Statistical Machine Learning Approaches in Medicine and Biomedical Sciences

Eric Waters
eswaters@asu.edu
School of Computing and Augmented Intelligence
Arizona State University
Tempe, Arizona, USA

Collin Wood
cjwood8@asu.edu
School of Computing and Augmented Intelligence
Arizona State University
Tempe, Arizona, USA

Cameron Anundson
canundso@asu.edu
School of Computing and Augmented Intelligence
Arizona State University
Tempe, Arizona, USA

Adam Nugroho
ainugroh@asu.edu
School of Computing and Augmented Intelligence
Arizona State University
Tempe, Arizona, USA

Kesudh Giri
kgiri1@asu.edu
School of Computing and Augmented Intelligence
Arizona State University
Tempe, Arizona, USA

Matteo Bergsagel
matteo.bergsagel@asu.edu
School of Computing and Augmented Intelligence
Arizona State University
Tempe, Arizona, USA

1 PAPER OVERVIEW

The paper "Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine" provides a thorough overview of the use of statistical and machine learning techniques in analyzing big data in medicine and behavioral science. The authors describe how different types of data, such as clinical, genomic, and environmental data, can be combined statistically to gain a more complete understanding of human physiology and disease.

The paper starts by introducing the concept of well-known regression analyses, such as linear and logistic regressions, which are commonly used in clinical data analyses. It then discusses modern statistical models, such as Bayesian networks, that have been developed to analyze more complex data. The authors also explain how to use modern statistical models to represent the interaction of clinical, genomic, and environmental data.

The paper emphasizes the importance of understanding the statistical tools available in medicine for analyzing large datasets. Data from biomedical and behavioral science is becoming larger and more complex with each passing year [34]. As a result, it is critical to be aware of this trend and understand the statistical tools available for analyzing these datasets.

The authors introduce big data in terms of clinical data, single nucleotide polymorphism (SNP), gene expression studies, and their interactions with the environment. They explain how these different types of data can be combined using statistical methods to gain a more comprehensive understanding of human physiology and disease.

The paper acknowledges some of the difficulties associated with using big data analysis in medicine, such as privacy concerns, ethical considerations, quality control issues, missing value imputation issues, and so on. It does, however, highlight several applications of big data analysis in medicine, including personalized medicine, drug discovery, disease diagnosis and prognosis prediction, and healthcare delivery optimization.

The authors describe how decision trees, random forests, support vector machines (SVMs), neural networks (NNs), and deep learning can be used for big data analysis in medicine. They also discuss the benefits and drawbacks of each technique.

The paper concludes with a detailed explanation of Bayesian networks and their applications in medical big data analysis. The authors describe how Bayesian networks can be used to model complex relationships among various types of data and make probabilistic predictions. They also go over the benefits and drawbacks of using Bayesian networks for big data analysis in medicine.

Overall, this paper provides valuable insights into the use of statistical and machine learning techniques in analyzing big data in medicine. It highlights the importance of understanding these tools for improving patient outcomes while acknowledging some challenges associated with using big data analysis in medicine. The authors also provide a detailed explanation

2 MACHINE LEARNING TECHNIQUES

2.1 Regression

Regression is a statistical technique to model how one or more input variables impact a desired output variable.

2.1.1 Linear Regression. The idea of linear regression is to fit a straight line to a set of data (see Figure 1) by minimizing the Sum of Squares Error (SSE). Traditionally, a regression model creates a line by discovering weights through the error minimization process. This iterative procedure, in discovering weights, eventually produces a line that can be thought of as a *best fit* for that particular data. Using this derived line, a continuous output variable can be predicted given one or more input variables [3]. Linear Regression is often thought of as the most basic mode of machine learning, and it is a common practice to deploy its abilities as a baseline for understanding if there are basic relationships amongst features.

