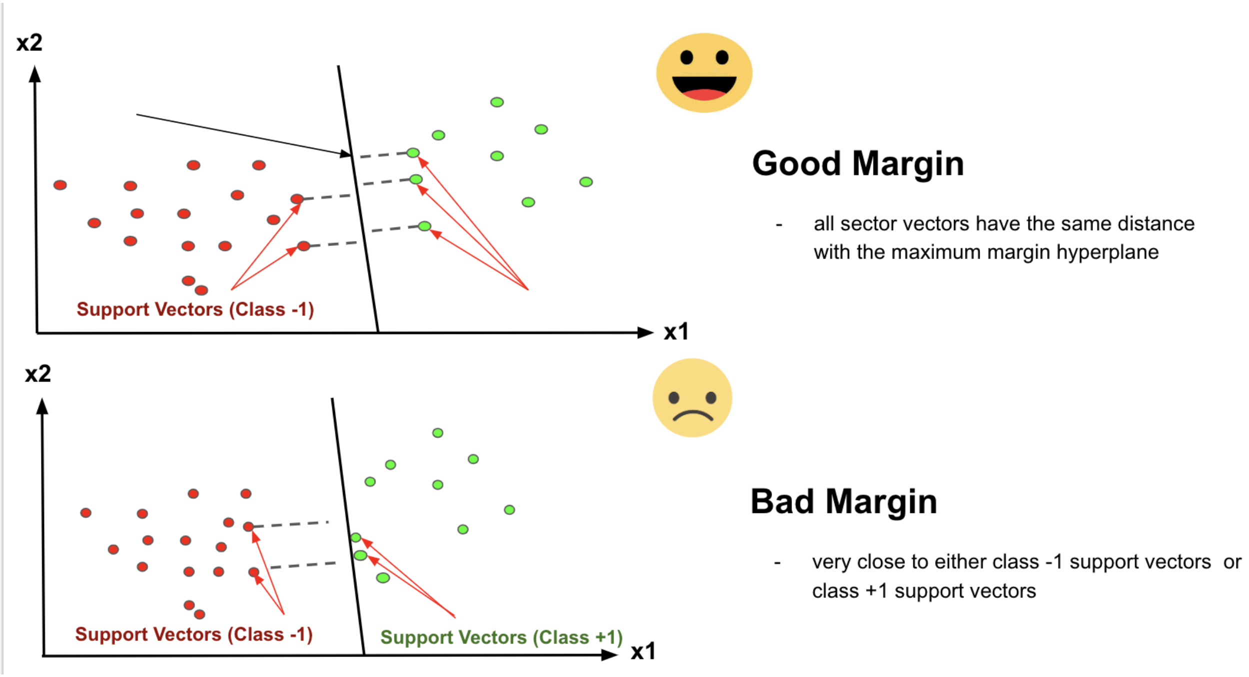


Figure 15: Random Forest

Extreme Gradient Boosting (XGBoost): Extreme Gradient Boosting (XGBoost) is a form of gradient boosting that takes more detailed approximations into account when determining the best model [58]. Gradient Boosting, like Random Forests [69] above, is an ensemble learning algorithm that generates a final model based on a series of individual models, typically decision trees. The gradient is used to minimize the loss function, like how neural networks [24] use gradient descent to optimize weights. It computes second-order gradients of the loss function to minimize loss and advanced regularization (L1 and L2) [58], which reduces over-fitting, and improves model generalization and performance. XGBoost is fast to interpret and can handle large-sized datasets well.

Support Vector Machine (SVM): Support vector machine (SVM) [73], is a common technique that can be used for classification, regression, or other tasks. The hypothesis used in SVM is that there is a hyperplane in the feature space which means that data are linearly separable. In the training process, SVM seeks to find this hyperplane, which separates two classes from each other. This hyperplane should have two features: (1) It must separate dataset in two classes; (2) This hyperplane must be in the middle of the two classes to have the highest margin from two classes. It is effective in high-dimensional spaces and can behave differently based on different mathematical functions known as the kernel. Linear, polynomial, radial basis function (RBF), sigmoid, etc., are the popular kernel functions used in SVM classifier [58]. However, when the data set contains more noise, such as overlapping target classes, SVM does not perform well. Moreover, this hypothesis is not practical.

In the following **Figure 16** we demonstrate SVM.


x2
x1
Support Vectors (Class -1)
x2
x1
Support Vectors (Class -1)
Support Vectors (Class +1)

Good Margin

- all sector vectors have the same distance with the maximum margin hyperplane

Bad Margin

- very close to either class -1 support vectors or class +1 support vectors

Figure 16: Support Vector Machine

K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) [74] which can be utilized in both classification as well as regression situation, is an “instance-based learning” or non-generalizing learning, also known as a “lazy learning” algorithm. KNN uses data and classifies new data points based on similarity measures (e.g., Euclidean distance function) [58]. In this method, we determine the class of the new sample as follows: first, we compare this sample with the training dataset to determine the k closest samples in the training set, called neighbors. In the next step, the class of this data sample is determined based on the majority voting of neighbors. In this method, k is a key parameter that indicates the number of the closest training samples in the feature space. It is quite robust to noisy training data, and accuracy depends on the data quality. The biggest issue with KNN is to choose the optimal number of neighbors to be considered.

In the following **Figure 17** we demonstrate a KNN.

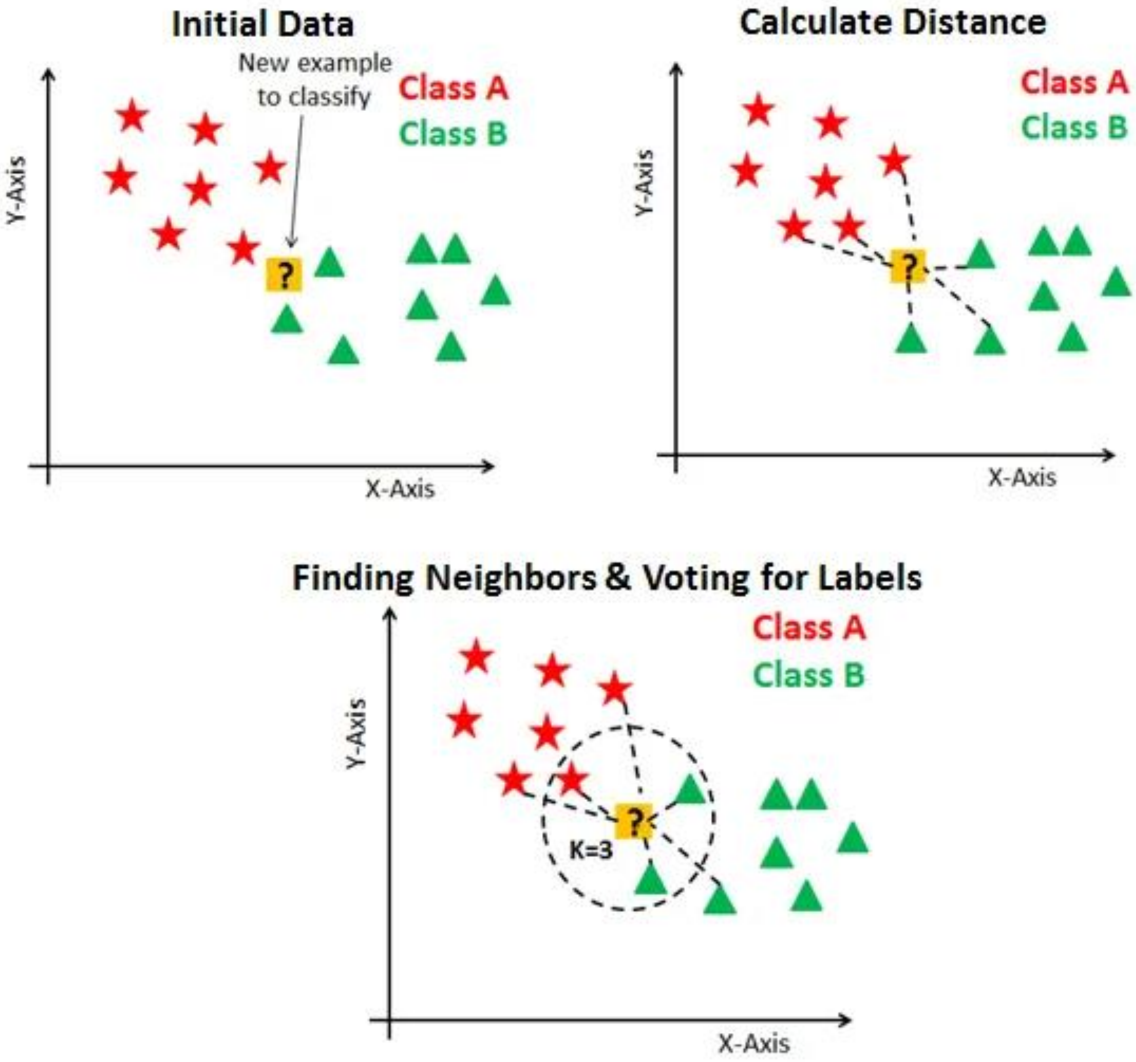


Figure 17: K- Nearest Neighbor

Artificial Neural Network (ANN): Artificial neural network includes input variables, output variables and weights. The network's behavior depends on the relationship between input and output variables [75,76]. ANNs consist of three layers- these layers include several processing units called neurons. The first layer is the input layer that receives raw data, and the second layer is also known as the hidden layer that performs a learning task. The third layer is also known as the output layer which depends on the learning process in the hidden layer as well as weights related to input units and hidden units.

In the following **Figure 18** we demonstrate a ANN.

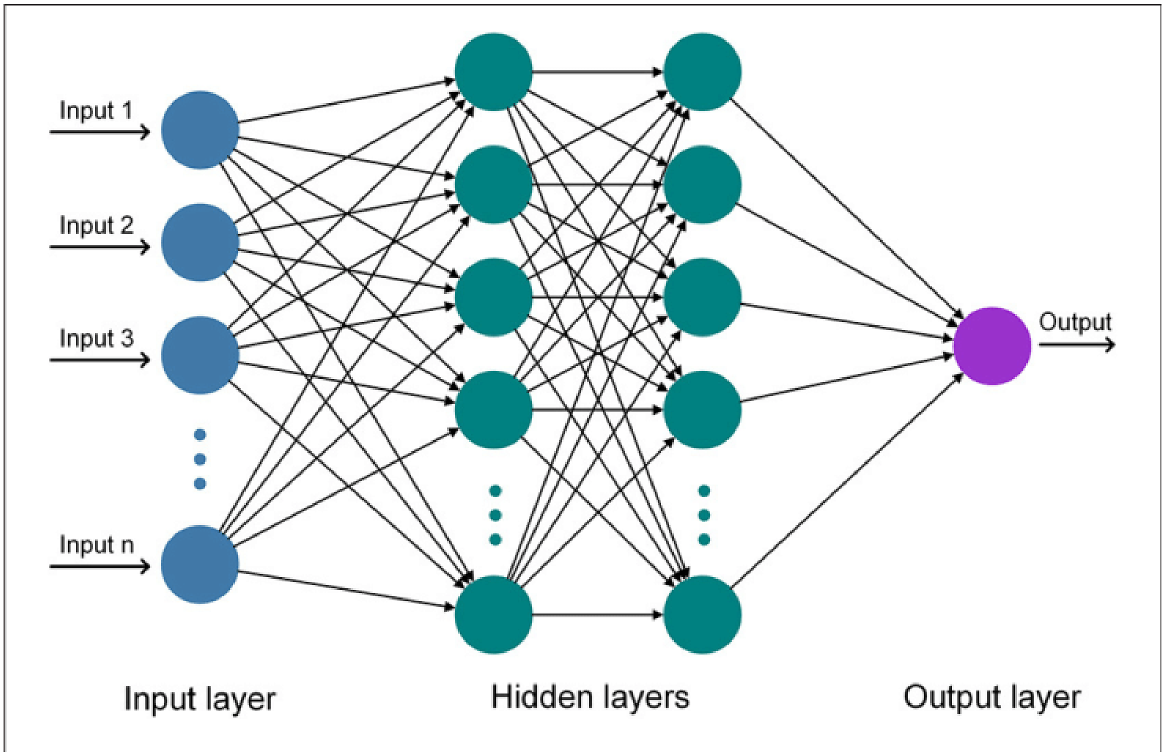


Figure 18: Artificial Neural Network

The designer determines the number of hidden layers and the number of neurons in each layer through trial and error. The most common method for training ANNs and modifying weights to get the lowest error is the back-propagation algorithm.

Deep Learning (DL): Deep learning is a subset of artificial neural networks (ANN)-based machine learning approaches with representation learning. Deep learning provides a computational architecture by combining several processing layers, such as input, hidden, and output layers, to learn from data [24,75]. The main advantage of deep learning over traditional machine learning methods is its better performance in several cases, particularly learning from large datasets [76, 77]. DL can work with labeled and unlabeled datasets and can be trained to achieve several goals [78]. The most common deep learning algorithms are: Multi-layer Perceptron (MLP), Convolutional Neural Network (CNN, or ConvNet), Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) [76].

In the following **Table-1** we summarize the advantages and disadvantages of the above-mentioned supervised methods.

Table 1: Advantages and Disadvantages of Supervised Learning Methods

Algorithm	Advantages	Disadvanatges
Naïve Bayes (NB)	Simple implementation, high computational & learning speed, high classification speed, managing overfitting, managing noisy data, and managing missing values	Assuming independence of features, lack of ability to manage features with high correlation
Decision tree (DT)	Simple understanding, high computational & learning speed, high classification speed, managing missing values	lack of ability to manage overfitting, low ability to manage noisy data & data with high correlation
Artificial neural network (ANN)	High flexibility, high accuracy & classification speed, manage data with high correlation, suitable for nonlinear and complex databases	Difficult implementation, low learning speed, inability to manage missing values, inability to manage noisy data, lack of ability to manage overfitting
Random forest (RF)	Ability to manage noisy data, high classification speed, suitable for large and heterogeneous databases	Difficult implementation, low learning speed & ability manage missing values, low ability to manage overfitting & data with high correlation
Deep learning (DL)	Suitable for large and high dimensional databases, high accuracy & classification speed, ability to manage noisy data & data with high correlation	Difficult implementation, low learning speed, inability to manage overfitting, low ability to manage missing values
Support vector machine (SVM)	Ability to manage data with linear separability and nonlinear separability, high accuracy, high classification speed, ability to manage data with high correlation	Assuming linear separability for dataset, low ability to manage overfitting, low learning speed, low ability in managing missing values, low ability to manage noisy data
K-nearest neighbor (KNN)	Simple algorithm, stable performance, high learning speed, ability to manage overfitting	High computational overhead, sensitivity to local data structures, low classification speed, low ability in managing missing values, inability to manage noisy data & data with high correlation

Unsupervised: Unsupervised learning analyzes unlabeled datasets without the need for human interference, i.e., a data-driven process [24]. In this technique, the dataset includes data samples whose relevant output is not clear [79]. This learning scheme tries to discover the data patterns and relationships in the data. In unsupervised learning, data are compared based on a similarity scale to be categorized in groups. This is widely used for extracting generative features, identifying meaningful trends and structures, groupings in results, and exploratory purposes. The most common unsupervised learning tasks are clustering, density estimation, feature learning, dimensionality reduction, finding association rules, anomaly detection, etc. **Figure 19** demonstrates an Unsupervised Learning technique.

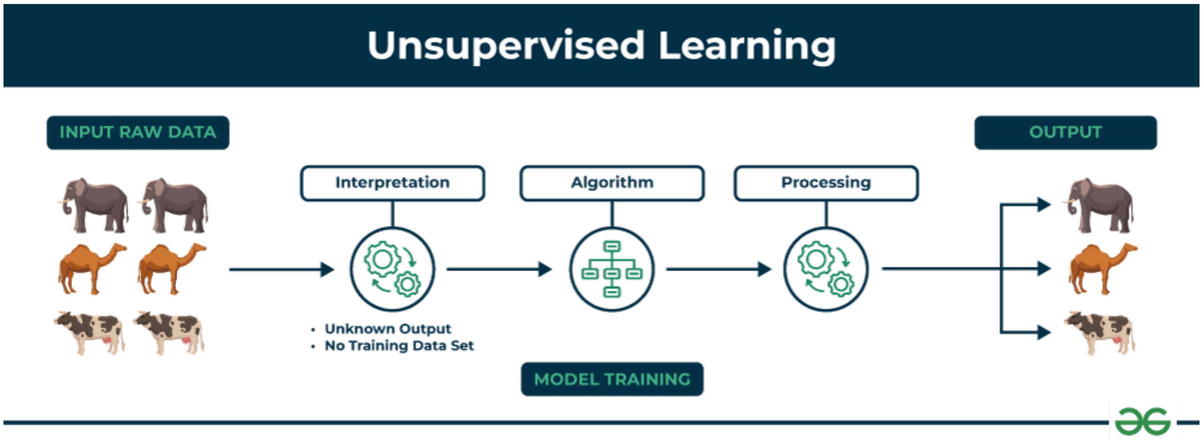


Figure 19: Unsupervised Learning

In the following, we introduce some unsupervised learning methods-

K-Means Clustering: K-means clustering [80] is a fast, robust, and simple algorithm that provides reliable results when data sets are well-separated from each other. The purpose of K-means is to group n data samples to k clusters, so that each cluster is known based on its center. This method is an iteration-based technique [81]. Initially, k random cluster centers are considered, and all data points are linked to the closest cluster center. When clusters are established, so that all the data points in the database belong to one of the clusters, a new center will be re-calculated in each cluster. This means that cluster centers are updated in each iteration. This algorithm is repeated until any cluster center does not change.

Hierarchical Clustering: This clustering scheme aims to group data points to clusters, so that cluster members (data points in a cluster) have the highest similarity to each other compared to data points in other clusters [81]. This process is carried out based on two techniques: top-to-down (Divisive clustering) and bottom-to-up (agglomerative clustering). In the divisive clustering, all data points are first placed in one group. Then, this group is divided into smaller groups. This process continues until each sample is placed in one group. In the agglomerative clustering, each sample is first placed in a cluster. Then, similar groups are merged to establish larger groups. This process continues until all data points are placed in one group. In the hierarchical clustering method, we need no previous information about the number of clusters. This scheme is simply implemented.

Fuzzy-C-means (FCM): It is a clustering method based on fuzzy logic. In this method, each sample can be in one or more clusters [81]. FCM determines clusters based on different similarity scales such as distance. Note that one or more similarity scales may be used in the clustering process and this issue depends on application or the dataset. The clustering process is repeated to find the best cluster centers. Like the K-means clustering method, FCM must be aware of the number of clusters.

In the following **Table-2** we summarize the advantages and disadvantages of the above-mentioned supervised methods.

Table 2: Advantages and Disadvantages of Unsupervised Learning

Algorithm	Advantages	Disadvanatges
K-means clustering	High clustering speed, suitable for small and large databases, easy understanding	Sensitivity to noisy data, low accuracy, requiring primary knowledge about the number of clusters
Hierarchical clustering	High accuracy & clustering speed, low sensitivity to noisy data, no need of primary knowledge about the number of clusters, easy implementation	Weak performance for large and small databases
Fuzzy-c-means (FCM)	Low sensitivity to noisy data, high accuracy	Requires primary knowledge about the number of clusters, low accuracy

Semi-Supervised Learning: Semi-supervised learning can be defined as a hybridization of the above-mentioned supervised and unsupervised methods, as it operates on both labeled and unlabeled data [24, 76]. Thus, it falls between learning “without supervision” and learning “with supervision”. In the real world, labeled data could be rare in several contexts, and unlabeled data are numerous, where semi-supervised learning is useful [61]. In this learning method, both labeled and unlabeled datasets are used in the learning process. Therefore, this technique requires a supervised learning algorithm to be trained on a labeled training set. Moreover, an unsupervised learning algorithm should be used to produce data samples with new labels [82,83]. These data samples are added to the labeled training set for the supervised learning algorithm.

Figure 20 shows the general framework of Semi-supervised learning.

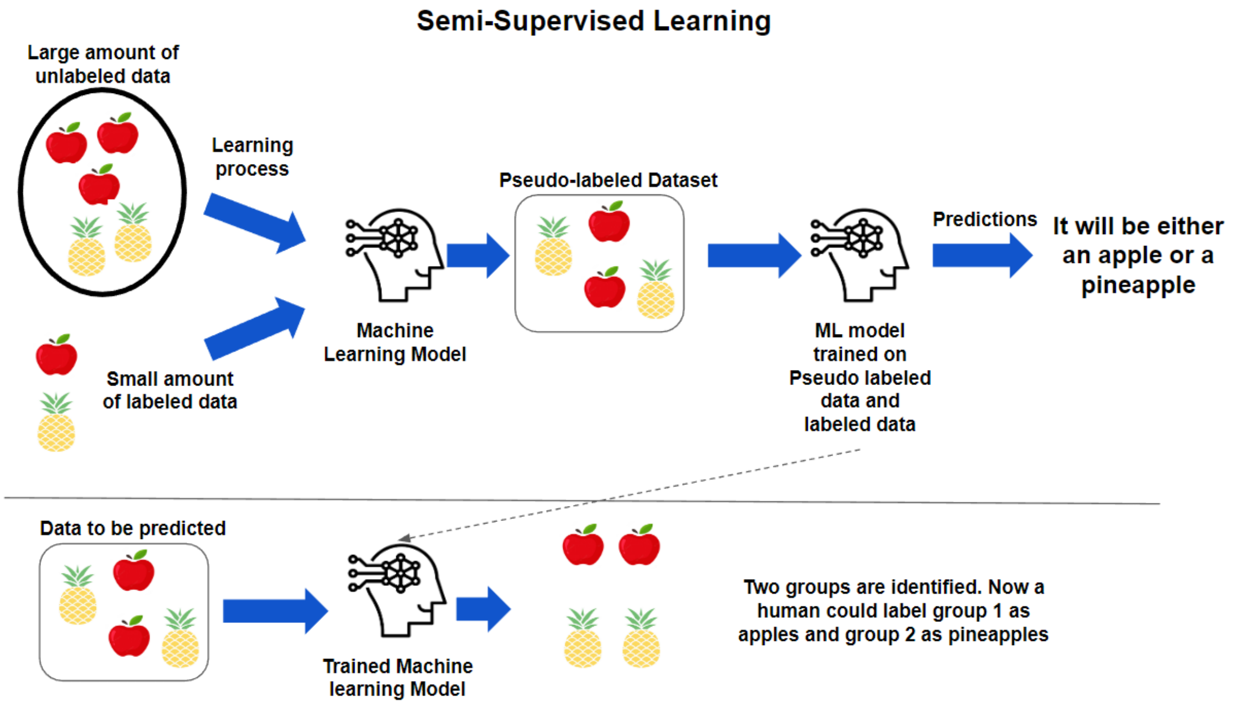

Semi-Supervised Learning
Large amount of unlabeled data
Learning process
Small amount of labeled data
Machine Learning Model
Pseudo-labeled Dataset
ML model trained on Pseudo labeled data and labeled data
Predictions
It will be either an apple or a pineapple
Data to be predicted
Trained Machine learning Model
Two groups are identified. Now a human could label group 1 as apples and group 2 as pineapples

Figure 20: Semi-Supervised Learning

Reinforcement Learning: Reinforcement learning is a type of machine learning algorithm that enables software agents and machines to automatically evaluate the optimal behavior in a particular context or environment to improve its efficiency [84], i.e., an environment-driven approach. A reinforcement learning-based model learns continuously through interaction with the environment and collects information to perform its activity [85].

Following **Figure 21** shows Reinforced Learning.

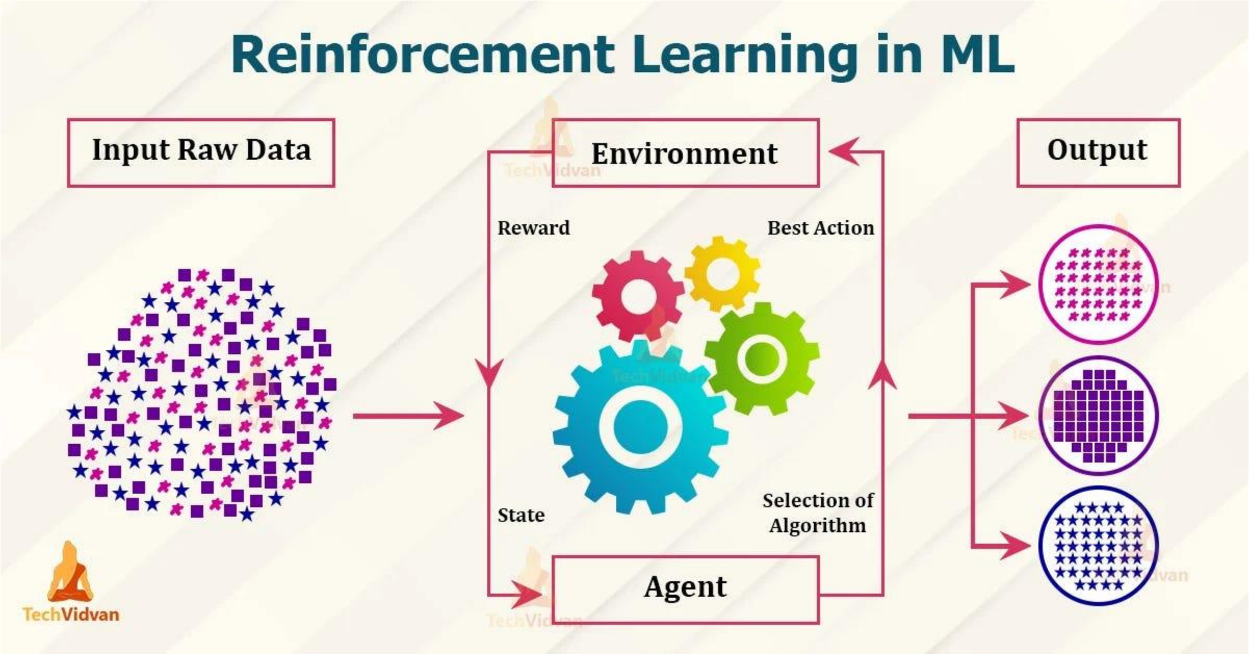


Figure 21: Reinforced Learning

In the following, we introduce the most important reinforcement learning methods.

Monte Carlo (MC) Methods: MC is an incremental episode-by-episode scheme [86,87]. MC-based methods are free-model which means that they do not require the complete environment model and learn based on experiences (i.e., they learn using interactions with the environment). MC can solve the reinforcement learning problem by averaging sample returns. Monte Carlo (MC) methods guarantee that appropriate sample returns are available because they are often used for episodic tasks which means that an experience must be divided into episodes. Ultimately, an action is selected, and all episodes will also stop. After an episode is terminated, values and policies are updated.

Q-Learning: It is known as an appropriate and popular algorithm in reinforcement learning. Q-Learning helps an agent to learn its best actions. In this method, there is a table called Q-Table. This table maintains action-state pairs and the corresponding values. In fact, action-state pairs are known as inputs in this table and the Q-value is its output. In Q-learning, the purpose is to maximize the Q-value [86,87].

Deep Reinforcement Learning (DRL): It is a combination of deep learning and reinforcement learning. This scheme can be used to solve many complex issues. It helps the agents to become more intelligent. This improves their ability to optimize the policy. Reinforcement learning is a machine learning technique, which can operate without any database. Therefore, in DRL, agents can first produce the dataset through interaction with the environment. Then, this database is used to train deep networks in DRL [86,87].

In the following **Table-3** we summarize the advantages and disadvantages of the above-mentioned semi- supervised methods.

Table 3: Advantages and Disadvantages of Semi-Supervised Methods

Algorithm	Advantages	Disadvantages
Monte Carlo (MC) methods	A free-model scheme	High variance of returns, low convergence speed, trapping in local optimism
Q-Learning	A free-model, off-policy, and forward learning scheme	No generalization capability, inability to predict the optimal amount for not observed situations.
Deep reinforcement learning (DRL)	Suitable for issues with high dimensions, the ability to approximate the unobserved situations, generalizability	Unstable model, rapid changes in the policy with a slight change in Q-Value

Types of Evaluation Methods

ML-based methods in healthcare or medicine are divided into two main categories based on evaluation schemes: simulation-based evaluation and practical implementation-based evaluation.

Simulation-Based Evaluation: Most ML-based models designed in healthcare use simulation tools to evaluate their performance because they are more available than practical implementation. To evaluate ML-based models, it is necessary to simulate this learning model using suitable simulation tools such as MATLAB, SAS, and R to determine its efficiency. We evaluate these learning models based on various evaluation scales. In general, evaluation criteria are divided into two main categories:

Discrimination Scales: These scales analyze the ability of an ML-based model for ranking or distinguishing between two classes. The most important discrimination scales are ROC, AU-ROC, F1-Score, Sensitivity, and Specificity. We introduce these scales in Section 3.

Calibration Scales: These scales determine how many predicted outcomes match actual outcomes. In the real world, these scales are very important because these scales analyze the expected profits or losses. For example, if the death risk caused by surgery is more than the death risk without surgery, the surgeon may not perform this surgery and abandon it.

Practical Implementation-Based Evaluation: ML-based models in medicine should be evaluated using their practical implementation because it allows us to analyze learning models in real environments. In practical implementation, we must evaluate the learning model in a real-time manner and continuously update this model and re-validate it. Some important scales during the practical implementation of learning models in healthcare include their generalizability for new data, user feedback, comparing model performance with an expert in the relevant area, and comparing model performance with other existing models.

Applications

ML-based methods in healthcare or medicine are divided into two main categories based on application: diagnosis and treatment.

Diagnosis: Machine learning can be used in this area to help physicians detect disease in the early stages and reduce the detection time. For example, machine learning can be used for improving medical images, analyzing laboratory results, detecting disease, identifying the degree of disease etc.

Treatment: Some ML-based methods can help with the treatment of diseases. For example, machine learning can be used to diagnose suitable doses, monitor the treatment procedure, and predict the progression of the disease. These methods reduce treatment costs, reduce costs related to drug production, improve the treatment procedure, save time to discover appropriate drugs.

APPLICATION OF ML-BASED TECHNIQUES IN HEALTHCARE

In this section we investigate some ML-based methods in medicine or healthcare.

Wu et al. presented a predictive model of heart failure using ML techniques [88]. The authors used three ML methods for prediction, logistic regression, support vector machine (SVM), and boosting. Comparative analysis was performed to investigate the performance of each method using 10-fold cross-validation. The aim of the paper was to provide the right diagnosis of heart failure at least six months before it occurs. Menden et al. developed ML models, particularly ANNs, to calculate the reaction of cancer cell lines to medical treatment, which measured throughout IC50 values [89]. Based on their study, the potential efficacy of thousands of drugs can be tested through silico, as anti-tumor agents depending on their formation.

Borisov et al. utilized three ML algorithms, namely, SVM, binary tree (BT), and random forest (RF), to predict the clinical effectiveness of cancer drugs by transferring attributes attained using the expression-based data from cell lines [90]. The aim of the paper was to present a suitable method for drug scoring and/or tailored medication and the algorithms were tested on different datasets of cancer-like diseases. Fakoor et al. used unsupervised (PCA) and deep learning methods (SoftMax Regression) on gene expression data to cope with the challenges of feature dimensionality and improve the diagnosis and classification of cancer types [91].

In study [92], four publicly accessible datasets were processed (i.e., datasets from the Dana-Farber Cancer Institute, the University of Michigan, the University of Toronto, and Brigham and Women's Hospital respectively). The k-nearest neighbor technique, naive Bayes with the assumption of a couple of normal attribute distribution, and distribution through histograms, SVM, and decision tree were used. The performance of ML techniques was assessed, and SVM showed the best results among all datasets. Akay, M. F. proposed breast cancer analysis based on the SVM combined with feature selection technique [93]. Experiments were carried out on different common datasets, including the Wisconsin breast cancer dataset (WBCD). Specificity, sensitivity, classification performance, positive, negative predictive values, and receiver running characteristic confusion matrix and curves are used to evaluate the efficacy of the proposed algorithm. The results revealed that the greatest classification accuracy is 99.5%, which is obtained from the SVM model containing five selected features.

In the study [94], using feature selection technique, the algorithms supplied data with normal dimensionality and provided precise results. In this paper, experiments were conducted using four distinct feature selection techniques and four classifiers on four datasets. Artificial NNs increase the classification efficiency of breast cancer when utilizing feature selection. Study [95] demonstrated the effectiveness of various statistical and ML techniques that were employed for the assessment of missing data. Imputation methods based on statistical techniques, e.g., mean, hot-deck and multiple imputation, and machine learning techniques, e.g., multi-layer perceptron (MLP), self-organisation maps (SOM) and k-nearest neighbour (KNN), were applied to data collected through the “El Álamo-I” project, and the results were then compared to those obtained from the listwise deletion (LD) imputation method. The accuracies of predictions on early cancer relapse were measured using artificial neural networks (ANNs), in which different ANNs were estimated using the data sets with imputed missing values. Study [96] shows microarray breast cancer data were utilized to classify the cases that applied ML systems. First, 8 different machine learning algorithms are applied to the data, without applying any feature selection methods. Then two different feature selection methods are applied. The results of the classifications are compared with each other and with the results of the first case. The methods applied are SVM, KNN, MLP, Decision Trees, Random Forest, Logistic Regression, Adaboost and Gradient Boosting Machines. After applying the two different feature selection methods with the best 50 features applied, SVM gave the best results.

Another prostate cancer diagnosing method using ML techniques was proposed by Hussain et al. [97]. Multi-ML techniques, such as SVM and Bayesian, were used to efficiently diagnose prostate cancer. Some feature extraction methods were also used for further efficiency enhancement. Chan et al. proposed a method combining feature selection and ML techniques to diagnose oral cancer [98]. Five feature selection methods have been proposed and experimented on the oral cancer prognosis dataset. In the second stage, the model with the features selected from each feature selection method is tested on the proposed four types of classifiers; these are namely, ANFIS, artificial neural network, support vector machine and logistic regression. A k-fold cross-validation is implemented on all types of classifiers due to the small sample size.

Fuery et al. [99] utilized SVM to analyze gene expressions in finding and classifying harmful tissues. Cho and Won [100] investigated studies that aimed to assess cancer classification and feature selection methods. In this study many features and classifiers were explored using three benchmark datasets to systematically evaluate the performances of the feature selection methods and machine learning classifiers. Three benchmark datasets are Leukemia cancer dataset, Colon cancer dataset and Lymphoma cancer data set. Pearson's and Spearman's correlation coefficients, Euclidean distance, cosine coefficient, information gain, mutual information and signal to noise ratio have been used for feature selection. Multi-layer perceptron, k-nearest neighbor, support vector machine and structure adaptive self–organizing map have been used for classification. Also, we have combined the classifiers to improve the performance of classification.

Cancer classification based on gene expression data using ML techniques has attracted much attention recently as a promising research field [101, 102, 103, 104].

Tan and Gilbert proposed the use of supervised ML techniques in correctly classifying cancerous and normal tissues from the gene expression profiles [105]. Classification tasks were performed using the C4.5 decision tree, after which they bagged and boosted decision trees on seven publicly available cancerous microarray. Guyon et al. used SVM based on recursive feature elimination in order to address the problem of selecting a small subset of genes, recorded on DNA micro-arrays, from broad patterns of gene expression data [106].

Jin et al. dealt with the high-dimensional problem using the Chi-square method for tag selection of the serial analysis of gene expression before classifying binary and multicategory cancer types [107]. Five different ML algorithms (C4.5, SVM, nearest neighbor, naive Bayes, and RIPPER) were used for classifying cancer types. Wang et al. proposed the use of a set of feature selection algorithms, namely, wrappers, correlation-based feature selection, and filters together with ML algorithms, such as naive Bayes, decision trees, and SVM, for the extraction of significant information in microarray data analysis [108].

Chen et al. demonstrated a novel supervised ML model based on Monte Carlo methods, local field, and SVM theory. The proposed model was applied to accurately find patterns in high-dimensional gene datasets of colon cancer [109]. Wang et al. proposed an ML-based model called SRL-RNN which uses reinforcement learning and recurrent neural network (RNN) [110].

The purpose of SRL-RNN is to solve the dynamic treatment regime (DTR) problem. Zhu et al. presented a semi-supervised learning method called TE-DLSTM to identify body activities using inertial sensors [111]. This method uses a deep long short-term network (DLSTM) to extract high-level features.

Zhai et al. [112] suggested a semi-supervised learning system using a two-dimensional convolutional neural network (CNN) to classify electrocardiogram (ECG). This learning issue classifies time series signals with unbalanced classes: normal beats, supraventricular ectopic beats (SVEB), and ventricular ectopic beats (VEB). The purpose of this scheme is to diagnose SVEB and VEB without labeling ECG data.

According to the World Health Organization (WHO) Coronavirus disease (COVID-19) is an infectious disease [113]. Recently, the ML learning techniques have become popular in the battle against COVID-19 [114, 115]. For the COVID-19 pandemic, the learning techniques are used to classify patients at high risk, their mortality rate, and other anomalies [116]. Studies [117, 118] show that ML techniques can also be used to better understand the virus's origin, COVID-19 outbreak prediction, as well as for disease diagnosis and treatment. Deep learning is also seen as a crucial technique for potential applications, particularly for COVID-19 pandemic [119, 120, 121].

