Classification of Cancer Types based on Gene Expression Data

Yinchao He
Computer Science & Engineering
University of Nebraska - Lincoln
Lincoln, NE, USA
yhe@huskers.unl.edu

Ryan Bockmon
Computer Science & Engineering
University of Nebraska - Lincoln
Lincoln, NE, USA
ryan.bockmon@huskers.unl.edu

Miracle Modey
Computer Science & Engineering
University of Nebraska - Lincoln
Lincoln, NE, USA
miracle.modey@huskers.unl.edu

Sarah Roscoe
Computer Science & Engineering
University of Nebraska - Lincoln
Lincoln, NE, USA
sroscoe@huskers.unl.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313559

***Abstract*—With the rapid increase of genomic data, analyzing this huge bioinformatics data becomes a new challenge. Therefore, computer-based processes and algorithms are much more powerful to use. For analysis of across tumor types, John N Weinstein declares the uncertain feasibility of applying characterization based on molecular changes to complement pathological analysis for classification of cancers. Hence, for this limitation, we compared four deep learning methods, K-Means, Support Vector Machine, Formal Concept Analysis, and Association rules, to classify cancers based on the gene expression cancer RNA-Seq data set with 801 cases and 20531 genes for each case. The results show that SVM has the highest accuracy (99.8% and 99.2%) out of all our objectives, which is followed by K-Means with 91.75%. The third highest overall result (83.1%) is the Formal Concept Analysis algorithm. Association rules have the lowest accuracy with 72.25%. This comparison supplies a good guide for the classification of cancer types based RNA-seq.**

Keywords:Cancer, RNA-Seq gene expression, Molecular changes, Clustering, Classification, Association rules.

I. Introduction

In the human genome, a lot of nucleotides are essential to diseases or other phenotypes [1]. Hence, genomic data about nucleotides can be a meaningful source to recognize diseases. Recently, healthcare professionals in the United Kingdom's national health service hospitals consider incorporating genome sequencing into clinical care for diagnoses of rare diseases and some cancers [2]. With the widespread use of genome sequencing, the number of whole-genome sequences is much larger than before, so analyzing these big data efficiently and accurately becomes a new challenge. Therefore, computer-based processes and algorithms such as Machine learning and data mining interest many bioinformaticians' attention. For analysis across tumor types, John N Weinstein states some limitations of his cancer analysis; for example, the uncertain effectiveness of using molecular profiles to categorize cancers and the uncertain feasibility of applying characterization based on molecular changes to complement pathological analysis for classification of cancers [3]. Therefore, we use four computer-based algorithms, such as K-Means, Support Vector Machine, Formal Concept Analysis, and Association rules, to classify cancers based on the gene expression cancer RNA-Seq Data Set, which is collected for detecting and analyzing molecular defects in different types of cancers by the TCGA Research Network[3], [4]. At the end, we find the best method to classify cancers by evaluating their accuracy. It could solve some limitations to extend pathology research from a molecular perspective.

II. Objective

Our planned objectives for this project will be as follows:

- Objective 1: Apply a multi-class support vector machine (SVM) to classify cancers based on expression data.
- Objective 2: Use frequent sets and association rules to detect the association between each gene. Convert the original data set into the binary matrix. Then, generate association rules by the Apriori algorithm. At the end, use the association rules to predict the gene expression.
- Objective 3: Apply a formal concept analysis algorithm [5]to analyze and classify the data set.
- Objective 4: Use the K-Means algorithm to find the similarity between each patient. Then, use these clusters to classify cancers based on RNA-Seq expression data.
- Objective 5: Compare and analyze these four methods' results to determine which method has the highest accuracy to classify cancers based on RNA-Seq expression data.

III. Related Work

Classifying gene expression data (GED) can help determine which genes are associated with diseases, or which situations can affect expression levels across multiple genes [6]. It is also important to understand the biological and statistical significance of various classification methods [7]. In this section, we will examine the related work of classifying gene expression data with SVM, frequent sets and association rules, K-Means, and FCA.

Various SVM techniques have been applied specifically to classify gene expression data. Feature selection techniques paired with SVM has been surveyed [8]. Particle swarm optimization has been used in classifying cancer gene expression data[9]. Using an SVM to support class-mean info, two types of leukemia have been classified [10]. Finally, non-negative matrix factorization, along with a weighted kernel width SVM, was tested with colon cancer and acute leukemia microarray data sets [11].

Association rules have been extracted to be used in conjunction with SVM to classify leukemia and lung cancer gene expression datasets [12]. Independently, rule pruning strategies proposed in [13] are used to discover the top k interesting patterns. This method was tested on cancer data sets. Generalized rule induction can develop association rules to determine the relationship between cancer classes (tested on leukemia data) and related genes [14]. In addition, fuzzy association rule mining can distinguish the significant genes in a data set [15].

K-Mean is a traditional method to find the similarity within gene expression datasets. K-means could cluster genome data to identify and classify cancer signatures [16]. What's more, 92 % average positive prediction accuracy was obtained with the K-means clustering approach using centroid initialization, distance measures, and split methods based on breast cancer Wisconsin (BCW) diagnostic dataset [17]. The method is also used to select informative features, which are then passed to the classifier for low redundancy [18].