2.1.2 Logistic Regression. Logistic regression is similar to linear regression in that it uses input variables to predict an output variable. However, it differs in that the output variable is a binary value. It decides the binary output variable by using Maximum Likelihood Estimation (MLE) to fit a sigmoid function that ranges from 0 to 1 (see Figure 2). This sigmoid function represents the probability of a given input data point belonging to the specified "true" class. If the predicted probability is greater than 0.5, it is typically assigned the

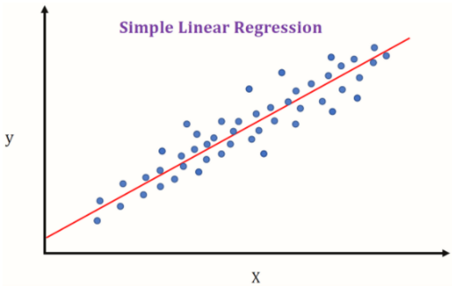

Figure 1: An example of Simple Linear Regression [25]

positive class label. Otherwise, it is assigned the negative class label. Because it outputs a binary value, it is often used for classification tasks [16].

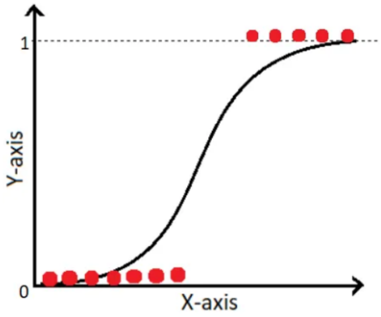

Figure 2: A sigmoid function used to separate two classes [26]

2.2 Bayesian Networks

Bayesian Networks take advantage of some interesting properties found in probability theory. Using Bayes' theorem yields the ability to make inferences about a system of random variables. To define a network that is useful for making Bayesian predictions, a directed acyclic graph is created, where the nodes in our system represent the random variables and the edges represent conditional relationships amongst the variables. To make use of the newly defined network, the random variables must have probability distributions indicating whether they are true or false, then the network can be used to make inferences on our network.

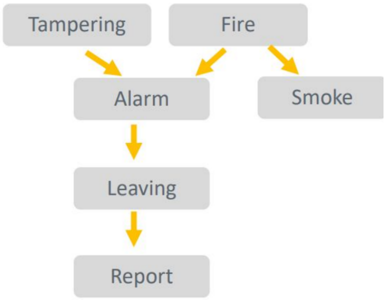

Figure 3: Fire Alarm Inferencing (Bayesian Network) [19]

Take this Fire Alarm Network as an example seen in Figure 3. In this network, the nodes are random variables representing different factors as a result of a fire starting or an alarm being tampered with, which are the *super parents* of our network. This means that all of the probabilities are either a result of tampering or fire being true. To make an inference on this system, it is sufficient to use Bayes' theorem to ask a question such as: *What is the probability that a report is true given that a fire did not occur?* These inferences can help us generate an understanding of our feature space and how likely certain scenarios are. In prediction, the Bayesian Network is used as a tool to discover the likelihood of an event, and if the probability is sufficiently higher than some threshold (depending on the use case), the event is predicted.

2.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a machine learning algorithm used for both classification and regression. It works by calculating the distance of an input data point to all other data points and finding the "k-nearest" data points (neighbors). For classification, the class label of the input data point is often the majority class label of its k-nearest neighbors. For regression, a continuous value can be predicted by averaging the specified target variable of the k-nearest neighbors [28]. A visualization of K-Nearest Neighbors can be seen in Figure 4.

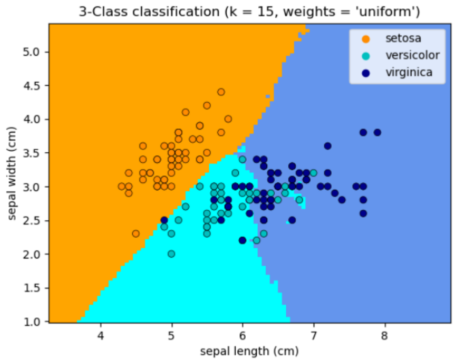

Figure 4: The KNN class boundaries for K=3 (Iris Dataset) [28]

2.4 Conditional Logistic Regression

In certain scenarios, a modification to logistic regression might yield more sufficient results, especially when using statistical tools to ensure the regression is performed *fairly*. Common issues with classical regression techniques include the lack of handling bias of different population samples, and in the case of highly critical systems like medicine, accounting for treatment as part of a case study requires configuration of the original logistic regression. In CLR (Conditional Logistic Regression), there are two main properties that separate it from the classical model: stratification and matching [1]. In particular, CLR uses these techniques to limit the amount of data being evaluated when performing the regression on certain conditions. In stratification, trial subjects are partitioned into sample populations by a factor other than 'received treatment'.