CONCLUSION

In this section, we will conclude the role of machine learning in healthcare by mentioning few key takeways:

- ML models can analyze large amounts of data and identify patterns and predictions that would be difficult or impossible for human analysts to detect.
- Machine learning has the potential to revolutionize the field of healthcare by enabling more accurate predictive analytics and diagnosis.
- Improving Data Quality: Data cleaning and preprocessing techniques to ensure the data used for machine learning is accurate, complete, and usable. Deploy suitable methods to handle missing values.
- Tackling Complex and High-dimensional Data: Healthcare data, such as medical images and time-series data, can be complicated and high-dimensional. Using various methods of ML to extract the significant features.
- Interpretable Models: Use interpretable models and visualization tools to help healthcare providers understand the predictions being made by machine learning models.
- Model Validation and Testing: Validate and test machine learning models to ensure they are accurate and reliable.
- Building a Multidisciplinary Team: Assemble a team with expertise in both the healthcare domain and machine learning to overcome the technical challenges of using machine learning in healthcare.

Thus, effectively processing the data and handling the diverse learning algorithms are important, for a machine learning-based solution and eventually building intelligent applications.

PREDICTIVE ANALYTICS

Under this section we present four different case studies of healthcare where machine learning has been applied with great effect.

Case Study- 1

CLASSIFICATION OF ERYTHEMETASQUAMOUS DERMATOSIS BY THE METHOD OF RANFDOM FOREST

Dwaipayan Mukhopadhyay, Dieudonne J. Phanord, Rohan J. Dalpatadu, Laxmi P. Gewali and Ashok K. Singh.

DOI:10.47363/JDMRS/2023(4)143

Introduction: Machine Learning (ML) methods have found wide applications in dermatology [1](Chan et al., 2020). Thomsen, Iversen, Titlestad & Winther [2] (2020) reviewed 2175 publications and found that the most common usage of ML methods was in the binary classification of malignant melanoma from images. Adamson and Smith [3] have a word of advice about usage of ML methods in diagnosis of skin diseases that inclusivity must be kept in mind for classification results to be accurate. Steele et al. [4] searched PubMed, Embase, and CENTRAL, and found that the performance of ML methods was variable, and overall accuracy measure was not a good measure for sub-group accuracy.

Erythematosquamous Dermatosis has symptoms of itchy skin or pruritus; possible causes for pruritus include an underlying medical condition, contact with an irritant or a reaction to a medication [5]. An important gene associated with this skin disease [6] (MalaCards) is IL22RA1 (Interleukin 22 Receptor Subunit Alpha 1). Back in 1998, Guvenir, Demiroz, and Ilter [7] introduced a new classifier called Voting Feature Intervals (VFI) in which each feature voted on a class, with the class getting the most votes declared as the predicted class value. The overall accuracy of VFI was reported to be 99.2%. Data for this article is given in Dua and Graff [8]. Singh, Sinha and Yadav [9] used logistic regression, support vector machine and K-Nearest neighbor classifiers on this dataset and computed accuracy measures. Rathore et el. [10] (2022) used the XGBoost model on this dataset. We will use this data set and apply the method of random forest which uses a bagging algorithm: a Random Forest (RF) model randomly selects features to use from the set of all features, grows a large number of trees, then uses majority vote to classify the response for each observation. We will also compute the accuracy-based importance measure for each feature, and then fit a reduced RF model using most important features.

Data: The dataset [8] has 366 subjects with measurements on 34 features and a categorical response variable. One of these features is age, a continuous variable; family. history is binary, eosinophils.infiltrate is (0,1,2) ordinal, and the remaining 31 features are all (0,1,2,3) ordinal. The response variable is categorical with 6 levels:

C1: psoriasis

C2: seboreic dermatitis

C3: lichen planus

C4: pityriasis rosea

C5: cronic dermatitis

C6: pityriasis rubra pilaris

There are 8 missing values in the Age column, which are all removed yielding the final dataset of 358 observations on 35 variables. This dataset is split into a 75% training set of 269 rows and a 25% test set of 89 rows; RF model is then fit to the training set, and accuracies are computed for both training and test sets separately.

Random Forest Classifier Method: The method of random forest is a decision-trees based supervised learning method for categorical or continuous response variable Y. It randomly selects a subset of observations and a subset of features at a time to fit a large number of decision trees to predict Y and then averages (mode for classification, mean for regression) these predicted Y values for the predicted Y. Random forest is one of the most accurate predictive methods [11] (Hastie, Tibshirani & Friedman, 2009) and it reduces overfitting [12] (Schonlau, M., & Zou, R. Y., 2020). All computations and graphs are done in the statistical software environment R [13]. Even though the R package [14] randomForest yields out of bag (OOB) accuracy, we report all accuracy measures for multi-class classification for both training and test datasets.

Accuracy Measures for Multi-Class Classification: Commonly used measures for multi-level classifiers (accuracy, precision, recall and F1 [15] are briefly described. These measures are calculated from the confusion matrix shown in Table 3, where (Ci,j = number of times true response of j get predicted as i; i, j = 1, 2, …, 6).

Table 4: Confusion Matrix for the 6-level Class

	OBSERVED					
PREDICTED	C_1	C_2	C_3	C_4	C_5	C_6
C_1	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$	$C_{1,4}$	$C_{1,5}$	$C_{1,6}$
C_2	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$	$C_{2,4}$	$C_{2,5}$	$C_{2,6}$
C_3	$C_{3,1}$	$C_{3,2}$	$C_{3,3}$	$C_{3,4}$	$C_{3,3}$	$C_{3,4}$
C_4	$C_{4,1}$	$C_{4,2}$	$C_{4,3}$	$C_{4,4}$	$C_{4,3}$	$C_{4,4}$
C_5	$C_{5,1}$	$C_{5,2}$	$C_{5,3}$	$C_{5,,4}$	$C_{5,5}$	$C_{5,6}$
C_6	$C_{6,1}$	$C_{6,2}$	$C_{6,3}$	$C_{6,,4}$	$C_{6,5}$	$C_{6\ ,6}$

The one vs all binary performance measures accuracy, precision, recall, F1 and the overall prediction accuracy [15, 16] are calculated from the following formulas:

$$\text{Overall prediction accuracy} = \frac{\sum_{j=1}^{6} C_{j,j}}{\sum_{i=1}^{6}\sum_{j=1}^{6} C_{j,j}} = \frac{\text{Sum of diagonal elements of the confusion matrix}}{\text{Sum of all elements of the confusion matrix}}$$

$$\text{Precision}_j \; = \; \frac{C_{j,j}}{\sum_{k=1}^{6} C_{j,k}} \; = \; \frac{j-th \text{ diagonal element of the confusion matrix}}{\text{Sum of j-th row of the confusion matrix}}$$

$$\text{Recall}_\text{j} \; = \; \frac{C_{j,j}}{\sum_{k=1}^{6} C_{k,j}} = \frac{j-th \text{ diagonal element of the confusion matrix}}{\text{Sum of j-th column of the confusion matrix}}$$

$$\text{F1}_j \; = \; \frac{2 \times \text{Precision}_j \times \text{Recall}_j}{(\text{Precision}_j + \text{Recall}_j)} = \text{the harmonic mean of Precision and Recall for class } j$$

Figure 22 : One vs All Binary Performance Measures

The Area Under the Curve (AUC) for each class is given by [15] (Molin et al 2021):

$$AUC_j = \frac{1}{2}\left(\frac{TP_j}{TP_j + FN_j} + \frac{TN_j}{TN_j + FP_j} \right), \; j\text{=1,2,...6.}$$

Figure 23: AUC Curve Formulae

where TPj = true positive, TNj = true negative, FPj = false positive and FNj = false negative for the j-th class, shown below as elements of the confusion matrix CMj for the j-th class (j = 1, …,6):

For multi-class classification problems, macro- and micro-averages of the above measures [15-17] are also included.

$$recall_{macro} = \frac{\sum_{j=1}^{6} recall_j}{6}$$

$$precision_{macro} = \frac{\sum_{j=1}^{6} precision_j}{6}$$

$$recall_{micro} = \frac{\sum_{j=1}^{6} TP_j}{\sum_{j=1}^{6} TP_j + \sum_{j=1}^{6} FN_j}$$

$$precision_{micro} = \frac{\sum_{j=1}^{6} TP_j}{\sum_{j=1}^{6} TP_j + \sum_{j=1}^{6} FN_j}$$

$$AUC_{macro} = \frac{\sum_{j=1}^{6} AUC_j}{6}$$

Figure 24: Performance Measures for Multi-Class Classification

There is no well-accepted multi-class Receiver Operating Characteristic Analysis [15, 17] and hence micro-averaged AUC's are not computed.

ML literature recommends splitting the original dataset into a training set and a test set [16] and reporting all performance metrics for both training and test sets.

Results: The package random Forest was used with 250 trees (parameter ntree = 250) to fit a random forest classifier to the training data set using all 34 features. Variable importance for each feature was computed using decrease in accuracy as the measure of feature importance. The RF model fitted to the training set was then used to predict the response for the test set, Tables 4 and 5 show the accuracy measures of the RF classifier for the training and test datasets, respectively.

Table 5: Confusion Matrix and Accuracy Measure of the RF Using All Features for the Training Set

Observed

Predicted	1	2	3	4	5	6	Recall	Precision	F1	AUC
1	84	0	0	0	0	0	1	1	1	1
2	0	38	0	0	0	0	1	1	1	1
3	0	0	57	0	0	0	1	1	1	1
4	0	0	0	35	0	0	1	1	1	1
5	0	0	0	0	38	0	1	1	1	1
6	0	0	0	0	0	17	1	1	1	1
Macro average							1	1	1	1
Micro average							1	1	1	

Table 6: Confusion Matrix and Accuracy Measure of the RF Using All Features for the Training Set

Observed

Predicted	1	2	3	4	5	6	Recall	Precision	F1	AUC
1	27	0	0	0	0	0	1.00	1.00	1.00	1.00
2	0	20	0	1	0	0	0.91	0.95	0.93	0.95
3	0	0	14	0	0	0	1.00	1.00	1.00	1.00
4	0	2	0	12	0	0	0.92	0.86	0.89	0.95
5	0	0	0	0	10	0	1.00	1.00	1.00	1.00
6	0	0	0	0	0	3	1.00	1.00	1.00	1.00
Macro average							1.00	0.95	1.00	0.98
Micro average							0.95	1.00	1.00	

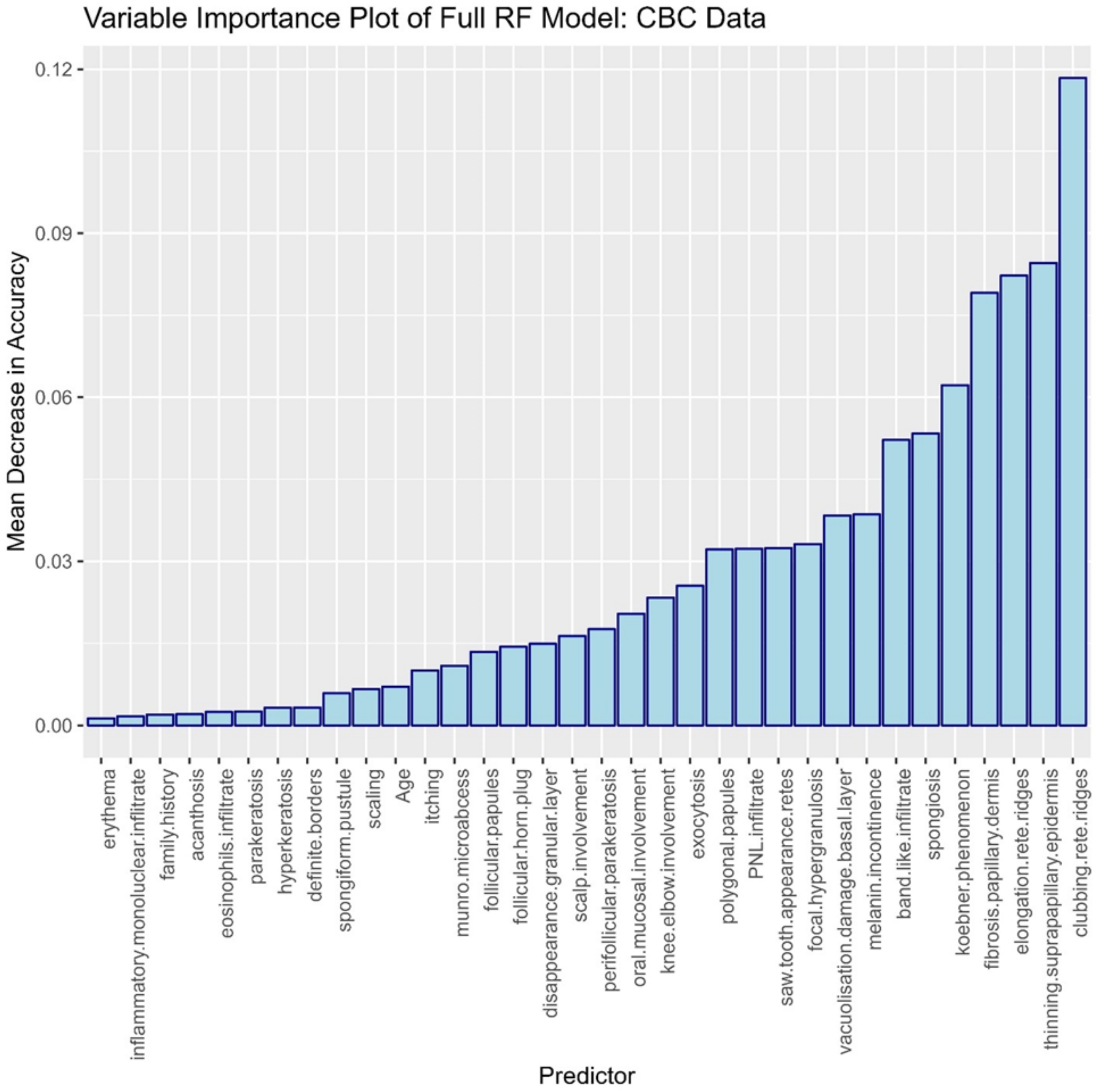


Figure 25: Variable Importance Plot for the Full RF Model

Figure 25, the variable importance plot of the Full RF model (i.e., the RF model with all features in the model), shows the mean decrease in prediction accuracy for each feature if the feature is removed from the model.

We next drop the bottom 17 features and fit the Reduced RF Model. Tables 6 and 7 show the confusion matrix and the prediction accuracy of the reduced RF classifier.

Table 7: Confusion Matrix and the Prediction Accuracy of the RF Classifier Using Best 17 Features for the Training Set

	Observed									
Predicted	1	2	3	4	5	6	Recall	Precision	F1	AUC
1	84	0	0	0	0	0	1	1	1	1
2	0	38	0	5	0	0	1	0.88	0.94	1
3	0	0	57	0	0	0	1	1	1	1
4	0	0	0	30	0	0	0.86	1	0.92	1
5	0	0	0	0	38	0	1	1	1	1
6	0	0	0	0	0	17	1	1	1	1
Macro average							0.97	0.96	0.97	1.00
Micro average							0.98	0.98	0.98	

Table 8: Confusion Matrix and the Prediction Accuracy of the RF Classifier Using Best 17 Features for the Test Set

	Observed									
Predicted	1	2	3	4	5	6	Recall	Precision	F1	AUC
1	27	0	0	0	0	0	1.00	1.00	1.00	1.00
2	0	20	0	1	0	0	0.91	0.95	0.93	0.95
3	0	0	14	0	0	0	1.00	1.00	1.00	1.00
4	0	2	0	12	0	0	0.92	0.86	0.89	0.95
5	0	0	0	0	10	0	1.00	1.00	1.00	1.00
6	0	0	0	0	0	3	1.00	1.00	1.00	1.00
Macro average							0.95	0.93	0.94	0.98
Micro average							0.95	0.95	0.95	

It can be seen from Tables 6 and 7 that for both training and test datasets, the RF classifier based on top 17 features is quite accurate.

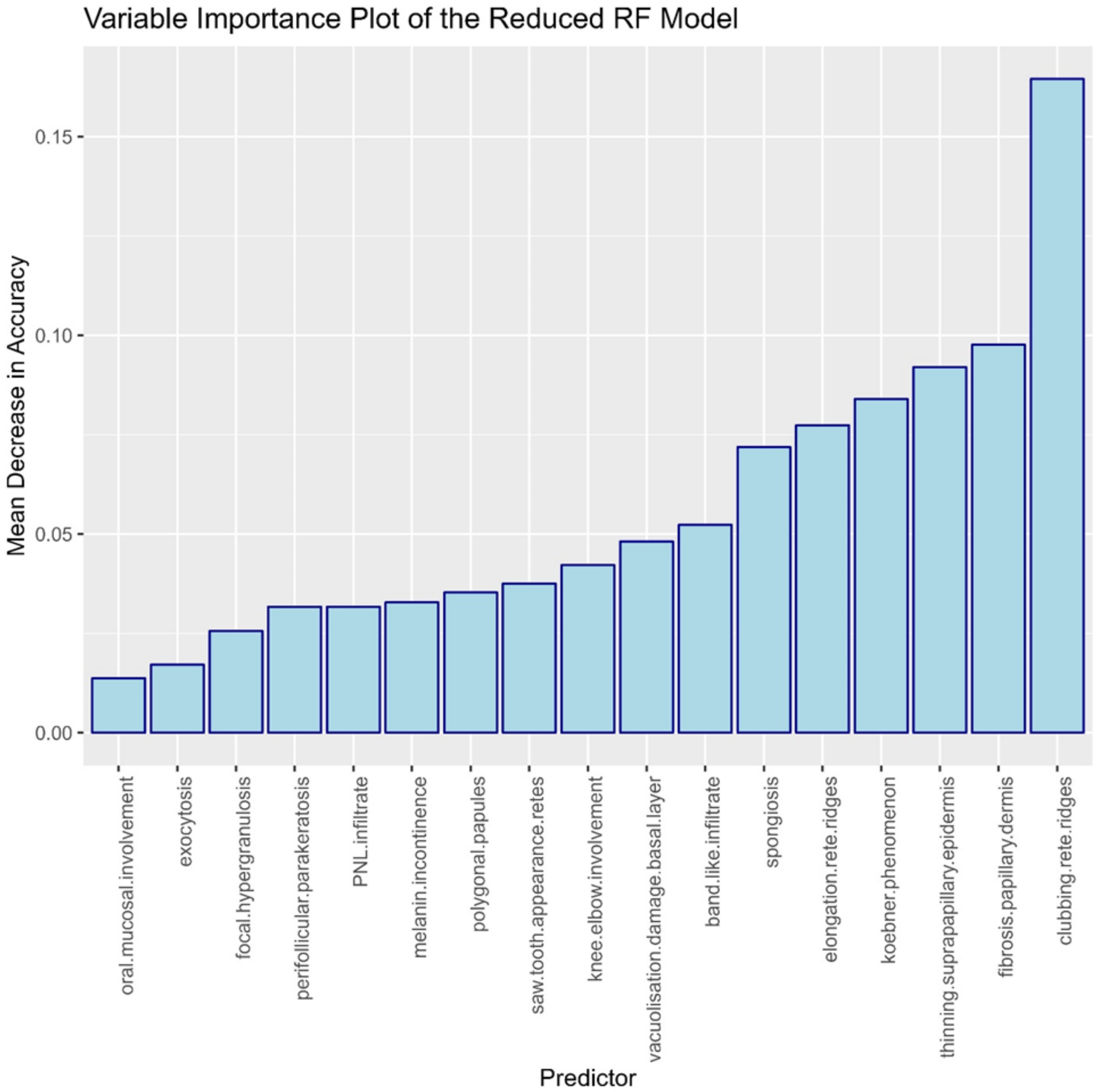


Figure 26: Variable Importance Plot for the Reduced RF Model

Conclusions: We have successfully demonstrated that the method of RF classifier is able to classify Erythematosquamous Dermatosis with high accuracy.

References:

[1] Chan, S., Reddy, V., Myers, B., Thibodeaux, Q., Brownstone, N., Liao, W.(2020). Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations. Dermatol Ther (Heidelb) (2020) 10:365–386 .https://doi.org/10.1007/s13555-020-00372-0

[2] Kenneth Thomsen, Lars Iversen, Therese Louise Titlestad & Ole Winther (2020). Systematic review of machine learning for diagnosis and prognosis in dermatology, Journal of Dermatological Treatment, 31:5, 496-510, DOI: 10.1080/09546634.2019.1682500

[3] Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 2018;154(11):1247–1248. doi:10.1001/jamadermatol.2018.2348

[4] Steele, L., Tan, X., Olabi, B., Gao, J., Tanaka, R. and Williams, H. (2022), Determining the clinical applicability of machine learning models through assessment of reporting across skin phototypes and rarer skin cancer types: a systematic review. J Eur Acad Dermatol Venereol. Accepted Author Manuscript. https://doi.org/10.1111/jdv.18814

[5] Cleveland Clinic (2022) https://my.clevelandclinic.org/health/diseases/11879-pruritus

[6]MalaCards.https://www.malacards.org/card/erythematosquamous_dermatosis

[7] Guvenir, H.A., Demiroz, G., Ilter, N. (1998). Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals, Artificial Intelligence in Medicine 13 (1998) 147–165.

[8] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[9] S. K. Singh, A. Sinha and S. Yadav, "Performance Analysis of Machine Learning Algorithms for Erythemato-Squamous Diseases Classification," 2022 IEEE International Conference on

[10] Abhishek Singh Rathore, Siddhartha Kumar Arjaria, Manish Gupta, Gyanendra Chaubey, Amit Kumar Mishra & Vikram Rajpoot (2022). Erythemato-Squamous Diseases Prediction and Interpretation Using Explainable AI, IETE Journal of Research, DOI: 10.1080/03772063.2022.21149537

[11] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Springer series in statistics, The elements of statistical learning: Data mining, inference, and prediction (2nd ed.) pp. 587–590.

[12] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3–29. https://doi.org/10.1177/1536867X20909688

[13] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[14] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[15] Molin, Nicole, Molin, Clifford, Dalpatadu, R.J., Singh, A. K. (2021). Prediction of Obstructive Sleep Apnea Using FFT of Overnight Breath Recordings (2021). Machine Learning with Applications, Volume 4, 15 June 2021, 100022 https://www.sciencedirect.com/science/article/pii/S2666827021000037

[16] Tutz, G. (2011). Regression for categorical data (pp. 210–214). Cambridge University Press.

[17] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45, 427–437.

Case Study- 2

A COMPARATIVE STUDY OF ACUTE ALCOHOLIC HEPATITIS VS. NON-ALCOHOLIC HEPATITIS PATIENTS FROM A COHORT WITH CHRONIC ALCOHOL DEPENDENCE

Kyaw Min Tun , Zahra Dossaji , Blaine L. Massey , Kavita Batra, Chun-Han Lo , Yassin Naga , Salman Mohammed , Abebe Muraga , Ahmad Gill , Dwaipayan Mukhopadhyay , Ashok Singh , Daisy Lankarani , Jose Aponte-Pieras and Gordon Ohning

https://doi.org/10.3390/genes14040780

Abstract: The rate of alcoholic hepatitis (AH) has risen in recent years. AH can cause as much as 40–50% mortality in severe cases. Successful abstinence has been the only therapy associated with long-term survival in patients with AH. Thus, it is crucial to be able to identify at-risk individuals in order to implement preventative measures. From the patient database, adult patients (age 18 and above) with AH were identified using the ICD-10 classification from

November 2017 to October 2019. Liver biopsies are not routinely performed at our institution. Therefore, patients were diag-nosed with AH based on clinical parameters and were divided into “probable” and “possible” AH. Logistic regression analysis was performed to determine risk factors associated with AH. A sub-analysis was performed to determine variables associated with mortality in AH patients. Among the 192 patients with alcohol dependence, there were 100 patients with AH and 92 patients without AH. The mean age was 49.3 years in the AH cohort, compared to 54.5 years in the non-AH cohort. Binge drinking (OR 2.698; 95% CI 1.079, 6.745; p = 0.03), heavy drinking (OR 3.169; 95% CI 1.348, 7.452; p = 0.01), and the presence of cirrhosis (OR 3.392; 95% CI 1.306, 8.811; p = 0.01) were identified as characteristics more commonly found in the AH cohort. Further, a higher inpatient mortality was seen in those with a probable AH diagnosis (OR 6.79; 95% CI 1.38, 44.9; p = 0.03) and hypertension (OR 6.51; 95% CI 9.49, 35.7; p = 0.02). A higher incidence of mortality was also noted among the non-Caucasian race (OR 2.72; 95% CI 4.92; 22.3; p = 0.29). A higher mortality rate despite a lower incidence of alcohol use among non-Caucasian patients may indicate healthcare disparities.

Keywords: alcohol use; alcoholic hepatitis; prevention

Introduction: Alcohol use is responsible globally for approximately 3 million deaths per year, ac-counting for 5.3% of all deaths, according to a 2022 report from the World Health Organization (WHO) [1]. Excessive alcohol use is also the third leading cause of preventable deaths in the United States (U.S.) and contributes to the development of acute alcoholic hepatitis (AH) [2]. Furthermore, it has been estimated that 10 to 15% of patients in the U.S. who chronically consume alcohol develop alcohol-associated liver disease (ALD) [2]. AH is a manifestation of ALD, a spectrum of liver injury that begins with steatosis and can potentially progress to acute alcoholic hepatitis, alcohol-associated cirrhosis, and AH with acute or acute-on-chronic liver failure [3].

AH is a clinical syndrome with a hallmark presentation that includes rapid onset of jaundice, hepatomegaly, ascites, encephalopathy, and generalized signs or symptoms in-cluding fever, abdominal pain, or muscle wasting [3,4]. However, individuals with AH may also present with only mild symptoms or non-specific laboratory abnormalities. Therefore, determining the incidence of AH can be an obstacle in part due to diagnostic challenges. Other factors, such as comorbidities and improper ICD (international classifi-cation of diseases) coding, can further undermine the accuracy of the AH incidence rate [3].

Nonetheless, the rate of AH has risen in recent years, particularly in the relatively younger population with an average age of 53 years [3,4]. In patients with severe AH, which can be determined by a Maddrey's discriminant function (MDF) value greater than 32, six-month mortality can be as high as 40% [5]. Previously reported risk factors for AH among individuals with alcohol use include female gender, high body mass index, genetic susceptibility, malnutrition, tobacco dependence, and concomitant liver diseases [2,3,5]. Liver biopsy in patients with AH displays distinct histopathological patterns consistent with hepatocellular injury, such as lobular inflammation, hepatocyte ballooning, micro- and macro-vesicular steatosis, and fibrosis [3]. While "definite" diagnosis of AH requires liver biopsy, AH can be clinically diagnosed and categorized as "probable" if there are no confounding factors or "possible" if there are potential confounding factors [3]. If the clin-ical and laboratory criteria are not met and/or if there is an alternative explanation for a patient's presentation, the patient is categorized as non-AH.

Once diagnosed, there are several stratification algorithms, such as MDF, used to predict disease severity and mortality. Successful abstinence from alcohol has been the only intervention or therapy associated with long-term survival in patients with AH [3–6]. Therefore, it is critical to identify at-risk individuals in order to provide personalized counseling on alcohol use disorder and implement preventative measures.

A retrospective comparative analytical study was performed to identify the risk factors associated with developing AH among patients with chronic alcohol dependence at a single tertiary institution located in the metropolitan area of Las Vegas, Nevada.

Methods: We performed a retrospective review of patient charts from the medical record data-base at a single tertiary academic county medical institution. The study, along with the waiver of informed consent, was approved by the institutional review board (IRB) at the University Medical Center of Southern Nevada (UMC), Las Vegas, Nevada (IRB number: UMC-2019-248, approved 10 November 2021). We first queried the UMC electronic medi-cal record database using the following criteria: adult patients (age 18 and older) who were admitted for elevated transaminases, elevated bilirubin, alcohol-induced liver injury including alcoholic hepatitis, or alcohol use disorder were identified using the ICD-10 classification (International Classification of Diseases, tenth revision) over a 35-month pe-riod extending from 1 November 2017 to 10 October 2019. We did not extend the study be-yond October 2019 as to avoid additional confounding factors and elevated aminotrans-ferases that can occur with the disease due to the 2019 coronavirus (COVID-19).

Selection Criteria and Measures: While there is a spectrum of diagnostic criteria for alcohol use disorder based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [7], patients were in-cluded in our study on initial screening if there was a diagnosis of alcohol use disorder on medical coding or reported use of alcohol for at least 6 months with less than 60 days of abstinence [3]. Afterwards, additional data was retrieved regarding the pattern of alcohol use. A patient was determined to be a binge drinker if there was consumption of 5 or more standard drinks in males or 4 or more standard drinks in females in a 2-h period as de-fined by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) [3,8]. A heavy drinking pattern was defined as consuming more than 4 standard drinks on any day, more than 14 standard drinks per week in males, more than 3 standard drinks on any day, or more than 7 standard drinks per week in females, in accordance with the definition from the NIAAA [8]. A standard drink in the United States is described by the NIAAA as containing 14 g of ethanol, as found in 5 fluid ounces of wine, 12 fluid ounces of beer, and 1.5 fluid ounces of distilled spirits such as vodka, hard liquor, and te-quila [9,10]. The alcohol percentage in each drink was approximated at 5%, 12%, and 40% in beer, wine, and hard liquor/distilled spirits, respectively, based on prior literature standards [10].

Liver biopsies are not routinely performed on patients with suspected AH at our in-stitution. Hence, the diagnosis of AH was made clinically, both from the ICD-10 coding and from a review of the charts by the authors. Patients were diagnosed with probable AH if all of the following criteria were met: onset of jaundice within the past 8 weeks; ongoing consumption of alcohol for 6 or more months with less than 60 days of abstinence before the onset of jaundice; an aspartate aminotransferase (AST)/alanine aminotransferase (ALT) ratio >1.5 with both values <400 IU/L; an AST > 50 IU/L; and a serum total bilirubin >3.0 mg/dL [3]. If some but not all criteria were satisfied, or if there was the presence of po-tential confounding factors including but not limited to ischemic hepatitis, cocaine use, drug-induced liver disease, and metabolic liver disease, or if alcohol use could not be as-sessed properly based on chart review, the patient was allocated to a possible AH category [3]. Exclusion criteria were age younger than 18, abstinence from alcohol for ≥60 days, outpatient status, and a diagnosis of neither probable nor possible AH. If there were mul-tiple hospitalizations for AH during the study period, only the latest encounter was in-cluded in our study. The AH cohort was also divided into two categories. Patients who were diagnosed with AH for the first time were labeled as first-time AH. On the other hand, patients who had previously had at least one documented episode of AH were clas-sified as having recurrent AH. Using ICD-10 classification, a patient's medical chart from several local hospitals was reviewed through an interconnected electronic health records system to determine whether the patient had a prior diagnosis of AH.

The hepatotoxicity profile of the home medications of the patients was assessed using a grading system from the database of the National Library of Medicine of the National In-stitutes of Health [11]. A 5-point scale was used to estimate the level of hepatotoxicity of a medication: A = well-known cause; B = highly likely cause; C = probable cause; D = possi-ble cause; E = unlikely cause or suspected but unproven cause [11]. If a medication was from category A, B, C, or D, drug-induced liver injury was determined to be a confounding factor, and the patient was classified as "Possible AH" if he or she met the diagnostic cri-teria otherwise. If a patient was taking several hepatotoxic medications, a decision was made to include or exclude him from AH based on the history, clinical symptoms, and la-boratory data available in the patient's chart since the laboratory anomalies could also be induced by the medications.

Data Collection: The data was collected from patient charts between 1 November 2017 and 10 October 2019. The data was divided into three categories: sociodemographic and behavioral histo-ry, clinical or medical characteristics, and hospital outcomes. Sociodemographic and behavioral history data included age, body mass index (BMI), gender, race, health insurance status, homelessness, prior history of AH, family history of alcohol use, duration of alco-hol use, drinking pattern, percentage of alcohol content in the type of drink reported, to-bacco use, and illicit drug use, including intravenous (IV) drug use. Clinical or medical information included: presence of encephalopathy, cirrhosis, ascites, use of hepatotoxic medications, presence of viral hepatitis, model for end-stage liver disease-sodium score (MELD-Na) at admission, MDF score at admission, liver biopsy report if performed, hy-pertension, hyperlipidemia, glycated hemoglobin (HgbA1c), human immunodeficiency virus (HIV) status, and whether treatment with glucocorticoids was required during the hospitalization for AH. Hospital outcomes such as disposition, inpatient mortality, and length of hospital stay were also documented. During the data collection, patients with possible or probable AH were labeled as such in their respective categories and were also listed under the umbrella category of AH.

Sample Size Justification and Power Analysis: Power was ascertained separately for t, chi-square, and multiple logistic regression by using Cohen's effect size conventions (effect size = 0.5 for t-tests; effect size = 0.3 for the chi-square test) [12–14]. For the logistic regression analysis, we utilized the formula pro-posed by Green et al. [146, $N \geq 50 + 8$ m], where 'm' corresponds to the number of predic-tors. The total number of predictors was 12, according to which N = 146 was deemed ap-propriate [15]. The total sample size estimated with a power of 0.80 was 128 and 143 for the t-test and chi-square test, respectively. The sample size with the greatest value (N = 146) was considered appropriate since it satisfies the minimum requirement of all the sta-tistical tests used.

Statistical Analysis: First, the data was recoded for running analytical operations. All assumptions, in-cluding normality and homogeneity of variance, were assessed. Categorical variables were represented as frequencies and proportions, whereas normally distributed continuous variables were represented by means and standard deviations. A square root transformation was applied to the non-normally distributed variables for the normal approxima-tion. The Chi-square/Fisher exact test was used for comparing the categorical groups. Ad-justed standardized residuals greater than 2 were considered significant cells for contin-gency tables larger than 2 × 2. Continuous outcomes among two groups (AH vs. non-AH, probable vs. possible AH) were compared using an independent-samples t-test or a Welch t-test. A multivariate logistic regression model was fit to generate adjusted odds ratios for the likelihood of alcoholic hepatitis as an outcome. Estimates of parameters were obtained through the maximum likelihood estimation method with 95% Wald’s confidence limits for the logistic model. The final model was selected based upon the Akaike Information Criterion (AIC) and the Schwarz Criterion (SC) [16]. Additional regression analyses were performed to generate an adjusted odds ratio for the likelihood of inpatient mortality as an outcome within the AH cohort.

For regression analyses, polytomous categorical variables were dummy coded to cal-culate accurate parameters. All tests were two-sided, and a p-value of < 0.05 was considered significant. The Statistical Package for Social Sci-ences for Windows, version 27.0 (SPSS, Chicago, IL, USA), and Statistical Analysis System (SAS 9.4, Cary, NC, USA) were used to analyze the data for multivariate logistic regression.

Results: There was a total of 298 patients who were admitted to our tertiary teaching hospital in Southern Nevada from 1 November 2017 to 10 October 2019 and met the initial screen-ing criteria. Of the 298 patients, 106 were determined to have no history of alcohol de-pendence and were subsequently excluded from the study. From the remaining cohort, 100 patients were diagnosed with AH and were listed under the AH cohort; 92 patients were determined not to have AH using the criteria mentioned previously and were categorized under the non-AH cohort. We performed a bivariate comparison of AH and non-AH pa-tients in three categories: socio-demographic/behavioral history, clinical or medical char-acteristics, and hospital outcomes (Tables 1–3). Patients with AH were slightly younger, with a mean age of 49.3 years compared to that of non-AH patients at 54.5 years ($p = 0.008$). BMI and gender distribution were similar between the AH and non-AH cohorts. A higher incidence of AH was observed among non-Hispanic whites compared to other rac-es ($p = 0.02$). Prior history of AH was correlated with a higher risk of developing AH ($p = 0.007$). Certain alcohol consumption patterns, such as binge drinking ($p < 0.001$), heavy drinking ($p < 0.001$), and the percentage of alcohol in the consumed beverage ($p = 0.002$), were also associated with AH. Other socio-economic factors such as health insurance sta-tus, homelessness, family history of alcohol use, tobacco use, and illicit drug use did not display a statistically significant correlation with AH (Table 9).

Table 9: Bivariate Comparisons of Socio-Demographic and Behavioral History of the Sample (N = 192).