Formal Concept Analysis (FCA) is an offshoot of a field called lattice theory in Modern Algebra. FCA allows for real-world data to be analyzed with the help of theoretical constructs [19]. Most often, FCA is used for clustering and classification, and many FCA applications include analysis of gene expression data and other biological data [20]. The task of clustering with FCA is aided by its natural structure of concepts or maximal subsets of the binary matrix containing all 1s. These subsets are called *concepts* and can be ordered hierarchically by set inclusion. The resulting concept lattice structure can provide key insight into how certain objects and attributes are associated with the original data set [21]. FCA can also be extended to work with Fuzzy data [22], thus extending its possibilities for application.

IV. Problem Statement

Our goal is to classify the cases of a gene expression data set into one of four types of cancer: Breast cancer (BRCA), Kidney renal clear cell carcinoma (KIRC), Colon adenocarcinoma (COAD), Lung adenocarcinoma (LUAD), and Prostate adenocarcinoma (PRAD). Another goal is to evaluate these classification models based on accuracy to prove the effectiveness of using molecular profiles to categorize cancers.

V. Data

The data used is part of the RNA-Seq (HiSeq) PANCAN data set. It is a random extraction of gene expressions of patients having different types of cancer diagnosis: breast cancer (BRCA), Kidney renal clear cell carcinoma (KIRC), Colon adenocarcinoma (COAD), Lung adenocarcinoma (LUAD), and Prostate adenocarcinoma (PRAD) [4]. Attributes are RNA-Seq gene expression levels measured by the Illumina HiSeq platform consisting of all real numbers. There are a total of 801 cases and 20531 genes for each case. In this data set, One hundred and forty-one cases have LUAD. Three hundred cases have BRCA. Seventy-eight cases have KIRC. One hundred thirty-four cases have PRAD. All of them show in figure 1.

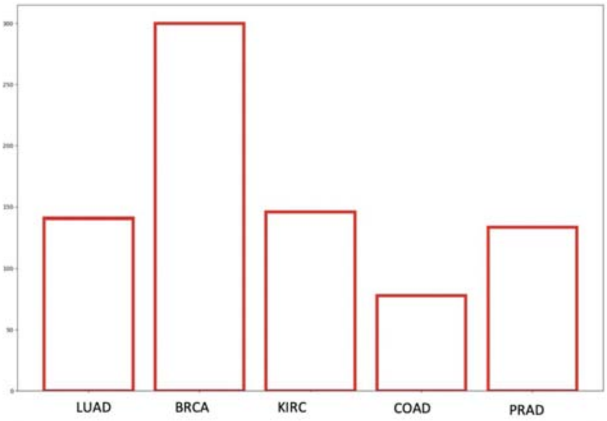

Fig. 1. cancer summary of cases

VI. Methods

A. Data preprocessing

There are a few different problems with this data set. The first problem is that there is a large number of attributes (over 20,000), and it would take a long time to train and validate any classification algorithm. To fix this, we remove any attribute that has missing data from any sample resulting in eliminating 5000 attributes. After removing all attributes that did not show up in each of the samples, we then used principal component analysis (PCA) to further reduce the number of attributes [23]. When running PCA, we decided to make two different data sets. One (with 11 PCs, in figure 2) was created because we found that 11 PCs was considered the optimal number of PCs to retain. The second (with 5 PCs, in the figure 3) was created because there is a total of 5 different classes and wanted to see if each class could be represented by an individual PC.

Besides, we also used the Univariate Selection method to select a subset of genes (5 genes, figure 4 and 11 genes, figure 5) to reduce the dimensionality but get comparable performance.

The second issue is that the data consists of real numbers; therefore, any conventional association algorithm such as Apriori will be unable to run on the data set [24]. To fix this issue, we binarize the US data set.

The final issue was that the original data format is in two documents, one containing expression data and the other containing class labels that correspond to each row in the original file. In order to run the data on the formal concept

