Study and comparison of algorithms for cancer prediction using gene expression data.

Contributors

- Amarjeet Kumar (UFID - 38817064),
- Ujwala Guttikonda (UFID - 57914323),
- Nikhil Yerra (UFID - 95453265)

Introduction:

The physical characteristics and symptoms of the patient are traditionally used in cancer classification, one of the crucial fields of medical sciences, to acquire a deeper understanding of the patient's situation. Studying the gene expression data is crucial since the cause and spread of cancer are more closely tied to genes.

Therefore, we categorize each kind of cancer into its subtypes in this study by looking at the data on gene expression for distinct cancer types. We collect samples from GEO datasets and use the standard deviation by mean ratio, ANOVA - F value, and 2 statistics to identify the best features. Then we ran the K Nearest Neighbors, Naive Bayes, and Random Forest classifiers. Then, for all possible combinations, we compare a classifier's accuracy to a feature selection method and look at a feature selection, better accuracy-producing classifier pair for a given dataset.

Analysis of Datasets

The Series Matrix Files of GEO datasets were used in the experiments; they were downloaded from https://www.ncbi.nlm.nih.gov/gds. With arrays of a matrix file designating the probe id and columns designating the samples, each dataset contains the genomic data for various samples.

Series Matrix Files Samples and PROBE IDs are the two components of a series matrix file. A PROBE ID is the same as the microarray chip's PROBE ID, which was used to create the file's data. The values from of the microarray are entered for each sample in the PROBE ID field. These numbers represent the levels of gene expression of the genes connected to a specific PROBE. This same series matrix file additionally includes the metadata that really is pertinent to the data set. We have taken five GEO datasets, which have been briefly explained below:

Breast Cancer Dataset (GSE27562):

The breast cancer dataset contains 54, 675 probe sets and 31 normal specimens, 57 premalignant cancer specimens, and 37 innocuous breast cancer samples. In order to conduct the experiments, we have removed 15 gastrointestinal tumor samples and 7 brain tumor samples from the datasets. Because there is no data to suggest that 15 cancer samples, we have also excluded those samples.

Lung Cancer Dataset (GSE19804):

GSE19804 is a lung cancer dataset with 120 samples, 60 tissue samples—60 from lung cancer and 60 from normal tissue—and 54, 675 probe sets.

Tumor dataset (GSE4290):

It contains 23 samples. 157 tumor samples include 26 astrocytomas, 50 oligodendrogliomas and 81 glioblastomas.

Pancreatic dataset (GSE59856)

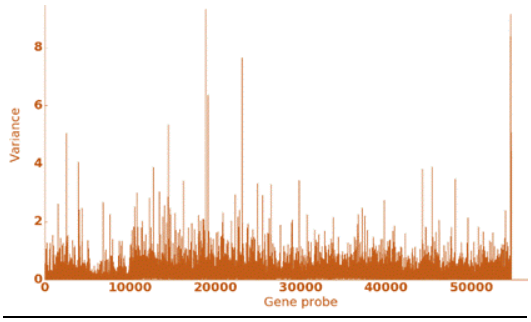
The dataset includes 150 healthy samples, 52 liver cancer samples, 50 colorectal cancer samples, 98 biliary tract specimens, 50 carcinoma samples, 50 esophageal samples, 50 pancreatic cancer samples, and 100 pancreatic cancer samples. There are also 2555 probe sets in this dataset.

Leukemia Dataset (GSE33315):

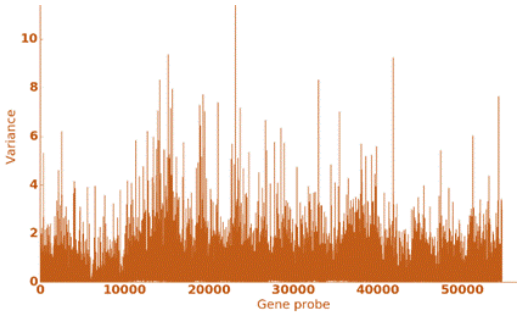
There are 7 multiple kinds of leukemia cancer in the dataset. With 22,283 probe sets, it contains 83 samples of T-ALL, 115 samples of type hyperdiploid, 40 specimens of TCF3-PBX1, 99 samples of ETV RUNX1, 30 specimens of MLL, 23 samples of PH, and 23 samples of hypodiploidy. Four CD34 and four CD10CD19 samples were disqualified because they were insufficient for classification. In order to concentrate on leukemia, we also exempted 153 samples with an unknown karyotype.