This ensures that the treatment vector doesn't *overpower* the other important features being measured as a contributor to the results. In matching, CLR reduces the bias in treatment effect analysis by comparing the results of individuals with certain characteristics, in medical use cases, this can potentially mean comparing against individuals in a similar demographic.

2.5 Multifactor Dimensionality Reduction

Dimensionality reduction refers to the process of transforming high-dimensional data into fewer dimensions in order to reduce complexity. A popular dimensionality reduction technique is Principal Component Analysis (PCA), which eliminates the least significant attributes of the data. Multifactor Dimensionality Reduction (MDR), on the other hand, takes a different approach to reducing dimensionality. Rather than eliminating features from consideration, it combines features in a strategic manner. MDR was specifically designed to detect gene-gene interactions [22, 34].

2.6 Polymorphism Interaction Analysis

Similar to Multifactor Dimensionality Reduction, Polymorphism Interaction Analysis (PIA) also reduces dimensionality by combining features in such a way that those combinations best predict the outcome. It primarily differs from MDR in that it uses the percentage of misclassified instances and the Gini index (a measure of feature importance) to score the interactions of the features. Furthermore, PIA uses ten-fold cross-validation, a method that splits the data into ten components and iteratively trains and tests the data on different subsets of the data in order to estimate the model's performance [20].

2.7 Support Vector Machines (SVMs)

Support Vector Machines [7] are a machine learning technique typically used for classification. They classify data points by creating a hyperplane, called the decision boundary, that maximizes the margin between the classes. The side of the hyperplane that a testing point falls on determines its class label.

2.7.1 Recursive Feature Elimination (RFE). In this extension to traditional SVMs, the training process will involve reducing the features iteratively until our model only considers the most important features for creating separable hyperplanes. This is done by initially training an SVM classifier using the entire feature space, and then ranking the features. Features are ranked based on the weights of the hyperplane, which is the solution to the SVM primal problem. After ranking all SVM combinations of features, we select the subset of the feature space which includes only the highest-ranking features (in other words we remove the worst-ranking feature from all future SVM training). This process can be repeated recursively until removing features doesn't increase the accuracy of our model or until there aren't any more features to eliminate [17].

2.7.2 Recursive Feature Addition (RFA). In recursive feature addition, we take a similar approach to RFE in that we are retraining an SVM classifier on many different feature spaces and comparing their outputs until an optimal subset is found. In this algorithm, we first start by training several SVMs to only look at one feature, and compare their outputs. We select the best SVM feature space (in the initial stage this feature space has only one dimension) and recursively add all combinations of features. Then the best ranking SVM feature space is selected and this process is recursively iterated until accuracy doesn't improve or a threshold is defined. The main difference between RFA and RFE is how we start the recursion, in RFA our final step produces a significantly larger feature space than that which we started with, while in RFE the final step usually results in a minimal feature space for producing accurate predictions [13].

2.7.3 With local search (local). SVM with local search takes the initial hyperplane obtained by the traditional SVM algorithm and attempts to optimize it by searching for a better solution in the vicinity of the original solution. It does this by slightly adjusting the original hyperplane and reevaluating its performance. If the new performance is superior to the original performance, the new hyperplane is kept and the algorithm continues until it converges on some specified criteria. This algorithm is often effective but also computationally expensive and unrealistic to run on massive datasets [6, 34].

2.7.4 Genetic Algorithm (GA). Like with other SVM enhancements, we start off with our default SVM and make a modification that finds the best feature space for training the model. This time, however, we start by selecting a population from our data and randomly selecting a set of features to train the model on. Then, in terms of genetic algorithms and genetic coding, we select the members with the highest fitness (fitness score is once again determined by the results of the hyperplane weights) and use this population to continue iterating. A genetic algorithm is deployed that uses the selected subset of features in the high-fitness population to determine which additional features to examine and include or exclude in the next generation of training. This is called *selection* in evolutionary terms and the selection function for a genetic algorithm greatly depends on the specific use case. In the Genetic Algorithm, there is more randomness inherent in the selection and generation phases, yet this provides the advantage of finding an optimal feature space without needing to recursively add and subtract a feature at a time. After the next generation of features is evaluated, we can then compute the fitness of our child population, it is trivial to stop iterating after our results converge regardless of running more generations [31].