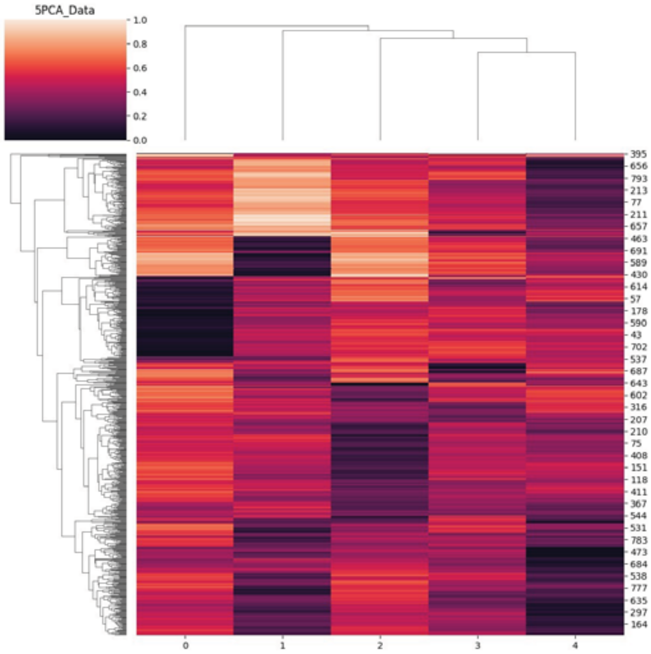

Fig. 2. Heatmap of 5PCA Data

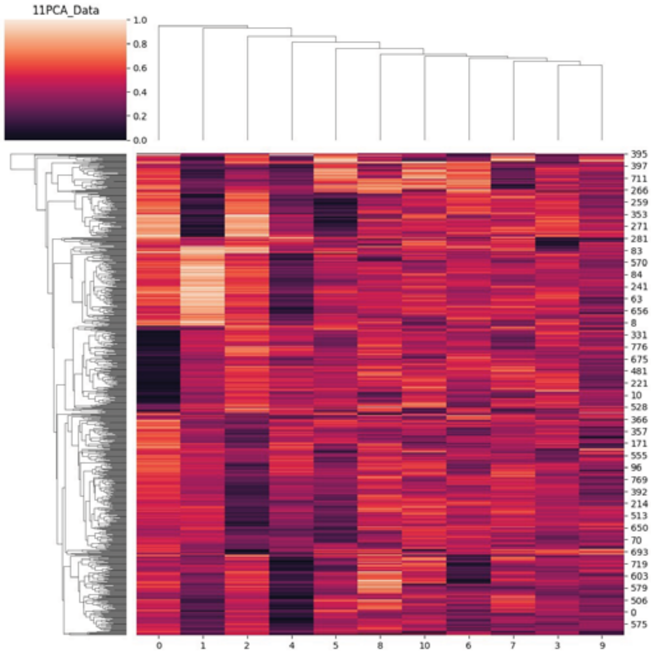

Fig. 3. Heatmap of 11PCA Data

analysis code, the labels of different cancer types (PRAD, LUAD, BRCA, COAD, KIRC) were converted to integer classes (0, 1, 2, 3, 4, respectively) and appended after the last column in the expression data file.

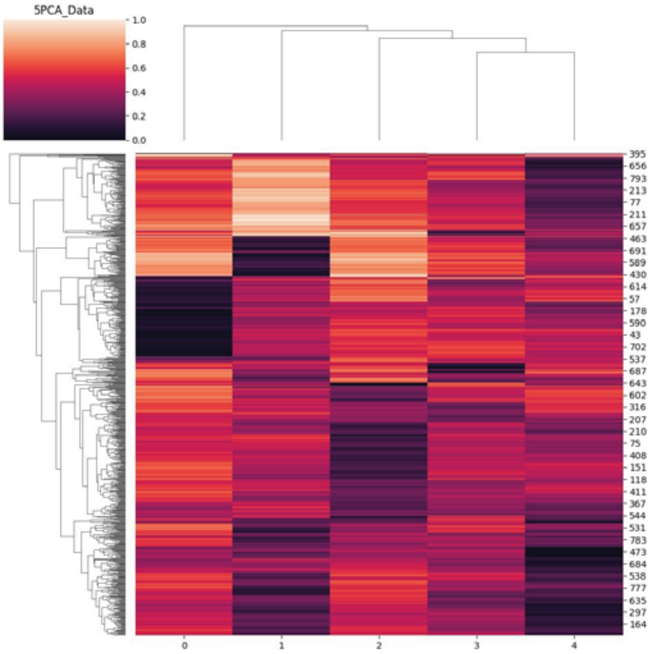

Fig. 4. Heatmap of 5US Data

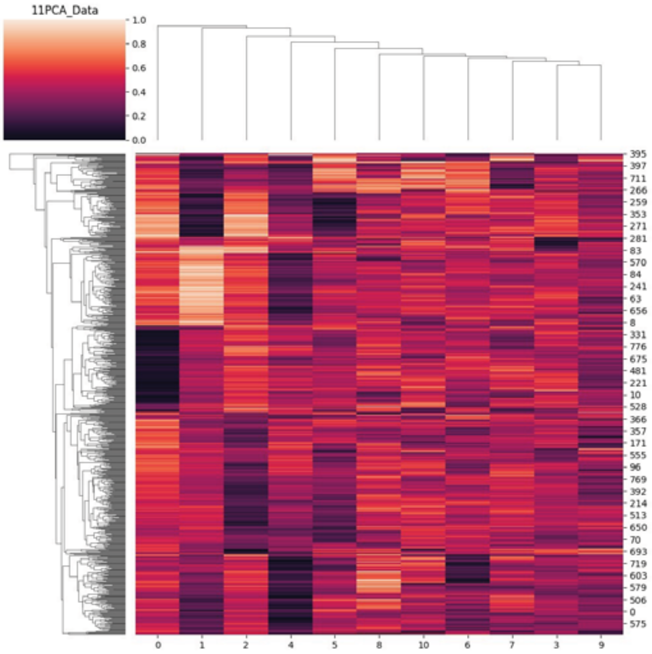

Fig. 5. Heatmap of 11US Data

B. SVM

SVMs are a supervised machine learning model that uses classification algorithms for two-group classification problems. A SVM takes in data points and outputs a hyperplane, a plane that best separates the classes and classifies the data points whether they fall on either side of that hyperplane. SVMs are inherently two-class classifiers; however, the algorithm can be extended to support multiple classes as is for our data set [25].