Variable	**Categories**	**AH,** **n (%)** **100 (52.1)**	**Non-AH, n (%),** **92 (47.9)**	***p*-Value**	**Test Statistics**	**Effect Size**
Age (Mean ± SD)	-	49.3 ± 12.0	54.5 ± 14.2	**0.008**	−2.701	−0.390
BMI (Mean ± SD)	-	27.9 ± 8.5	27.6 ± 8.1	0.7	0.288	0.042
Gender	Male	64 (64.0)	63 (68.5)	0.5	0.429	0.047
	Female	36 (36.0)	29 (31.5)			
Race	White	76 (76.0)	52 (56.5)	**0.02**	9.435	0.222
	Black	13 (13.0)	24 (26.1)			
	Hispanic	10 (10.0)	12 (13.0)			
	Other	1(1.0)	4 (4.3)			
Health insurance	Public	59 (59.0)	60 (65.2)	0.3	2.472	0.113
	Private	22 (22.0)	22 (23.9)			
	Uninsured	19 (19.0)	10 (10.9)			
Homelessness	Yes	17 (17.0)	18 (19.6)	0.6	0.212	0.033
	No	83 (83.0)	84 (80.4)			
Prior history of AH	Yes	29 (29.0)	12 (13.0)	**0.007**	7.264	0.195
	No	71 (71.0)	80 (87.0)			
Family history of alcohol use	Yes	10 (10.0)	5 (5.4)	0.2	1.387	0.085
	No	90 (90.0)	87 (94.6)			
Duration of alcohol use in years (Mean ± SD)	-	16.0 ± 11.6	18.5 ± 11.4	0.2	−1.292	−0.217
Binge drinking	Yes	51 (51.0)	15 (16.3)	**< 0.001**	25.570	0.365
	No	49 (49.0)	77 (83.7)			

Heavy drinking	Yes	74 (74.0)	33 (35.9)	**< 0.001**	28.238	0.383
	No	26 (26.0)	59 (64.1)			
% of Alcohol (Mean ± SD)	-	24.1 ± 17.0	16.6 ± 15.5	**0.002**	3.105	0.462
Tobacco use	Yes	47 (47.0)	53 (57.6)	0.1	2.161	0.106
	No	53 (53.0)	39 (42.4)			
Pack years (Mean ± SD)	-	21.5 ± 19.7	24.5 ± 24.7	0.5	−0.653	−0.132
IV drug use	Yes	6 (6.0)	5 (5.4)	0.9	0.028	0.012
	No	94 (94.0)	87 (94.6)			
Non-IV drug use	Yes	42 (42.0)	42 (45.7)	0.6	0.260	0.037
	No	58 (58.0)	50 (54.3)			

Notes: Other race includes Asian, Pacific Islanders, Native American and Alaska Native; Public in-surance includes Medicare, Medicaid and VA; p values less than 0.05 are considered statistically sig-nificant and are bolded in the table. Data are represented as frequencies and proportions unless stated otherwise. Among those who had previous history of AH, number of episodes varied from 1 (min.) to 4 (max.).

A higher incidence of AH was noted in patients with underlying liver diseases ($p < 0.001$), cirrhosis ($p < 0.001$), a high MELD-Na score ($p < 0.001$), or those who presented with ascites ($p < 0.001$). For AH patients, the mean MDF score was 22.1 ± 8.58 and the mean MELD-Na score was 20.9 ± 8.7 (Table 2).

The most frequently seen etiology of chronic liver disease in our sample was alco-hol-related (N = 60 (60%) in the AH cohort and N = 25 (25.17%) in the non-AH cohort). The remaining medical characteristics, such as viral hepatitis, hypertension, or hyperlipidem-ia, did not display a statistically significant relationship with AH (Table 2). Out of the total 192 patients, only 70 (36.5%) had HgbA1c information available. Therefore, the presence or absence of diabetes/prediabetes could not be accurately determined for all patients from the sample and was not included in the analysis. There were only three records of con-comitant positive viral hepatitis among the AH cohort; all three incidences were due to chronic Hepatitis C virus infection (HCV). In the non-AH cohort, there were 15 patients with a positive viral hepatitis panel. Thirteen patients were tested positive for HCV, one for chronic hepatitis B virus (HBV), and one for both HCV and chronic HBV. Lastly, the diagnosis of AH did not have a statistically significant impact on disposition, inpatient mortality, or length of hospital stay (Table 10).

Table 10: Bivariate Comparisons of Clinical or Medical Characteristics of the Sample (N = 192).

Variable	**Categories**	**AH**	**Non-AH**	***p*-Value**	**Test Statistics**	**Effect Size**
Encephalopathy	Yes	20 (20.0)	15 (16.3)	0.5	0.439	0.048
	No	80 (80.0)	77 (83.7)			
Cirrhosis	Yes	34 (34.0)	12 (13.0)	< **0.001**	11.551	0.245
	No	66 (66.0)	80 (87.0)			
Ascites	Yes	42 (42.0)	10 (10.9)	< **0.001**	23.514	0.350
	No	58 (58.0)	82 (89.1)			
Underlying liver disease	Yes	84 (84.0)	51 (55.4)	< **0.001**	18.731	0.312
	No	16 (16.0)	41 (44.6)			
Taking hepatotoxic medications	Yes	24 (24.0)	34 (37.0)	0.05	3.815	0.141
	No	76 (76.0)	58 (63.0)			
Hepatitis panel	Positive	3 (3.0)	15 (16.3)	**0.002**	9.983	0.228
	Negative	97 (97.0)	77 (83.7)			
MELD-Na score at admission (Mean ± SD)	–	20.9 ± 8.7	13.9 ± 7.4	< **0.001**	4.972	0.849
Maddrey’s discriminant function score at admission (Mean ± SD)	–	22.1 ± 8.58	N/A *	N/A	N/A	N/A
Hypertension	Yes	44 (44.0)	53 (57.6)	0.06	3.550	0.136
	No	56 (56.0)	39 (42.4)			
Hyperlipidemia	Yes	27 (27.0)	31 (33.7)	0.3	1.019	0.073
	No	73 (73.0)	61 (66.3)			
HIV	Yes	1 (1.0)	4 (4.3)	0.2	2.117	0.105
	No	99 (99.0)	88 (95.7)			
Treatment with glucocorticoids	Yes	19 (19.0)	9 (9.8)	0.07	3.268	0.130
	No	81 (81.0)	83 (90.2)			

Notes: Maddrey's discriminant function score was applicable only to those with AH and thus were not included in the analysis for non-AH patients. *p* values < 0.05 are considered statistically significant and are bolded in the table. Data are represented as frequencies and proportions unless stated otherwise. Some categories may not add to 100% due to missing data.

Table 11: Bivariate Comparisons of the Hospital Outcomes (N = 192).

Variable	**Categories**	**AH**	**Non-AH**	***p*-Value**	**Test Statistics**	**Effect Size**
Disposition	Home	61 (61.0)	59 (64.1)	0.9	0.604	0.056
	Facilities	13 (13.0)	11 (12.0)			
	Death	8 (8.0)	5 (5.4)			
	AMA	6 (6.0)	6 (6.5)			
	Others	12 (12.0)	11 (12.0)			
Length of hospital stay (Mean ± SD)	-	6.25 ± 1.18	6.91 ± 1.46	0.4	−0.804	−0.116

Among the AH cohort, there were 43 patients (43%) with probable AH and 57 pa-tients (57%) with possible AH. There were only 6 liver biopsies available; therefore, a defi-nite diagnosis of AH could not be made in most patients and was not included as a sub-category. In addition, the AH cohort was also divided into first-time AH and recurrent AH. There were 71 patients (71%), who were diagnosed with AH for the first time, and 29 patients (29%), who had recurrent AH. In the latter group, there were 19 patients (65.52%), 4 patients (13.80%), 5 patients (17.24%), and 1 patient (3.45%) who, respectively, had one, two, three, and four episodes of AH prior to the index presentation.

A multivariate logistic regression analysis between the AH and non-AH cohorts was then performed on the overall cohort, using the development of AH as the outcome. Binge drinking was associated with a higher risk of developing AH (odds ratio [OR], 2.698; 95% confidence interval [CI], 1.079–6.745; p = 0.03), as was heavy drinking (OR, 3.169; 95% CI, 1.348–7.452; p = 0.01) (Table 11).

The presence of cirrhosis was also associated with a great-er likelihood of developing concurrent AH (OR, 3.392; 95% CI, 1.306–8.811; p = 0.01). The presence of cirrhosis also predisposes patients to AH (OR, 3.392; 95% CI, 1.306–8.811; p = 0.01). Other variables were not statistically significant.

Table 12: Predictors or Risk Factors of AH (Multivariate Logistic Regression).

Variable(s)	**Odds Ratios Estimate**	**95% Confidence Limits**		***p*-Value**
Age	0.985	0.956	1.015	0.33
Gender	0.594	0.276	1.276	0.18
Race, White vs. Non-Hispanic Black	1.406	0.519	3.806	0.50
Hispanic vs. Non-Hispanic Black	0.867	0.225	3.338	0.84
Other race vs. Non-Hispanic Black	0.076	0.005	1.092	0.06
Insurance, Public vs. uninsured	0.352	0.117	1.062	0.06
Insurance, Private vs. uninsured	0.517	0.15	1.788	0.30
Body mass index	0.999	0.949	1.052	0.96
Prior history of Alcohol Hepatitis (Yes vs. No)	1.539	0.608	3.898	0.36
Binge drinking (Yes vs. No)	**2.698**	1.079	6.745	**0.03**
Heavy drinking (Yes vs. No)	**3.169**	1.348	7.452	**0.01**
Hypertension (Yes vs. No)	0.56	0.256	1.224	0.15
Hyperlipidemia (Yes vs. No)	0.873	0.38	2.007	0.75
Underlying liver disease (Yes vs. No)	2.026	0.81	5.067	0.13
Cirrhosis (Yes vs. No)	**3.392**	1.306	8.811	**0.01**

A forest plot with the OR estimates for the likelihood of AH with respect to particular variables is demonstrated in Figure 27.

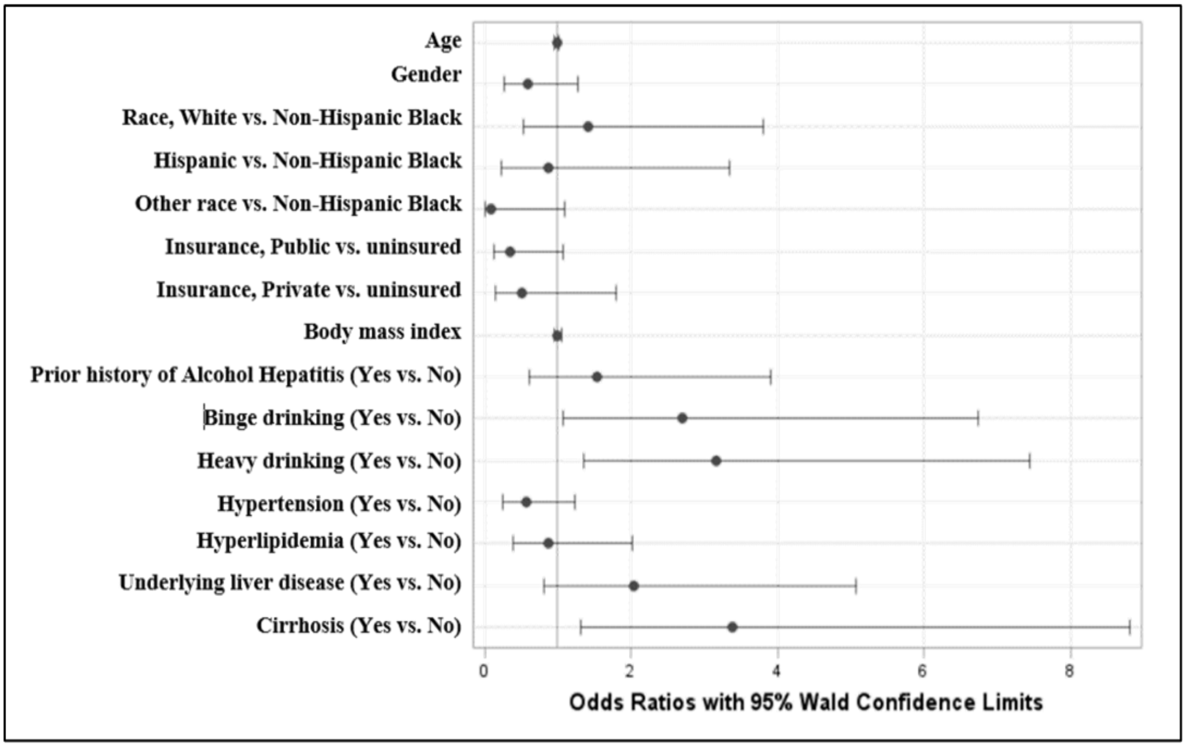


Figure 27: Forest Plot Showing Odds Ratio Estimates for Likelihood of Alcoholic Hepatitis

A logistic regression analysis between probable and possible AH groups was also performed within the AH cohort, examining the outcome of inpatient mortality. The results are shown in Table 5. Patients with probable AH had a higher risk of inpatient mortality compared to those with possible AH (OR, 6.79; 95% CI, 1.38–44.9; *p* = 0.03). Concomitant hypertension was also associated with a higher probability of inpatient mortality amongst AH patients (OR, 6.51, 95%; CI, 1.49–35.7; *p* = 0.02).

Table 13: Predictors or Risk Factors of Inpatient Mortality (Logistic Regression).

Variable(s)	**Odds Ratios Estimate**	**95% Confidence Limits**		***p*-Value**
Age	0.979	0.913	1.05	0.54
Gender	0.412	0.088	1.80	0.24
Race, White vs. Non-White	2.72	0.492	22.3	0.29
Probable AH vs. Possible AH	**6.79**	1.38	44.9	**0.03**
Insurance, Public vs. uninsured	0.815	0.113	7.39	0.84
Insurance, Private vs. uninsured	2.73	0.324	28.8	0.36
Body mass index	0.988	0.893	1.08	0.8
Prior history of Alcohol Hepatitis (Recurrent AH vs. First Time AH)	2.44	0.574	10.9	0.23
Binge drinking (Yes vs. No)	0.515	0.094	2.72	0.43
Heavy drinking (Yes vs. No)	0.665	0.077	5.46	0.7
Hypertension (Yes vs. No)	6.51	0.949	35.7	**0.02**
Hyperlipidemia (Yes vs. No)	0.189	0.021	1.05	0.08
Cirrhosis (Yes vs. No)	1.18	0.262	5.19	0.83

Discussion: Alcoholic hepatitis falls under the spectrum of alcohol-associated liver diseases. The rate of alcohol consumption and incidence of AH, as well as binge and heavy drinking patterns, have been rising in the U.S. in the past few decades [17,18]. For instance, the proportion of patients born between 1945 and 1965 who were admitted to 169 medical centers in the U.S. with a primary diagnosis of AH increased from 26% to 31% from the year 2000 to 2011 [17].

Additionally, a study from 2003 reported that alcohol consumption was responsible for 44% of all deaths among liver disease patients [2].

Our study demonstrated that binge and heavy drinking lead to a higher risk of developing alcoholic hepatitis, which is consistent with prior studies [4–6]. The alcohol content within the consumed beverage is also a crucial variable. Our results were consistent with those from prior studies that showed heavy drinking is correlated with the development of AH [4,8,18,19].

Moreover, even a single episode of binge drinking can lead to increased levels of serum endotoxin (lipopolysaccharide) and 16S ribosomal DNA, which are markers of dysbiosis and translocation of the gut microbiome to the bloodstream [20]. The endotoxin subsequently causes increased levels of inflammatory markers, which can induce a dysregulated immune response that in turn increases the risk for AH [20].

However, only about 6 to 20% of individuals with a heavy drinking pattern develop AH [4]. Therefore, other risk factors such as gender, genetic predisposition, race, and type of beverage also contribute to the risk of developing AH. For example, although it was not noted in our study, it has been previously demonstrated that women can develop alcohol-related liver injury at lower levels of alcohol consumption [3,4,19]. A possible explanation offered for this relationship is the higher level of serum endotoxin in women compared to men during alcohol intake [20]. In comparison to non-Hispanic Whites, African Americans and Asian Americans/Pacific Islanders also have lower hospitalization rates due to AH, whereas higher hospitalization rates have been observed amongst Hispanics and Native Americans [21,22]. Similar results were seen in our study, where the majority of AH patients were non-Hispanic whites. Nevertheless, we discovered that non-Caucasian Americans have a higher rate of mortality compared to their Caucasian counterparts (OR 2.72; 95% CI: 0.492–22.3; $p = 0.29$). The higher mortality rate despite a lower rate of hospitalizations among the non-Caucasian American demographic may be indicative of disparities in healthcare access.

Our data revealed a higher incidence of AH in patients with cirrhosis than in their non-cirrhotic counterparts. This association may be explained by the impaired metabolism of alcohol due to defective hepatic function, leading to an increased buildup of lipopolysaccharide endotoxin and subsequent activation of inflammatory cytokines [20]. In addition, the presence of cirrhosis at the time of admission may suggest a prolonged history of alcohol use or frequent at-risk alcohol use patterns such as binge drinking or heavy drinking.

We identified probable AH and hypertension as two characteristics associated with inpatient mortality among our AH cohort. In patients with probable AH, patients display more binge or heavy drinking patterns and present with more severe laboratory abnormalities, signifying a higher degree of hepatic injury [3,18]. Hypertension has not been previously demonstrated in the literature to be related to inpatient mortality in alcoholic hepatitis. However, hypertension can sometimes be suggestive of underlying cardiovascular disease, and alcohol intake has been demonstrated to have a J- or U-shaped relationship with cardiovascular ailment, indicating that while an inverse correlation with total mortality is seen in individuals with light alcohol consumption (2–4 drinks per day for men and 1–2 drinks per day for women), excessive alcohol consumption may be associated with cardiovascular complications and mortality [23–25]. More specifically, two meta-analyses have concluded that hypertension is correlated with >20 g of alcohol intake per day in women and >30 g of alcohol intake per day in men [26,27]. This suggests that hypertension may be indicative of excessive alcohol use and precede subsequent cardiovascular damage via increased oxidative stress and imbalances in neurohormonal pathways

[25]. Given the lack of longitudinal follow-up in our study, it was not possible to infer the relationship between alcohol intake and cardiovascular disease from our data.

Another principle that implicates hypertension in the development of liver disease is via the renin-angiotensin system (RAS) [28]. It has been established that the classical RAS axis produces angiotensin II, which can induce a pro-oxidant, pro-inflammatory, and fibrogenic effect on the liver [28]. Conversely, the counter-regulator RAS axis generates angiotensin 1–2, which negates the action of angiotensin II as an anti-oxidant and anti-fibrogenic agent [28]. Angiotensin-converting enzyme inhibitors or angiotensin receptor blockers inhibit the production of angiotensin II and have been shown to be beneficial in the treatment of chronic liver diseases. However, further clinical trials are required to determine their efficacy and safety profile in patients with alcoholic liver disease. Among the 100 patients with AH in our sample, there were 84 patients with underlying liver disease, of which hepatic steatosis was the most common (N = 71). We acknowledge that in patients with metabolic syndrome, it is not possible to distinguish between alcoholic-related liver disease and non-alcoholic fatty liver disease (NAFLD) even with a biopsy [3]. Hence, markers of the metabolic syndrome such as diabetes (A1c $\geq$ 6.5), dyslipidemia, and/or BMI $\geq$ 25 were included as confounding variables, and patients with such features were categorized as having a possible AH due to the degree of alcohol use. Thus, among 71 patients with hepatic steatosis, there were 43 patients with possible AH and 28 patients with probable AH. Only 3 patients, all of whom tested positive for the hepatitis C virus, were noted to have viral hepatitis data; therefore, a bivariate or logistic regression analysis could not be performed. Prior data in the literature has noted that patients with hepatitis C who have a heavy drinking pattern tend to develop a higher stage of fibrosis and viremia than their non-AH counterparts [4].

As a county catchment hospital, a large percentage of our patients were insured by public sectors such as Medicare and Medicaid; there were no insurance-specific differences noted.

Our study excluded patients starting from the beginning of the COVID-19 pandemic to minimize confounding variables. COVID-19 has been reported to cause varying degrees of liver enzyme abnormalities, which would have complicated data interpretation in our retrospective study when compared to patients in the pre-pandemic era [29,30]. Additionally, the pandemic has led to an overall increase in alcohol consumption and the incidence of AH. In a regional study [from Fresno, California] by Sohal et al., a 69% increase in AH-related hospitalization was noted after implementation of stay-at-home orders [31]. More specifically, there was a 100% increase in hospitalization of patients under 40 years old and a 125% increase in female patients; only a 34% rise was noted in males [31]. It is hypothesized that the younger individuals and females experienced a higher burden of economic, social, and psychological stressors from the pandemic, which led to increased alcohol use.

AH can cause as high as 40–50% mortality in severe cases, which is indicated by a MDF score >32 [4]. Currently, abstinence from alcohol remains the sole management recommendation associated with long-term survival [3,4,18,19]. Treatment with prednisolone in severe cases of AH correlates with a reduction in 28-day mortality but not long-term survival [19]. Pentoxifylline, a phosphodiesterase inhibitor, is no longer used due to a lack of associated short-term or long-term survival benefit based on the data from the STOPAH trial (steroids or pentoxifylline for alcoholic hepatitis trial) [19]. There is yet no generalized consensus or validated effectiveness for other novel approaches, including vitamin E, N-acetylcysteine, anti-tumor necrosis factor-alpha, granulocyte-colony stimulating factor, and fecal microbiota transplantation [4,18,19].

Our study was conducted at a hospital that serves a metropolitan area with a high rate of alcohol consumption. Nevada is recognized as a state with one of the highest estimates of binge drinking, especially among individuals aged 18–34 [32]. Among the several risk factors for AH that have been validated in the literature, our study emphasizes certain patterns, such as binge drinking and heavy drinking, that may be prevalent in other similar metropolitan settings. The information from our study will allow healthcare professionals to customize their approach to addressing alcohol dependence in the community. Furthermore, our study demonstrated that probable AH and underlying hypertension are correlated with increased inpatient mortality. Consequently, a multidisciplinary approach can be designed in a timely fashion to implement preventative measures in patients at higher risk or from disadvantaged socioeconomic backgrounds.

We acknowledge several limitations in our study. First, although we have extensively adjusted for demographic, lifestyle, and clinical risk factors for AH, as with all observational studies, we cannot rule out the possibility of residual confounding. Second, our cohort was comprised of patients presenting to a single tertiary center; therefore, our findings may not be generalizable to milder AH patients. Third, alcohol intake was self-reported, leading to the possibility of recall bias as well as inaccurate quantification of alcohol intake, especially in those with an extensive history of alcohol consumption. The percentage of alcohol in the beverages consumed by the patients was estimated into three categories for the purpose of analysis, which could lead to overgeneralization. Fourth, since liver biopsies were not routinely performed for the diagnosis of AH at our institution, a diagnosis of “definite AH” could not be made in most patients. However, a 2020 practice guideline from the American Association for the Study of Liver Diseases states that AH can be diagnosed clinically based on history, presenting symptoms, and laboratory criteria [3]. Additionally, the use of biopsy in AH is usually limited to clinical trials, may not be routinely available in all clinical settings, and is further limited by inter-pathologist variability. Lastly, due to the lack of sufficient information available, a correlation between diabetes and AH, if any, could not be investigated.

Conclusions: Our study demonstrated that the incidence of AH was higher in younger patients. We also demonstrated that binge drinking, and/or heavy drinking correlate with the development of alcoholic hepatitis. A higher incidence of AH was also found in patients with cirrhosis. Hypertension and probable AH were also correlated with increased inpatient mortality. Higher mortality rates and a lower hospitalization rate among African American patients may be reflective of healthcare disparities in our metropolitan county hospital. Our results further suggest the presence of a strong relationship between cardiovascular disease and the inflammatory state induced by alcohol consumption. Abstinence remains the only treatment that can lead to long-term survival. The data from our study can be used to better identify patients at risk of developing alcoholic hepatitis. Interventions such as motivational counseling, timely guidance to community resources, and closer monitoring may be beneficial in this patient population. Further studies that investigate cardiovascular impairment and AH, such as whether adequate treatment of cardiovascular diseases lowers the risk of AH, may be warranted.

Author Contributions: Conceptualization and Methodology: K.M.T., J.A.-P., D.L., and G.O.: data retrieval: K.M.T., Z.D., B.M., Y.N., S.M., A.M., and A.G.; formal analysis: K.B., D.M., and A.S.; writing—original draft preparation: K.M.T., Z.D., B.M., K.B., and S.M.; writing—review and editing: K.M.T., Z.D., B.M., C.-H.L., and J.A.-P.; supervision—G.O.; Z.D., and B.M. share

equal second co-authorship. All authors have read and agreed to the published version of the manuscript.

Funding: The research did not receive any funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of the University Medical Center of Southern Nevada (UMC-2019-248, approved 10 November 2021).

Informed Consent Statement: Informed consent was waived due to the retrospective nature of the study. Given the large number of patients, it would be impracticable to do the study if informed consent were required. The research involved no more than minimal risk to the subjects. The waiver or alteration would not adversely affect the rights and welfare of the subjects. The data was reported in a de-identified fashion.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to HIPAA.

Acknowledgments: Z.D. and B.M. share equal second co-authorship.

Conflicts of Interest: The authors declare no conflicts of interest.

References:

1. Alcohol. World Health Organization. Available online: https://www.who.int/news-room/fact-sheets/detail/alcohol (accessed on 13 September 2022).
2. Basra, S.; Anand, B.S. Definition, epidemiology and magnitude of alcoholic hepatitis. *World J Hepatol*. **2011**, *3*, 108–113. https://doi.org/10.4254/wjh.v3.i5.108.

3. Crabb, D.W.; Im, G.Y.; Szabo, G.; Mellinger, J.L.; Lucey, M.R. Diagnosis and Treatment of Alcohol-Associated Liver Diseases: 2019 Practice Guidance from the American Association for the Study of Liver Diseases. *Hepatology* **2020**, *71*, 306–333. https://doi.org/10.1002/hep.30866.

4. Yeluru, A.; Cuthbert, J.A.; Casey, L.; Mitchell, M.C. Alcoholic Hepatitis: Risk Factors, Pathogenesis, and Approach to Treatment. *Alcohol Clin. Exp. Res*. **2016**, *40*, 246–255. https://doi.org/10.1111/acer.12956.

5. Lucey, M.R.; Mathurin, P.; Morgan, T.R. Alcoholic hepatitis. *N. Engl. J. Med*. **2009**, *360*, 2758–2769. https://doi.org/10.1056/NEJMra0805786.

6. Jinjuvadia, R.; Liangpunsakul, S. Translational Research and Evolving Alcoholic Hepatitis Treatment Consortium. Trends in Alcoholic Hepatitis-related Hospitalizations, Financial Burden, and Mortality in the United States. *J. Clin. Gastroenterol*. **2015**, *49*, 506–511. https://doi.org/10.1097/MCG.0000000000000161.

7. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*; American Psychiatric Association: Washington, DC, USA, 2022. https://doi.org/10.1176/appi.books.9780890425787.

8. Drinking Levels Defined. National Institute on Alcohol Abuse and Alcoholism. Available online: https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking (accessed on 18 September 2022).

9. Rethinking Drinking Homepage—Niaaa. National Institute on Alcohol Abuse and Alcoholism. Available online: https://www.rethinkingdrinking.niaaa.nih.gov/ (accessed on 18 September 2022).

10. Lemmens, P.H. The alcohol content of self-report and 'standard' drinks. *Addiction* **1994**, *89*, 593–601. https://doi.org/10.1111/j.1360-0443.1994.tb03336.x.

11. Livertox—NCBI Bookshelf. Available online: https://www.ncbi.nlm.nih.gov/books/NBK547852/ (accessed on 18 September 2022).

12. Length, R.V. Some Practical Guidelines for Effective Sample Size Determination. *Am. Stat.* **2001**, *55*, 187–193.

13. Faul, F.; Erdfelder, E.; Buchner, A.; Lang, A.-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* **2009**, *41*, 1149–1160.

14. Faul, F.; Erdfelder, E.; Lang, A.-G.; Buchner, A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **2007**, *39*, 175–191.

15. Green, S.B. How Many Subjects Does It Take To Do A Regression Analysis. *Multivar. Behav. Res.* **1991**, *26*, 499–510.

16. Ludden, T.M.; Beal, S.L.; Sheiner, L.B. Comparison of the Akaike Information Criterion, the Schwarz criterion and the F test as guides to model selection. *J. Pharmacokinet. Biopharm.***1994**, *22*, 431–445. https://doi.org/10.1007/BF02353864.

17. Nguyen, T.A.; De Shazo, J.P.; Thacker, L.R.; Puri, P.; Sanyal, A.J. The Worsening Profile of Alcoholic Hepatitis in the United States. *Alcohol Clin. Exp. Res*. **2016**, *40*, 1295–1303. https://doi.org/10.1111/acer.13069.

18. Aday, A.W.; Mitchell, M.C.; Casey, L.C. Alcoholic hepatitis: Current trends in management. *Curr. Opin. Gastroenterol*. **2017**, *33*, 142–148. https://doi.org/10.1097/MOG.0000000000000359.

19. Thursz, M.R.; Forrest, E.H.; Ryder, S.; STOPAH investigators. Prednisolone or Pentoxifylline for Alcoholic Hepatitis. *N. Engl. J. Med*. **2015**, *373*, 282–283. https://doi.org/10.1056/NEJMc1506342.

20. Bala, S.; Marcos, M.; Gattu, A.; Catalano, D.; Szabo, G. Acute binge drinking increases serum endotoxin and bacterial DNA levels in healthy individuals. *PLoS ONE* **2014**, *9*, e96864. https://doi.org/10.1371/journal.pone.0096864.

21. Shirazi, F.; Singal, A.K.; Wong, R.J. Alcohol-associated Cirrhosis and Alcoholic Hepatitis Hospitalization Trends in the United States. *J. Clin. Gastroenterol*. **2021**, *55*, 174–179. https://doi.org/10.1097/MCG.0000000000001378.

22. Liangpunsakul, S. Clinical characteristics and mortality of hospitalized alcoholic hepatitis patients in the United States. *J. Clin. Gastroenterol*. **2011**, *45*, 714–719. https://doi.org/10.1097/MCG.0b013e3181fdef1d.

23. Goel, S.; Sharma, A.; Garg, A. Effect of Alcohol Consumption on Cardiovascular Health. *Curr; Cardiol; Rep*. **2018**, *20*, 19. https://doi.org/10.1007/s11886-018-0962-2.

24. Di Castelnuovo, A.; Costanzo, S.; Bagnardi, V.; Donati, M.B.; Iacoviello, L.; de Gaetano, G. Alcohol dosing and total mortality in men and women: An updated meta-analysis of 34 prospective studies. *Arch. Intern. Med*. **2006**, *166*, 2437–2445. https://doi.org/10.1001/archinte.166.22.2437. PMID: 17159008.

25. Grønbaek, M.; Johansen, D.; Becker, U.; Hein, H.O.; Schnohr, P.; Jensen, G.; Vestbo, J.; Sørensen, T.I. Changes in alcohol intake and mortality: A longitudinal population-based study. *Epidemiology* **2004**, *15*, 222–228. https://doi.org/10.1097/01.ede.0000112219.01955.56.

26. Briasoulis, A.; Agarwal, V.; Messerli, F.H. Alcohol consumption and the risk of hypertension in men and women: A systematic review and meta-analysis. *J. Clin. Hypertens.* **2012**, *14*, 792–798.

27. Taylor, B.; Irving, H.M.; Baliunas, D.; Roerecke, M.; Patra, J.; Mohapatra, S.; Rehm, J. Alcohol and hypertension: Gender differences in dose response relationships determined through systematic review and meta-analysis. *Addiction* **2009**, *104*, 1981–1990.

28. Simões, E.; Silva, A.C.; Miranda, A.S.; Rocha, N.P.; Teixeira, A.L. Renin angiotensin system in liver diseases: Friend or foe? *World J. Gastroenterol.* **2017**, *23*, 3396–3406. https://doi.org/10.3748/wjg.v23.i19.3396.

29. Moon, A.M.; Barritt, A.S., 4th. Elevated Liver Enzymes in Patients with COVID-19: Look, but Not Too Hard. *Dig. Dis. Sci.* **2021**, *66*, 1767–1769. https://doi.org/10.1007/s10620-020-06585-9.

30. Cascella, M.; Rajnik, M.; Aleem, A.; Dulebohn, S.C.; Di Napoli, R. Features, Evaluation, and Treatment of Coronavirus (COVID-19). 2022 Oct 13. In *StatPearls*; StatPearls Publishing: Treasure Island FL, USA, 2022.

31. Sohal, A.; Khalid, S.; Green, V.; Gulati, A.; Roytman, M. The Pandemic Within the Pandemic: Unprecedented Rise in Alcohol-related Hepatitis During the COVID-19 Pandemic. *J. Clin. Gastroenterol.* **2022**, *56*, e171–e175. https://doi.org/10.1097/MCG.0000000000001627.

32. Nelson, D.E.; Naimi, T.S.; Brewer, R.D.; Bolen, J.; Wells, H.E. Metropolitan-area estimates of binge drinking in the United States. *Am. J. Public Health.* **2004**, *94*, 663–671. https://doi.org/10.2105/ajph.94.4.663.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Case Study- 3

ML CLASSIFICATION OF CANCER TYPES USING HIGH DIMENSIONAL GENE ESPRESSION MICROARRAY DATA

Dwaipayan Mukhopadhyay, Dieudonne J. Phanord, Rohan J. Dalpatadu, Laxmi P. Gewali
and Ashok K. Singh.

Abstract: Cancer is a disease caused by the abnormal growth of cells in different parts of body is one of the top causes of death globally. Microarray gene expression data plays a critical role in the identification and classification of cancer tissues. Due to recent advancements in Machine Learning (ML) techniques, researchers are analyzing gene expression data using a variety of such techniques to model the progression rate & treatment of cancer patients with great effect.

But high dimensionality alongside the presence of highly correlated columns in gene expression datasets leads to computational difficulties. This paper aims to propose the use of ML classification techniques- Linear Discriminant Analysis (LDA) & Random Forest (RF) for classifying five types of cancer (breast cancer, kidney cancer, colon cancer, lung cancer and prostate cancer) based on high dimensional microarray gene expression data. Principal component analysis (PCA) was used for dimensionality reduction, and principal component scores of the raw data for classification. Six distinct categorization performance measures were used to evaluate these approaches; RF method provided us with higher accuracy than LDA method. The method and results of this article should be helpful to researchers who are dealing with many genes in microarray data.

Keywords: Principal Components Analysis; Linear Discriminant Analysis, Random Forest, Precision, Recall, F1, AUC, macro-averaged AUC, micro-averaged AUC.

Introduction: Cancer is a disease which can start almost anywhere in the human body, in which some of the body's trillion cells grow uncontrollably and spread to other parts of the body. There are over 200 types of cancer such as colon, liver, ovarian and breast etc. [1, 2]. In 2023, 1,958,310 new cancer cases and 609,820 cancer deaths were projected in the United States. [3]. This prompts a clear understanding of the underlying mechanism and characteristics of this potentially fatal disease alongside identifying the most significant genes responsible for it.

Cancer can alter the gene expression profile of the body cells. Therefore, microarray data is utilized in clinical diagnosis to recognize down or up the regulated gene expression, which is the reason for generating new biomarkers, and leading to cancer disease [4]. Microarray data analysis has been a popular approach for diagnosing cancer, and DNA microarray is a technology used to collect data on large numbers of various gene expressions at the same time [5,6]. The classification and identification of gene expression using DNA microarray data is an effective tool for cancer diagnosis and prognosis for specific cancer subtypes. Gene expression analysis can assure medical experts whether a patient suffers from cancer within a relatively shorter time than traditional methods. Recently, its analysis has emerged as an important means for addressing the fundamental challenges associated with cancer diagnosis and drug discovery [7,8]. Analysis of gene expression data involves the identification of informative genes, [9] and [10] demonstrates that cancer classification can be improved by identifying informative genes which in turn can be used to accurately predict new sample classes.

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to "self-learn" or obtain information from training data; recognize patterns in data and develop their own predictions, improving over time without being explicitly programmed [11]. Medical researchers and clinicians are utilizing several ML techniques on medical data sets to construct an intelligent diagnosis system [12]. Massive volume of data is being generated in the medical industry thanks to the digital revolution in information technology. ML techniques are highly suited for analyzing these massive data sets, and multiple algorithms have been used to diagnose various diseases [13,14,15]. Numerous research has been done to classify cancer using microarray gene expression data. Golub et al. [16] suggested a strategy based on expression profiles generated by microarrays. According to ML theory, classification outcomes are dependent on the features of the input set, the training algorithm, and the system's capacity to adapt to the original data. It is necessary to evaluate the behavior of various classifiers on provided data.

Recently, several classification approaches were created in the ML domain, and many of them were utilized in cancer classification [17]. However, there are several difficulties possible to face in the microarray classification process like (a) The microarray genes expression data constitutes many highly correlated genes for just a small sample size. The small number of cancer samples compared with the number of features can degrade the performance of the classifier and increase the risk of over-fitting. (b) Various uncertainties associated with the process of acquiring microarray data, for example, fabrication, image processing etc., resulting in unexplained fluctuation in the data. (c) The majority of genes in the microarray date are redundant for classifying diverse tissue types [18,19].

The earliest detection of cancer is among the most efficient approaches to reduce cancer-related death [20,21,22,23]. The microarray's primary characteristic is its greater number of genes (p) in comparison to the number of tissues (n) [24]. In most gene expression studies selection of relevant genes to differentiate between patients with and without cancer is a common task [25,26,27,28,29,30]. Due to overestimation & various linearity issues it is difficult to categorize high-dimensional microarray data (p > n) using statistical approaches [31,32]. There is no single optimal method to examine microarray data, with its continually evolving analysis methods [33]. Various supervised and unsupervised ML techniques have also been adopted to identify the most significant genes [34,35,36].In microarray gene expression analysis, gene selection or feature selection (FS) is utilized to improve cancer classification performance while using fewer samples, eliminate undesired & repetitive attributes from data and ultimately counter the curse of dimensionality by identifying the most informative genes to enhance disease prediction accuracy

[37,38]. ML and dimensionality reduction techniques also perform exceptionally well at classifying biologic data [39, 40, 41]. Hence it may be beneficial to use feature selection methods which can address the challenges arising from high data dimensionality and small sample size.

The remainder of this paper is structured in the following manner. Section 2 discusses the related work. Section 3 presents the materials and methods. In Section 4, we present the experimental results. Finally, in Section 5, we conclude the paper giving a discussion.