3 FINDINGS OF THE PAPER

Traditional problem-solving approaches in the case of data analysis for Bio-Informatics, involving clinical data, Gene-Gene interactions, and Gene-Environment interactions involve usage of linear equations on the modeled variables [18]. This approach has its advantages and disadvantages, the primary issue being the modeling of causality between the modeled variables.

With multiple methods developed over the years to analyze Clinical, Gene-Gene, and Gene-Environment data, a few of which have been discussed in the previous section, the optimal method to represent causality between variables and produce predictive models is Causal Bayesian Networks (Causal BNs) [23].

Causal BNs have demonstrated the ability to extract relationships between modeled variables based on causality, even on extremely large data sets with a high number of variables to be modeled. This is due to the fact that Causal BNs have a significant number of advantages over Traditional Bayesian Networks that allow them to model relationships between variables optimally when causality is involved. This allows them to produce predictive models from clinical and genome data with high levels of accuracy [33]. The most important advantages Causal BNs hold over Traditional BNs are:

- Ability to predict interventions in the network
- Allowance for cycles in the network, which translates into the ability to handle feedback loops
- Handling of unobserved variables that in turn affect other variables in the network by adding them into the network as new nodes

The usability of Causal BNs can be explained using an example network.

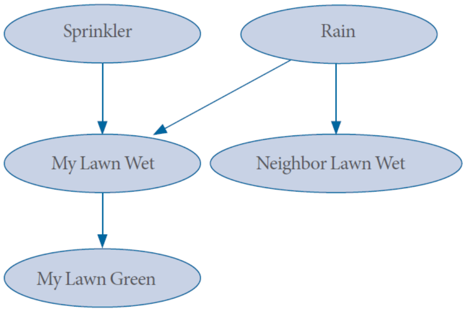

Figure 5: An example Bayesian Network

In the example given in Figure 5, it can be derived that Lawns can become wet either through Rain or by turning on a sprinkler. It can also be noted that the neighbor does not own a sprinkler system. When our lawn is wet, there is a possibility that it might be green.

There are a few types of sub-networks that help model causality between variables by a direct connection, with the parent node directly influencing its children. There are three main sub-network types, which we can derive from the example in Figure 5, namely Converging Arcs, Diverging Arcs, and Serial Arcs.

In Converging Arcs, as shown in Figure 6, multiple parent nodes converge into a child node, in this case, either a Sprinkler (Node A) or Rain (Node B) may cause our lawn to become wet (Node C). Thus, A and B become dependent on the case that C occurs.

In Diverging Arcs, as shown in Figure 7, a single parent node influences multiple children, in this case, Rain (Node A) may cause our lawn to become wet (Node B) and/or cause the neighbor's lawn to become wet (Node C). Thus, B and C become independent of the case that A occurs.

In Serial Arcs, as shown in Figure 8, a parent node influences its child node, which in turn influences its child node. In this case Sprinkler (Node A) may cause our lawn to become wet (Node B) which in turn may cause our lawn to become green (Node C). Thus, A and C become independent of the case that B occurs.