Data was randomly split into two subsets: subset one was used to train the SVM, while the other subset was to test the overall accuracy of the model. The training set consisted of 50% of the data, and the testing set consisted of the other 50%. Because we created different data sets based on different data reduction criteria, we wanted to compare how well the SVM did at classifying the different data sets. The evaluation was based on the overall accuracy of the model, and time it took to train the model.

C. Frequent Sets and Association Rules

We use frequent sets and association rules to classify different types of tumors. If certain attributes (gene expression data) appear more frequently together, they will form frequent sets and association rules. We want to see if gene expression data could predict different types of cancer based on their frequent sets and association rules created from the data sets.

Two Univariate Selection data sets (with 5 and 11 attributes) were classified with this code. We select 80% data of one tumor type data set and use the Apriori algorithm to mine frequent sets and association rules. In the end, we use 20% data to test the classified model. Eventually, we calculate the accuracy to evaluate the method.

D. K-Means

K-Means clustering is a method of partitioning a data set into k distinct clusters where each instance or data point belongs only to one group. A K-Means clustering algorithm takes in k number of clusters and each data point to the closest cluster or centroid. The number of clusters, K, could be determined by the Elbow method with the sum of squared errors (SSE). The algorithm keeps iterating until the assignment of instances or data points to clusters isn't changing. Therefore, the final result satisfies this formula:

$$\min\sum_{j=1}^{k}\sum_{i=1}^{n}||x_i - c_j||^2$$

k: the number of clusters, n: the number of cases, x_i: case i, c_j: centroid of cluster j

For this objective, we used the elbow method to calculate the felicitous number of clusters firstly. Then, we ran the K-Means algorithm on the full data set and compared the accuracy to the ground truth.

E. FCA

Our classification of the gene expression data set was accomplished with the code developed by a member of our research group for her Master's thesis [5]. The code was originally developed for image classification with Fuzzy FCA but can take any real-valued input and classify it based on the data given. The classification is accomplished by constructing the concept lattice from the data set. Objects (records) are classified together if they are in the same concept, with a certain probability from the Normal distribution. The program then checks against the actual class to see if the guess was correct. Each of the PCA data sets (with 5 and 11 PCAs) were classified with this code, as well as the 5 US data set. The algorithm was run on the Holland Computing Center's Crane partition.

TABLE I
SVM PCA5 Data Set Confusion Matrix, Accuracy: 94.0%

	PRAD	LUAD	BRCA	KIRC	COAD
PRAD	70	0	0	0	0
LUAD	1	59	0	4	7
BRCA	1	2	141	2	0
KIRC	2	1	3	70	0
COAD	0	1	0	0	36

F. Evaluation method

We use accuracy to evaluate the performance of classification models. Therefore, the model with the highest accuracy is the best method to classify cancers. Accuracy is defined as the quotient of the number of correct predictions and the total number of predictions, which could be converted to this formula:

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

TP: True Positive, TN:True Negative, FP:False Positive, FN:False Negative

VII. Results

In this section, we discuss the results for each of our objectives. We discuss overall trends and limitations in the following sections.

A. SVM

A SVM was run on three different reduced data sets. The first data set was the reduced data set before PCA was run on it. The total accuracy of that model was 99.8%, with a run time of 10 seconds. The second data set was the reduced data set with 11 PCs. The total accuracy of that model was 99.2%, with a run time of less than one second. The final data set was the reduced data set with 5 PCs. The total accuracy of that model was 94.0% with a run time of less than 1 second. Results from this model can be found in Table 1.

B. K-Means

The K-Means method was run on two data sets. In figure 6, the Elbow method shows that the optimal number of clusters was 5. The first data set was the 5-PCA, which produced and accuracy score of 90.75%. The second data set was the 11-PCA, which produced an accuracy score of 91.75%. The results for running the K-Means algorithm on the data sets 5-PCA, and 11-PCA are shown in Tables II and III respectively.