Initial data analysis

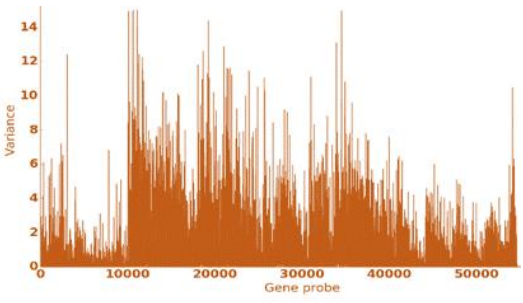
We analyzed to analysis the data before proceeding to preprocessing and training operations and we obsered the variance of each and every dataset as follows:



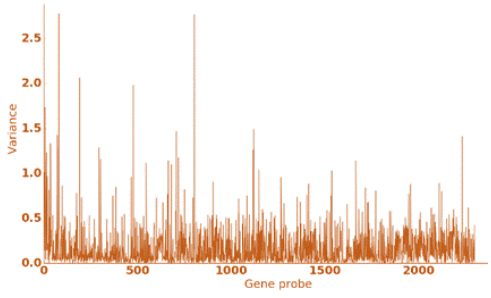
Variance of gene expression level in breast cancer



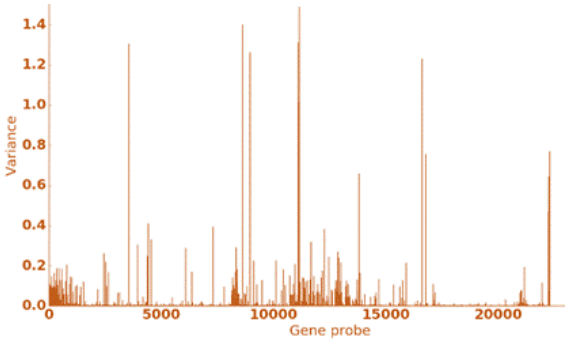
Variance of gene expression level in Lung cancer



Variance of gene expression level in tumor



Variance of gene level in pancreatic cancer



Variance of gene level in leukemia cancer

Classification Models

We used the below-mentioned methods for classification: K-Nearest Neighbors, Naive Bayes, and Random Forest.

K-Nearest Neighbors: One of the basic classification techniques is K-nearest neighbors. We record the category of the instances in the training set when we train the algorithm. We identify the Closest k neighborhood of such a group in terms of shortest distance from any of this sample when trying to predict the class of a new sample using the trained model. The category of a new sample can be determined using any distance metric. Euclidean distance has been used as a distance metric in our experiments, with K=1.

Naive Bayes: The posterior probability of a provided sample can be calculated by the Naive Bayes classifier using the Bayes theorem. It bases its calculation on the idea of feature independence. In our situation, since there is a high likelihood of genes being related to one another, this is rarely the case. For a hypothesis *h* on a given data *D*, the Bayes theorem to calculate the probability of *h* given the event *D* has occurred is:

$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

where *P(h)* and *P(D)* are independent probabilities of *h* and *D* respectively. *P(D|h)* is the conditional probability of *D* given *h*. The classifier used, makes use of Gaussian
Distribution.

Random Forest: A classification algorithm made up of numerous decision trees is called the random forest. It tries to build an uncorrelated forest of trees whose estimation by committee is much more accurate than that of any individual tree by using bagging and highlight unpredictability once building each individual tree.

Feature selection

As can be seen from the characterization of the sets of data above, these same gene expression sets of data have high - dimensional in proportion to the number of observations in the dataset. However, not all features are necessary for classification because some of them hold data that is redundant and others include data that is unrelated to classification. This will not only speed up the training and testing of the model and increase its accuracy as a classifier.

There are several methods for selecting or extracting features. For our experiments, we used three statistical techniques, which are described below. The F1 score is used to gauge how accurate each one of these classifiers is.

Standard deviation by the mean ratio: The approach described above is simplified in this way. Since most classes overlap even for features with the highest variance, as we saw above in the particular instance the of pancreatic cancer dataset, the category cross-correlation score may not produce satisfactory results in this situation. As a result, to choose the significant features, we merely use variability by an average ratio of the features. The idea that any feature showing increased variance in relation to its mean would be having significant information about the dataset, regardless of the classes, determines the significance of the features. This method chooses the k features with the maximum variance by mean ratio. We have changed the value of k in our studies from 100 to 500.