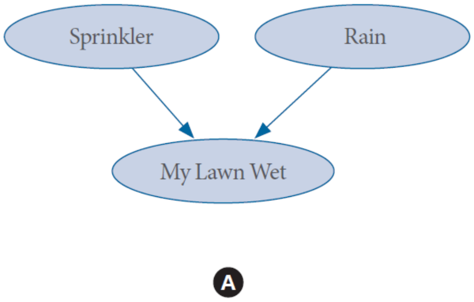

Figure 6: Sub-network Type A: Converging Arcs

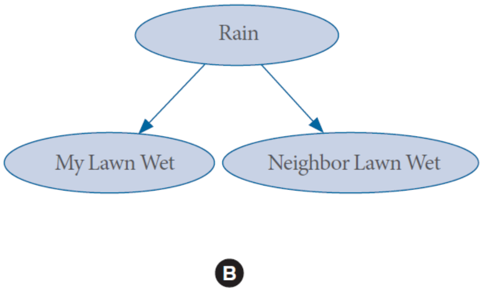

Figure 7: Sub-network Type B: Diverging Arcs

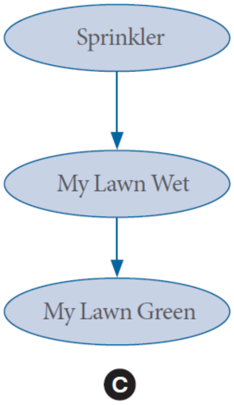

Figure 8: Sub-network Type C: Serial Arcs

The usage of these subnetwork types allows Causal BNs to express causality in intuitive ways with a myriad of combinations possible. This provides to be useful when large datasets are involved, proving to be invaluable for Bio-Informatical Data Analysis. Causal BNs as a statistical tool are therefore optimal to model complex clinical parameters, Gene-Gene interactions, and Gene-Environment interactions from large datasets of clinical or genomic data.

4 APPLICATIONS

There are many applications for the machine learning techniques presented. Four are highlighted in this paper: clinical data, gene expression data, single nucleotide polymorphisms (SNPs), and epigenetic regulation of the genome.

4.1 Clinical Data

For applications with clinical data, linear regression and logistic regression were highlighted. Linear regression was successfully used to analyze the association between implementing electronic health record and ED visits, hospitalizations, and office visits for patients with diabetes mellitus [27]. A study also performed research comparing initial tumor size and reduction rate when treated with targeted agents. By using both univariate and multivariate linear regression, the study determined that the rate of tumor size reduction was correlated to the initial size for individual tumors [35].

There were also several applications for logistic regression. One application was found in studying the effectiveness of a multidisciplinary surgical safety checklist. Implementation of the checklist was analyzed using logistic regression and found to reduce surgical complication and mortality [8]. Logistic regression also was used in a study of children diagnosed with urinary anomalies. The study found that there were certain risk factors that were associated with an increased risk of kidney anomalies, including gestational diabetes, preexisting diabetes, and maternal renal disease [29]. In another case, logistic regression was used to characterize the most common medical problems in in-flight medical emergencies. This study also analyzed the type of assistance that was provided onboard [24].

4.2 Gene Expression Data

Gene expression data was also highlighted as a field that could have useful applications of machine learning technology. Specifically, gene clustering analysis within large gene expression datasets has been shown to be a useful application of the k-Nearest Neighbor algorithm [10, 15, 21, 30]. Bayesian networks also have a useful application within gene expression data analysis for modeling causality. These are useful for performing inference and can be used to model gene expression level regulation [34].

4.3 Single Nucleotide Polymorphisms

Single nucleotide polymorphisms (SNPs) are a DNA variation due to altering a single nucleotide in a genome. They must occur within at least 1% of the population and are the most common DNA variation. Interaction between SNPs is believed to play a significant part of the development of complex diseases. Logistic regression has an application here for linking SNPs to disease outcome [34]. In addition, multifactor dimensionality reduction (MDR) and polymorphism interaction analysis (PIA) have been found to have utility in determining SNP combinations that have disease-predicting interactions [11, 12, 14]. Finally, support vector machines (SVMs) have been used to determine interaction among SNPs [5].

4.4 Epigenetic Regulation of the Genome

Epigenetic regulation of the genome is another field in which these machine learning techniques have potential applications. This regulation takes place when methyl groups are added to cytosines in the DNA. The effect of this is that DNA expression can be changed without modifications to the DNA. This process was recently shown to take place in fully differentiated cells [4]. As a result, these interactions must be considered when studying gene expression. Bayesian models have application here, and they were used to study how DNA methylation patterns are transmitted across cell division [9].

5 POSSIBLE EXTENSIONS

The integration of machine learning approaches in medicine and biomedical sciences holds great promise for improving our understanding of disease mechanisms, identifying new therapeutic targets, and developing personalized treatments. There are several possible extensions of statistical machine learning approaches in medicine and biomedical sciences that go beyond those covered in the paper. Three examples are the Naive Bayes classifier, the Decision Tree, and the Random Forest.

5.1 Naive Bayes Classifier

Naive Bayes is a popular classification algorithm in statistical machine learning that is widely used in medicine and biomedical sciences for various applications, such as disease diagnosis, drug discovery, and medical image analysis.

The Naive Bayes classifier is based on Bayes' theorem, which states that the probability of a hypothesis H given some observed evidence E is proportional to the probability of the evidence given the hypothesis multiplied by the prior probability of the hypothesis, divided by the probability of the evidence:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

In the case of classification, we can use Bayes' theorem to calculate the probability of each class given some observed features or variables. The "naive" assumption in Naive Bayes is that the features are independent of each other given the class label, which simplifies the calculation of probabilities.