Related Work: ML can assist in automating intelligent processes, increasing development efficiency and accuracy, and lowering costs [42]. Over the years ML-based classifiers have been widely used in classification of cancer sub-types. Several studies tried to assess whether ML can help in oncology care, by investigating the applications of ML in cancer risk stratification, diagnoses, and medication development [17,43,44,45]. According to those studies, ML can help in cancer prediction and diagnosis by analyzing pathology profiles and imaging studies.

BRCA (Breast Cancer gene) genes produce proteins that help repair damaged DNA and are referred to as tumor suppressor genes since certain changes in these genes can cause cancer [46]. People born with a certain variant of BRCA tend to develop cancer at early ages. Chang, Dalpatadu, Phanord and Singh [47] fitted a Bayesian Logistic Regression model for prediction of breast cancer using the Wisconsin Diagnosis Breast Cancer (WDBC) data set [48] which was downloaded from the UCI Machine Learning Repository; precision, recall and F1-measures of 0.93, 0.89, and 0.91 were reported for the training data, and 0.87, 0.91, 0.89 for the test data, respectively.HER2 protein accelerates breast cancer cell growth and HER2 positive patients when treated with medicines which attack the HER2 protein. Gene expression patterns of HER2 are quite complex and pose a challenge to pathologists. Cordova et al. (2023) developed a new interpretable ML method in immunohistochemistry for accurate HER2 classification and obtained high precision (0.97) and high accuracy (0.89) using immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) data [49].

Kidney renal cell carcinoma (KIRC) is the most prevalent type of kidney cancer, with a survival rate of less than 5 years and 338,000 estimated number of new cases each year [50]. ICD profile of KIRC. Wang et al. (2023) correlated the immunogenic cell death (ICD) of KIRK with the heterogeneity and therapeutic complexity which is useful for developing optimal immunotherapy strategy for KIRC patients [51].

A common cancerous tumor in the digestive track is colon adenocarcinoma (COAD) and is commonly associated with fatty acids [52]; diagnosis of COAD is difficult as there are hardly any early symptoms. Li et al. (2017) used a genetic algorithm and the k-nearest neighbors clustering method to determine genes which can accurately classify samples as well as class subtypes for a TCGA RNA-seq dataset of 9066 cancer patients and 602 normal samples [53].

Lung adenocarcinoma (LUAD) is a common form of lung cancer which also gets detected in the middle/late stages and therefore is hard to treat [54]. Yang et al. (2022) used a dataset of gene expression profiles from 515 tumor samples and 59 normal tissues and split the dataset into two significantly different clusters; they further showed that using age, gender, pathological stages, and risk score as predictors of LUAD increased the prediction accuracy measures [55]. Liu, Lei, Zhang, and Wang (2022) used cluster analysis on enrichment scores of 12 stemness signatures to identify three LUAD subtypes, St-H, St-M and St-L for six different datasets [56].

Prostate adenocarcinoma (PRAD) is common in elderly men, and patients suffering from PRAD typically have good prognosis [57]. Khosravi et al. (2021) used Deep Learning ML models on an MRI dataset from 400 subjects with suspected prostate cancer combined with histological data and reported high accuracies [58].

PCA is an exploratory multivariate statistical technique for simplifying complex data sets [59, 60, 61]. It has been used in a wide range of biomedical problems, including the analysis of microarray data in search of outlier genes [62], analysis of other types of expression data [63, 64] as well as cancer classification [65]. AK Oladejo, TO Oladele, YK Saheed (2018) presented two methods of dimension reduction: feature extraction (FE) and FS; one-way Anova for FE and PCA was utilized for FS [66]. The Support vector machine (SVM) and k-nearest neighbor (K-NN) were used for the classification of leukemia genome data. The obtained results gave an accuracy of 90% for SVM and 81.67% for K-NN.

MO Adebiyi, MO Arowolo, MD Mshelia, OO Olugbara (2022) applied the machine learning algorithms of RF and the SVM with the feature extraction method of LDA to the Wisconsin Breast Cancer Dataset [67]. The SVM with LDA and RF with LDA yielded accuracy results of 96.4% and 95.6% respectively. Evidence from this study shows that better prediction is crucial and can benefit from machine learning methods. This research has validated the use of feature extraction in predicting a diagnostic system for breast cancer when compared to the existing literature.

Ak, Muhammet Fatih (2020) utilized the Wisconsin Breast Cancer Dataset [48] for the comparison of most of the major machine-learning procedures for detection and diagnosis [69]. Supervised learning-decision tree, RF, multilayer perception, SVM, and linear regression (LR) were compared in both the classification and regression categories. The results revealed that under the classification algorithm, the SVM provides high accuracy; however, under the regression methodology, multilayer perception regression delivers reduced errors. Díaz-Uriarte, Ramón (2006) investigated the implication of RF for classification of microarray data (including multi-class problems) and propose a new method of gene selection in classification problems based on RF [70]. The study used simulated and nine microarray data sets and demonstrated that random forest has comparable performance to other classification methods, including diagonal discriminant analysis (DLDA), KNN, and SVM, and that the new gene selection procedure yields very small sets of genes without compromising predictive accuracy.

AC Tan, D Gilbert (2003) classified cancer using gene expression data using three distinct tree-based supervised ML techniques [71]. Seven different categories of cancer data were classified using bagged and boosted decision trees (DT) alongside C4.5 DT. The bagging DT outperforms the other two. A Sharma, S Imoto, S Miyano, V Sharma (2012) proposed a Null space-based feature selection method for gene expression data in terms of supervised classification. [72]. Scatter matrices-generated null space information were utilized as a feature selection method in removing the duplicate gene expressions. After effectively lowering the dimension of the features, classification was performed using three different types of classifiers: SVM, naïve Bayes (NB), and LDA.

Degroeve, De Baets, Van de Peer and Rouz´e (2002) created a balanced train and set by randomly selecting 1000 positive instances and 1000 negative and created a test data with 281 positive and 7505 negative instances and another test data set with 281 positive and 7643 negative instances; they used SVM classifier, a NB classifier, and a traditional method for feature selection for predicting splice site and obtained improved performance. Precision obtained for these datasets ranged in 93-98% range, but the recall and F1-measures were in 25-49% range

[73]. Peng, Li and Liu (2006) compared various methods of gene selection over four microarray gene expression datasets and showed that the hybrid method works well on the four datasets [74].

Sharma and Paliwal (2008) used Gradient LDA method for three small microarray gene expression datasets: acute leukemia, small round blue-cell tumor (SRBCT) and lung adenocarcinoma and have obtained higher accuracies than some competing methods [75]. Bar-Joseph, Gitter and Simon (2012) provided a discussion of how time-series gene expression data is used for identification of activated genes in biological processes and describe how basic patterns lead to gene expression programs [76]. Cho et al. (2004) proposed a modified kernel Fisher discriminant analysis (KFDA) for the analysis of the hereditary breast cancer dataset [77]. The KFDA classifier employed the mean-squared-error as the gene selection criterion. D Huang (2009) evaluated the classification performance of LDA, prediction analysis for microarrays (PAM), shrinkage centroid regularized discriminant analysis (SCRDA), shrinkage linear discriminant analysis (SLDA) and shrinkage diagonal discriminant analysis (SDDA) by applying these methods to six public cancer gene expression datasets [78].

Dwivedi (2018) used the method of Artificial Neural Network (ANN) for classification of acute cases of lymphoblastic leukemia and myeloid leukemia and reported over 98% overall classification accuracy [79]. Sun et al. (2019) used the genome deep learning method to analyze 6,083 samples from the Whole Exon Sequencing mutations with 12 types of cancer and 1991 non-cancerous samples from the 1000 Genome Project and obtained overall classification accuracies ranging in 70% - 97% [80]. A survey of feature selection literature for gene expression microarray data analysis based on a total of 132 research articles [81] was conducted by Alhenawi, Al-Sayyed, Hudaib and Mirjalili (2022). Khatun et al. (2023) developed an ensemble rank-based feature selection method (EFSM) and a weighted average voting scheme to overcome the problems posed by high dimensionality of microarray gene expression data [82]. They obtained overall classification accuracies of 100% (leukemia), 95% (colon cancer), and 94.3% for the 11-tumor dataset. Osama, Shaban and Ali (2023) have provided a review of ML methods for cancer classification of microarray gene expression data; data pre-processing and feature selection methods including filter, wrapper, embedded, ensemble, and hybrid algorithms [83].

Kabir et al. (2023) compared two different dimension reduction techniques—PCA, and autoencoders for the selection of features in a prostate cancer classification analysis. Two machine learning methods—neural networks and SVM—were further used for classification. The study showed that the classifiers performed better on the reduced dataset [84]. Another study Adiwijaya et al. (2018) utilized PCA dimension reduction method that includes the calculation of variance proportion for eigenvector selection followed by the classification methods, SVM and Levenberg-Marquardt Backpropagation (LMBP) algorithm. Based on the tests performed, the classification method using LMBP was more stable than SVM [10].

Kharya, S., D. Dubey, and S. Soni (2013) compared the accuracy of the SVM, ANN, Naive Bayes classifier, and AdaBoost tree to identify a potent model for breast cancer prediction as observational research [85]. PCA was used to reduce dimensionality. The study found that, when compared to techniques like decision trees, regression trees, and so on, ANN came out to be the one with the most reliable approach in making real-time predictions and prognoses. Rana et al (2015) used machine learning classification algorithms, which use stored historical data to learn from and forecast new input categories, benign and malignant tumors [86]. According to this study, the random forest model demonstrated the highest accuracy of 96% to detect different cancers.

Based on previous research, the general scheme in the process of classification of microarray data for the detection of cancer can be conducted via preprocessing the data and dimensionality reduction followed by cancer classification.

Materials and Methods: In this article, we have used the Linear Discriminant Analysis (LDA) classifier [87] and the random forest (RF) classifier [88] on an 801 rows x 20531 columns (genes) dataset of patients with five cancer types: BRCA, KIRC, COAD, LUAD and PRAD; the dataset has no missing values. Variables in this dataset are RNA-Seq gene expression levels measured by illumina HiSeq platform. The variables are dummy named gene XX. This dataset (gene expression cancer RNA-Seq) was downloaded from the UCI Machine Learning Repository [89]. The statistical software package R (2023) was used for all data analyses and visualizations [90]. We computed Principal Component (PC) scores [91] of the data and performed the 5-level classification on an increasing number of PC's and obtained excellent classification results using just the first two components PC1 and PC2. Five cancer types are described below.

We will next provide brief descriptions of the methods of data analysis and the common measures of accuracy used in multi-level classification.

Principal Components Analysis (PCA): PCA is a dimension-reduction technique which creates new and uncorrelated linear combinations of original variables (principal components); the values of the principal components are called PC-Scores and can be used in place of the original variables for further analyses such as Multiple Linear Regression (MLR) or Discrimination and Classification. Using PC-Scores instead of original variables as predictors eliminates the problem of multicollinearity. PCA was performed using the correlation matrix which normalizes the input variables.

Linear Discriminant Analysis (LDA): LDA is itself a dimension-reduction technique which is used for separating a dataset into 2 or more subgroups, and for classification of new data into these subgroups. LDA is typically one of the methods used for multi-level classification problems. The LDA method involves computing separating hyperplanes for classification purposes [92] (pp. 587–590). We used the function prcompfast of the R-package Morpho to first perform PCA of the gene expression microarray dataset at hand and the PC-Scores were used as input variables for LDA. All computations were performed on a Windows 10 PC with AMD Ryzen Threadpiper1950X 16-Core Processor and 128 GB usable RAM.

Random Forest (RF): The RF method is a decision-tree based method that can be used for classification (categorical response) or regression (continuous response) problems. It randomly selects a subset of rows (samples) and a subset of columns (features) at a time and fits decision trees a very large number of times to predict Y and then uses a voting mechanism to predict Y values. Random forest is known to be highly accurate [93].

Training and Test Datasets: In ML literature, it is common practice to randomly split the available dataset into Training and Test datasets and report the accuracy measures of prediction for both datasets. Typically, higher accuracy measures are obtained for the training set than the test set. The entire raw dataset was used to compute PC-Scores by using the fast-PCA method of the R-package Morpho. A dataset of 801 rows and 25 PC-scores was created, and then this dataset of PC-scores was randomly split into an 80% training set and 20% test set. The LDA and RF methods were used on the training set of PC-Scores and the accuracy measures given below were computed for both training and test sets.

Accuracy Measures for Multi-Level Classification: All accuracy measures are computed from the confusion matrix which is a cross-tabulation of observed *Y* and predicted Y values.

Overall Accuracy (OA) = sum of diagonal elements of CM/sum of all elements of CM

Precision_j = j-th diagonal element of CM/sum of j-th column of CM

Recall_j = j-th diagonal element of CM/sum of j-th row of CM

F1_j = harmonic mean of Precision_j and Recall_j

The following accuracy measures are computed for each level by calculating the one vs all binary confusion matrices:

Area Under the Curve (AUC)

Macro- and micro-averages of AUC

Explanations of the accuracy measures and computational details are provided in [52].

Results: PCA was run on the entire 801 rows x 20531 genes data set, and trial-and-error showed that just the first two principal components were sufficient for classification purposes. The genes with highest absolute loadings are shown in Table 9.

A scatterplot of the first two PC-scores for the entire dataset is shown in Figure 1. A clear separation between BRCA and KIRC cancer sub-types with some overlap between COAD, LUAD and PRAD is seen in Figure 28.

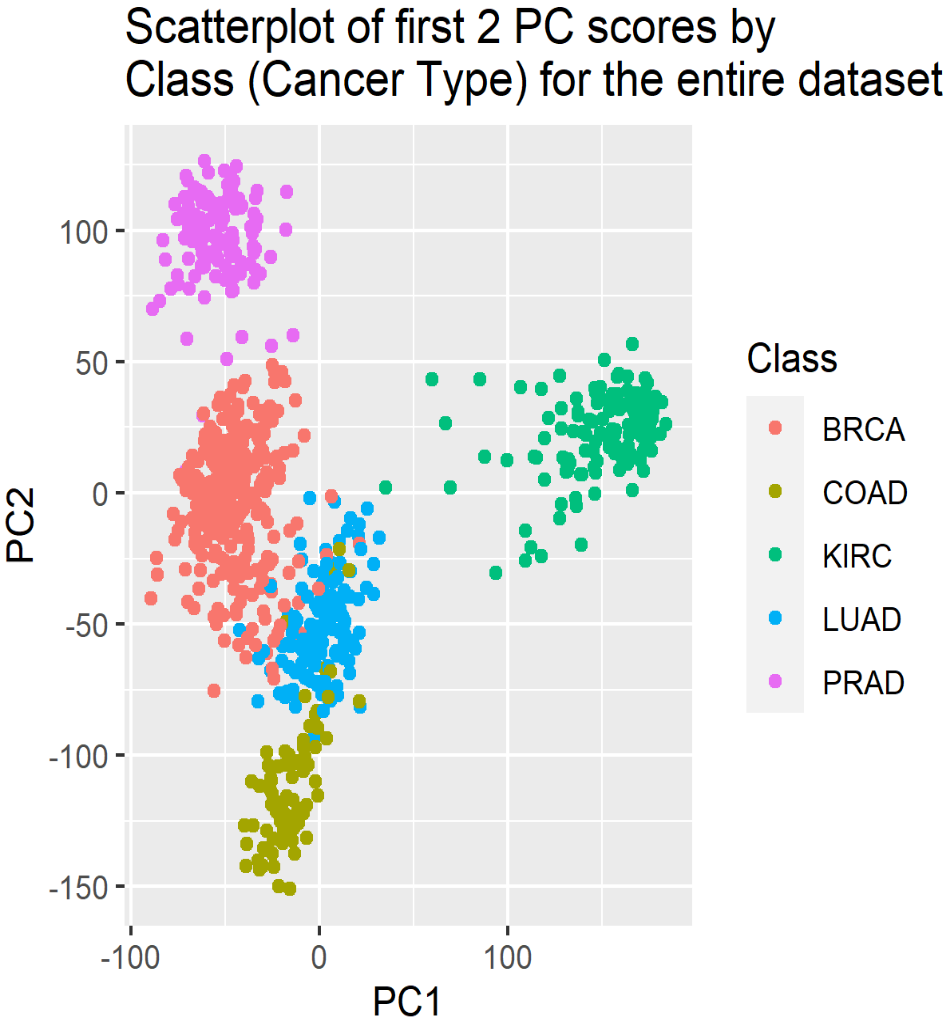
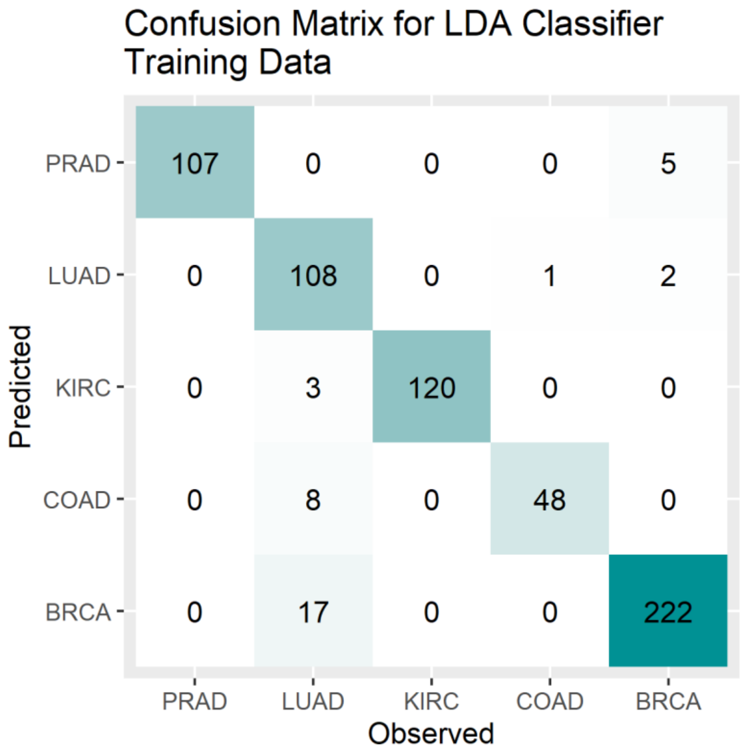


Figure 28: Scatterplot of PC2 vs PC1 for the Entire Data

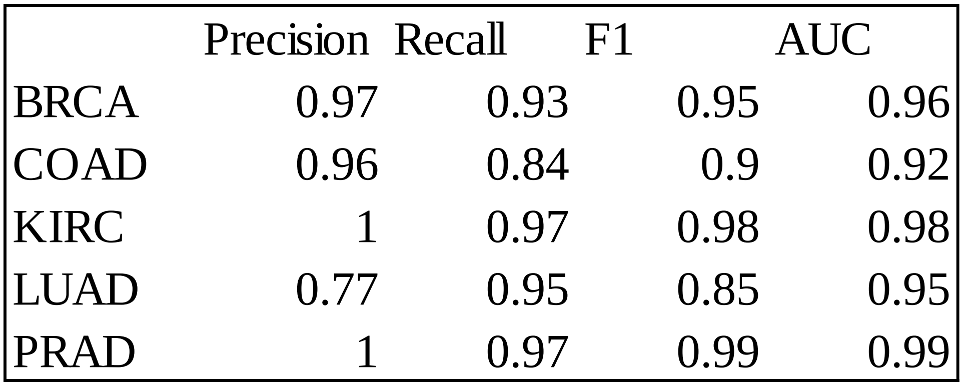
Figures 29-32 show plots of the confusion matrices for the LDA and RF classifiers for training and test sets, respectively. Figures 33 - 40 show that all measures of multi-level accuracy are high for both training and test datasets and both LDA and RF methods.

Accuracy Measures for the LDA Classifier for Training Data:



	PRAD	LUAD	KIRC	COAD	BRCA
PRAD	107	0	0	0	5
LUAD	0	108	0	1	2
KIRC	0	3	120	0	0
COAD	0	8	0	48	0
BRCA	0	17	0	0	222

Figure 29: Confusion Matrix Plot for the LDA Classifier – Training Data



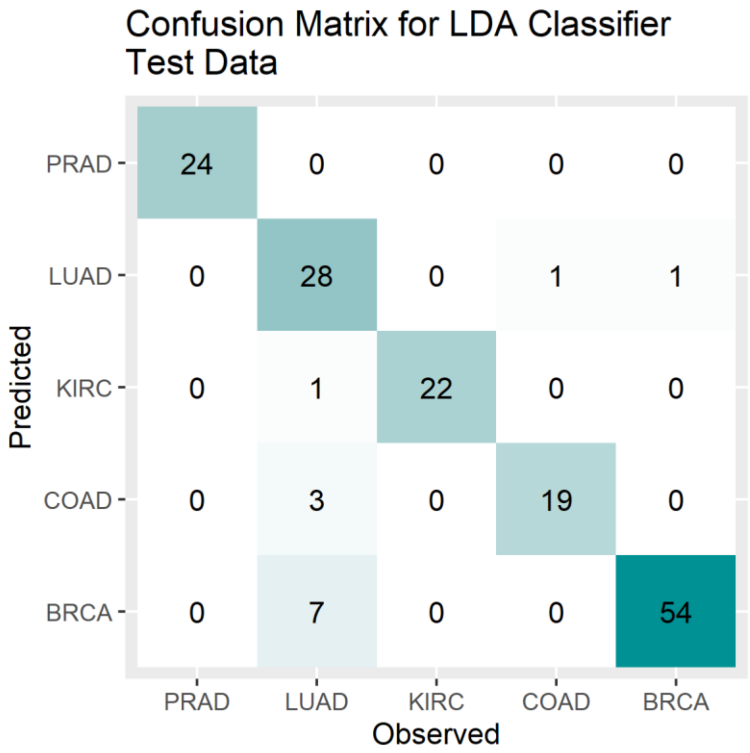
	Precision	Recall	F1	AUC
BRCA	0.97	0.93	0.95	0.96
COAD	0.96	0.84	0.9	0.92
KIRC	1	0.97	0.98	0.98
LUAD	0.77	0.95	0.85	0.95
PRAD	1	0.97	0.99	0.99

Figure 30: Precision, Recall, F1 and AUC Measures for the LDA Classifier – Training Data

Macro average AUC	0.94	0.93	0.94	0.95
Micro average AUC	0.94	0.94	0.94	na
OA	0.94			
na: no micro-averaged AUC exists in the ML literature				

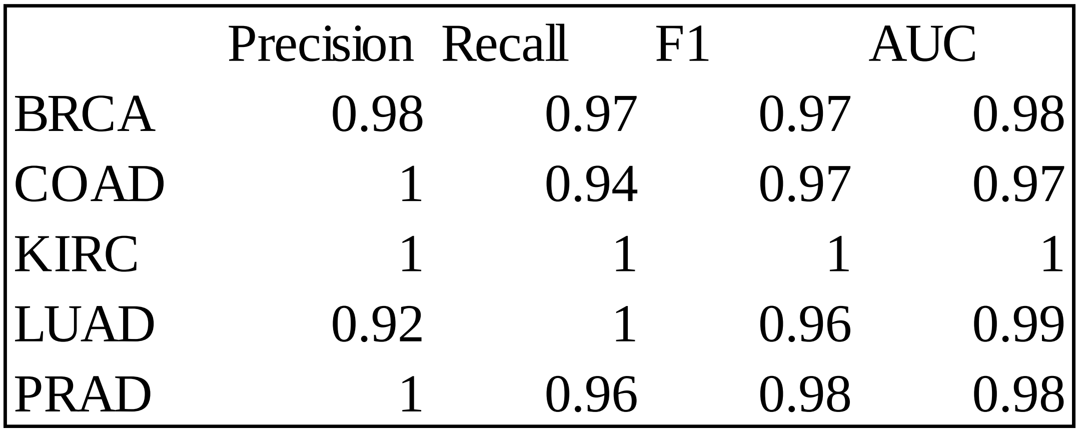
Figure 31: Macro and Micro-Averaged AUC Measures for the LDA Classifier – Training Data

Accuracy Measures for the LDA Classifier for Test Data:



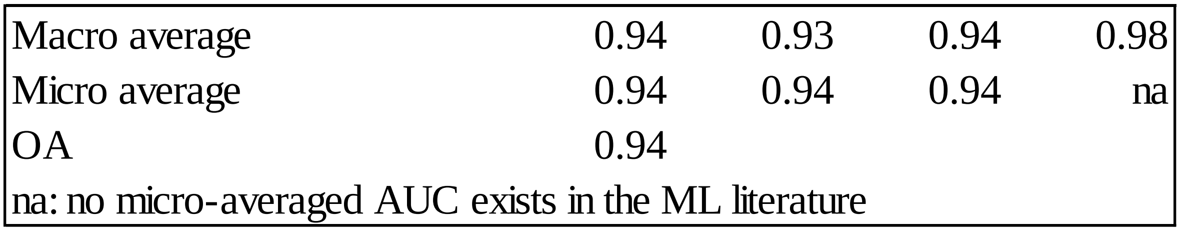
	PRAD	LUAD	KIRC	COAD	BRCA
PRAD	24	0	0	0	0
LUAD	0	28	0	1	1
KIRC	0	1	22	0	0
COAD	0	3	0	19	0
BRCA	0	7	0	0	54

Figure 32: Confusion Matrix Plot for the LDA Classifier – Test Data



	Precision	Recall	F1	AUC
BRCA	0.98	0.97	0.97	0.98
COAD	1	0.94	0.97	0.97
KIRC	1	1	1	1
LUAD	0.92	1	0.96	0.99
PRAD	1	0.96	0.98	0.98

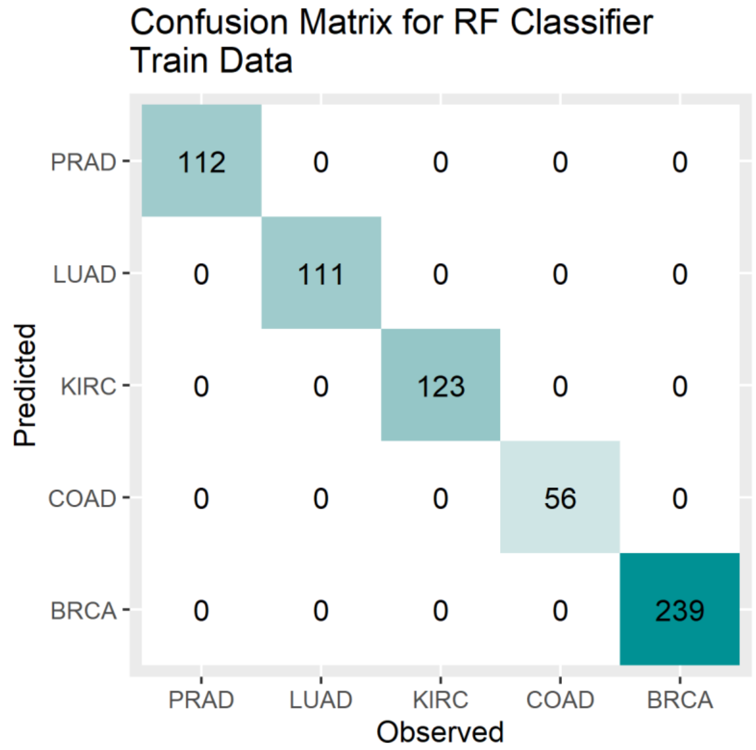
Figure 33: Confusion Matrix Plot and Accuracy Measures for the LDA Classifier – Test Data



Macro average	0.94	0.93	0.94	0.98
Micro average	0.94	0.94	0.94	na
OA	0.94			
na: no micro-averaged AUC exists in the ML literature				

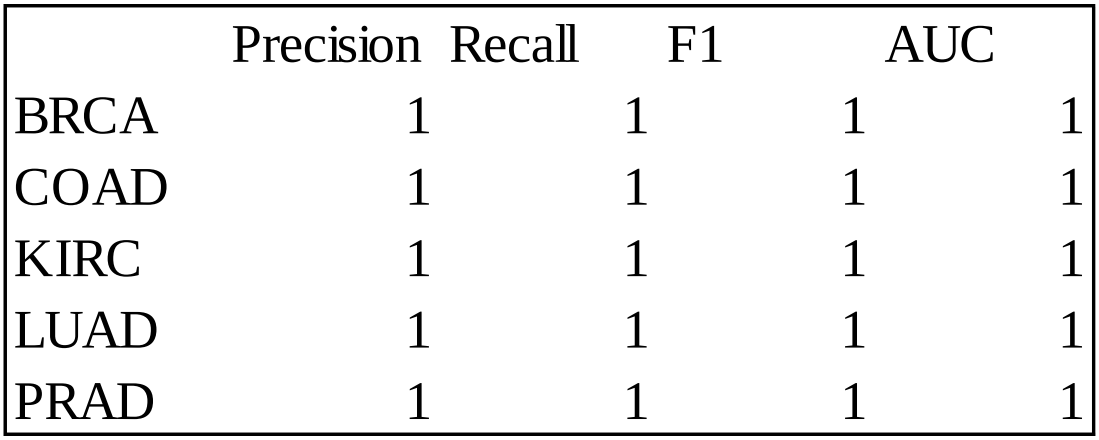
Figure 34: Macro and Micro-Averaged AUC for the LDA Classifier – Test Data

Accuracy Measures for the RF Classifier for Training Data:



Predicted \ Observed	PRAD	LUAD	KIRC	COAD	BRCA
PRAD	112	0	0	0	0
LUAD	0	111	0	0	0
KIRC	0	0	123	0	0
COAD	0	0	0	56	0
BRCA	0	0	0	0	239

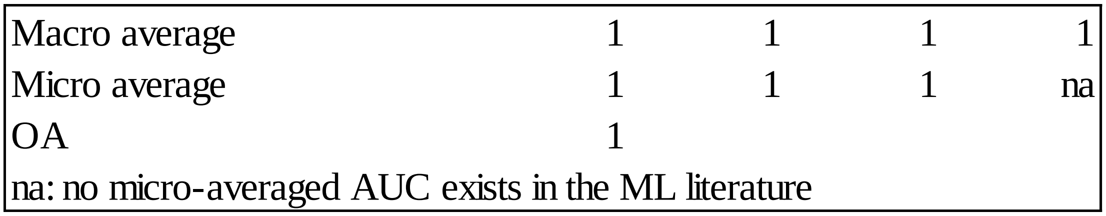
Figure 35: Confusion Matrix Plot for the RF Classifier – Training Data



	Precision	Recall	F1	AUC
BRCA	1	1	1	1
COAD	1	1	1	1
KIRC	1	1	1	1
LUAD	1	1	1	1
PRAD	1	1	1	1

Figure 36: Precision, Recall, F1 and AUC Measures for the RF Classifier – Training Data

Table 6: Macro and Micro averaged AUC for the RF Classifier – Training Data



Macro average	1	1	1	1
Micro average	1	1	1	na
OA	1			
na: no micro-averaged AUC exists in the ML literature				

Figure 37: Macro and Micro Averaged AUC for the RF Classifier – Training Data

Accuracy Measures for the RF Classifier for Test Data:

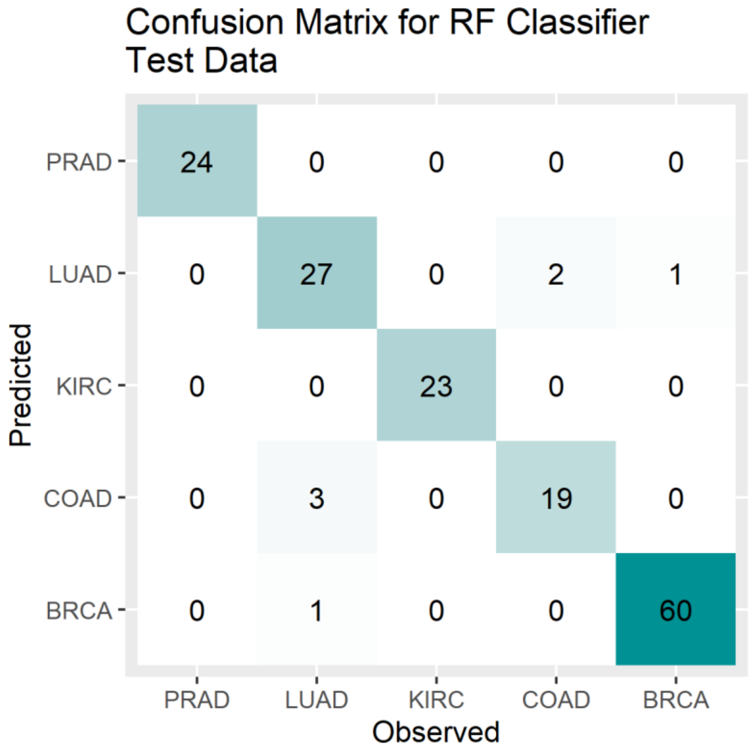
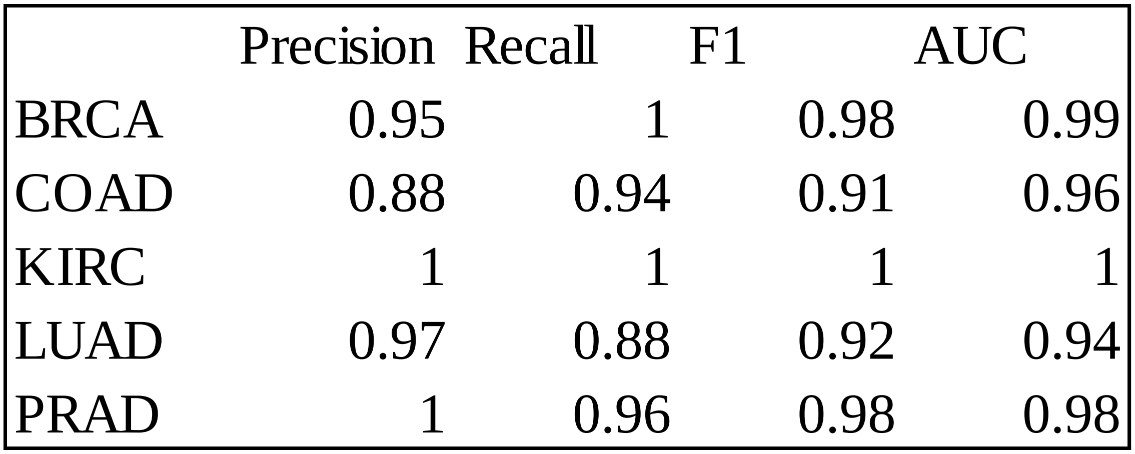
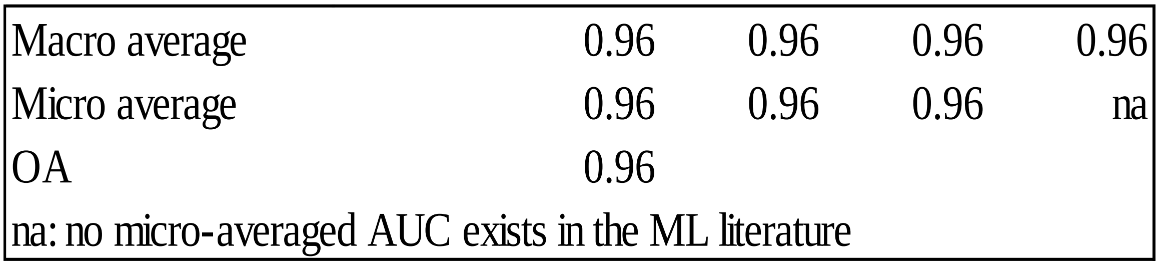


Figure 38: Confusion Matrix Plot for the RF Classifier – Test Data



	Precision	Recall	F1	AUC
BRCA	0.95	1	0.98	0.99
COAD	0.88	0.94	0.91	0.96
KIRC	1	1	1	1
LUAD	0.97	0.88	0.92	0.94
PRAD	1	0.96	0.98	0.98