C. FCA

The Fuzzy FCA classification algorithm was run on the reduced data sets with 5 PC's, 11 PC's, and 5 genes using univariate selection. The algorithm was run on the Holland Computing Center's Crane partition. The results were better on the smaller data sets, which both produced roughly 83%

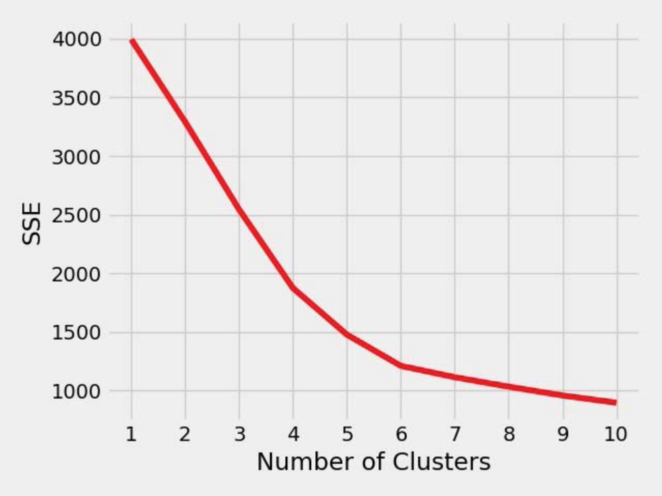

Fig. 6. the result of the elbow method

TABLE II
K-MEANS ON 5-PCA CONFUSION MATRIX

	PRAD	LUAD	BRCA	KIRC	COAD
PRAD	132	1	1	0	1
LUAD	0	136	0	0	5
BRCA	0	45	247	0	0
KIRC	0	1	0	145	0
COAD	0	11	0	0	67

accuracy, than on the larger data set, which has roughly 67% accuracy (Table IV). There is no accuracy difference between 5-PCA and 5-US.

D. Frequent Sets and Association Rules

The method was run on two different reduced data sets, 11 PCAs and 5 PCAs, with three different values of minimum support and confidence, 0.4, 0.6, 0.8.

For 5 PCAs data set, in figure 7 and figure 8, when minimum support was set to 0.4 and minimum confidence was set to 0.6, the accuracy was 17.25% (Table V). When minimum support was set to 0.6 and minimum confidence was set to 0.6, the accuracy was 22.125% (Table VI). When minimum support set to 0.8 and minimum confidence set to 0.6, the accuracy was 46.25% (Table VII). When minimum support was set to 0.6 and minimum confidence was set to 0.4, the accuracy was 22.125%. When minimum support was set to 0.6 and minimum confidence was set to 0.6, the accuracy was 22.125% and when minimum support was set to 0.6 and minimum confidence was set to 0.8, the accuracy was 22.125%.

TABLE III
K-MEANS ON 11-PCA CONFUSION MATRIX

	PRAD	LUAD	BRCA	KIRC	COAD
PRAD	134	1	0	0	0
LUAD	0	141	0	0	0
BRCA	0	56	244	0	0
KIRC	0	1	0	145	0
COAD	0	8	0	0	70

TABLE IV
RESULTS FOR THE FORMAL CONCEPT ANALYSIS CLASSIFICATION ALGORITHM WITH THE REDUCED DATA SETS.

Data set	**Accuracy**
5 PC's	83.125%
5 US	83.125%
11 PC's	67.875%

TABLE V
MEAN - FREQUENT SETS PCA5 DATA SET CONFUSION MATRIX, MIN_SUPPORT = 0.4, MIN_CONFI = 0.6, ACCURACY : 17.25%

	PRAD	LUAD	BRCA	KIRC	COAD
PRAD	16	131	7	9	131
LUAD	0	46	0	14	4
BRCA	3	335	50	27	15
KIRC	0	3	0	25	0
COAD	0	40	0	0	1

For 11 PCAs data set, in figure 9 and figure 10, with minimum support set to 0.4 and minimum confidence set to 0.6, the accuracy was 31.0% (Table VIII). When minimum support was set to 0.6 and minimum confidence was set to 0.6, the accuracy was 49.125% (Table IX). When minimum support was set to 0.8 and minimum confidence was set to 0.6, the accuracy was 72.25% (Table X). When minimum support was set to 0.6 and minimum confidence was set to 0.4, the accuracy was 49.125% and when minimum support was set to 0.6 and minimum confidence was set to 0.8, the accuracy was 49.125%.

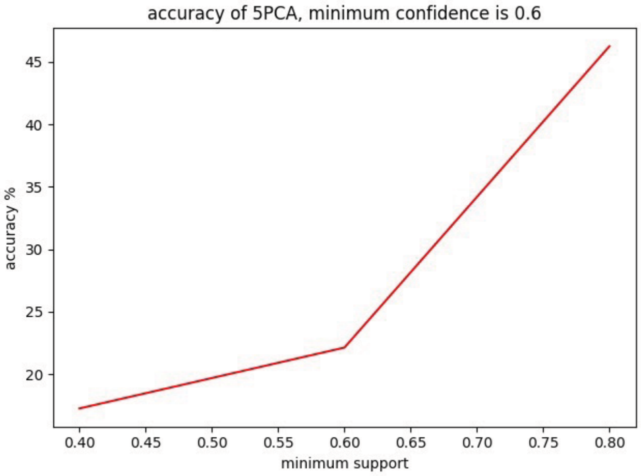

Fig. 7. Accuracy of Frequent Sets and Association Rules, minimum confidence is 0.6, 5PCA

VIII. DISCUSSION

According to figure 11, SVM has the highest accuracy (99.8% and 99.2%) out of all our objectives. Training a SVM model with the reduced data set resulted in the highest accuracy of 99.8% but also resulted in the longest run time,