For example, in medical diagnosis, we may want to classify a patient as having a certain disease or not based on their symptoms and medical history, such as in the privacy-preserving classification of breast cancer data as benign or malignant [32]. We can use Naive Bayes to calculate the probability of each class (disease or no disease) given the observed symptoms and medical history. We first estimate the prior probabilities of each class based on the frequency of the disease in the population. Then, we calculate the conditional probabilities of each feature given each class, based on the frequency of each symptom and medical history in patients with and without the disease. Finally, we use Bayes' theorem to calculate the posterior probabilities of each class given the observed features and classify the patient as having the class with the highest probability.

One advantage of Naive Bayes is that it is computationally efficient and requires a small amount of training data compared to other classification algorithms. It also performs well in high-dimensional feature spaces, which is often the case in biomedical data analysis. However, the naive assumption of feature independence may not always hold in practice, and the performance of Naive Bayes can be affected by imbalanced class distribution or irrelevant features.

5.2 Decision Tree

A decision tree is a powerful statistical machine learning approach widely used in medicine and biomedical sciences. It is a non-parametric method that makes predictions based on a series of binary decisions or splits, which divide the dataset into smaller subsets until a terminal node or leaf is reached. Each split is based on a particular feature or attribute of the data, and the decision is made by selecting the feature that maximizes the information gain or minimizes the impurity measure of the data.

In medicine and biomedical sciences, decision trees are commonly used in areas such as clinical decision-making, diagnosis, prognosis, and treatment planning. For instance, decision trees have been used to predict the risk of developing certain diseases based on patient characteristics, such as age, sex, family history, lifestyle factors, and biomarkers. They have also been used to identify the most effective treatment options for specific patient subgroups based on their clinical and molecular profiles.

One of the advantages of decision trees is their ability to handle both categorical and continuous variables, as well as missing data, without the need for data normalization or transformation. Additionally, decision trees can handle nonlinear relationships between variables and are easily interpretable, making them useful for generating insights and guiding clinical decision-making. They are also relatively computationally efficient, making them suitable for large datasets with complex structures.

However, decision trees can be sensitive to small changes in the dataset and prone to overfitting, particularly when the number of features is large. To address these issues, ensemble methods such as random forests and boosting can be used to improve the accuracy and robustness of the models.

In conclusion, decision trees are a valuable tool in statistical machine learning approaches in medicine and biomedical sciences. They offer a flexible and interpretable framework for making predictions and generating insights from complex and heterogeneous datasets. However, careful attention must be paid to avoid overfitting and to optimize the performance of the models.

5.3 Random Forest

Random Forest is a popular machine learning algorithm that can be used in various applications, including medicine and biomedical sciences. It is a type of ensemble learning method that combines multiple decision trees to improve the predictive performance of the model.

In medicine, Random Forest has been used in various areas such as disease diagnosis, prognosis, and drug discovery. For example, Random Forest has been used to predict the risk of developing diseases such as cancer, diabetes, and heart disease. It has also been used to identify potential drug targets, predict the efficacy of drugs, and classify medical data using feature ranking [2].

In biomedical sciences, Random Forest has been used in areas such as genomics, proteomics, and metabolomics. For example, Random Forest has been used to predict protein-protein interactions, classify gene expression data, and identify biomarkers for diseases.

Random Forest is a versatile algorithm that can handle both categorical and continuous variables, and it can also handle missing data. It is also robust to overfitting, which is a common problem in machine learning.

Overall, Random Forest is a powerful tool in statistical machine learning approaches for medicine and biomedical sciences, and it has the potential to provide valuable insights and predictions that can improve patient outcomes and advance scientific research.

REFERENCES

[1] n.d.. Conditional logistic regression. https://people.stat.sc.edu/hoyen/Stat705/Notes/Lecture12.pdf. Retrieved April 11, 2023.

[2] M. Z. Alam and M. S. Rahman. 2019. A random forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked* 15 (2019), 100180.

[3] Rebecca Bevans. 2022. Simple linear regression: An easy introduction and examples. https://www.scribbr.com/statistics/simple-linear-regression/

[4] Adrian Bird. 2007. Perceptions of epigenetics. *Nature* 447, 7143 (May 2007), 396–398.