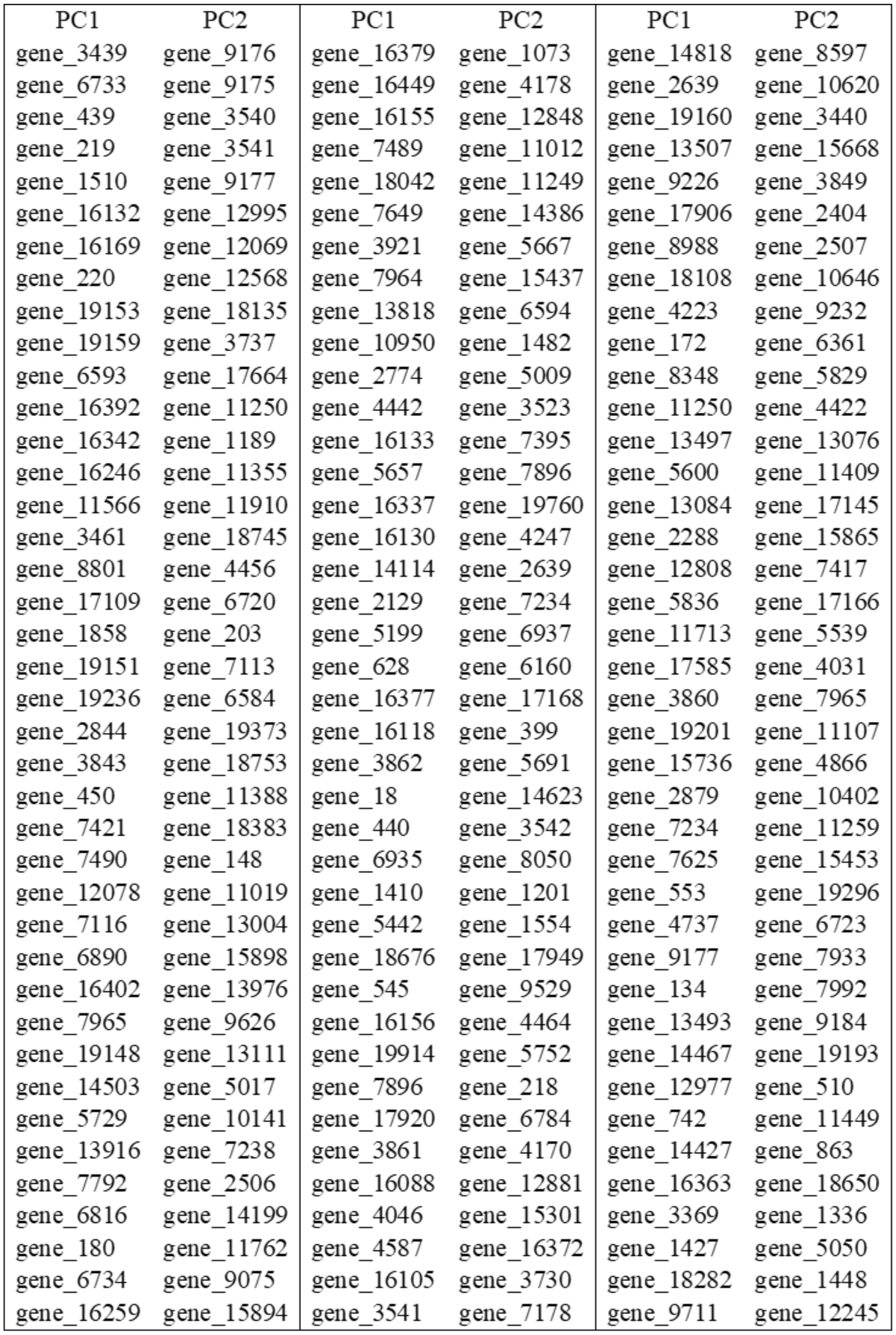
Figure 39: Precision, Recall, F1 and AUC Measures for the RF Classifier – Test Data



Macro average	0.96	0.96	0.96	0.96
Micro average	0.96	0.96	0.96	na
OA	0.96			
na: no micro-averaged AUC exists in the ML literature				

Figure 40: Macro and Micro Averaged AUC for the RF Classifier – Test Data

In Figure 41 we provide the variables (genes) with high absolute loadings on the first two PC-scores; such a table can be very useful for selection of features (genes)



PC1	PC2	PC1	PC2	PC1	PC2
gene_3439	gene_9176	gene_16379	gene_1073	gene_14818	gene_8597
gene_6733	gene_9175	gene_16449	gene_4178	gene_2639	gene_10620
gene_439	gene_3540	gene_16155	gene_12848	gene_19160	gene_3440
gene_219	gene_3541	gene_7489	gene_11012	gene_13507	gene_15668
gene_1510	gene_9177	gene_18042	gene_11249	gene_9226	gene_3849
gene_16132	gene_12995	gene_7649	gene_14386	gene_17906	gene_2404
gene_16169	gene_12069	gene_3921	gene_5667	gene_8988	gene_2507
gene_220	gene_12568	gene_7964	gene_15437	gene_18108	gene_10646
gene_19153	gene_18135	gene_13818	gene_6594	gene_4223	gene_9232
gene_19159	gene_3737	gene_10950	gene_1482	gene_172	gene_6361
gene_6593	gene_17664	gene_2774	gene_5009	gene_8348	gene_5829
gene_16392	gene_11250	gene_4442	gene_3523	gene_11250	gene_4422
gene_16342	gene_1189	gene_16133	gene_7395	gene_13497	gene_13076
gene_16246	gene_11355	gene_5657	gene_7896	gene_5600	gene_11409
gene_11566	gene_11910	gene_16337	gene_19760	gene_13084	gene_17145
gene_3461	gene_18745	gene_16130	gene_4247	gene_2288	gene_15865
gene_8801	gene_4456	gene_14114	gene_2639	gene_12808	gene_7417
gene_17109	gene_6720	gene_2129	gene_7234	gene_5836	gene_17166
gene_1858	gene_203	gene_5199	gene_6937	gene_11713	gene_5539
gene_19151	gene_7113	gene_628	gene_6160	gene_17585	gene_4031
gene_19236	gene_6584	gene_16377	gene_17168	gene_3860	gene_7965
gene_2844	gene_19373	gene_16118	gene_399	gene_19201	gene_11107
gene_3843	gene_18753	gene_3862	gene_5691	gene_15736	gene_4866
gene_450	gene_11388	gene_18	gene_14623	gene_2879	gene_10402
gene_7421	gene_18383	gene_440	gene_3542	gene_7234	gene_11259
gene_7490	gene_148	gene_6935	gene_8050	gene_7625	gene_15453
gene_12078	gene_11019	gene_1410	gene_1201	gene_553	gene_19296
gene_7116	gene_13004	gene_5442	gene_1554	gene_4737	gene_6723
gene_6890	gene_15898	gene_18676	gene_17949	gene_9177	gene_7933
gene_16402	gene_13976	gene_545	gene_9529	gene_134	gene_7992
gene_7965	gene_9626	gene_16156	gene_4464	gene_13493	gene_9184
gene_19148	gene_13111	gene_19914	gene_5752	gene_14467	gene_19193
gene_14503	gene_5017	gene_7896	gene_218	gene_12977	gene_510
gene_5729	gene_10141	gene_17920	gene_6784	gene_742	gene_11449
gene_13916	gene_7238	gene_3861	gene_4170	gene_14427	gene_863
gene_7792	gene_2506	gene_16088	gene_12881	gene_16363	gene_18650
gene_6816	gene_14199	gene_4046	gene_15301	gene_3369	gene_1336
gene_180	gene_11762	gene_4587	gene_16372	gene_1427	gene_5050
gene_6734	gene_9075	gene_16105	gene_3730	gene_18282	gene_1448
gene_16259	gene_15894	gene_3541	gene_7178	gene_9711	gene_12245

Figure 41: Significant Genes With Highest Absolute Loadings on the First Two PC-Scores

Discussion: We have demonstrated successful application of PCA for dimensionality reduction on a dataset with a very large number of genes collected from a much smaller number of subjects. PCA results showed that the first 50 components (PC) cumulatively explained 71% of all variability present in the 801 × 20532 gene expression data, with the first two PC's explaining only 26% of total variability. The first two PC's, however, were sufficient for classification of cancer sub-types with high accuracy. This can be seen from the plot of the first two components by of cancer sub-type. LDA was able to classify each of the five cancer-subtypes with high accuracies except for LUAD which had a precision of 77% for the training set. The RF method was able to classify each sub-type with very high accuracy. The PCA loadings on 20532 genes were sorted in order of magnitude and genes (features) important for classification were identified. Our results are not generalizable, but the proposed classification method should be very helpful to researchers and clinicians working with gene expression microarray data of very high dimensionality. It should be noted that high accuracy is achieved by the LDA and the RM classifiers using just the first two PC-Scores even though only 26% of variability was explained by the first two components.

Limitations of the Study: PCA as used in the present study only works for continuous variables and should not be used when the variables are categorical (nominal or ordinal). The method presented here is generalizable to continuous variables in other similar datasets, but the results are not generalizable.

Author Contributions: Conceptualization, DM. and AKS; methodology, DM. and AKS.; validation, DM, DP, RD, LG, AS.; formal analysis, DM, AK, RD.; investigation, DM.; DM, AK, DP, LG; .; data curation, DM.; writing—original draft preparation, DM.; writing—review and editing, DM.; DM, AK, DP, LG.; visualization, DM.; supervision, AKS and DP.; project administration, AKS and DP... All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: "Not applicable" since secondary data is used.

Data Availability Statement: The microarray gene expression data is available at

https://archive.ics.uci.edu/dataset/401/gene+expression+cancer+rna+seqmicroarray-202

Conflicts of Interest: All authors declare no conflicts of interest.

References:

1. Alladi, Subha Mahadevi, Vadlamani Ravi, and Upadhyayula Suryanarayana Murthy. "Colon cancer prediction with genetic profiles using intelligent techniques." *Bioinformation* 3, no. 3 (2008): 130. 10.6026/97320630003130.

2. Alon, Uri, Naama Barkai, Daniel A. Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J. Levine. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays." *Proceedings of the National Academy of Sciences* 96, no. 12 (1999): 6745-6750. https://doi.org/10.1073/pnas.96.12.6745.

3. Siegel, Rebecca L., Kimberly D. Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. "Cancer statistics, 2023." *Ca Cancer J Clin* 73, no. 1 (2023): 17-48. DOI:10.3322/caac.21763.

4. Slonim, Donna K. "From patterns to pathways: gene expression data analysis comes of age." *Nature genetics* 32, no. 4 (2002): 502-508. https://doi.org/10.1038/ng1033.

5. Harrington, Christina A., Carsten Rosenow, and Jacques Retief. "Monitoring gene expression using DNA microarrays." *Current opinion in Microbiology* 3, no. 3 (2000): 285-291. https://doi.org/10.1016/S1369-5274(00)00091-6.

6. Schena, Mark, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* 270, no. 5235 (1995): 467-470. DOI: 10.1126/science.270.5235.467.

7. Brewczyński, Adam, Beata Jabłońska, Agnieszka Maria Mazurek, Jolanta Mrochem-Kwarciak, Sławomir Mrowiec, Mirosław Śnietura, Marek Kentnowski, Zofia Kołosza, Krzysztof Składowski, and Tomasz Rutkowski. "Comparison of selected immune and hematological parameters and their impact on survival in patients with HPV-related and HPV-unrelated oropharyngeal Cancer." *Cancers* 13, no. 13 (2021): 3256. https://doi.org/10.3390/cancers13133256.

8. Munkácsy, Gyöngyi, Libero Santarpia, and Balázs Győrffy. "Gene expression profiling in early breast cancer—patient stratification based on molecular and tumor microenvironment features." *Biomedicines* 10, no. 2 (2022): 248. https://doi.org/10.3390/biomedicines10020248.

9. Siang, Tan Ching, Ting Wai Soon, Shahreen Kasim, Mohd Saberi Mohamad, Chan Weng Howe, Safaai Deris, Zalmiyah Zakaria, Zuraini Ali Shah, and Zuwairie Ibrahim. "A review of

cancer classification software for gene expression data." *International Journal of Bio-Science and Bio-Technology* 7, no. 4 (2015): 89-108. http://dx.doi.org/10.14257/ijbsbt.2015.7.4.10.

10. Adiwijaya, Wisesty Untari, E. Lisnawati, Annisa Aditsania, and Dana Sulistiyo Kusumo. "Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification." *Journal of Computer Science* 14, no.

11. Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

12. Sidey-Gibbons, Jenni AM, and Chris J. Sidey-Gibbons. "Machine learning in medicine: a practical introduction." *BMC medical research methodology* 19 (2019): 1-18. https://doi.org/10.1186/s12874-019-0681-4.

13. Erickson, Bradley J., Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L. Kline. "Machine learning for medical imaging." *radiographics* 37, no. 2 (2017): 505-515. https://doi.org/10.1148/rg.2017160130.

14. Mahmood, Nasir, Saman Shahid, Taimur Bakhshi, Sehar Riaz, Hafiz Ghufran, and Muhammad Yaqoob. "Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach." *Medical & Biological Engineering & Computing* 58 (2020): 2631-2640. https://doi.org/10.1007/s11517-020-02245-2.

15. Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." *Artificial Intelligence in medicine* 23, no. 1 (2001): 89-109. https://doi.org/10.1016/S0933-3657(01)00077-X.

16. Golub, Todd R., Donna K. Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P. Mesirov, Hilary Coller et al. "Molecular classification of cancer: class discovery and class

prediction by gene expression monitoring." *science* 286, no. 5439 (1999): 531-537. DOI: 10.1126/science.286.5439.531.

17. Kourou, Konstantina, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. "Machine learning applications in cancer prognosis and prediction." *Computational and structural biotechnology journal* 13 (2015): 8-17. https://doi.org/10.1016/j.csbj.2014.11.005.

18. Wang, Xujing, Martin J. Hessner, Yan Wu, Nirupma Pati, and Soumitra Ghosh. "Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction." *Bioinformatics* 19, no. 11 (2003): 1341-1347. https://doi.org/10.1093/bioinformatics/btg154.

19. Mohamad, Mohd Saberi, Sigeru Omatu, Michifumi Yoshioka, and Safaai Deris. "An approach using hybrid methods to select informative genes from microarray data for cancer classification." In *2008 Second Asia International Conference on Modelling & Simulation (AMS)*, pp. 603-608. IEEE, 2008. DOI: 10.1109/AMS.2008.71

20. Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 71, no. 3 (2021): 209-249. https://doi.org/10.3322/caac.21660.

21. Reid, Alison, Nick de Klerk, and Arthur W. Musk. "Does exposure to asbestos cause ovarian cancer? A systematic literature review and meta-analysis." *Cancer epidemiology, biomarkers & prevention* 20, no. 7 (2011): 1287-1295. https://doi.org/10.1158/1055-9965.EPI-10-1302.

22. Ünver, Halil Murat, and Enes Ayan. "Skin lesion segmentation in dermoscopic images with combination of YOLO and grabcut algorithm." *Diagnostics* 9, no. 3 (2019): 72. https://doi.org/10.3390/diagnostics9030072.

23. Maniruzzaman, Md, Md Jahanur Rahman, Benojir Ahammed, Md Menhazul Abedin, Harman S. Suri, Mainak Biswas, Ayman El-Baz, Petros Bangeas, Georgios Tsoulfas, and Jasjit S. Suri. "Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms." *Computer methods and programs in biomedicine* 176 (2019): 173-193. https://doi.org/10.1016/j.cmpb.2019.04.008.

24. Kalina, Jan. "Classification methods for high-dimensional genetic data." *Biocybernetics and Biomedical Engineering* 34, no. 1 (2014): 10-18. https://doi.org/10.1016/j.bbe.2013.09.007.

25. Lee, Jae Won, Jung Bok Lee, Mira Park, and Seuck Heun Song. "An extensive comparison of recent classification tools applied to microarray data." *Computational Statistics & Data Analysis* 48, no. 4 (2005): 869-885. https://doi.org/10.1016/j.csda.2004.03.017.

26. Yeung, Ka Yee, Roger E. Bumgarner, and Adrian E. Raftery. "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data." *Bioinformatics* 21, no. 10 (2005): 2394-2402. https://doi.org/10.1093/bioinformatics/bti319.

27. Jirapech-Umpai, Thanyaluk, and Stuart Aitken. "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes." *BMC bioinformatics* 6 (2005): 1-11. https://doi.org/10.1186/1471-2105-6-148.

28. Hua, Jianping, Zixiang Xiong, James Lowey, Edward Suh, and Edward R. Dougherty. "Optimal number of features as a function of sample size for various classification

rules." *Bioinformatics* 21, no. 8 (2005): 1509-1515. https://doi.org/10.1093/bioinformatics/bti171.

29. Li, Yi, Colin Campbell, and Michael Tipping. "Bayesian automatic relevance determination algorithms for classifying gene expression data." *Bioinformatics* 18, no. 10 (2002): 1332-1339. https://doi.org/10.1093/bioinformatics/18.10.1332.

30. Díaz-Uriarte, Ramón. "Supervised methods with genomic data: a review and cautionary view." *Data analysis and visualization in genomics and proteomics* (2005): 193-214. DOI:10.1002/0470094419.

31. Piao, Yongjun, Minghao Piao, Kiejung Park, and Keun Ho Ryu. "An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data." *Bioinformatics* 28, no. 24 (2012): 3306-3315. https://doi.org/10.1093/bioinformatics/bts602.

32. Chen, Kun-Huang, Kung-Jeng Wang, Kung-Min Wang, and Melani-Adrian Angelia. "Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data." *Applied Soft Computing* 24 (2014): 773-780. https://doi.org/10.1016/j.asoc.2014.08.032.

33. Akay, Mehmet Fatih. "Support vector machines combined with feature selection for breast cancer diagnosis." *Expert systems with applications* 36, no. 2 (2009): 3240-3247. https://doi.org/10.1016/j.eswa.2008.01.009.

34. Brahim-Belhouari, Sofiane, and Amine Bermak. "Gaussian process for nonstationary time series prediction." *Computational Statistics & Data Analysis* 47, no. 4 (2004): 705-712. https://doi.org/10.1016/j.csda.2004.02.006.

35. Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 68, no. 6 (2018): 394-424. https://doi.org/10.3322/caac.21492.

36. Cai, Jie, Jiawei Luo, Shulin Wang, and Sheng Yang. "Feature selection in machine learning: A new perspective." *Neurocomputing* 300 (2018): 70-79. https://doi.org/10.1016/j.neucom.2017.11.077.

37. Jain, Anil, and Douglas Zongker. "Feature selection: Evaluation, application, and small sample performance." *IEEE transactions on pattern analysis and machine intelligence* 19, no. 2 (1997): 153-158. DOI: 10.1109/34.574797.

38. Wang, Yu, Igor V. Tetko, Mark A. Hall, Eibe Frank, Axel Facius, Klaus FX Mayer, and Hans W. Mewes. "Gene selection from microarray data for cancer classification—a machine learning approach." *Computational biology and chemistry* 29, no. 1 (2005): 37-46. https://doi.org/10.1016/j.compbiolchem.2004.11.001.

39. Liu, Kun-Hong, Muchenxuan Tong, Shu-Tong Xie, and Vincent To Yee Ng. "Genetic programming based ensemble system for microarray data classification." *Computational and mathematical methods in medicine* 2015 (2015). DOI: 10.1155/2015/193406.

40. Bhonde, Swati B., and Jayashree R. Prasad. "Performance analysis of dimensionality reduction techniques in cancer detection using microarray data." *Asian Journal For Convergence In Technology (AJCT) ISSN-2350-1146* 7, no. 1 (2021): 53-57. https://doi.org/10.33130/AJCT.2021v07i01.012.

41. Sun, Xiaoxiao, Yiwen Liu, and Lingling An. "Ensemble dimensionality reduction and feature gene extraction for single-cell RNA-seq data." *Nature communications* 11, no. 1 (2020): 5853. https://doi.org/10.1038/s41467-020-19465-7.

42. Rowe, Raymond C., and Ronald J. Roberts. "Artificial intelligence in pharmaceutical product formulation: knowledge-based and expert systems." *Pharmaceutical Science & Technology Today* 1, no. 4 (1998): 153-159. https://doi.org/10.1016/S1461-5347(98)00042-X.

43**.** Yu, Chaoran, and Ernest Johann Helwig. "The role of AI technology in prediction, diagnosis and treatment of colorectal cancer." *Artificial Intelligence Review* 55, no. 1 (2022): 323-343. https://doi.org/10.1007/s10462-021-10034-y.

44. Kumar, Yogesh, Surbhi Gupta, Ruchi Singla, and Yu-Chen Hu. "A systematic review of artificial intelligence techniques in cancer prediction and diagnosis." *Archives of Computational Methods in Engineering* 29, no. 4 (2022): 2043-2070. https://doi.org/10.1007/s11831-021-09648-w.

45. McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back et al. "International evaluation of an AI system for breast cancer screening." *Nature* 577, no. 7788 (2020): 89-94. https://doi.org/10.1038/s41586-019-1799-6.

46. Mersch, Jacqueline, Michelle A. Jackson, Minjeong Park, Denise Nebgen, Susan K. Peterson, Claire Singletary, Banu K. Arun, and Jennifer K. Litton. "Cancers associated with BRCA 1 and BRCA 2 mutations other than breast and ovarian." *Cancer* 121, no. 2 (2015): 269-275. https://doi.org/10.1002/cncr.29041.

47. Chang, Michael, Rohan J. Dalpatadu, Dieudonne Phanord, and Ashok K. Singh. "Breast cancer prediction using bayesian logistic regression." *Biostatistics and Bioinformatics* 2, no. 3 (2018): 1-5. https://doi.org/10.47739/2475-9465/1039.

48. Wolberg, William, W. Street, and Olvi Mangasarian. "Breast cancer wisconsin (diagnostic)." UCI Machine Learning Repository 414 (1995): 415. DOI: 10.24432/C5DW2B.

49. Cordova, Claudio, Roberto Muñoz, Rodrigo Olivares, Jean-Gabriel Minonzio, Carlo Lozano, Paulina Gonzalez, Ivanny Marchant, Wilfredo González-Arriagada, and Pablo Olivero. "HER2 classification in breast cancer cells: A new explainable machine learning application for immunohistochemistry." *Oncology Letters* 25, no. 2 (2023): 1-9. https://doi.org/10.3892/ol.2022.13630.

50. Hu, Fuyan, Wenying Zeng, and Xiaoping Liu. "A gene signature of survival prediction for kidney renal cell carcinoma by multi-omic data analysis." *International journal of molecular sciences* 20, no. 22 (2019): 5720. https://doi.org/10.3390/ijms20225720.

51. Wang, Licheng, Yaru Zhu, Zhen Ren, Wenhuizi Sun, Zhijing Wang, Tong Zi, Haopeng Li et al. "An immunogenic cell death-related classification predicts prognosis and response to immunotherapy in kidney renal clear cell carcinoma." *Frontiers in Oncology* 13 (2023): 1147805. https://doi.org/10.3389/fonc.2023.1147805.

52. Yue, Fu-Ren, Zhi-Bin Wei, Rui-Zhen Yan, Qiu-Hong Guo, Bing Liu, Jing-Hui Zhang, and Zheng Li. "SMYD3 promotes colon adenocarcinoma (COAD) progression by mediating cell proliferation and apoptosis." *Experimental and Therapeutic Medicine* 20, no. 5 (2020): 1-1. https://doi.org/10.3892/etm.2020.9139.

53. Li, Yuanyuan, Kai Kang, Juno M. Krahn, Nicole Croutwater, Kevin Lee, David M. Umbach, and Leping Li. "A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data." *BMC genomics* 18 (2017): 1-13. https://doi.org/10.1186/s12864-017-3906-0.

54. Liu, Yangyang, Lu Liang, Liang Ji, Fuquan Zhang, Donglai Chen, Shanzhou Duan, Hao Shen, Yao Liang, and Yongbing Chen. "Potentiated lung adenocarcinoma (LUAD) cell growth, migration and invasion by lncRNA DARS-AS1 via miR-188-5p/KLF12 axis." *Aging (Albany NY)* 13, no. 19 (2021): 23376. DOI: 10.18632/aging.203632.

55. Yang, Jian, Zhike Chen, Zetian Gong, Qifan Li, Hao Ding, Yuan Cui, Lijuan Tang et al. "Immune landscape and classification in lung adenocarcinoma based on a novel cell cycle checkpoints related signature for predicting prognosis and therapeutic response." *Frontiers in Genetics* 13 (2022): 908104. https://doi.org/10.3389/fgene.2022.908104.

56. Liu, Qian, Jiali Lei, Xiaobo Zhang, and Xiaosheng Wang. "Classification of lung adenocarcinoma based on stemness scores in bulk and single cell transcriptomes." *Computational and Structural Biotechnology Journal* 20 (2022): 1691-1701. https://doi.org/10.1016/j.csbj.2022.04.004.

57**.** Zhao, Xin, Daixing Hu, Jia Li, Guozhi Zhao, Wei Tang, and Honglin Cheng. "Database mining of genes of prognostic value for the prostate adenocarcinoma microenvironment using the cancer gene atlas." *BioMed research international* 2020 (2020). https://doi.org/10.1155/2020/5019793.

58. Khosravi, Pegah, Maria Lysandrou, Mahmoud Eljalby, Qianzi Li, Ehsan Kazemi, Pantelis Zisimopoulos, Alexandros Sigaras et al. "A deep learning approach to diagnostic classification of

prostate cancer using pathology–radiology fusion." *Journal of Magnetic Resonance Imaging* 54, no. 2 (2021): 462-471. https://doi.org/10.1002/jmri.27599.

59. Basilevsky, Alexander T. *Statistical factor analysis and related methods: theory and applications*. John Wiley & Sons, 2009. DOI:10.1002/9780470316894

60. Everitt, Brian, and Graham Dunn. *Applied multivariate data analysis*. Vol. 2. London: Arnold, 2001. DOI:10.1002/9781118887486.

61. Pearson, Karl. "LIII. On lines and planes of closest fit to systems of points in space." *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2, no. 11 (1901): 559-572. https://doi.org/10.1080/14786440109462720.

62. Hilsenbeck, Susan G., William E. Friedrichs, Rachel Schiff, Peter O'Connell, Rhonda K. Hansen, C. Kent Osborne, and Suzanne AW Fuqua. "Statistical analysis of array expression data as applied to the problem of tamoxifen resistance." *Journal of the National Cancer Institute* 91, no. 5 (1999): 453-459. https://doi.org/10.1093/jnci/91.5.453.

63. Vohradsky, Jiří, Xin-Ming Li, and Charles J. Thompson. "Identification of procaryotic developmental stages by statistical analyses of two-dimensional gel patterns." *Electrophoresis* 18, no. 8 (1997): 1418-1428. https://doi.org/10.1002/elps.1150180817.

64. Craig, J. C., J. H. Eberwine, J. A. Calvin, B. Wlodarczyk, G. D. Bennett, and R. H. Finnell. "Developmental expression of morphoregulatory genes in the mouse embryo: an analytical approach using a novel technology." *Biochemical and molecular medicine* 60, no. 2 (1997): 81-91. https://doi.org/10.1006/bmme.1997.2576.

65. Liu, JingJing, WenSheng Cai, and XueGuang Shao. "Cancer classification based on microarray gene expression data using a principal component accumulation method." *Science China Chemistry* 54 (2011): 802-811. https://doi.org/10.1007/s11426-011-4263-5.

66. Oladejo, Ayomikun Kubrat, Tinuke Omolewa Oladele, and Yakub Kayode Saheed. "Comparative evaluation of linear support vector machine and K-nearest neighbour algorithm using microarray data on leukemia cancer dataset." *Afr. J. Comput. ICT* 11, no. 2 (2018): 1-10. https://afrjcict.net/2017/08/29/african-journal-of-computing-ict/.

67. Adebiyi, Marion Olubunmi, Micheal Olaolu Arowolo, Moses Damilola Mshelia, and Oludayo O. Olugbara. "A linear discriminant analysis and classification model for breast cancer diagnosis." *Applied Sciences* 12, no. 22 (2022): 11455. https://doi.org/10.3390/app122211455.

68. Ak, Muhammet Fatih. "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications." In *Healthcare*, vol. 8, no. 2, p. 111. MDPI, 2020. https://doi.org/10.3390/healthcare8020111.

69. Díaz-Uriarte, Ramón, and Sara Alvarez de Andrés. "Gene selection and classification of microarray data using random forest." *BMC bioinformatics* 7 (2006): 1-13. https://doi.org/10.1186/1471-2105-7-3.

70. Tan, Aik Choon, and David Gilbert. "Ensemble machine learning on gene expression data for cancer classification." (2003). URL: http://bura.brunel.ac.uk/handle/2438/3013.

71. Sharma, Alok, Seiya Imoto, Satoru Miyano, and Vandana Sharma. "Null space based feature selection method for gene expression data." *International Journal of Machine Learning and Cybernetics* 3 (2012): 269-276. https://doi.org/10.1007/s13042-011-0061-9.

72. Degroeve, Sven, Bernard De Baets, Yves Van de Peer, and Pierre Rouzé. "Feature subset selection for splice site prediction." *Bioinformatics* 18, no. suppl_2 (2002): S75-S83.7. 10.1093/bioinformatics/18.suppl_2.s75

73. Peng, Yanxiong, Wenyuan Li, and Ying Liu. "A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification." *Cancer informatics* 2 (2006): 117693510600200024. https://doi.org/10.1177/117693510600200024.

74. Sharma, Alok, and Kuldip K. Paliwal. "Cancer classification by gradient LDA technique using microarray gene expression data." *Data & Knowledge Engineering* 66, no. 2 (2008): 338-347. https://doi.org/10.1016/j.datak.2008.04.004.

75. Bar-Joseph, Ziv, Anthony Gitter, and Itamar Simon. "Studying and modelling dynamic biological processes using time-series gene expression data." *Nature Reviews Genetics* 13, no. 8 (2012): 552-564. https://doi.org/10.1038/nrg3244.

76. Cho, Ji-Hoon, Dongkwon Lee, Jin Hyun Park, and In-Beum Lee. "Gene selection and classification from microarray data using kernel machine." *FEBS letters* 571, no. 1-3 (2004): 93-98. https://doi.org/10.1016/j.febslet.2004.05.087.

77. Huang, Desheng, Yu Quan, Miao He, and Baosen Zhou. "Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data." *Journal of experimental & clinical cancer research* 28 (2009): 1-8. https://doi.org/10.1186/1756-9966-28-149.

78. Dwivedi, Ashok Kumar. "Artificial neural network model for effective cancer classification using microarray gene expression data." *Neural Computing and Applications* 29 (2018): 1545-1554. https://doi.org/10.1007/s00521-016-2701-1.

79. Sun, Yingshuai, Sitao Zhu, Kailong Ma, Weiqing Liu, Yao Yue, Gang Hu, Huifang Lu, and Wenbin Chen. "Identification of 12 cancer types through genome deep learning." *Scientific reports* 9, no. 1 (2019): 17256. https://doi.org/10.1038/s41598-019-53989-3.

80. Alhenawi, Esra'A., Rizik Al-Sayyed, Amjad Hudaib, and Seyedali Mirjalili. "Feature selection methods on gene expression microarray data for cancer classification: A systematic review." *Computers in Biology and Medicine* 140 (2022): 105051. https://doi.org/10.1016/j.compbiomed.2021.105051.

81. Khatun, R., Akter, M., Islam, M.M., Uddin, M.A., Talukder, M.A., Kamruzzaman, J., Azad, A.K.M., Paul, B.K., Almoyad, M.A.A., Aryal, S. and Moni, M.A., 2023. Cancer classification utilizing voting classifier with ensemble feature selection method and transcriptomic data. *Genes*, *14*(9), p.1802. https://doi.org/10.3390/genes14091802.

82. Osama, Sarah, Hassan Shaban, and Abdelmgeid A. Ali. "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review." *Expert Systems with Applications* 213 (2023): 118946. https://doi.org/10.1016/j.eswa.2022.118946.

83. Kabir, Md Faisal, Tianjie Chen, and Simone A. Ludwig. "A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction." *Healthcare Analytics* 3 (2023): 100125. https://doi.org/10.1016/j.health.2022.100125.

84. Kharya, S., D. Dubey, and S. Soni. "Predictive machine learning techniques for breast cancer detection." *International journal of computer science and information Technologies* 4, no. 6 (2013): 1023-8.

85. Rana, Mandeep, Pooja Chandorkar, Alishiba Dsouza, and Nikahat Kazi. "Breast cancer diagnosis and recurrence prediction using machine learning techniques." *International journal of research in Engineering and Technology* 4, no. 4 (2015): 372-376. DOI:10.15623/ijret.2015.0404066

86. Johnson, Richard Arnold, and Dean W. Wichern. "Applied multivariate statistical analysis." (2002). https://books.google.com/books?id=gFWcQgAACAAJ.

87. Genuer, Robin, Jean-Michel Poggi, Robin Genuer, and Jean-Michel Poggi. *Random forests*. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-56485-8_3

88. Frank, Andrew. "UCI machine learning repository." *http://archive. ics. uci. edu/ml* (2010). DOI: 10.24432/C5R88H.

89. Team, R. Core. "R: A language and environment for statistical computing. R Foundation for Statistical Computing." *(No Title)* (2013). https://www.R-project.org/

90. Jolliffe, Ian T., and Jorge Cadima. "Principal component analysis: a review and recent developments." *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374, no. 2065 (2016): 20150202. https://doi.org/10.1098/rsta.2015.0202.

91. Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. New York: springer, 2009. https://doi.org/10.1007/978-0-387-21606-5.

92. Molin, Nicole L., Clifford Molin, Rohan J. Dalpatadu, and Ashok K. Singh. "Prediction of obstructive sleep apnea using Fast Fourier Transform of overnight breath recordings." *Machine Learning with Applications* 4 (2021): 100022. https://doi.org/10.1016/j.mlwa.2021.100022.

Case Study- 4

USING SUPERVISED LEARNING TECHNIQUES TO PREDICT VAPING QUITTING BEHAVIOUR AMONG YOUNG ADULTS IN THE UNITED STATES

Dwaipayan Mukhopadhyay, Manoj Sharma, Kavita Batra, Ashok Singh

Abstract: The substantial rise in electronic cigarettes, known as e-cigarette or vaping, has become a significant public health concern especially among young people. Quitting vaping is a difficult problem to solve, people try to quit at least once a year with little success. Hence in this study we aim to discover the potential factors behind the initiation and then possibly sustain the desire to quit smoking using Linear Discriminant Analysis (LDA) & Random Forest (RF) method. A dataset of 619 observations including two response variables, Initiation and Sustenance were utilized to fit both RF and LDA after splitting the dataset into an 80% training set and 20% test set. RF method gave us significantly better accuracy measures compared to that of LDA method in both test and training dataset.

Introduction: Daily usage of electronic nicotine products (ie, e-cigarettes, vaping) was reported as 11.7% and current usage being 25% in high school students, as per studies conducted in 2019 [5]. Youngsters vaping are at risk for nicotine addiction, toxicant exposure, and potential transition to cigarettes. [1,2]. It is necessary to assess initiation in quitting and sustaining attempts in this population to guide treatment development.

Some studies have shown most people who have tried vaping do not continue to use the device in the long run [13, 14]. It is therefore imperative to identify the minute group of users who are likely to become long-term vapers as this may indicate vaping dependency which could lead to chronic health effects. Prior studies [14-18] have suggested a set of characteristics that may be unique to current vapers, including younger age, females as well as initiating vaping due to lower cost as well as certain flavors. Adolescents are quite easily attracted to e-cigarette flavors and are more likely to sustain vaping usage assuming it to have lower risks implying that individual level variables are correlated [19]. Hence it is difficult to seclude a set of independent predictors of current vaping using just regression. Owing to these circumstances' multicollinearity became a pivotal issue, which should be handled carefully prompting the usage of more advanced statistical techniques.

Environments of tobacco research are increasingly complicated and advanced analytical tools are required to tackle big volumes of data [28]. Supervised machine learning is one such technique with increasing popularity in health research [6-8], [20-22]. Attenuating model overfitting, high accuracy alongside robust predictions makes machine learning a very convenient alternative compared to traditional regression [12].

Another appealing factor is being able to make distinctive and meaningful associations in a flexible and exploratory manner simultaneously skipping usual distributional assumptions [29,30].

Applications of machine learning in tobacco research are emerging in recent years [9-11], [23-26],[28,29] and vaping [3-4]. Studies in some other fields of tobacco research have shown classification trees demonstrating decent performance in the status of smoking cessation status [20] and cohesion to nicotine replacement therapy [21].

Linear discriminant analysis [30,31] is another supervised algorithm providing better results in a multi-class classification task with known class labels. [32] shows the usage of linear discriminant analysis in studying situational features associated with having or not having the urge to smoke while attempting to quit.

The problem at hand, as explained in the Data and Methods Section, is a 5-level classification problem. There are two statistical classification methods which can be applied in such a situation: multinomial logistic regression and the Linear Discriminant Analysis (LDA). Multinomial logistic regression is not being used since LDA is shown to yield better accuracy than multinomial logistic regression when the number of levels of the response variable is greater than 4 [33].

Data and Methods: The dataset used in this study has 619 observations on a total of 31 variables including two response variables, Initiation and Sustenance:

Initiation = Initiation of quitting vaping

Sustenance = intent to sustain vaping quitting behavior

There are a total of eight constructs, created from 29 available features in the dataset:

BC_Overall = BC1+BC2+BC3+BC4+BC5 (Construct of behavioral confidence)

PE_Overall = PE1+PE2+PE3+PE4+PE5 (Construct of changes in the physical environment)

ET_Overall = ET1+ET2+ET3 (Construct of emotional transformation)

PC_Overall = PC1+PC2+PC3 (Construct of practice for change)

SE_Overall = SE1+SE2+SE3 (Construct of changes in the social environment)

A_Sum = A1+A2+A3+A4+A5 (sum of all advantages)

D_Sum = D1+D2+D3+D4+D5 (sum of all disadvantages)

PD = A_Sum - D_Sum

LDA is a latent variable supervised learning method used for dimensionality reduction and robust classification; the LDA features do not depend upon multivariate normality of the features, giving LDA the robustness property. LDA projects the data on a lower-dimension space which maximizes the distance between response levels or classes. The linear discriminants yield the maximum ratio of between-class variance to within-class variance [34].

Random Forest (RF) is a supervised machine learning method for classification and regression which uses bagging (bootstrap aggregation) for variance reduction of a decision tree. RF randomly selects subsets of features and subsets of data for training of decision trees, and in the case of classification, RF uses a voting method to predict with the final class for the response variable [35].

In this study, the LDA classifier, and the decision-tree based Random Forest (RF) classifier are used to predict the response variables Initiation and Sustenance as functions of the eight predictors or feature shown above.

The entire 619x10 dataset consisting of the 2 responses and 8 features was split into an 80% training set and 20% test set. Both LDA and RF models were fitted to the 2 response variables separately, and accuracies were computed for both training and test data sets.