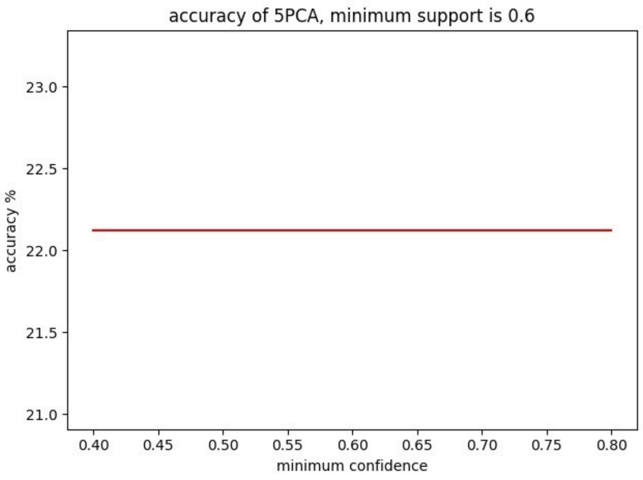

Fig. 8. Accuracy of Frequent Sets and Association Rules, minimum support is 0.6, 5PCA

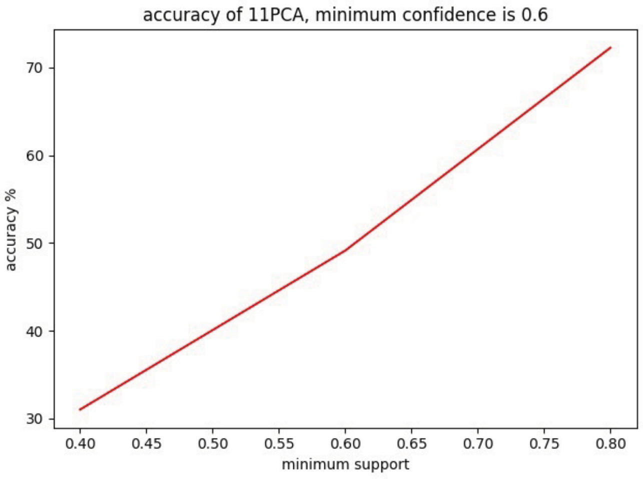

Fig. 9. Accuracy of Frequent Sets and Association Rules, minimum confidence is 0.6, 11PCA

more than ten times longer than the other data sets. The best SVM model based on total accuracy and total run time would be the one created from the 11 PCA data set. It had high accuracy (99.2%) and less than a second training time. Since SVMs finds hyperplanes that splits the data into groups based on classes, the only way that it could have done so well is that the original data set is set up in such a way that it is easily dividable based on different types of cancers based on the gene expression data that was collected.

TABLE VI
MEAN - FREQUENT SETS PCA5 DATA SET CONFUSION MATRIX, MIN_SUPPORT = 0.6, MIN_CONFI = 0.6, ACCURACY : 22.125%

	PRAD	LUAD	BRCA	KIRC	COAD
PRAD	16	131	7	0	131
LUAD	0	46	0	24	46
BRCA	3	225	50	27	225
KIRC	0	3	0	25	3
COAD	0	40	0	0	40

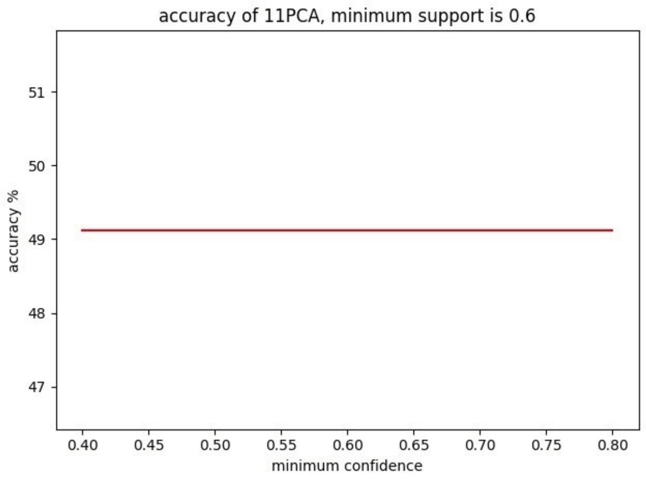

Fig. 10. Accuracy of Frequent Sets and Association Rules, minimum support is 0.6, 11PCA

TABLE VII
MEAN - FREQUENT SETS PCA5 DATA SET CONFUSION MATRIX, MIN_SUPPORT = 0.8, MIN_CONFI = 0.6, ACCURACY : 46.25%

	PRAD	LUAD	BRCA	KIRC	COAD
PRAD	16	131	8	9	131
LUAD	0	46	1	14 ‘	46
BRCA	3	225	243	27“	225
KIRC	0	3	1	25	3
COAD	0	40	0	0	40

Our second-best accuracy was K-Means with 91.75% on the 11-PCA data set, and 90.75% on the 5-PCA data set. The high accuracy could be due to the unsupervised nature of the clustering.