[5] Shyh-Huei Chen, Jielin Sun, Latchezar Dimitrov, Aubrey R Turner, Tamara S Adams, Deborah A Meyers, Bao-Li Chang, S Lilly Zheng, Henrik Grönberg, Jianfeng Xu, and Fang-Chi Hsu. 2008. A support vector machine approach for detecting gene-gene interaction. *Genet. Epidemiol.* 32, 2 (Feb. 2008), 152–167.

[6] Shyh-Huei Chen, Jielin Sun, Latchezar Dimitrov, Aubrey R. Turner, Tamara S. Adams, Deborah A. Meyers, Bao-Li Chang, S. Lilly Zheng, Henrik Grönberg, Jianfeng Xu, and et al. 2008. A support vector machine approach for detecting gene-gene interaction. *Genetic Epidemiology* 32, 2 (2008), 152–167. https://doi.org/10.1002/gepi.20272

[7] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 3 (1995), 273–297. https://doi.org/10.1007/bf00994018

[8] Eefje N. de Vries, Hubert A. Prins, Rogier M.P.H. Crolla, Adriaan J. den Outer, George van Andel, Sven H. van Helden, Wolfgang S. Schlack, M. Agnès van Putten, Dirk J. Gouma, Marcel G.W. Dijkgraaf, Susanne M. Smorenburg, and Marja A. Boermeester. 2010. Effect of a Comprehensive Surgical Safety System on Patient Outcomes. *New England Journal of Medicine* 363, 20 (2010), 1928–1937. https://doi.org/10.1056/NEJMsa0911535 arXiv:https://doi.org/10.1056/NEJMsa0911535 PMID: 21067384.

[9] Audrey Qiuyan Fu, Diane P Genereux, Reinhard Stöger, Charles D Laird, and Matthew Stephens. 2010. Statistical inference of transmission fidelity of DNA methylation patterns over somatic cell divisions in mammals. *Ann. Appl. Stat.* 4, 2 (2010), 871–892.

[10] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 5439 (Oct. 1999), 531–537.

[11] S N Goodman. 1999. Probability at the bedside: the knowing of chances or the chances of knowing? *Ann. Intern. Med.* 130, 7 (April 1999), 604–606.

[12] Jiang Gui, Angeline S Andrew, Peter Andrews, Heather M Nelson, Karl T Kelsey, Margaret R Karagas, and Jason H Moore. 2011. A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility. *Ann. Hum. Genet.* 75, 1 (Jan. 2011), 20–28.

[13] T. Hamed, S. C. Kremer, and R. Dara. 2017. Recursive feature addition: A novel feature selection technique. https://www.researchgate.net/publication/332875563_Recursive_Feature_Addition_a_Novel_Feature_Selection_Technique_Including_a_Proof_of_Concept_in_Network_Security. Retrieved April 11, 2023.

[14] H He, W S Oetting, M J Brott, and S Basu. 2010. Pair-wise multifactor dimensionality reduction method to detect gene-gene interactions in a case-control study. *Hum. Hered.* 69, 1 (2010), 60–70.

[15] R Herwig, A J Poustka, C Müller, C Bull, H Lehrach, and J O'Brien. 1999. Large-scale clustering of cDNA-fingerprinting data. *Genome Res.* 9, 11 (Nov. 1999), 1093–1105.

[16] IBM. [n. d.]. What is logistic regression? https://www.ibm.com/topics/logistic-regression Accessed: April 10, 2023.

[17] J. de Jong. 2017. SVM with recursive feature elimination in R. https://johanndejong.wordpress.com/2016/01/17/svm-with-recursive-feature-elimination/. Retrieved April 10, 2023.

[18] Peter Kraft and David J. Hunter. 2009. The challenge of assessing complex gene–environment and gene–gene interactions. In *Human Genome Epidemiology: Building the evidence for using genetic information to improve health and prevent disease.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195398441.003.0009 arXiv:https://academic.oup.com/book/0/chapter/297763024/chapter-ag-pdf/44508164/book_34832_section_297763024.ag.pdf

[19] J. Lee. 2023. Bayesian Networks.