The accuracy measures of a multi-class classifier are computed from the confusion matrices for training and test data. The accuracy measures for a multi-class classifier are Overall Accuracy, Average Precision, Recall, F1 measures, and approximate Areas Under the Curve (AUC) as explained in [36].

$C_{i,j}$ = number of times true response of level j get predicted as i; i, j = 0,1, …, 4).

The performance measures accuracy, precision, recall, F1 and the overall prediction accuracy [36] are given by:

Precision, Recall, and F1 measures for each category , $j = 0, 2, ..., 4$ are defined below:

$$\text{Precision}_{\text{j}} \;=\; \frac{C_{j,j}}{\sum_{k=0}^{4} C_{j,k}}$$

$$\text{Recall}_{\text{j}} \;=\; \frac{C_{j,j}}{\sum_{k=1}^{4} C_{k,j}}$$

$$\text{F1}_{j} \;=\; \frac{2 \times \text{Precision}_{\text{j}} \times \text{Recall}_{\text{j}}}{(\text{Precision}_{\text{j}} + \text{Recall}_{\text{j}})} = \text{harmonic mean of Precision and Recall}$$

$$\text{Overall prediction accuracy} = \frac{\sum_{j=0}^{4} C_{j,j}}{\sum_{i=0}^{4}\sum_{j=0}^{4} C_{i,j}}$$

Figure 42: Performance Measures for Classification Problem

Area Under the Curve (AUC) is a measure of accuracy for binary classification and can be computed from binary vonfusion matrix (CM) for each class *j=0,i,...,4* from the following formula

$$AUC_j = \frac{1}{2}\left(\frac{TP_j}{TP_j + FN_j} + \frac{TN_j}{TN_j + FP_j} \right)$$

Figure 43: AUC Measure Formulae

where $TP\ j$ = true positive, $TN\ j$ = true negative, $FP\ j$ = false positive and $FN\ j$ = false negative for the jth class, as shown in the confusion matrix CM_j for class j (j = 1, ...,4):

$$\begin{array}{cc} & \textit{Observed} \\ CM_j = & \begin{bmatrix} \Pr edicted & C_j = 1 & C_j = 0 \\ C_j = 1 & TP_j & FP_j \\ C_j = 0 & FN_j & TN_j \end{bmatrix} \end{array}$$

Figure 44: Confusion Matrix Formulae

Micro- and Macro- averages [36] of the above measures are also included.

The Overall Accuracy (OA) is calculated from the confusion matrix by using the formula

OA = sum of diagonal elements of CM/ sum of all elements of CM

Results: Summary statistics of the 2 response variables and the 8 predictors are shown in Table 14.

Table 14 : Summary Statistics

Variable	n	mean	sd	median	min	max
Initiation	619	1.89	1.32	2	0	4
Sustenance	619	1.74	1.35	2	0	4
PD	619	3.94	6.01	3	-16	20
BC_Overall	619	8.84	6.02	9	0	20
PE_Overall	619	10.09	5.56	10	0	20
ET_Overall	619	5.98	3.55	6	0	12
PC_Overall	619	5.77	3.43	6	0	12
SE_Overall	619	6.05	3.40	6	0	12

The feature variables are summarized visually as well, with Figures 45,46 showing box plots of the two responses and the features. Figure 3 shows the frequencies (counts) of the two response variables in 5 categories 0, 1, …, 4.

Figure 45 shows that the variability in Sustenance is higher than that in Initiation.

Figure 46 shows that there are enough observations in each level of the two response variables, which implies that classification into the 5 classes will not suffer from having too few observations in any one class.

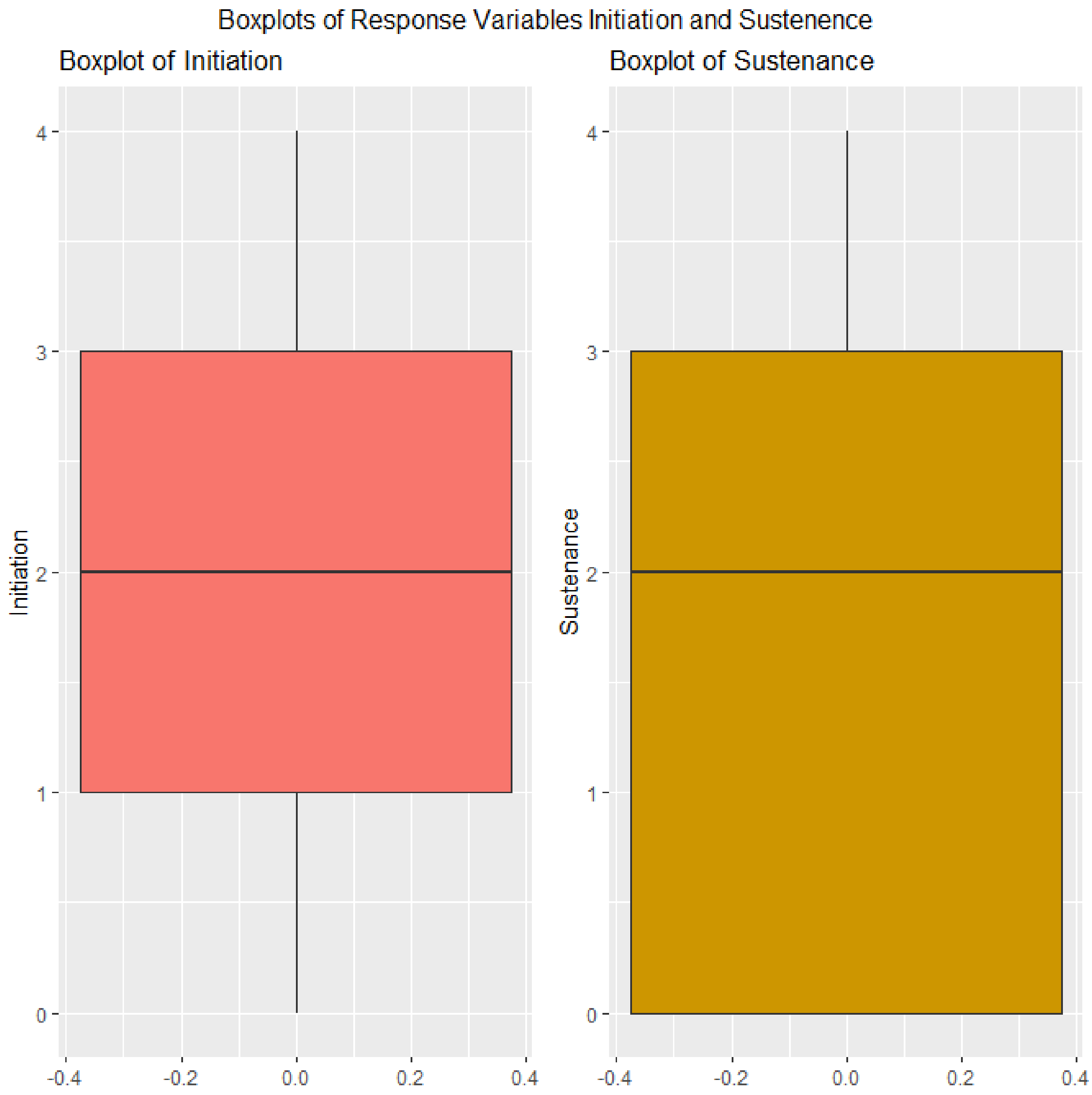


Figure 45 : Boxplots of the Response Variables Initiation and Sustenance

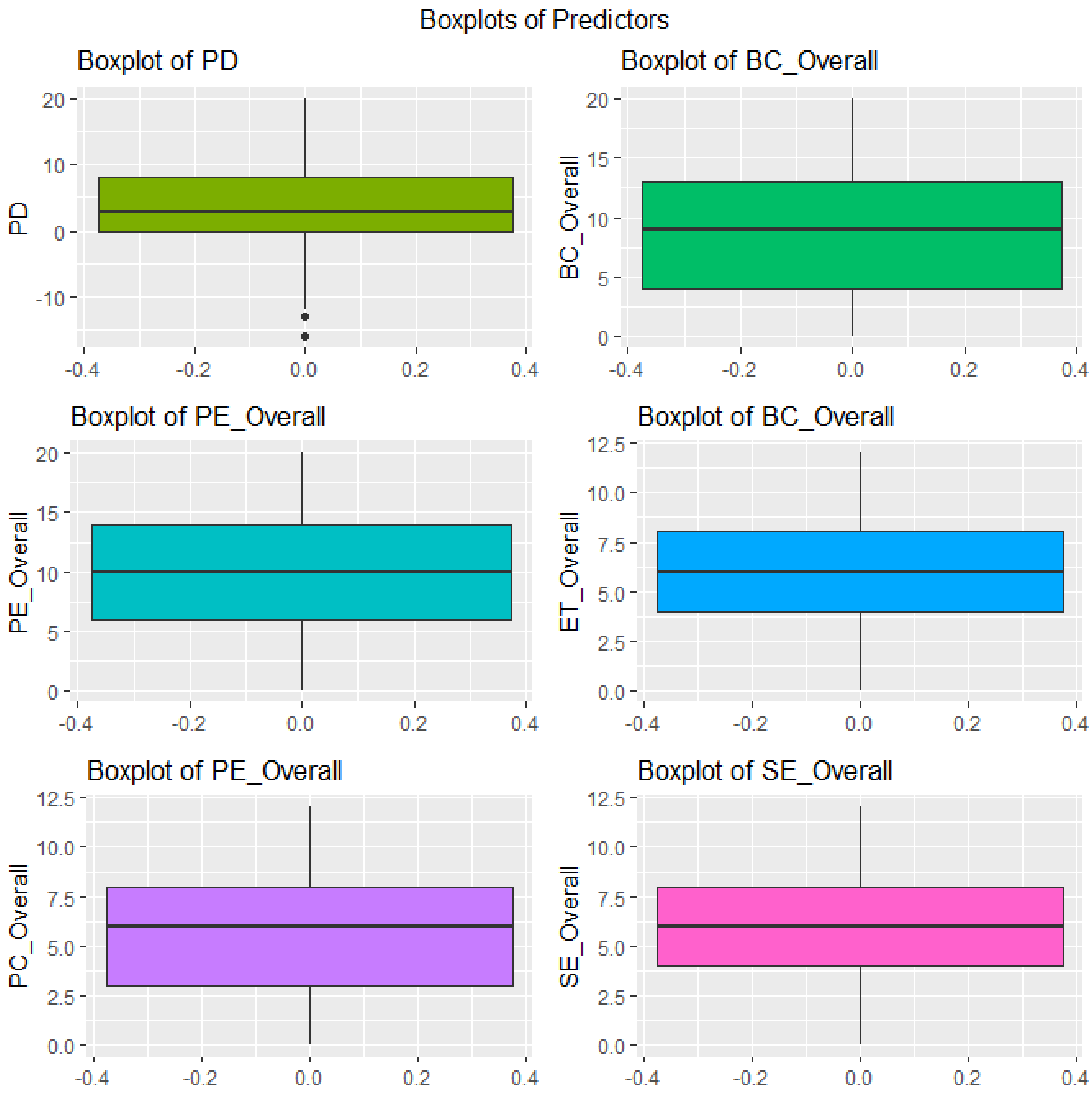


Figure 46: Boxplots of the Predictors

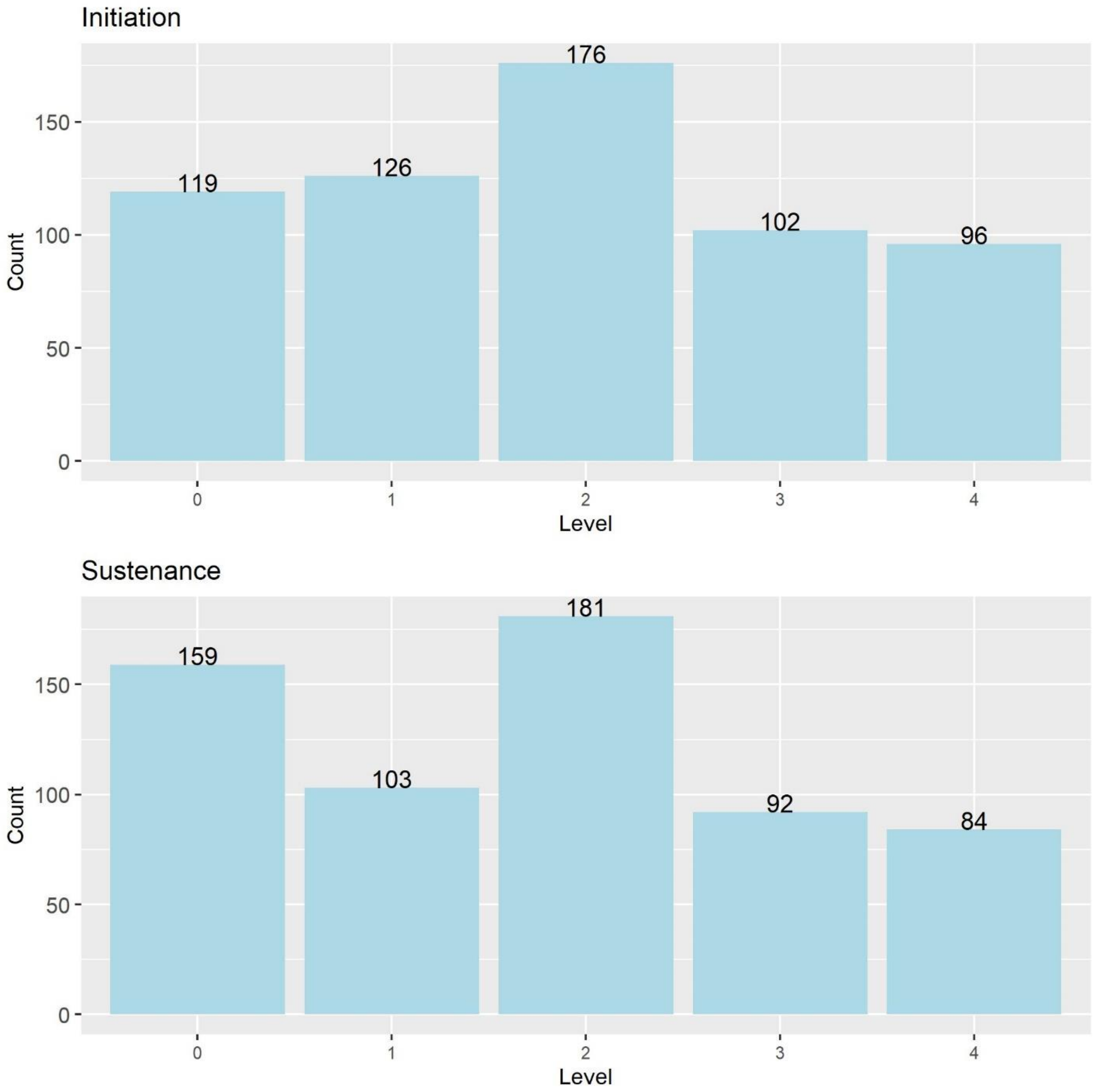


Figure 47: Bar Charts of the Level Counts for Initiation and Sustenance

LDA Classification Results: Figures 4-5 show LDA score plots for Initiation and Sustenance, respectively. In both cases, the first two linear discriminants explain more than 99% of total variability in the data. Figures 4 and 5 also show that separation between classes is not clear. This is reflected in low values of accuracy measures of the LDA classifier.

Accuracy Measures of LDA Classification for Initiation: As mentioned earlier, all accuracy measures are computed from confusion matrix of classification. The confusion matrices of LDA for the training and test sets for prediction of Initiation are shown in Tables 14 and 15. The LDA qualifier does not have good overall accuracy for either of the two response variables.

Table 15: Confusion Matrix for the LDA Classifier (Train Data)

	Observed				
Predicted	0	1	2	3	4
0	69	33	16	7	3
1	1	0	0	0	0
2	19	61	101	50	24
3	0	2	2	5	1
4	10	5	14	19	54

Overall Accuracy for LDA Classifier for Training Set = 0.46

Table 16: Confusion Matrix for LDA Classifier (Test Data)

	Observed				
Predicted	0	1	2	3	4
0	13	14	9	0	0
1	0	0	0	0	0
2	6	10	30	16	3
3	0	0	0	2	1
4	1	1	4	3	10

Overall Accuracy for LDA Classifier for Test Set = 0.45

The binary accuracy for class k (k=0,1,2,3,4) are calculated from the full confusion matrices as follows:

Table 17 : Confusion Matrix

Class	Observed Class k	Observed Class Other
k	$N_{k,k}$	$N_{k,Other}$
Other	$N_{Other,,k}$	$N_{Other,Other}$

Binary Overall Accuracy of Class k = $(N_{k,k}+ N_{Other,Other})//(N_{k,k}+ N_{k,Other} +N_{Other,Other}+ N_{Other,Other})$

The binary confusion matrices for each class are computed from the CM of Tables 14 and 15 above, and binary overall accuracies of LDA classifier for both training and test sets are calculated; these values are shown in Table 18.

Table 18: Binary Accuracies of LDA Classifier for Both Training and Test Sets for Initiation

	0	1	2	3	4
Training	82.06%	79.44%	62.50%	83.67%	84.68%
Test	75.61%	79.67%	60.98%	83.74%	89.43%

Table 19: Macro and Micro Averages of Precision, Recall and F1 and AUC's for Each Class for Training Set for Initiation for LDA Classifier

Micro-macro average		Class	AUC
macro.Precision	39.29%	0	77.42%
macro.Recall	43.53%	1	49.87%
macro.F1	41.30%	2	66.76%
micro.Precision	46.17%	3	52.48%
micro.Recall	46.17%	4	77.13%
micro.F1	46.17%	AUC.macro	64.73%

Table 20: Macro and Micro Averages of Precision, Recall and F1 and AUC's for Each Class for Test Set for Initiation for LDA Classifier

Micro-macro average		Class	AUC
macro.Precision		0	71.33%
macro.Recall	43.14%	1	50.00%
macro.F1		2	63.01%
micro.Precision	44.72%	3	54.27%
micro.Recall	44.72%	4	81.59%
micro.F1	44.72%	AUC.macro	64.04%

Table 21: Binary Accuracies of LDA Classifier for Both Training and Test Sets for Initiation

	0	1	2	3	4
Training	76.01%	83.67%	60.48%	84.27%	88.31%
Test	81.30%	78.86%	62.60%	86.18%	90.24%

Table 22: Macro and Micro Averages of Precision, Recall and F1 and AUC's for Each Class for Training Set for Sustenance for LDA Classifier

Micro-macro average		Class	AUC
macro.Precision	34.11%	0	73.84%
macro.Recall	39.70%	1	49.76%
macro.F1	36.70%	2	62.94%
micro.Precision	46.37%	3	50.19%
micro.Recall	46.37%	4	76.20%
micro.F1	46.37%	AUC.macro	62.59%

Table 23: Macro and Micro Averages of Precision, Recall and F1 and AUC's for Each Class for Test Set for Sustenance for LDA Classifier

Micro-macro average	42.71%	Class	AUC
macro.Precision		0	78.29%
macro.Recall	42.71%	1	48.99%
macro.F1		2	67.07%
micro.Precision	49.59%	3	50.00%
micro.Recall	49.59%	4	77.22%
micro.F1	49.59%	AUC.macro	64.32%

Accuracy Measures of LDA Qualifier: For LDA qualifier, Tables 4 and 7 show that binary accuracy for each class, other than class 2, is quite high for both of the responses for both training and test data sets. . Tables 5 and 8 show that micro- and macro- averages for Initiation fall in the range 39% - 46% for Initiation and 35% - 47% range for Sustenance; the AUC-values lie in 50%-71% range for Initiation, and 49% - 78% range for Sustenance Tables 19 and 22.

Table 24: Binary Accuracies of RF Classifier for Both Training and Test Sets for Initiation

	0	1	2	3	4
Training	95.36%	95.97%	93.35%	96.98%	96.98%
Test	84.55%	71.54%	59.35%	76.42%	87.80%

Table 25: Macro and Micro Averages of Precision, Recall and F1 and AUC's for Each Class for Training Set for Initiation for RF Classifier

Micro-macro average		Class	AUC
macro.Precision	90.32%	0	94.07%
macro.Recall	89.04%	1	92.31%
macro.F1	89.68%	2	92.83%
micro.Precision	89.31%	3	91.73%
micro.Recall	89.31%	4	94.77%
micro.F1	89.31%	AUC.macro	93.14%

Table 26: Macro and Micro Averages of Precision, Recall and F1 and AUC's for Each Class for Test Set for Initiation for RF Classifier

Micro-macro average	39.42%	Class	AUC
macro.Precision	42.37%	0	70.63%
macro.Recall	40.84%	1	52.35%
macro.F1	39.84%	2	55.31%
micro.Precision	39.84%	3	57.42%
micro.Recall	39.84%	4	80.67%
micro.F1	39.42%	AUC.macro	63.27%

Table 27: Binary Accuracies of RF Classifier for Both Training and Test Set for Sustenance

	0	1	2	3	4
Training	90.32%	89.72%	84.88%	92.34%	93.15%
Test	80.49%	75.61%	67.48%	80.49%	88.62%

Table 28: Macro and Micro Averages of Precision, Recall and F1 and AUC's for Each Class for Training Set for Sustenance for RF Classifier

Micro-macro average		Class	AUC
macro.Precision	77.01%	0	88.86%
macro.Recall	71.61%	1	73.36%
macro.F1	74.21%	2	85.78%
micro.Precision	75.20%	3	80.15%
micro.Recall	75.20%	4	84.48%
micro.F1	75.20%	AUC.macro	82.53%

Table 29: Macro and Micro Averages of Precision, Recall and F1 and AUC's for Each Class for Test Set for Sustenance for RF Classifier

Micro-macro average	41.79%	Class	AUC
macro.Precision	44.46%	0	76.37%
macro.Recall	43.08%	1	50.13%
macro.F1	46.34%	2	64.63%
micro.Precision	46.34%	3	61.51%
micro.Recall	46.34%	4	73.43%
micro.F1	41.79%	AUC.macro	65.21%

Accuracy Measures of RF Qualifier: For RF qualifier, Tables 23 and 26 show that binary accuracy for each class is higher than that for the corresponding LDA qualifier for both responses for training are around 90% and fall in the range 72% - 77% for Sustenance. Tables 24 and 27 show that, for Initiation, micro- and macro- averages are close to 90% for training and fall in 72% - 77% range for test set; the AUC-values are larger than 90% range for training data, and 50% - 77% range for Sustenance in Tables 19 and 22.

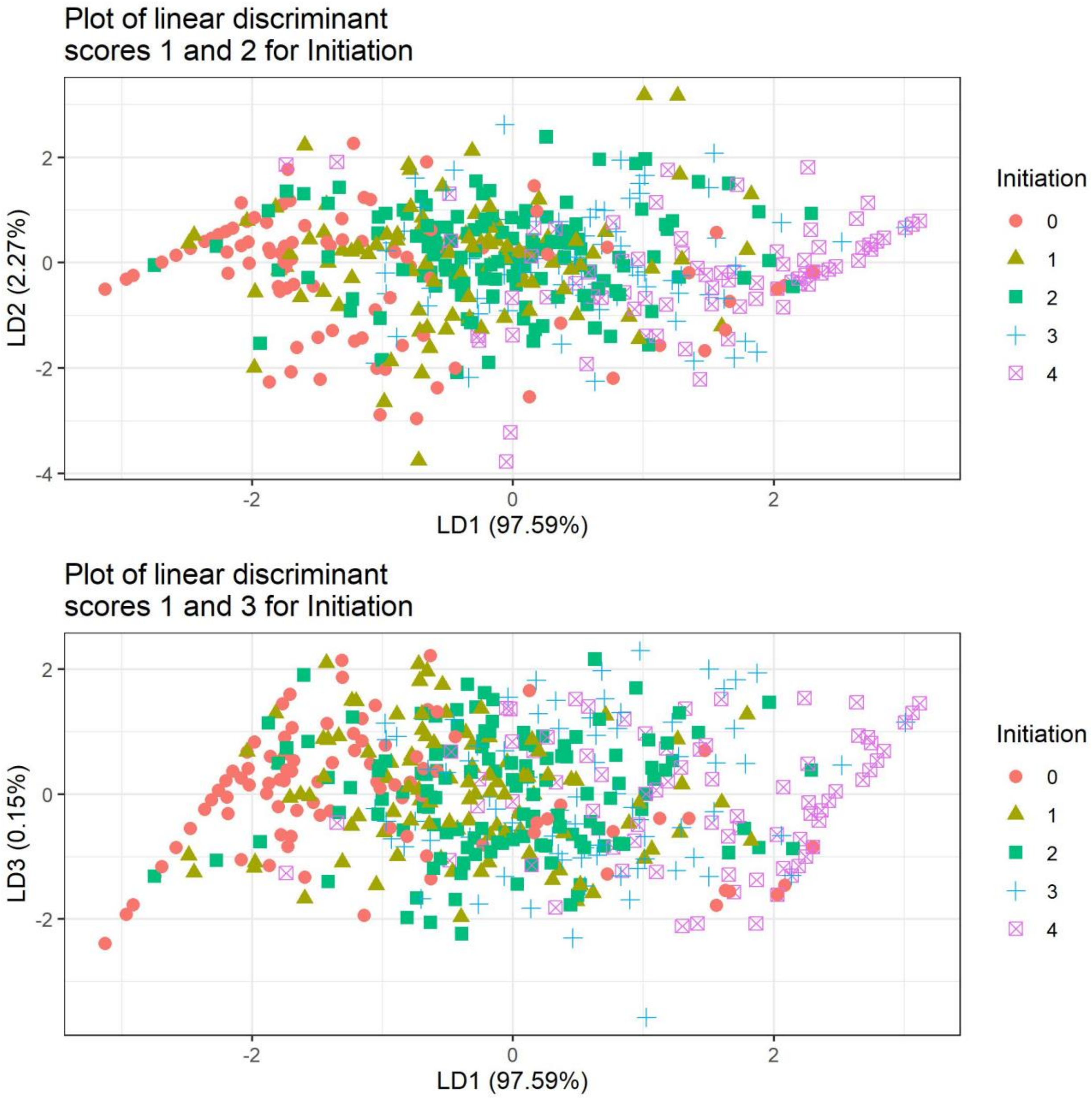
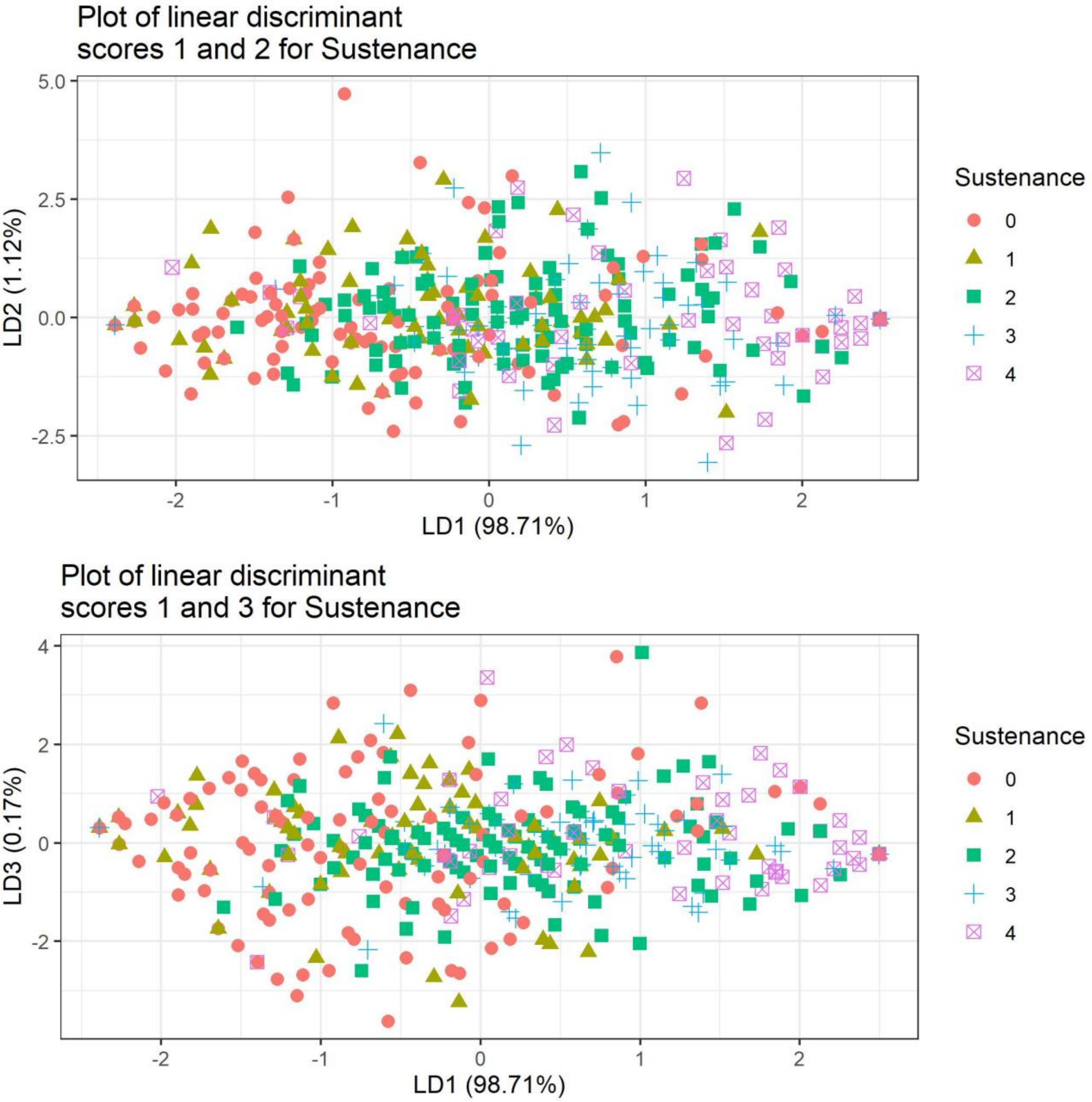


Figure 48 : LDA Score Plots for Initiation



Plot of linear discriminant scores 1 and 2 for Sustenance

Plot of linear discriminant scores 1 and 3 for Sustenance

Figure 49: LDA Score Plots for Sustenance

Figure 50, variable importance plot of the RF model for Initiation, shows that BC_Overall and PE_Overall are the important feature for prediction of Initiation. The situation is different for Sustenance with the 3 features ET_Overall, PC_Overall and SE_Overall being equally important for predicting Sustenance.

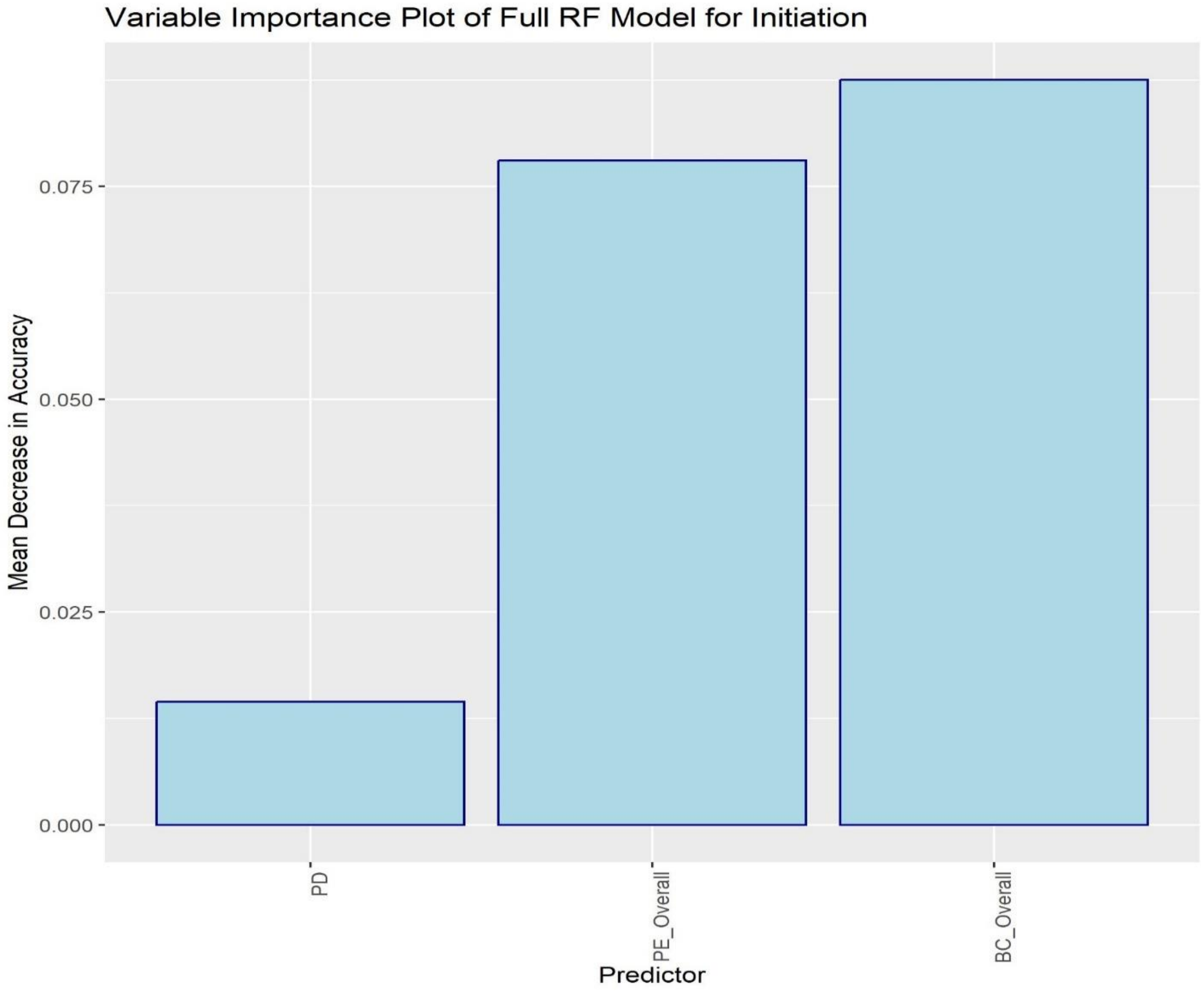


Figure 50: Variable Importance Plot for the Random Forest (RF) Model for Initiation

77

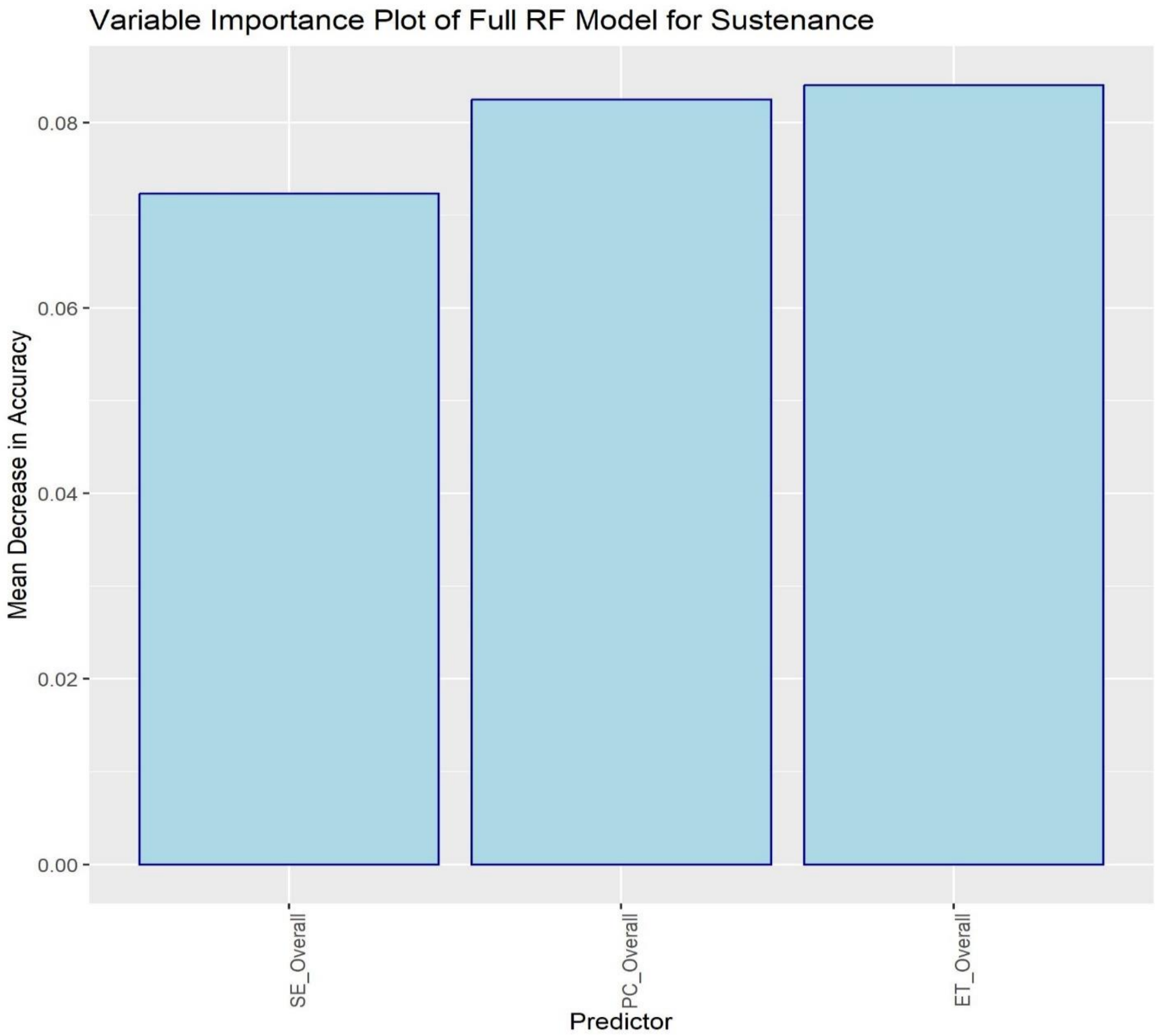


Figure 51: Variable Importance Plot for the Random Forest (RF) Model for Sustenance

References:

1. US Department of Health and Human Services (2016). *E-cigarette Use Among Youth and Young Adults: A Report of the Surgeon General*. US Department of Health and Human Services;.