The third highest overall result (83.1%) is the Formal

TABLE VIII
MEAN - FREQUENT SETS PCA11 DATA SET CONFUSION MATRIX, MIN_SUPPORT = 0.4, MIN_CONFI = 0.6, ACCURACY : 31%

	PRAD	LUAD	BRCA	KIRC	COAD
PRAD	36	131	0	0	2
LUAD	0	54	0	0	11
BRCA	0	191	112	2	4
KIRC	0	1	0	23	1
COAD	0	54	0	0	23

TABLE IX
MEAN - FREQUENT SETS PCA11 DATA SET CONFUSION MATRIX, MIN_SUPPORT = 0.6, MIN_CONFI = 0.6, ACCURACY : 49.125%

	PRAD	LUAD	BRCA	KIRC	COAD
PRAD	36	131	0	0	2
LUAD	0	107	0	0	11
BRCA	0	227	204	2	4
KIRC	0	3	0	23	1
COAD	0	77	0	0	23

TABLE X
MEAN - FREQUENT SETS PCA11 DATA SET CONFUSION MATRIX, MIN_SUPPORT = 0.8, MIN_CONFI = 0.6, ACCURACY : 72.25%

	PRAD	LUAD	BRCA	KIRC	COAD
PRAD	36	131	0	0	2
LUAD	0	107	0	1	36
BRCA	0	227	265	2	7
KIRC	0	3	0	99	2
COAD	0	77	0	0	71

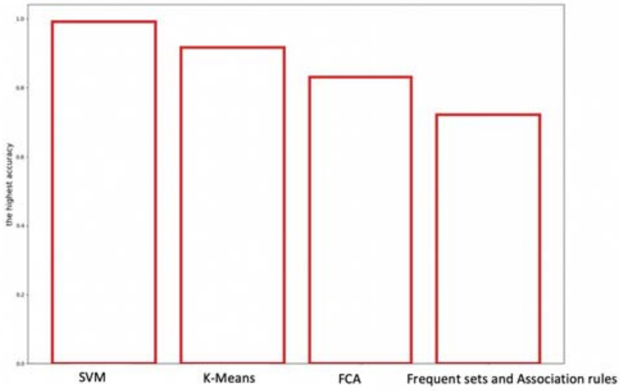

Fig. 11. Four Models' accuracy rank

Concept Analysis algorithm from the 5 PCA data set. The higher performance in the smaller data set is in line with the algorithm's requirements/limitations, as a smaller initial matrix can lead to a smaller number of concepts generated. This is beneficial because the classification process does not become intractable due to the amount of potential concepts to be generated.

Training a ”Frequent Sets and Association Rules” model with 11 USs data sets, and the minimum support is 0.8 and minimum confidence is 0.6, has the highest accuracy, 72.25%. Training the model with 5 USs data sets, and the minimum support is 0.4 and minimum confidence is 0.6, has the lowest accuracy, 17.25%. Moreover, with the increase of minimum support, the accuracy enhances. However, with the increase of minimum confidence, the accuracy does not change.

In this experiment, for the ”Frequent Sets and Association Rules” classification, the accuracy increase with the increase of the minimum support. However, the accuracy does not change with the change of the minimum confidence. The reason is that the frequent sets generated by higher minimum support mean occur in the type of tumor more. Moreover, association rules hardly ever exist in most types of tumors, so it cannot be a criterion of classification.

Overall, the methods with the largest resultant accuracy appear to be using an SVM, clustering with K-Means or using the FCA algorithm with a small data set. Using association rules can have a mid-range accuracy, but with low minimum support, the accuracy sharply decreases. Using the FCA algorithm for a large data set is also ill-advised, as the larger data set exponentially hinders the lattice construction.

IX. LIMITATIONS

In this section, we identify some limitations both overall in the classification process, as well as limitations that are individual to the specific algorithms.

One possible limitation to classification is that while some algorithms were able to classify the data with high accuracy, the data set only consisted of data from patients that had some form of cancer. It would be more beneficial if the data set consisted of patients that did not have any form of cancer.

There are three limitations to ”frequent sets and association rules” classification. The first one is that strong association rules may not exist in most types of tumors, so association rules cannot be a classify criteria. Another one is that because the largest size of the frequent item is very small, the frequent set usually exists in several types of tumors, which reduces accuracy a lot; for instance, A is the most frequent and largest-size itemset for BRCA, but a lot of COAD patients' mRNA-Seq also has the same itemset. The last one is that the Apriori algorithm is computationally expensive, so it cannot calculate on data sets with too many attributes.

A limitation of the FCA classification is that the code system was originally designed to classify images. A major limitation of using FCA is that large data sets are problematic. This is due to the way concepts are constructed. Without a constraint to limit concept discovery, as the size of the data set increases, the number of concepts increases exponentially. We created a reduced data set with 25 PCs to test if there was any additional accuracy information, but the algorithm could not be run on this data set due to system time and memory constraints.