[20] Leah E Mechanic, Brian T Luke, Julie E Goodman, Stephen J Chanock, and Curtis C Harris. 2008. Polymorphism interaction analysis (PIA): A method for investigating complex gene-gene interactions. *BMC Bioinformatics* 9, 1 (2008). https://doi.org/10.1186/1471-2105-9-146

[21] G S Michaels, D B Carr, M Askenazi, S Fuhrman, X Wen, and R Somogyi. 1998. Cluster analysis and data visualization of large-scale gene expression data. *Pac. Symp. Biocomput.* (1998), 42–53.

[22] Alison A Motsinger and Marylyn D Ritchie. 2006. Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene - gene interactions in human genetics and pharmacogenomics studies. *Human Genomics* 2, 5 (2006). https://doi.org/10.1186/1479-7364-2-5-318

[23] Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[24] Drew C. Peterson, Christian Martin-Gill, Francis X. Guyette, Adam Z. Tobias, Catherine E. McCarthy, Scott T. Harrington, Theodore R. Delbridge, and Donald M. Yealy. 2013. Outcomes of Medical Emergencies on Commercial Airline Flights. *New England Journal of Medicine* 368, 22 (2013), 2075–2083. https://doi.org/10.1056/NEJMoa1212052 arXiv:https://doi.org/10.1056/NEJMoa1212052 PMID: 23718164.

[25] Benjamin Obi Tayo Ph.D. 2020. Linear Regression Basics for absolute beginners. https://pub.towardsai.net/linear-regression-basics-for-absolute-beginners-68ed9ff980ae

[26] Khushwant Rai. 2020. The math behind logistic regression. https://medium.com/analytics-vidhya/the-math-behind-logistic-regression-c2f04ca27bca

[27] Mary Reed, Jie Huang, Richard Brand, Ilana Graetz, Romain Neugebauer, Bruce Fireman, Marc Jaffe, Dustin W Ballard, and John Hsu. 2013. Implementation of an outpatient electronic health record and emergency department visits, hospitalizations, and office visits among patients with diabetes. *JAMA* 310, 10 (Sept. 2013), 1060–1065.

[28] Scikit-learn. [n. d.]. Nearest Neighbors. https://scikit-learn.org/stable/modules/neighbors.html Accessed: April 10, 2023.

[29] Margarett Shnorhavorian, Rachel Bittner, Jonathan L Wright, and Stephen M Schwartz. 2011. Maternal Risk Factors for Congenital Urinary Anomalies: Results of a Population-based Case-control Study. *Urology* 78, 5 (Nov. 2011), 1156–1161.

[30] P T Spellman, G Sherlock, M Q Zhang, V R Iyer, K Anders, M B Eisen, P O Brown, D Botstein, and B Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9, 12 (Dec. 1998), 3273–3297.

[31] Z. Tao, L. Huiling, W. Wenwen, and Y. Xia. 2018. Ga-SVM based feature selection and parameter optimization in hospitalization expense modeling. *Applied Soft Computing* (November 13 2018). https://www.sciencedirect.com/science/article/pii/S1568494618306264 Retrieved April 10, 2023.

[32] A. Wood, V. Shpilrain, K. Najarian, and D. Kahrobaei. 2019. Private naive bayes classification of personal biomedical data: Application in cancer data analysis. *Computers in Biology and Medicine* 105 (2019), 144–150.

[33] Changwon Yoo and Gregory F. Cooper. 2004. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *Artificial Intelligence in Medicine* 31, 2 (2004), 169–182. https://doi.org/10.1016/j.artmed.2004.01.018 Data Mining in Genomics and Proteomics.

[34] Changwon Yoo, Luis Ramirez, and Juan Liuzzi. 2014. Big Data Analysis using modern statistical and machine learning methods in medicine. *International Neurourology Journal* 18, 2 (2014), 50. https://doi.org/10.5213/inj.2014.18.2.50

[35] Takeshi Yuasa, Shinji Urakami, Shinya Yamamoto, Junji Yonese, Kenji Nakano, Makoto Kodaira, Shunji Takahashi, Kiyohiko Hatake, Kentaro Inamura, Yuichi Ishikwa, and Iwao Fukui. 2011. Tumor size is a potential predictor of response to tyrosine kinase inhibitors in renal cell cancer. *Urology* 77, 4 (April 2011), 831–835.