2. National Academies of Sciences, Engineering, and Medicine (2018). *Public Health Consequences of E-cigarettes*. The National Academies Press.

3. Fu R, Mitsakakis N, Chaiton M. (2021). A machine learning approach to identify correlates of current e-cigarette use in Canada. Explor Med. ;2: 74–85.

4. Romijnders KAGJ, Pennings JLA, van Osch L, de Vries H, Talhout R .(2019). A combination of factors related to smoking behavior, attractive product characteristics, and socio-cognitive factors are important to distinguish a dual user from an exclusive e-cigarette user. Int J Environ Res Public Health. ;16: 4191. pmid:31671505

5. Miech R, Johnston L, O'Malley PM, Bachman JG, Patrick ME.(2017-19). Trends in adolescent vaping, . *N Engl J Med*. 2019;381(15):1490-1491. doi:10.1056/NEJMc1910739.

6. Morgenstern JD, Buajitti E, O'Neill M, Piggott T, Goel V, Fridman D, et all. (2019). Predicting population health with machine learning: a scoping review. BMJ Open. ;10: e037860. pmid:33109649

7. Mak KK, Lee K, Park C. (2019). Applications of machine learning in addiction studies: A systematic review. Psychiatry Res.;275: 53–60. pmid:30878857

8. Fu R, Kundu A, Mitsakakis N, Elton-Marshall T, Wang W, Hill S, et al. 2021 [cited 31 Aug 2021] Machine learning applications in tobacco research: a scoping review. Tob Control.. pmid:34452986

9. Coughlin LN, Tegge AN, Sheffer CE, Bickel WK. (2020). A machine-learning approach to predicting smoking cessation treatment outcomes. Nicotine Tob Res. ;22: 415–422. pmid:30508122

10. Dumortier A, Beckjord E, Shiffman S, Sejdić E. (2016). Classifying smoking urges via machine learning. Comput Methods Programs Biomed;137: 203–213. pmid:28110725

11. Suchting R, Hébert ET, Ma P, Kendzor DE, Businelle MS. (2019). Using elastic net penalized Cox proportional hazards regression to identify predictors of imminent smoking lapse. Nicotine Tob Res.;21: 173–179. pmid:29059349

12. Morgenstern JD, Buajitti E, O'Neill M, Piggott T, Goel V, Fridman D, et al. (2020).Predicting population health with machine learning: a scoping review. BMJ Open.;10: e037860. pmid:33109649

13. Shiplo S, Czoli CD, Hammond D. (2015). E-cigarette use in Canada: prevalence and patterns of use in a regulated market. BMJ Open.;5:e007971.

14. Delnevo CD, Giovenco DP, Steinberg MB, Villanti AC, Pearson JL, Niaura RS, et al. (2016).Patterns of electronic cigarette use among adults in the United States. Nicotine Tob Res.;18:715-9.

15. Bold KW, Kong G, Cavallo DA, Camenga DR, Krishnan-Sarin S. (2016). Reasons for trying e-cigarettes and risk of continued use. Pediatrics.;138: e20160895.

16. Camara-Medeiros A, Diemert L, O'Connor S, Schwartz R, Eissenberg T, Cohen JE. (2020). Perceived addiction to vaping among youth and young adult regular vapers. Tob Control.;[Epub ahead of print].

17. Notley C, Ward E, Dawkins L, Holland R. (2018). The unique contribution of e-cigarettes for tobacco harm reduction in supporting smoking relapse prevention. Harm Reduct J. ;15:31.

18. Landry RL, Groom AL, Vu TT, Stokes AC, Berry KM, Kesh A, et al. (2019). The role of flavors in vaping initiation and satisfaction among U.S. adults. Addict Behav. ;99:106077.

19. Montreuil A, MacDonald M, Asbridge M, Wild TC, Hammond D, Manske S, et al. (2017). Prevalence and correlates of electronic cigarette use among Canadian students: cross-sectional findings from the 2014/15 Canadian Student Tobacco, Alcohol and Drugs Survey. CMAJ Open. ;5:E460-7.

20. Mukhopadhyay, D., Phanord, D. D., Dalpatadu, R. J., Gewali, L. P., & Singh, A. K. (2023). Classification of Erythematosquamous Dermatosis by the Method of Random Forest. Journal of Dermatology Research Reviews & Reports, 4(1), 1-6.

21. Mak KK, Lee K, Park C. (2019). Applications of machine learning in addiction studies: a systematic review. Psychiatry Res.;275:53-60.

22. Sekercioglu N, Fu R, Kim SJ, Mitsakakis N. (2020). Machine learning for predicting long-term kidney allograft survival: a scoping review. Ir J Med Sci.; [Epub ahead of print].

23. Singh A, Katyan H. (2019). Classification of nicotine-dependent users in India: a decision-tree approach. J Public Health.;27:453-9.

24. Reps JM, Rijnbeek PR, Ryan PB. (2019). Supplementing claims data analysis using self-reported data to develop a probabilistic phenotype model for current smoking status. J Biomed Inform.;97:103264.

25. Kim N, McCarthy DE, Loh WY, Cook JW, Piper ME, Schlam TR, et al. (2019). Predictors of adherence to nicotine replacement therapy: machine learning evidence that perceived need predicts medication use. Drug Alcohol Depend.;205:107668.

26. Suchting R, Hébert ET, Ma P, Kendzor DE, Businelle MS. (2019). Using elastic net penalized cox proportional hazards regression to identify predictors of imminent smoking lapse. Nicotine Tob Res.;21:173-9.

27. Andueza, A.; Del Arco-Osuna, M.Á.; Fornés, B.; González-Crespo, R.; Martín-Álvarez, J.-M. (2023). Using the Statistical Machine Learning Models ARIMA and SARIMA to Measure the Impact of COVID-19 on Official Provincial Sales of Cigarettes in Spain. Int. J. Interact. Multimed. Artif. Intell., 8, 73–87

28. Fu, R.; Shi, J.; Chaiton, M.; Leventhal, A.M.; Unger, J.B.; Barrington-Trimis, J.L.(2022). A Machine Learning Approach to Identify Predictors of Frequent Vaping and Vulnerable Californian Youth Subgroups. Nicotine Tob. Res., 24, 1028–1036.

29. Shi, J.; Fu, R.; Hamilton, H.; Chaiton, M. (2022). A Machine Learning Approach to Predict E-Cigarette Use and Dependence among Ontario Youth. Health Promot. Chronic Dis. Prev. Can., 42, 21–28.

30. Duda RO, Hart PE, Stork DG.(2000) Pattern Classification. Wiley;.

31. Fukunaga K. (2013). Introduction to Statistical Pattern Recognition. Academic Press;.

32. Dumortier A, Beckjord E, Shiffman S, Sejdić E. (2016). Classifying smoking urges via machine learning. Comput Methods Programs Biomed. Dec;137:203-213. doi: 10.1016/j.cmpb.2016.09.016. Epub 2016 Sep 23. PMID: 28110725; PMCID: PMC5289882.

33. El-Habil, Abdalla (2014). A Comparative Study between Linear Discriminant Analysis and Multinomial Logistic Regression. DOI:10.35552/0247-028-006-008.

34. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning, Second Edition, Springer, N.Y. pp. 106-119.

35. Ashok K. Singh, Myongjee Yoo, & and Rohan J. Dalpatadu (2018). Determinants of Customer Satisfaction at the San Francisco International Airport. J Tourism Hospit, Vol. 8 Iss. 1 No: 397, pp. 1-8.

36.Molin, Nicole, Molin, Clifford, Dalpatadu, R.J., Singh, A. K. (2021). Prediction of

Obstructive Sleep Apnea Using FFT of Overnight Breath Recordings (2021).

Machine Learning with Applications, Volume 4, 15 June 2021, 100022

https://www.sciencedirect.com/science/article/pii/S2666827021000037

APPENDIX A – R CODES FOR ALL CASE STUDIES

```
setwd("F:/AK128_May 13 2015/AKS/AKS 2023/Grads/Dwaipayan/TCGA-PANCAN-
HiSeq-801x20531")
D <- readRDS(file="cancer_data.rds")
dim(D)  #  801 20532

#install.packages("Morpho")
library(Morpho)
pcafast <- prcompfast(D[,2:20532])

names(pcafast)
dim(pcafast$x)
head(pcafast$x)
summary(pcafast)
nrow(pcafast$roration)
str(pcafast$rotation)
# num [1:20531, 1:801] 0.00014 -0.00308 -0.00378 -0.00181 -0.00265 ...
# - attr(*, "dimnames")=List of 2
#  ..$ : chr [1:20531] "gene_0" "gene_1" "gene_2" "gene_3" ...
#  ..$ : chr [1:801] "PC1" "PC2" "PC3" "PC4" ...

write.csv(pcafast$rotation[,c(1,2,3)],"PC1_2.csv")

table(D$Class)
#BRCA COAD KIRC LUAD PRAD
# 300   78  146  141  136

P <- cbind.data.frame(D$Class, pcafast$x)
names(P)[1] <- "Class"
#
===================================================================
===============
set.ssed(31371)
# Split data into Training and Test sets
n.row <- nrow(P)
n.test <- trunc(0.2*n.row)
test.I <- sample(1:n.row,n.test,replace=FALSE)
P.test <- P[test.I,]
P.train <- P[-test.I,]
dim(P.test) # 160 802
dim(P.train)# 641 802
dim(P) # 801 802
#
===================================================================
==============================
#==================================================================
================================
# 01/05/24
#==================================================================
================================

library(MASS) # for LDA
library(ggplot2) # for all graphs
library(gridExtra) # for combining graphs
library(randomForest) # for random forest
#library(e1071) # for SVM and Naive Bayes

# ------------------------------------------------------------


###############################################################
# ==================================================================
# function to compute recall, precision, F1 for 5-level classification
PRF.multi <- function(CM)
{
A <- matrix(NA,nrow=5,ncol=3)
R1 <- CM[1,1]/sum(CM[,1]) # diag1/sum(column1)
R2 <- CM[2,2]/sum(CM[,2]) # diag2/sum(column2)
R3 <- CM[3,3]/sum(CM[,3]) # diag3/sum(column3)
R4 <- CM[4,4]/sum(CM[,4]) # diag4/sum(column4)
R5 <- CM[5,5]/sum(CM[,5]) # diag4/sum(column5)

P1 <- CM[1,1]/sum(CM[1,]) # diag1/sum(row1)
P2 <- CM[2,2]/sum(CM[2,]) # diag2/sum(row2)
P3 <- CM[3,3]/sum(CM[3,]) # diag3/sum(row3)
P4 <- CM[4,4]/sum(CM[4,]) # diag4/sum(row4)
P5 <- CM[5,5]/sum(CM[5,]) # diag4/sum(row5)


F1.1 <- 2*R1*P1/(P1+R1)
F1.2 <- 2*R2*P2/(P2+R2)
F1.3 <- 2*R3*P3/(P3+R3)
F1.4 <- 2*R4*P4/(P4+R4)
F1.5 <- 2*R5*P5/(P5+R5)
A[1,] <- c(P1,R1,F1.1)
A[2,] <- c(P2,R2,F1.2)
A[3,] <- c(P3,R3,F1.3)
A[4,] <- c(P4,R4,F1.4)
A[5,] <- c(P5,R5,F1.5)


colnames(A) <- c("Precision","Recall","F1")
rownames(A) <- c("BRCA","COAD","KIRC","LUAD","PRAD")
return(A)
}

# ========================================================
# Overall accuracy of multi-class classifier
OA <- function(CM)
{
result <- sum(diag(CM))/sum(CM)
result
}
#
================================================================
==
#
================================================================
==
# function to compute binary confusion matrices from 5-class CM
#
#                    Observed
# Pred       BRCA      COAD     KIRC      LUAD      PRAD
#    BRCA    CM[1,1]    CM[1,2]  CM[1,3]    CM[1,4]    CM[1,5]
#    COAD    CM[2,1]    CM[2,2]  CM[2,3]    CM[2,4]    CM[2,5]
#    KIRC    CM[3,1]    CM[3,2]  CM[3,3]    CM[3,4]    CM[3,5]
#    LUAD    CM[4,1]    CM[4,2]  CM[4,3]    CM[4,4]    CM[4,5]
#    PRAD    CM[5,1]    CM[5,2]  CM[5,3]    CM[5,4]    CM[5,5]

## binary CM
#Pred         C    not.C
# C          TP   FP
#not.C       FN   TN
#
```

```
TP <- vector()
FP <- vector()
FN <- vector()
TN <- vector()

binary.cm <- function(cm)
{
TP[1] <- cm[1,1]
FN[1] <- sum(cm[,1])-cm[1,1] # sum of column 1
FP[1] <- sum(cm[1,])-cm[1,1] # sum of row 1
TN[1] <- sum(cm)-(TP[1]+FP[1]+FN[1])

b1 <- c(TP[1],FN[1],FP[1],TN[1])
B1 <- matrix(b1, nrow = 2, ncol = 2)
#B1 <- as.data.frame(B1)

TP[2] <- cm[2,2]
FN[2] <- sum(cm[,2])-cm[2,2] # # sum of column 2
FP[2] <- sum(cm[2,])-cm[2,2] # sum of row 2
TN[2] <- sum(cm)-(TP[2]+FP[2]+FN[2])

b2 <- c(TP[2],FN[2],FP[2],TN[2])
B2 <- matrix(b2, nrow = 2, ncol = 2)

TP[3] <- cm[3,3]
FN[3] <- sum(cm[,3])-cm[3,3]
FP[3] <- sum(cm[3,])-cm[3,3]
TN[3] <- sum(cm)-(TP[3]+FP[3]+FN[3])

b3 <- c(TP[3],FN[3],FP[3],TN[3])
B3 <- matrix(b3, nrow = 2, ncol = 2)

TP[4] <- cm[4,4]
FN[4] <- sum(cm[,4])-cm[4,4]
FP[4] <- sum(cm[4,])-cm[4,4]
TN[4] <- sum(cm)-(TP[4]+FP[4]+FN[4])

b4 <- c(TP[4],FN[4],FP[4],TN[4])
B4 <- matrix(b4, nrow = 2, ncol = 2)

TP[5] <- cm[5,5]
FN[5] <- sum(cm[,5])-cm[5,5]
FP[5] <- sum(cm[5,])-cm[5,5]
```

```
TN[5] <- sum(cm)-(TP[5]+FP[5]+FN[5])

b5 <- c(TP[5],FN[5],FP[5],TN[5])
B5 <- matrix(b5, nrow = 2, ncol = 2)


# BRCA COAD KIRC LUAD PRAD

colnames(B1) <- c("Obs.BRCA","Obs.Other")
rownames(B1) <- c("Pred.BRCA","Pred.Other")

colnames(B2) <- c("Obs.COAD","Obs.Other")
rownames(B2) <- c("Pred.COAD","Pred.Other")

colnames(B3) <- c("Obs.KIRC","Obs.Other")
rownames(B3) <- c("Pred.KIRC","Pred.Other")

colnames(B4) <- c("Obs.LUAD","Obs.Other")
rownames(B4) <- c("Pred.LUAD","Pred.Other")

colnames(B5) <- c("Obs.PRAD","Obs.Other")
rownames(B5) <- c("Pred.PRAD","Pred.Other")

B12345 <- cbind.data.frame(B1,B2,B3,B4,B5)
rownames(B12345) <- c("Class","Other")
B12345
}
#
============================================================
==
# === above function computes binary CM's from multi-class CM =====

# -----------------------------------------------
# Precision, Recall, F1 for each class from binarized confusion matrix
# input BIN = output of binary.cm

PRF.binaries <- function(BIN)
{
CM1 <- BIN[,1:2]
CM2 <- BIN[,3:4]
CM3 <- BIN[,5:6]
CM4 <- BIN[,7:8]
CM5 <- BIN[,9:10]
```

```
PR1 <- CM1[1,1]/(CM1[1,1]+CM1[1,2])
PR2 <- CM2[1,1]/(CM2[1,1]+CM2[1,2])
PR3 <- CM3[1,1]/(CM3[1,1]+CM3[1,2])
PR4 <- CM4[1,1]/(CM4[1,1]+CM4[1,2])
PR5 <- CM5[1,1]/(CM5[1,1]+CM5[1,2])


Recall1 <- CM1[1,1]/(CM1[1,1]+CM1[2,1])
Recall2 <- CM2[1,1]/(CM2[1,1]+CM2[2,1])
Recall3 <- CM3[1,1]/(CM3[1,1]+CM3[2,1])
Recall4 <- CM4[1,1]/(CM4[1,1]+CM4[2,1])
Recall5 <- CM5[1,1]/(CM5[1,1]+CM5[2,1])

F11 <- 2*PR1*Recall1/(PR1+Recall1)
F12 <- 2*PR2*Recall2/(PR2+Recall2)
F13 <- 2*PR3*Recall3/(PR3+Recall3)
F14 <- 2*PR4*Recall4/(PR4+Recall4)
F15 <- 2*PR5*Recall5/(PR5+Recall5)

temp <-
c(PR1,Recall1,F11,PR2,Recall2,F12,PR3,Recall3,F13,PR4,Recall4,F14,PR5,Recall5,F15
)
result <- matrix(temp,nrow=5,ncol=3, byrow=TRUE)
colnames(result) <- c("Precision","Recall","F1")
rownames(result) <- c("BRCA","COAD","KIRC","LUAD","PRAD")
result
}


## -----------------------------------------------

macro_micro.avg <- function(BIN)
{
CM1 <- BIN[,1:2]
CM2 <- BIN[,3:4]
CM3 <- BIN[,5:6]
CM4 <- BIN[,7:8]
CM5 <- BIN[,9:10]

PR1 <- CM1[1,1]/(CM1[1,1]+CM1[1,2])
PR2 <- CM2[1,1]/(CM2[1,1]+CM2[1,2])
PR3 <- CM3[1,1]/(CM3[1,1]+CM3[1,2])
```

```
PR4 <- CM4[1,1]/(CM4[1,1]+CM4[1,2])
PR5 <- CM5[1,1]/(CM5[1,1]+CM5[1,2])


Recall1 <- CM1[1,1]/(CM1[1,1]+CM1[2,1])
Recall2 <- CM2[1,1]/(CM2[1,1]+CM2[2,1])
Recall3 <- CM3[1,1]/(CM3[1,1]+CM3[2,1])
Recall4 <- CM4[1,1]/(CM4[1,1]+CM4[2,1])
Recall5 <- CM5[1,1]/(CM5[1,1]+CM5[2,1])

F11 <- 2*PR1*Recall1/(PR1+Recall1)
F12 <- 2*PR2*Recall2/(PR2+Recall2)
F13 <- 2*PR3*Recall3/(PR3+Recall3)
F14 <- 2*PR4*Recall4/(PR4+Recall4)
F15 <- 2*PR5*Recall5/(PR5+Recall5)

macro.Prec <- (PR1+PR2+PR3+PR4+PR5)/5
macro.Recall <- (Recall1+Recall2+Recall3+Recall4+Recall5)/5
macro.F1 <- 2*macro.Prec*macro.Recall/(macro.Prec+macro.Recall)

micro.Prec_Num <- CM1[1,1]+CM2[1,1]+CM3[1,1]+CM4[1,1]+CM5[1,1]
micro.Prec_Den <- CM1[1,1]+CM2[1,1]+CM3[1,1]+CM4[1,1]+CM5[1,1] +
              CM1[1,2]+CM2[1,2]+CM3[1,2]+CM4[1,2]+CM5[1,2]
micro.Prec <- micro.Prec_Num/micro.Prec_Den

micro.Recall_Num <- CM1[1,1]+CM2[1,1]+CM3[1,1]+CM4[1,1]+CM5[1,1]
micro.Recall_Den <- CM1[1,1]+CM2[1,1]+CM3[1,1]+CM4[1,1]+CM5[1,1]+
               CM1[2,1]+CM2[2,1]+CM3[2,1]+CM4[2,1]+CM5[2,1]
micro.Recall <- micro.Recall_Num/micro.Recall_Den
micro.F1 <- 2*micro.Prec*micro.Recall/(micro.Prec+micro.Recall)

c(macro.Prec,macro.Recall,macro.F1,micro.Prec,micro.Recall,micro.F1)
}
# ========================================================
# ========================================================
# ------------- approximate AUC from confusion matrix -----
# 12/8/2020
# AUC from confusion matrix
# TP <- CM[1,1]
# TN <- CM[2,2]
# FP <- CM[1,2]
# FN <- CM[2,1]
# AUC.1 <- TP/(TP+FN)
# AUC.2 <- TN/(TN+FP)
# AUC <- (AUC.1+AUC.2)/2

AUC.macro <- function(CM1, CM2, CM3, CM4, CM5)
{
TP1 <- CM1[1,1]
TN1 <- CM1[2,2]
FP1 <- CM1[1,2]
FN1 <- CM1[2,1]
AUC1.1 <- TP1/(TP1+FN1)
AUC1.2 <- TN1/(TN1+FP1)
AUC1 <- (AUC1.1+AUC1.2)/2

TP2 <- CM2[1,1]
TN2 <- CM2[2,2]
FP2 <- CM2[1,2]
FN2 <- CM2[2,1]
AUC2.1 <- TP2/(TP2+FN2)
AUC2.2 <- TN2/(TN2+FP2)
AUC2 <- (AUC2.1+AUC2.2)/2

TP3 <- CM3[1,1]
TN3 <- CM3[2,2]
FP3 <- CM3[1,2]
FN3 <- CM3[2,1]
AUC3.1 <- TP3/(TP3+FN3)
AUC3.2 <- TN3/(TN3+FP3)
AUC3 <- (AUC3.1+AUC3.2)/2

TP4 <- CM4[1,1]
TN4 <- CM4[2,2]
FP4 <- CM4[1,2]
FN4 <- CM4[2,1]
AUC4.1 <- TP4/(TP4+FN4)
AUC4.2 <- TN4/(TN4+FP4)
AUC4 <- (AUC4.1+AUC4.2)/2

TP5 <- CM5[1,1]
TN5 <- CM5[2,2]
FP5 <- CM5[1,2]
FN5 <- CM5[2,1]
AUC5.1 <- TP5/(TP5+FN5)
AUC5.2 <- TN5/(TN5+FP5)
AUC5 <- (AUC5.1+AUC5.2)/2

#c(AUC1,AUC2,AUC3,AUC4)
result <- (AUC1+AUC2+AUC3+AUC4+AUC4)/5
c(AUC1,AUC2,AUC3,AUC4,AUC5,result)
}
# ==============================================================
# ==============================================================
# as.formula(paste("y ~ x1 + x2", "x3", sep = "+"))

formula2 <- as.formula(Class ~  PC1+PC2)
formula2

formula3 <- as.formula(Class ~  PC1+PC2+PC3)
formula3

formula5 <- as.formula(Class ~  PC1+PC2+PC3+PC4+PC5)
formula5

formula10 <- as.formula(Class ~
PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10)
formula10

# ==============================================================
head(D[,1])
head(D[,2])
head(D[,20532])
head(D[,20533])

#Start here  12/27/23  8:54 AM
################ Classification using raw data #################
# =============== using linear DA on raw data ================
LDA.raw <- lda(Class~., data=D)
# Error: protect(): protection stack overflow

# ######################### Random Forest on Raw Data ##########
#library(randomForest)
set.seed(1137311)

P.train$Class <- as.factor(P.train$Class)
P.test$Class <- as.factor(P.test$Class)
set.seed(11713)
rf0 <- randomForest(Class~., ntree = 250,importance = TRUE, data=D)
```

```
#Error: protect(): protection stack overflow

# ============================================================
####################### 5-level Classification ############################
#LDA Training Set
#using linear DA on 2 PC-Scores
# -------------------------------------------------------------
LDA2 <- lda(formula2,data=P.train)
prop.LDA2 = LDA2$svd^2/sum(LDA2$svd^2)
prop.LDA2 <- round(20*prop.LDA2,2)
prop.LDA2  #  22.47 29.28 20.74 14.41
sum(prop.LDA2)

p2.train <- predict(LDA2, P.train)$class
table(p2.train)

CM.LDA2.train <- table(p2.train,as.factor(P.train$Class))
CM.LDA2.train

# -----------------------------------------------------------
# Plot confusion matrix for LDA, training set
cm.LDA_train <- confusionMatrix(p2.train,as.factor(P.train$Class), dnn = c("Predicted",
"Observed"))

plt <- as.data.frame(cm.LDA_train$table)
plt$Predicted <- factor(plt$Predicted, levels=rev(levels(plt$Predicted)))

P.LDA_train <- ggplot(plt, aes(Predicted,Observed, fill= Freq)) +
      geom_tile() + geom_text(aes(label=Freq)) +
      scale_fill_gradient(low="white", high="#009194") +
      labs(x = "Observed",y = "Predicted") +
      scale_y_discrete(labels=c("BRCA", "COAD", "KIRC", "LUAD", "PRAD")) +
      scale_x_discrete(labels=c("PRAD", "LUAD", "KIRC", "COAD", "BRCA"))

P.LDA_train <- P.LDA_train + ggtitle("Confusion Matrix for LDA Classifier\nTraining
Data")+
           theme(legend.position = "none")
P.LDA_train
ggsave("P.LDA_train.bmp", width = 4, height = 4, dpi=300)

# "BRCA", "COAD", "KIRC", "LUAD", "PRAD"
# "PRAD", "LUAD", "KIRC", "COAD", "BRCA"
# -----------------------------------------------------------
#    p2     BRCA COAD KIRC LUAD PRAD
#    BRCA  223    0    0    3    3
#    COAD    0   52    0    2    0
#    KIRC    0    0  115    0    0
#    LUAD   17   10    4  102    0
#    PRAD    0    0    0    0  110


OA(CM.LDA2.train) # 0.94

library(ggplot2)
ggplot(data = P) + geom_point(aes(PC1, PC2, color = Class))+
                 ggtitle("Scatterplot of first 2 PC scores by \nClass (Cancer Type) for the
entire dataset")
ggsave("scoreplot.bmp", width = 4, height = 4, dpi=300)

# 1. Output CM.LDA2.train
write.table(CM.LDA2.train, "output_LDA.csv",
         append = TRUE,
         sep = ",",
         col.names = TRUE,
         row.names = FALSE,
         quote = FALSE)

#-----------------------------------------------------------

# -----------------------------------------------------------
#ggsave("LDA Score Plots.jpg",p12, dpi=300)
#ggsave("../results/LDA Score Plots.jpg",p12, dpi=300)
#dev.off()
# ------------------------------------------------------------
# -LDA: calculate confusion matrices for training set for each of 4 classes ---------
OA.LDA2.train <- as.data.frame(round(OA(CM.LDA2.train),2))
colnames(OA.LDA2.train)[1] <- "Overall_accuracy.LDA_train"

# 2. Output OA.LDA2.train

write.table(OA.LDA2.train, "output_LDA.csv",
         append = TRUE,
         sep = ",",
         col.names = TRUE,
         row.names = FALSE,
        quote = FALSE)


binCM.LDA2.train <- binary.cm(CM.LDA2.train)
binCM.LDA2.train <- as.data.frame(binCM.LDA2.train)

# 3. Output binCM.LDA2.train binary CM matrices

write.table(binCM.LDA2.train, "output_LDA.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
#------------------------------------------------------------------------
# Precision, Recall, F1 for each class from binarized confusion matrix
# input BIN = output of binary.cm

# 4. Output PRF.binaries computed from Binary CM matrices

PRF.LDA2.train <- round(PRF.binaries(binCM.LDA2.train),2)

write.table(PRF.LDA2.train, "output_LDA.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)

#------------------------------------------------------------------------
# 5. Output macro-micro averages computed from Binary CM matrices

mac.mic.LDA2.train <- as.data.frame(macro_micro.avg(binCM.LDA2.train))
colnames(mac.mic.LDA2.train) <- "LDA2.train_Average.PRF"
rownames(mac.mic.LDA2.train) <-
c("macro.Precision","macro.Recall","macro.F1","micro.Precision","micro.Recall","micro
.F1")

write.table(mac.mic.LDA2.train, "output_LDA.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)

binCM.LDA2.train.1 <- binCM.LDA2.train[,1:2]
binCM.LDA2.train.2 <- binCM.LDA2.train[,3:4]
binCM.LDA2.train.3 <- binCM.LDA2.train[,5:6]
binCM.LDA2.train.4 <- binCM.LDA2.train[,7:8]
binCM.LDA2.train.5 <- binCM.LDA2.train[,9:10]
#------------------------------------------------------------------------
# binCM.RF.train.1
AUCmacro.LDA2.train <-
as.data.frame(AUC.macro(binCM.LDA2.train.1,binCM.LDA2.train.2,binCM.LDA2.trai
n.3,binCM.LDA2.train.4,binCM.LDA2.train.5))
colnames(AUCmacro.LDA2.train)[1] <- "LDA_AUC"
#BRCA COAD KIRC LUAD PRAD
rownames(AUCmacro.LDA2.train) <-
c("AUC.BRCA","AUC.COAD","AUC.KIRC","AUC.LUAD","AUC.PRAD","AUC.mac
ro")

# 6. Output AUC computed from Binary CM matrices

write.table(AUCmacro.LDA2.train, "output_LDA.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
# --------------------------------------------------------------------
# --------------------------------------------------------------------
#
===============================================================
=======
# performance measures iof LDA2 for test set
p2.test <- predict(LDA2, P.test)$class
table(p2.test)
CM.LDA2.test <- table(p2.test,P.test$Class)
CM.LDA2.test

# 1. Output CM.LDA2.test
write.table(CM.LDA2.test, "output_LDA.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = FALSE,
        quote = FALSE)

#-------------------------------------------------------------

# -------------------------------------------------------------
#ggsave("LDA Score Plots.jpg",p12, dpi=300)
#ggsave("../results/LDA Score Plots.jpg",p12, dpi=300)
#dev.off()
# -------------------------------------------------------------
# -LDA: calculate confusion matrices for test set for each of 4 classes ---------
OA.LDA2.test <- as.data.frame(round(OA(CM.LDA2.test),2))
colnames(OA.LDA2.test)[1] <- "Overall_accuracy.LDA_test"

# 2. Output OA.LDA2.test

write.table(OA.LDA2.test, "output_LDA.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = FALSE,
        quote = FALSE)


binCM.LDA2.test <- binary.cm(CM.LDA2.test)
binCM.LDA2.test <- as.data.frame(binCM.LDA2.test)

# 3. Output binCM.LDA2.test binary CM matrices

write.table(binCM.LDA2.test, "output_LDA.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
#----------------------------------------------------------------------
# Precision, Recall, F1 for each class from binarized confusion matrix
# input BIN = output of binary.cm

# 4. Output PRF.binaries computed from Binary CM matrices

PRF.LDA2.test <- round(PRF.binaries(binCM.LDA2.test),2)

write.table(PRF.LDA2.test, "output_LDA.csv",
         append = TRUE,
         sep = ",",
         col.names = TRUE,
         row.names = TRUE,
         quote = FALSE)
# ----------------------------------------------------------
# Plot confusion matrix for LDA, test set
cm.LDA_test <- confusionMatrix(p2.test,as.factor(P.test$Class), dnn = c("Predicted",
"Observed"))

plt <- as.data.frame(cm.LDA_test$table)
plt$Predicted <- factor(plt$Predicted, levels=rev(levels(plt$Predicted)))

P.LDA_test <- ggplot(plt, aes(Predicted,Observed, fill= Freq)) +
        geom_tile() + geom_text(aes(label=Freq)) +
        scale_fill_gradient(low="white", high="#009194") +
        labs(x = "Observed",y = "Predicted") +
        scale_y_discrete(labels=c("BRCA", "COAD", "KIRC", "LUAD", "PRAD")) +
        scale_x_discrete(labels=c("PRAD", "LUAD", "KIRC", "COAD", "BRCA"))

P.LDA_test <- P.LDA_test + ggtitle("Confusion Matrix for LDA Classifier\nTest Data")+
              theme(legend.position = "none")
P.LDA_test
ggsave("P.LDA_test.bmp", width = 4, height = 4, dpi=300)

# "BRCA", "COAD", "KIRC", "LUAD", "PRAD"
# "PRAD", "LUAD", "KIRC", "COAD", "BRCA"
# ----------------------------------------------------------
#---------------------------------------------------------------------
# 5. Output macro-micro averages computed from Binary CM matrices

mac.mic.LDA2.test <- as.data.frame(macro_micro.avg(binCM.LDA2.test))
colnames(mac.mic.LDA2.test) <- "LDA2.test_Average.PRF"
rownames(mac.mic.LDA2.test) <-
c("macro.Precision","macro.Recall","macro.F1","micro.Precision","micro.Recall","micro
.F1")

write.table(mac.mic.LDA2.test, "output_LDA.csv",
         append = TRUE,
         sep = ",",
         col.names = TRUE,
         row.names = TRUE,
         quote = FALSE)
```

```
binCM.LDA2.test.1 <- binCM.LDA2.test[,1:2]
binCM.LDA2.test.2 <- binCM.LDA2.test[,3:4]
binCM.LDA2.test.3 <- binCM.LDA2.test[,5:6]
binCM.LDA2.test.4 <- binCM.LDA2.test[,7:8]
binCM.LDA2.test.5 <- binCM.LDA2.test[,9:10]
#-----------------------------------------------------------------------

AUCmacro.LDA2.test <-
as.data.frame(AUC.macro(binCM.LDA2.test.1,binCM.LDA2.test.2,binCM.LDA2.test.3,
binCM.LDA2.test.4,binCM.LDA2.test.5))
colnames(AUCmacro.LDA2.test)[1] <- "LDA_AUC"
#BRCA COAD KIRC LUAD PRAD
rownames(AUCmacro.LDA2.test) <-
c("AUC.BRCA","AUC.COAD","AUC.KIRC","AUC.LUAD","AUC.PRAD","AUC.mac
ro")

# 6. Output AUC computed from Binary CM matrices

write.table(AUCmacro.LDA2.test, "output_LDA.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
# -------------------------------------------------------------------
#
=============================================================
====
# ######################### Random Forest ############################
#library(randomForest)
set.seed(1137311)
# ======================================================
# Using RF on PC1, PC2 (010624)

P.train$Class <- as.factor(P.train$Class)
P.test$Class <- as.factor(P.test$Class)
#set.seed(11713)
rf1 <- randomForest(formula2, ntree = 250,importance = TRUE, data = P.train)
rf1
#       OOB estimate of  error rate: 5.93%
#Confusion matrix:
#     BRCA COAD KIRC LUAD PRAD class.error
```

```
#BRCA  231  0   0   9   0 0.03750000
#COAD    0 52   0  10   0 0.16129032
#KIRC    0  0 117   1   1 0.01680672
#LUAD   10  5   0  92   0 0.14018692
#PRAD    2  0   0   0 111 0.01769912


RF.predict.train <- predict(rf1, P.train, type="class")
df.RF.train <- cbind.data.frame(RF.predict.train,P.train$Class)
colnames(df.RF.train) <- c("Prediction","Observed")
CM.RF.train <- table(RF.predict.train,P.train$Class)

# Plot confusion matrix for RF, train set
cm.RF_train <- confusionMatrix(RF.predict.train,as.factor(P.train$Class), dnn =
c("Predicted", "Observed"))

plt <- as.data.frame(cm.RF_train$table)
plt$Predicted <- factor(plt$Predicted, levels=rev(levels(plt$Predicted)))

P.RF_train <- ggplot(plt, aes(Predicted,Observed, fill= Freq)) +
      geom_tile() + geom_text(aes(label=Freq)) +
      scale_fill_gradient(low="white", high="#009194") +
      labs(x = "Observed",y = "Predicted") +
      scale_y_discrete(labels=c("BRCA", "COAD", "KIRC", "LUAD", "PRAD")) +
      scale_x_discrete(labels=c("PRAD", "LUAD", "KIRC", "COAD", "BRCA"))

P.RF_train <- P.RF_train + ggtitle("Confusion Matrix for RF Classifier\nTrain Data")+
           theme(legend.position = "none")
P.RF_train
ggsave("P.RF_train.bmp", width = 4, height = 4, dpi=300)

# "BRCA", "COAD", "KIRC", "LUAD", "PRAD"
# "PRAD", "LUAD", "KIRC", "COAD", "BRCA"
# -----------------------------------------------------------

# =============== RF Accuracy Measures
===================================

# 1. Output CM.RF.train
write.table(CM.RF.train, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = FALSE,
        quote = FALSE)

#----------------------------------------------------------

OA.RF.train <- as.data.frame(round(OA(CM.RF.train),2))
colnames(OA.RF.train)[1] <- "Overall_accuracy.LDA_train"

# 2. Output OA.RF.train

write.table(OA.RF.train, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = FALSE,
        quote = FALSE)


binCM.RF.train <- binary.cm(CM.RF.train)
binCM.RF.train <- as.data.frame(binCM.RF.train)

# 3. Output binCM.RF.train binary CM matrices

write.table(binCM.RF.train, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
#-------------------------------------------------------------------
# Precision, Recall, F1 for each class from binarized confusion matrix
# input BIN = output of binary.cm

# 4. Output PRF.binaries computed from Binary CM matrices

PRF.RF.train <- round(PRF.binaries(binCM.RF.train),2)

write.table(PRF.RF.train, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
```

```
#----------------------------------------------------------------------
# 5. Output macro-micro averages computed from Binary CM matrices
binCM.RF.train.1 <- binCM.RF.train[,1:2]
binCM.RF.train.2 <- binCM.RF.train[,3:4]
binCM.RF.train.3 <- binCM.RF.train[,5:6]
binCM.RF.train.4 <- binCM.RF.train[,7:8]
binCM.RF.train.5 <- binCM.RF.train[,9:10]


mac.mic.RF.train <- as.data.frame(macro_micro.avg(binCM.RF.train))
colnames(mac.mic.RF.train) <- "RF.train_Average.PRF"
rownames(mac.mic.RF.train) <-
c("macro.Precision","macro.Recall","macro.F1","micro.Precision","micro.Recall","micro
.F1")

write.table(mac.mic.RF.train, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
#----------------------------------------------------------------------

AUCmacro.RF.train <-
as.data.frame(AUC.macro(binCM.RF.train.1,binCM.RF.train.2,binCM.RF.train.3,binCM.
RF.train.4,binCM.RF.train.5))
colnames(AUCmacro.RF.train)[1] <- "LDA_AUC"
#BRCA COAD KIRC LUAD PRAD
rownames(AUCmacro.RF.train) <-
c("AUC.BRCA","AUC.COAD","AUC.KIRC","AUC.LUAD","AUC.PRAD","AUC.mac
ro")

# 6. Output AUC computed from Binary CM matrices

write.table(AUCmacro.RF.train, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
# ----------------------------------------------------------------
# ----------------------------------------------------------------
```

```
RF.predict.test <- predict(rf1, P.test, type="class")
df.RF.test <- cbind.data.frame(RF.predict.test,P.test$Class)
colnames(df.RF.test) <- c("Prediction","Observed")
CM.RF.test <- table(RF.predict.test,P.test$Class)


# ================ RF Accuracy Measures
=================================

# 1. Output CM.RF.test
write.table(CM.RF.test, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = FALSE,
        quote = FALSE)

#----------------------------------------------------------
# Plot confusion matrix for RF, test set
cm.RF_test <- confusionMatrix(RF.predict.test,as.factor(P.test$Class), dnn =
c("Predicted", "Observed"))

plt <- as.data.frame(cm.RF_test$table)
plt$Predicted <- factor(plt$Predicted, levels=rev(levels(plt$Predicted)))

P.RF_test <- ggplot(plt, aes(Predicted,Observed, fill= Freq)) +
       geom_tile() + geom_text(aes(label=Freq)) +
       scale_fill_gradient(low="white", high="#009194") +
       labs(x = "Observed",y = "Predicted") +
       scale_y_discrete(labels=c("BRCA", "COAD", "KIRC", "LUAD", "PRAD")) +
       scale_x_discrete(labels=c("PRAD", "LUAD", "KIRC", "COAD", "BRCA"))

P.RF_test <- P.RF_test + ggtitle("Confusion Matrix for RF Classifier\nTest Data")+
            theme(legend.position = "none")
P.RF_test
ggsave("P.RF_test.bmp", width = 4, height = 4, dpi=300)
#-------------------------------------------------------------
OA.RF.test <- as.data.frame(round(OA(CM.RF.test),2))
colnames(OA.RF.test)[1] <- "Overall_accuracy.LDA_test"

# 2. Output OA.RF.test
```

```
write.table(OA.RF.test, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = FALSE,
        quote = FALSE)


binCM.RF.test <- binary.cm(CM.RF.test)
binCM.RF.test <- as.data.frame(binCM.RF.test)

# 3. Output binCM.RF.test binary CM matrices

write.table(binCM.RF.test, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
#-----------------------------------------------------------------------
# Precision, Recall, F1 for each class from binarized confusion matrix
# input BIN = output of binary.cm

# 4. Output PRF.binaries computed from Binary CM matrices

PRF.RF.test <- round(PRF.binaries(binCM.RF.test),2)

write.table(PRF.RF.test, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)

#-----------------------------------------------------------------------
# 5. Output macro-micro averages computed from Binary CM matrices
binCM.RF.test.1 <- binCM.RF.test[,1:2]
binCM.RF.test.2 <- binCM.RF.test[,3:4]
binCM.RF.test.3 <- binCM.RF.test[,5:6]
binCM.RF.test.4 <- binCM.RF.test[,7:8]
binCM.RF.test.5 <- binCM.RF.test[,9:10]
```

```
mac.mic.RF.test <- as.data.frame(macro_micro.avg(binCM.RF.test))
colnames(mac.mic.RF.test) <- "RF.test_Average.PRF"
rownames(mac.mic.RF.test) <-
c("macro.Precision","macro.Recall","macro.F1","micro.Precision","micro.Recall","micro
.F1")

write.table(mac.mic.RF.test, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
#-----------------------------------------------------------------------

AUCmacro.RF.test <-
as.data.frame(AUC.macro(binCM.RF.test.1,binCM.RF.test.2,binCM.RF.test.3,binCM.RF.
test.4,binCM.RF.test.5))
colnames(AUCmacro.RF.test)[1] <- "LDA_AUC"
#BRCA COAD KIRC LUAD PRAD
rownames(AUCmacro.RF.test) <-
c("AUC.BRCA","AUC.COAD","AUC.KIRC","AUC.LUAD","AUC.PRAD","AUC.mac
ro")

# 6. Output AUC computed from Binary CM matrices

write.table(AUCmacro.RF.test, "output_RF.csv",
        append = TRUE,
        sep = ",",
        col.names = TRUE,
        row.names = TRUE,
        quote = FALSE)
# -----------------------------------------------------------------

#
=================================================================
=======


rm(list = ls())
#####################################################
######## AH and non AH combined data ############

#### Selecting first 100 rows #########
```

```
###### AH with only 1's ###############

attach(AH_Non_AH_combined)

mydata <- AH_Non_AH_combined[1:100, ]

######## Creating a new data frame with only selected variables ##########
#######################################################################
####

ls(mydata) ###list of all variables###

myvars <- c("Age" , "Gender" , "Race" , "Binge_drnk" , "Hypertension" , "BMI",
"Hvy_drnk" , "Hyperlipidemia" , "Underlying_liver_disease" , "Cirrhosis_present" ,
"Prbble _or_Pssbl_AH",  "Prior_hx_AH" , "Primary_Insurance" , "Mortality"  )  ###
selecting specific variables####
newdata <- mydata[myvars] #### data frame with only selected variables suitable for
regression analysis#####

newdata$Race[newdata$Race == "hispanic"] <- "Hispanic"  ### Combining levels ###
newdata$Race[newdata$Race == "white" | newdata$Race == "American Indian"] <-
"White"  ### Combining levels ###


newdata$Gender[newdata$Gender == "f"] <- "F"  ### Combining levels ###

newdata$`Prbble _or_Pssbl_AH`[newdata$`Prbble _or_Pssbl_AH` == "possible"] <-
"Possible"   ### Combining levels ###
newdata$`Prbble _or_Pssbl_AH`[newdata$`Prbble _or_Pssbl_AH` == "probable"] <-
"Probable"   ### Combining levels ###

newdata$Primary_Insurance[newdata$Primary_Insurance == "medicaid" |
newdata$Primary_Insurance == "VA" | newdata$Primary_Insurance == "Medicare" |
newdata$Primary_Insurance == "Medicaid"]  <- "Public"   ### Combining levels ###
newdata$Primary_Insurance[newdata$Primary_Insurance == "private"]  <- "Private"
### Combining levels ###
newdata$Primary_Insurance[newdata$Primary_Insurance == "uninsured"]  <-
"Uninsured"    ### Combining levels ###


###################################################################
##########  Making and changing factor variables for categorical ones
#################
```

```
newdata$Gender <- factor(newdata$Gender, levels = c("F", "M"))

newdata$Race <- factor(newdata$Race, levels = c("African American", "Hispanic" ,
"White"))
levels(newdata$Race) <- c('Non-Hispanic Black', 'Hispanic','White')



newdata$Binge_drnk <- factor(newdata$Binge_drnk, levels = c("0", "1"))
levels(newdata$Binge_drnk) <- c('No', 'Yes')

newdata$Hypertension <- factor(newdata$Hypertension, levels = c("0", "1"))
levels(newdata$Hypertension) <- c('No', 'Yes')

newdata$Hyperlipidemia <- factor(newdata$Hyperlipidemia, levels = c("0", "1"))
levels(newdata$Hyperlipidemia) <- c('No', 'Yes')

newdata$Hvy_drnk <- factor(newdata$Hvy_drnk, levels = c("0", "1"))
levels(newdata$Hvy_drnk) <- c('No', 'Yes')

newdata$Underlying_liver_disease <- factor(newdata$Underlying_liver_disease, levels =
c("0", "1"))
levels(newdata$Underlying_liver_disease) <- c('No', 'Yes')

newdata$Cirrhosis_present <- factor(newdata$Cirrhosis_present, levels = c("0", "1"))
levels(newdata$Cirrhosis_present) <- c('No', 'Yes')

newdata$`Prbble _or_Pssbl_AH` <- factor(newdata$`Prbble _or_Pssbl_AH`, levels =
c("Possible", "Probable"))

newdata$Prior_hx_AH <- factor(newdata$Prior_hx_AH, levels = c("0", "1"))
levels(newdata$Prior_hx_AH) <- c('First time AH', 'Recurrent AH')

newdata$Primary_Insurance <- factor(newdata$Primary_Insurance, levels = c("Private",
"Public" , "Uninsured"))
levels(newdata$Primary_Insurance) <- c('Private', 'Public' , 'Uninsured')

newdata$Primary_Insurance <-  relevel(newdata$Primary_Insurance, ref = "Uninsured")

newdata$Mortality <- factor(newdata$Mortality, levels = c("0", "1"))
```

```
##############################################################################
#######
############# Logistic Regression Assumption check ###########
################################################################

logistic_model <- glm(Mortality ~ . - Prior_hx_AH , family = binomial(), newdata) ##
Run the model##
summary(logistic_model)

logistic_model2 <- glm(Mortality ~ . - `Prbble _or_Pssbl_AH` , family = binomial(),
newdata) ## Run the model##
summary(logistic_model2)

logistic_model3 <- glm(Mortality ~ . , family = binomial(), newdata) ## Run the model##
summary(logistic_model3)

##### Odds ratio & CI######


 tidy(logistic_model, conf.int=TRUE, exp=T)
tidy(logistic_model2, conf.int=TRUE, exp=T)
tidy(logistic_model3, conf.int=TRUE, exp=T)



######### Assumptions Check #######
install.packages("tidyverse")
library(tidyverse)
library(broom)
theme_set(theme_classic())

# Predict the probability (p)
probabilities <- predict(logistic_model, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "pos", "neg")
head(predicted.classes)

##### Logistic Regression Diagnostics ##########


####### Multicollinearity ########
```

```
    library(car)
    vif(logistic_model)

    ###### Influential observations #######
    plot(logistic_model, which = 4, id.n = 3)
    library(broom)
    library(dplyr)

    # Extract model results##
    model.data <- augment(logistic_model) %>%
      mutate(index = 1:n())

    #The data for the top 3 largest values, according to the Cookâ□□s distance, can be
    displayed as follow##

    model.data %>% top_n(3, .cooksd)

    ### Plot standarized results ####
    library(ggplot2)
    ggplot(model.data, aes(index, .std.resid)) +
      geom_point(aes(color = Mortality), alpha = .5) +
      theme_bw()

    ### Filter potential influential data points####

    model.data %>%
      filter(abs(.std.resid) > 3)
```

```
######################################################


    setwd("E:/AK128_May 13 2015/AKS/AKS 2022/Research/Dr Manoj Sharma")
    library(foreign)
    D <- read.spss('Vaping_Recoded_Cleaned_Dataset.sav', reencode='utf-8')
    head(D)
    dim(D)
    names(D)
    summary(D)
    write.csv(D, "Vaping_Recoded_Cleaned_Dataset.csv")

    names(D)
    summary(D)

    # The research questions can be answered by considering
```

```
#Y1 = Initiation and Y2 =  Sustenance as two response variables
#KB: Correct!

# Y1 using predictors PD,  BC_Overall ,  PE_Overall
#KB: Correct!

# Y2 using predictors ET_Overall,  PC_Overall, SE_Overall
# KB: Correct!

XY <- c("Initiation", "Sustenance", "PD", "BC_Overall","PE_Overall",
             "ET_Overall","PC_Overall","SE_Overall")
D1 <- D[XY]
write.csv(D1,"Responses and Predictors.csv")

lm1 <- lm(Initiation~PD + BC_Overall + PE_Overall, data = D1)
summary(lm1)
library(car)
vif(lm1)
#        PD BC_Overall PE_Overall
#  1.091885   2.420286   2.468544

df1 <- as.data.frame(lm1$residuals)
colnames(df1)[1] <- "residuals"
library(ggplot2)
shapiro.test(df1$residuals) # W = 0.99325, p-value = 0.007012

P1 <- ggplot(df1)+stat_qq(aes(sample=residuals)) +
  geom_qq_line(aes(sample=residuals))+
  geom_text(aes(x=0.5, y=3, label="Shapiro-test p-value = 0.01"))+
  ggtitle("Test of normality of residuals from the Linear Model\nInitiation~PD +
BC_Overall + PE_Overall")


lm2 <- lm(Sustenance~ET_Overall+PC_Overall+SE_Overall, data = D1)
summary(lm2)
#library(car)
vif(lm2)
#        ET_Overall PC_Overall SE_Overall
#  2.698432   2.938788   1.593225

df2 <- as.data.frame(lm2$residuals)
colnames(df2)[1] <- "residuals"
```

```
library(ggplot2)
shapiro.test(df2$residuals) # W = 0.99357, p-value = 0.009669

P2 <- ggplot(df2)+stat_qq(aes(sample=residuals)) +
  geom_qq_line(aes(sample=residuals))+
  geom_text(aes(x=0.5, y=3, label="Shapiro-test p-value = 0.01"))+
  ggtitle("Test of normality of residuals from the Linear
Model\nSustenance~ET_Overall+PC_Overall+SE_Overall")

library(gridExtra)
grid.arrange(P1,P2,nrow=2)
#
=======================================================================
==========
# Support Vector Regression
library(e1071)
Formula1 <- Initiation~PD + BC_Overall + PE_Overall

Formula2 <- Sustenance~ET_Overall+PC_Overall+SE_Overall


# SVM performance can be improved further by tuning the SVM
# perform a grid search to tune(optimize) SVM HYPERPARAMETERS
tune.svm1 <- tune(svm, Formula1,
          kernel = "radial", data=D1,
          type="eps-regression",
          ranges = list(epsilon = seq(0,1,0.1), cost = 2^(2:9)))
print(tune.svm1)
# Draw the tuning graph
plot(tune.svm1)
# ----------------------------


#################################################

################################################
```