X. CONCLUSION

This project classified real-valued gene expression data with four different methods: support vector machine, frequent sets and association rules, formal concept analysis, and K-Means to cluster and classify. SVM was run with multi-class support for the best accuracy out of all our methods. Frequent sets and association rules has reasonable results for a higher value of minimum support. Our FCA algorithm performs better in a smaller data set rather than a larger one. K-Means proved to be also very effective as it had the second-best accuracy next to the SVM.

One potential area for future research is that if there was data available for both patients with and without cancer, this model might be used to predict whether a patient has cancer. Furthermore, a model may be able to predict not only the presence of cancer but also which type of cancer may be present.

REFERENCES

[1] H. YiFei, G. Brad, and S. Adam, "Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data," Mar. 2017.

[2] S. Saskia, C, H. Melissa, P. Christine, S. Beverly, L. Celine, and C. Lyn, S, "Delivering genome sequencing in clinical practice: an interview study with healthcare professionals involved in the 100 000 genomes project," Sept. 2019.

[3] W. John, N, C. Eric, A, M. Gordon, B, M. S. Kenna, R, O. Brad, A, E. Kyle, S. Ilya, S. Chris, and S. Joshua, M, "The cancer genome atlas pan-cancer analysis project," Sept. 2013.

[4] "UCI Machine Learning Repository: Gene expression cancer RNA-Seq Data Set." https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq.

[5] N. R. Gotoor, "Image classification ssing fuzzy FCA," *Embargoed Master's Theses*, Dec. 2019.

[6] D. Jiang, C. Tang, and A. Zhang, "Cluster analysisfor gene expression data: A survey." https://www.computer.org/csdl/journal/tk/2004/11/k1370/13rRUy2YLTf, Nov. 2004.

[7] Y. Lu and J. Han, "Cancer classification using gene expression data," *Information Systems*, vol. 28, pp. 243–268, June 2003.

[8] G. V. S. George and V. C. Raj, "Review on feature selection techniques and the impact of svm for cancer classification using gene expression profile," *International Journal of Computer Science and Engineering Survey*, vol. 2, pp. 16–27, Aug. 2011.

[9] E. Alba, J. GarciaNieto, L. Jourdan, and E. Talbi, "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms," in *2007 IEEE Congress on Evolutionary Computation*, pp. 284–290, Sept. 2007.

[10] X. Zhang and H. Ke, "ALL/AML cancer classification by gene expression data using SVM and CSVM approach," *Genome Informatics*, vol. 11, pp. 237–239, 2000.

[11] N. Yuvaraj and P. Vivekanandan, "An efficient SVM based tumor classification with symmetry Non-negative matrix factorization using gene expression data," in *2013 International Conference on Information Communication and Embedded Systems*, pp. 761–768, Feb. 2013.

[12] M. Wang, X. Su, F. Liu, and R. Cai, "A cancer classification method based on association rules," in *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 1094–1098, May 2012.

[13] R. Cai, Z. Hao, W. Wen, and H. Huang, "Kernel based gene expression pattern discovery and its application on cancer classification," *Neurocomputing*, vol. 73, pp. 2562–2570, Aug. 2010.

[14] L. Huang, "An integrated method for cancer classification and rule extraction from microarray data," *J Biomed Sci*, vol. 16, p. 25, Feb. 2009.

[15] H. Mahmoodian, M. Hamiruce Marhaban, R. Abdulrahim, R. Rosli, and I. Saripan, "Using fuzzy association rule mining in cancer classification," *Australas Phys Eng Sci Med*, vol. 34, pp. 41–54, Apr. 2011.

[16] Z. Kakushadze and W. Yu, "K-means and cluster models for cancer signatures," Sept. 2017.

[17] A. Dubey, U. Gupta, and S. Jain, "Analysis of k-means clustering approach on the breast cancer wisconsin dataset," June 2016.

[18] S. Hengpraprohm and P. Chongstitvatana, "Selecting informative genes from microarray data for cancer classification with genetic programming classifier using k-means clustering and snr ranking," in *2007 Frontiers in the Convergence of Bioscience and Information Technologies*, pp. 211–218, Oct. 2007.

[19] R. Wille, "Restructuring lattice theory: An approach based on hierarchies of concepts," in *Formal Concept Analysis* (S. Ferré and S. Rudolph, eds.), Lecture Notes in Computer Science, pp. 314–339, Springer Berlin Heidelberg, 2009.

[20] K. Raza, "Formal Concept Analysis for Knowledge Discovery from Biological Data," *International Journal of Data Mining and Bioinformatics*, vol. 18, no. 4, p. 281, 2017.

[21] B. Ganter and R. Wille, "Formal concept analysis: mathematical foundations," Dec. 2012.

[22] R. Belohlavek, "What is a fuzzy concept lattice? II," in *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing* (S. O. Kuznetsov, D. Ślezak, D. H. Hepting, and B. G. Mirkin, eds.), vol. 6743, pp. 19–26, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[23] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[24] R. Agrawal, "Fast algorithms for mining association rules," p. 13.

[25] Z. Nianyin, Q. Hong, W. Zidong, L. Weibo, Z. Hong, and L. Yurong, "A new switching-delayed-pso-based optimized svm algorithm for diagnosis of alzheimer's disease," Dec. 2018.