BIBLIOGRAPHY

1. Turing, A. M. (2009). Computing machinery and intelligence (pp. 23-65). Springer Netherlands.
2. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
3. Elsebakhi, E., Lee, F., Schendel, E., Haque, A., Kathireason, N., Pathare, T., ... & Al-Ali, R. (2015). Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms. Journal of Computational Science, 11, 69-81.
4. Bashir, S., Qamar, U., Khan, F. H., & Naseem, L. (2016). HMV: A medical decision support framework using multi-layer classifiers for disease prediction. Journal of Computational Science, 13, 10-25.
5. Wati, D. A. R., & Abadianto, D. (2017, November). Design of face detection and recognition system for smart home security application. In 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 342-347). IEEE.
6. Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., & Kerr, J. (2014). Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. Frontiers in public health, 2, 36.
7. Omrani, H. (2015). Predicting travel mode of individuals by machine learning. Transportation research procedia, 10, 840-849.

8. Siddiqui, M. K., Morales-Menendez, R., Huang, X., & Hussain, N. (2020). A review of epileptic seizure detection using machine learning classifiers. Brain informatics, 7(1), 5.

9. Woldaregay, A. Z., Årsand, E., Botsis, T., Albers, D., Mamykina, L., & Hartvigsen, G. (2019). Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes. Journal of medical Internet research, 21(5), e11030.

10. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. nature, 542(7639), 115-118.

11. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.

12. Kaouk, J. H., Garisto, J., Eltemamy, M., & Bertolo, R. (2019). Robot-assisted surgery for benign distal ureteral strictures: step-by-step technique using the SP® surgical system. BJU international, 123(4), 733-739.

13. Tang, J., Liu, R., Zhang, Y. L., Liu, M. Z., Hu, Y. F., Shao, M. J., ... & Zhang, W. (2017). Application of machine-learning models to predict tacrolimus stable dose in renal transplant recipients. Scientific reports, 7(1), 42192.

14. Chen, Y., Liu, Q., & Guo, D. (2020). Emerging coronaviruses: genome structure, replication, and pathogenesis. Journal of medical virology, 92(4), 418-423.

15. Rao, S. R., DesRoches, C. M., Donelan, K., Campbell, E. G., Miralles, P. D., & Jha, A. K. (2011). Electronic health records in small physician practices: availability, use, and

perceived benefits. Journal of the American Medical Informatics Association, 18(3), 271-275.

16. Tian, L., Zhang, D., Bao, S., Nie, P., Hao, D., Liu, Y., ... & Wang, H. (2021). Radiomics-based machine-learning method for prediction of distant metastasis from soft-tissue sarcomas. Clinical radiology, 76(2), 158-e19.
17. Lanfranco, A. R., Castellanos, A. E., Desai, J. P., & Meyers, W. C. (2004). Robotic surgery: a current perspective. Annals of surgery, 239(1), 14-21.
18. Cao, L. (2017). Data science: a comprehensive overview. ACM Computing Surveys (CSUR), 50(3), 1-42.
19. Sarker, I. H., Hoque, M. M., Uddin, M. K., & Alsanoosy, T. (2021). Mobile data science and intelligent apps: concepts, AI-based modeling and research directions. Mobile Networks and Applications, 26(1), 285-303.
20. Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Computer Science, 2(3), 173.
21. Kubat, M. (2017). An introduction to machine learning. Springer.
22. Belciug, S., Gorunescu, F., Belciug, S., & Gorunescu, F. (2020). Era of intelligent systems in healthcare. Intelligent Decision Support Systems—A Journey to Smarter Healthcare, 1-55.
23. Alpaydin, E. (2020). Introduction to machine learning. MIT press.
24. Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

25. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2005, June). Practical machine learning tools and techniques. In Data mining (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.

26. Sarker, I. H. (2021). Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective. SN Computer Science, 2(3), 154.

27. Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). Machine learning: algorithms and applications. Crc Press.

28. Chen, P. H. C., Liu, Y., & Peng, L. (2019). How to develop machine learning models for healthcare. Nature materials, 18(5), 410-414.

29. Shouval, R., Fein, J. A., Savani, B., Mohty, M., & Nagler, A. (2021). Machine learning and artificial intelligence in haematology. British journal of haematology, 192(2), 239-250.

30. Rahmani, A. M., Ali, S., Yousefpoor, M. S., Yousefpoor, E., Naqvi, R. A., Siddique, K., & Hosseinzadeh, M. (2021). An area coverage scheme based on fuzzy logic and shuffled frog-leaping algorithm (sfla) in heterogeneous wireless sensor networks. Mathematics, 9(18), 2251.

31. Lee, S. W., Ali, S., Yousefpoor, M. S., Yousefpoor, E., Lalbakhsh, P., Javaheri, D., ... & Hosseinzadeh, M. (2021). An energy-aware and predictive fuzzy logic-based routing scheme in flying ad hoc networks (FANETs). IEEE Access, 9, 129977-130005.

32. Tao, W., Concepcion, A. N., Vianen, M., Marijnissen, A. C., Lafeber, F. P., Radstake, T. R., & Pandit, A. (2021). Multiomics and machine learning accurately predict clinical response to adalimumab and etanercept therapy in patients with rheumatoid arthritis. Arthritis & Rheumatology, 73(2), 212-222.

33. Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Panahiazar, M., ... & Sarrafzadegan, N. (2019). A database for using machine learning and data mining techniques for coronary artery disease diagnosis. Scientific data, 6(1), 227.
34. Golsorkhtabar, M., Nia, F. K., Hosseinzadeh, M., & Vejdanparast, Y. (2010, July). The novel energy adaptive protocol for heterogeneous wireless sensor networks. In 2010 3rd International Conference on Computer Science and Information Technology (Vol. 2, pp. 178-182). IEEE.
35. Nikravan, M., Movaghar, A., & Hosseinzadeh, M. (2018). A lightweight defense approach to mitigate version number and rank attacks in low-power and lossy networks. Wireless Personal Communications, 99, 1035-1059.
36. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion, 50, 71-91.
37. Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., ... & Nenadic, G. (2019). Machine learning methods for wind turbine condition monitoring: A review. Renewable energy, 133, 620-635.
38. Dhal, P., & Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. Applied Intelligence, 52(4), 4543-4581.
39. Tiwari, S. R., & Rana, K. K. (2021). Feature selection in big data: Trends and challenges. Data Science and Intelligent Applications: Proceedings of ICDSIA 2020, 83-98.

40. Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In Feature extraction: foundations and applications (pp. 1-25). Berlin, Heidelberg: Springer Berlin Heidelberg.

41. Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., & Hu, J. (2020). Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Computational Materials Science, 171, 109203.

42. Xu, Z., Qin, W., Tang, Q., & Jiang, D. (2014). Energy-efficient cognitive access approach to convergence communications. Science China Information Sciences, 57, 1-12.

43. Mandal, I. (2017). Machine learning algorithms for the creation of clinical healthcare enterprise systems. Enterprise Information Systems, 11(9), 1374-1400.

44. Feldman, K., Faust, L., Wu, X., Huang, C., & Chawla, N. V. (2017). Beyond volume: The impact of complex healthcare data on the machine learning pipeline. In Towards Integrative Machine Learning and Knowledge Extraction: BIRS Workshop, Banff, AB, Canada, July 24-26, 2015, Revised Selected Papers (pp. 150-169). Springer International Publishing.

45. Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (2020). Machine learning testing: Survey, landscapes and horizons. IEEE Transactions on Software Engineering, 48(1), 1-36.

46. Javaheri, D., Hosseinzadeh, M., & Rahmani, A. M. (2018). Detection and elimination of spyware and ransomware by intercepting kernel-level system routines. IEEE Access, 6, 78321-78332.

47. Mesbahi, M. R., Rahmani, A. M., & Hosseinzadeh, M. (2017). Highly reliable architecture using the 80/20 rule in cloud computing datacenters. Future Generation Computer Systems, 77, 77-86.

48. Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). Machine learning: algorithms and applications. Crc Press.

49. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT press.

50. Sarker, I. H., Abushark, Y. B., Alsolami, F., & Khan, A. I. (2020). Intrudtree: a machine learning based cyber security intrusion detection model. Symmetry, 12(5), 754.

51. Sarker, I. H., Abushark, Y. B., & Khan, A. I. (2020). Contextpca: Predicting context-aware smartphone apps usage based on machine learning techniques. Symmetry, 12(4), 499.

52. Liu, H., & Motoda, H. (Eds.). (1998). Feature extraction, construction and selection: A data mining perspective (Vol. 453). Springer Science & Business Media.

53. Sarker, I. H., Alqahtani, H., Alsolami, F., Khan, A. I., Abushark, Y. B., & Siddiqui, M. K. (2020). Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling. Journal Of Big Data, 7, 1-23.

54. Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In Feature extraction: foundations and applications (pp. 1-25). Berlin, Heidelberg: Springer Berlin Heidelberg.

55. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2005, June). Practical machine learning tools and techniques. In Data mining (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.

56. John, G. H., & Langley, P. (2013). Estimating continuous distributions in Bayesian classifiers. arXiv preprint arXiv:1302.4964.

57. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

58. Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In Feature extraction: foundations and applications (pp. 1-25). Berlin, Heidelberg: Springer Berlin Heidelberg.

59. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6), 417.

60. LIII, K. P. On lines and planes of closest fit to systems of points in space., 1901, 2. DOI: https://doi. org/10.1080/14786440109462720, 559-572.

61. Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). Machine learning: algorithms and applications. Crc Press.

62. Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big data, 7, 1-29.

63. Sarker, I. H. (2019). A machine learning based robust prediction model for real-life mobile phone data. Internet of Things, 5, 180-193.

64. Cessie, S. L., & Houwelingen, J. V. (1992). Ridge estimators in logistic regression. Journal of the Royal Statistical Society Series C: Applied Statistics, 41(1), 191-201.

65. Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

66. Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.

67. Breiman, L. (2017). Classification and regression trees. Routledge.

68. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

69. Sarker, I. H., Kayes, A. S. M., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data, 6(1), 1-28.

70. Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

71. Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. Neural computation, 9(7), 1545-1588.

72. Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. Neural computation, 13(3), 637-649.

73. Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine learning, 6, 37-66.

74. Villarrubia, G., De Paz, J. F., Chamoso, P., & De la Prieta, F. (2018). Artificial neural networks used in optimization problems. Neurocomputing, 272, 10-16.

75. Shrestha, Y. R., Krishna, V., & von Krogh, G. (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. Journal of Business Research, 123, 588-603.

76. Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big data, 7, 1-29.

77. Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. Ieee access, 6, 35365-35381.

78. Piccialli, F., Di Somma, V., Giampaolo, F., Cuomo, S., & Fortino, G. (2021). A survey on deep learning in medicine: Why, how and when?. Information Fusion, 66, 111-137.

79. Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). Supervised and unsupervised learning for data science. Springer Nature.

80. MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

81. Celebi, M. E., & Aydin, K. (Eds.). (2016). Unsupervised learning algorithms (Vol. 9, p. 103). Cham: Springer.

82. Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W., & Liu, J. (2021). Air quality predictions with a semi-supervised bidirectional LSTM neural network. Atmospheric Pollution Research, 12(1), 328-339.

83. Bull, L. A., Worden, K., & Dervilis, N. (2020). Towards semi-supervised and probabilistic classification in structural health monitoring. Mechanical Systems and Signal Processing, 140, 106653.

84. Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of artificial intelligence research, 4, 237-285.

85. Xu, X., Zuo, L., & Huang, Z. (2014). Reinforcement learning algorithms with function approximation: Recent advances and applications. Information sciences, 261, 1-31.

86. Coronato, A., Naeem, M., De Pietro, G., & Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. Artificial Intelligence in Medicine, 109, 101964.

87. Uprety, A., & Rawat, D. B. (2020). Reinforcement learning for iot security: A comprehensive survey. IEEE Internet of Things Journal, 8(11), 8693-8706.

88. Wu, J., Roy, J., & Stewart, W. F. (2010). Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Medical care, 48(6), S106-S113.

89. Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS one, 8(4), e61318.

90. Borisov, N., Tkachev, V., Suntsova, M., Kovalchuk, O., Zhavoronkov, A., Muchnik, I., & Buzdin, A. (2018). A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency. Cell Cycle, 17(4), 486-491.

91. Fakoor, R., Ladhak, F., Nazi, A., & Huber, M. (2013, June). Using deep learning to enhance cancer diagnosis and classification. In Proceedings of the international conference on machine learning (Vol. 28, pp. 3937-3949).

92. Podolsky, M. D., Barchuk, A. A., Kuznetcov, V. I., Gusarova, N. F., Gaidukov, V. S., & Tarakanov, S. A. (2016). Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. Asian Pacific journal of cancer prevention, 17(2), 835-838.

93. Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), 3240-3247.

94. Lindqvist, N., & Price, T. (2018). Evaluation of feature selection methods for machine learning classification of breast Cancer.

95. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning

methods in a real breast cancer problem. Artificial intelligence in medicine, 50(2), 105-115.

96. Turgut, S., Dağtekin, M., & Ensari, T. (2018, April). Microarray breast cancer data classification using machine learning methods. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-3). IEEE.

97. Hussain, L., Ahmed, A., Saeed, S., Rathore, S., Awan, I. A., Shah, S. A., ... & Awan, A. A. (2018). Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies. Cancer Biomarkers, 21(2), 393-413.

98. Chang, S. W., Abdul-Kareem, S., Merican, A. F., & Zain, R. B. (2013). Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC bioinformatics, 14, 1-15.

99. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10), 906-914.

100. Cho, S. B., & Won, H. H. (2003, January). Machine learning in DNA microarray analysis for cancer classification. In Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003-Volume 19 (pp. 189-198).

101. Chen, H., Zhao, H., Shen, J., Zhou, R., & Zhou, Q. (2015, June). Supervised machine learning model for high dimensional gene data in colon cancer detection. In 2015 IEEE International Congress on Big Data (pp. 134-141). IEEE.

102. Ghanat Bari, M., Ung, C. Y., Zhang, C., Zhu, S., & Li, H. (2017). Machine learning-assisted network inference approach to identify a new class of genes that coordinate the functionality of cancer networks. Scientific reports, 7(1), 6993.

103. Ayyad, S. M., Saleh, A. I., & Labib, L. M. (2019). Gene expression cancer classification using modified K-Nearest Neighbors technique. Biosystems, 176, 41-51.

104. Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H., & Khan, M. M. (2019). A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data. IEEE Access, 7, 22086-22095.

105. Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification.

106. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46, 389-422.

107. Jin, X., Xu, A., Bie, R., & Guo, P. (2006). Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In Data Mining for Biomedical Applications: PAKDD 2006 Workshop, BioDM 2006, Singapore, April 9, 2006. Proceedings (pp. 106-115). Springer Berlin Heidelberg.

108. Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., & Mewes, H. W. (2005). Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry, 29(1), 37-46.

109. Chen, H., Zhao, H., Shen, J., Zhou, R., & Zhou, Q. (2015, June). Supervised machine learning model for high dimensional gene data in colon cancer detection. In 2015 IEEE International Congress on Big Data (pp. 134-141). IEEE.

110. Wang, L., Zhang, W., He, X., & Zha, H. (2018, July). Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2447-2456).

111. Zhu, Q., Chen, Z., & Soh, Y. C. (2018). A novel semisupervised deep learning method for human activity recognition. IEEE Transactions on Industrial Informatics, 15(7), 3821-3830.

112. Zhai, X., Zhou, Z., & Tin, C. (2020). Semi-supervised learning for ECG classification without patient-specific labeled data. Expert Systems with Applications, 158, 113411.

113. World health organization: WHO. http://www.who.int/.

114. Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., & Singh, R. P. (2020). Significant applications of machine learning for COVID-19 pandemic. Journal of Industrial Integration and Management, 5(04), 453-479.

115. Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons & Fractals, 139, 110059.

116. Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., & Singh, R. P. (2020). Significant applications of machine learning for COVID-19 pandemic. Journal of Industrial Integration and Management, 5(04), 453-479.

117. Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., ... & Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. Algorithms, 13(10), 249.

118. Jamshidi, M., Lalbakhsh, A., Talla, J., Peroutka, Z., Hadjilooei, F., Lalbakhsh, P., ... & Mohyuddin, W. (2020). Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. Ieee Access, 8, 109581-109595.

119. Alakus, T. B., & Turkoglu, I. (2020). Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solitons & Fractals, 140, 110120.

120. Oh, Y., Park, S., & Ye, J. C. (2020). Deep learning COVID-19 features on CXR using limited training data sets. IEEE transactions on medical imaging, 39(8), 2688-2700.

121. Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Deep Learning applications for COVID-19. Journal of big Data, 8(1), 1-54.

CURRICULUM VITAE

DWAIPAYAN MUKHOPADHYAY

dmuk90@gmail.com• www.linkedin.com/in/dwai90

EDUCATION

2018- 2024 **Doctor of Philosophy in Mathematical Sciences (PhD)**

University of Nevada (UNLV), Las Vegas

2014 – 2016 **Master of Science in Mathematics (M.S)**

University of New Orleans, New Orleans

RELEVANT COURSEWORK

Probability & Statistics
Linear Algebra & Models
Theory of Estimation
Testing of Hypothesis
Operation Research

Multivariate Analysis
Time Series & Regression Analysis
Real Analysis
Stochastic Process
Statistical Quality Control & Applied Statistics

Econometrics
Reliability, Availability & Maintenance of Engineering System
Financial Mathematics
Survival Analysis of Statistical Data
Managerial Economics

RELEVANT WORK EXPERIENCE

2018 -2024 **Teaching Assistant**

UNLV, Las Vegas (Mathematical Science Dept.)

- Led recitation classes, Calculus-I (one semester), Calculus- II (second semester).
- Facilitated hour-long interactive tutorials with students in the Mathematics Tutor Centre.

- Taught College Algebra for 4+ years
- Taught Applied Statistics for Biological Sciences, Statistical methods - I, Fundamentals of College Mathematics and Finite Mathematics courses.

CERTIFICATIONS AND PROFESSIONAL EXAMS

- Passed Society of Actuaries Exam FM (Financial Mathematics).
- Passed Society of Actuaries Exam P (Probability).
- Base Programmer for SAS 9.
- Mathematics Department (University of New Orleans) scholarship.

COMPUTER SKILLS AND COMPETENCES

- Expert in most of the Microsoft Office programs.
- Proficient in Statistical Analysis System, a software suite used for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics in academic projects.
- Experience with R, programming language conducted to validate data, statistical computing and graphics development, forecasting and model fitting in academic projects.
- Experience with Econometric Views, statistical package utilized for general statistical and econometric analysis such as cross-sectional data analysis, time series estimation and forecasting in academic projects.
- Experience with Minitab, statistical package used for general statistical analysis in academic projects.
- Limited experience with Visual Basic for Applications (VBA), programming language utilized during university courses.
- Limited experience with Structured Query Language (SQL), programming language conducted during university courses.
- Limited experience with C, programming language conducted during university courses.
- Exposure to MySQL, open-source relational data base management system during university courses.
- Exposure to Microsoft Access, database management system during university courses.

RESEARCH & DATA ANALYTICS EXPEREINCES

University of Nevada, Las Vegas (Department of Mathematical Sciences) | Fall 2018 – Present

Graduate Researcher

• Drafted a research paper using statistical software R to classify different cancer types conducting Machine Learning classification on a high dimensional gene expression microarray data, determining possible genes causing cancer disease. Statistical techniques - Linear Discriminant Analysis, Random Forest, Principal Component Analysis, Confusion matrix,Area under the Curve (AUC),Macro- and micro-averages of AUC etc.

• Conducted statistical software R to draft a research paper investigating the potential factors behind the intent to initiate & sustain vaping quitting behavior in a sample population of 18-24 years old. Statistical techniques- Linear Discriminant Analysis, Random Forest,Confusion Matrix, Average Precision, Recall, F1, Areas under the Curve etc.

• Published a research paper utilizing Random Forest, a Machine Learning classification method to help in prediction of the skin disease Erythematosquamous Dermatosis. Statistical techniques- Random Forest,Confusion Matrix, Average Precision, Recall, F1 measures, and approximate Areas Under the Curve etc. Statistical software -R

• Collaborated with a diverse team to publish a paper determining the potential factors behind the cause of Alcoholic Hepatitis disease, which will be able to identify at-risk individuals in order to implement preventative measures. Statistical techniques- Logistic Regression, Odds Ratio, Forest Plot. Statistical software -R

• Collaborated with researchers to publish a paper on systematic review and meta-analysis to evaluate the effect of cystic fibrosis transmembrane conductance regulator (CFTR) modulators on liver enzymes. Statistical software -R.

University of New Orleans (Department of Mathematics) | Fall 2014 – Spring 2016

Graduate Researcher

• Conducted statistical software package R on Wisconsin breast cancer data for rigorous variable selection methods to find potential factors behind breast cancer.Statistical techniques used - Logistic Regression, Wald test, Information Value &Weight of Evidence, Confusion Matrix, Lift Curve, K- fold cross validation, Receiver Operating Characteristic Curve etc.

• Utilized statistical software package SAS to analyze data, generate output consisting of life tables and Survival model.Conclusion is drawn regarding graft survival after kidney transplant. Statistical techniques used - Life table analysis, Cox Proportional Hazard or Accelerated Failure

Time Model), Goodness of Fit test using Schoenfeld & Cox-Snell Residuals, Akaike Information Criterion etc.

PUBLICATIONS

Peer Reviewed

- Mukhopadhyay, Dwaipayan, et al. ”Classification of Erythematosquamous Dermatosis by the Method of Random Forest.” Journal of Dermatology Research Reviews & Reports 4.1 (2023): 1-6.

- Tun, Kyaw Min, et al. ”A Comparative Study of Acute Alcoholic Hepatitis vs. Non-Alcoholic Hepatitis Patients from a Cohort with Chronic Alcohol Dependence.” Genes 14.4 (2023): 780.

- Tun, Kyaw Min, et al. ”S1415 Effect of Cystic Fibrosis Transmembrane Conductance Regulator Modulators on Liver Enzymes Among Patients With Cystic Fibrosis: A Systematic Review and Meta-Analysis.” Official journal of the American College of Gastroenterology— ACG 118.10S (2023): S1079.

Non-Peer Reviewed

- Liu B, Ananda M, Mukhopadhyay D (2023). GEC: Generalized Exponentiated Composite Distributions . R package

version 0.1.0, < https://CRAN.R-project.org/package=GEC >.