Analysis of Variance (ANOVA) F-value: We determine the Analysis of variance F value when using this feature selection technique. Analysis of variance employs conventional, standardized language. The definitional equation of sample variance is $s2=1n^{-1}\sum(y_i-\bar{y})^2$. where the sum of squares concepts are variances from the sample mean, the divisor is known as the freedom degrees (DF), the sum is known as the total of the squares (SS), and the result is known as the mean square (MS). The analysis of variance (ANOVA) calculates 3 different variances. The distinction between both the variability of observational data and the variability of implies is accounted for by multiplying the deviations of the treatment implies from the ballroom mean by the number of measurements included in each treatment.

χ^2 statistics: Under this feature selection technique, the expected and observed values of a gene from across Size n are used to calculate the significance of a given feature using the 2 statistics. Let O represent the sample's observed value for a particular gene, and let E represent the expected value. Gene expression values are discretized into a number of intervals, and the 2 value is determined for each interval that use the equation below and then added up at the end. We compute the 2 statistics over the N samples for each gene. We can use the 2 value provided by this equation to recognize or reject a genotype. So, we choose the top k genes.

$$\chi 2 = \sum \frac{(O-E)^2}{E}$$

Results

We analyzed the above-mentioned 5 datasets with the three machine learning algorithms, respectively KNN, Naive Bayes, and Random Forest. The accuracy of the methods has been compared with chai, Anova, and standard mean accuracy feature selection parameters. We created five tables representing each database. Every table has three entries representing three feature selections and accuracy. Data has been randomized while keeping the 80% testing and 20% training ratio so that the accuracy will vary. To handle this problem, 10 iterations have been used, and the final accuracy is the average of these individual iterated accuracies. This led to a holistic comparison analysis.

Result accuracy for a classifier v/s feature selection

Breast cancer dataset

<500 Genes	Naïve Bayes	KNN	Random Forest
Original features	(54, 53958)	(68, 53958)	(60, 53958)
χ^2	(62, 200)	(52, 300)	(64, 200)
Mean ratio of std	(60, 300)	(65, 400)	(67, 400)
ANOVA F-classifier	(64, 100)	(76, 100)	(57, 100)

Lung cancer dataset

<500 Genes	Naïve Bayes	KNN	Random Forest
Original features	(94, 53958)	(86, 53958)	(92, 53958)
χ^2	(95, 500)	(97, 100)	(95, 200)
Mean ratio of std	(93, 300)	(95, 400)	(94, 400)
ANOVA F-classifier	(96, 100)	(96, 100)	(92, 500)

Tumor Dataset

<500 Genes	Naïve Bayes	KNN	Random Forest
Original features	(100, 53958)	(100, 53684)	(89,53826)
χ^2	(98, 400)	(100, 100)	(94, 500)
Mean ratio of std	(98, 300)	(100, 200)	(89, 500)
ANOVA F-classifier	(100, 300)	(100, 200)	(92, 200)

Pancreatic Dataset

<500 Genes	Naïve Bayes	KNN	Random Forest
Original features	(66, 2100)	(65, 2100)	(67, 2100)
χ^2	(63, 200)	(68, 200)	(65, 500)
Mean ratio of std	(54, 200)	(57, 200)	(63, 500)
ANOVA F-classifier	(68, 200)	(72, 200)	(65, 200)

Leukemia Dataset

<500 Genes	Naïve Bayes	KNN	Random Forest
Original features	(86,21573)	(72,21573)	(82, 21573)
χ^2	(92, 400)	(82, 100)	(88, 400)
Mean ratio of std	(76, 500)	(71, 500)	(76, 400)
ANOVA F-classifier	(93, 100)	(86, 500)	(86, 100)

In the Breast cancer, ANOVA-F, and KNN feature selection-classifier duo produced the best accuracy. Compared to the initial number of characteristics, the accuracy has significantly improved (by about 10%). (For KNN). Additionally, 100 chosen features in each classification have the maximum accuracy. Even for KNN, which had the highest accuracy for 100 features, adding more features did not increase the accuracy for other classifiers.

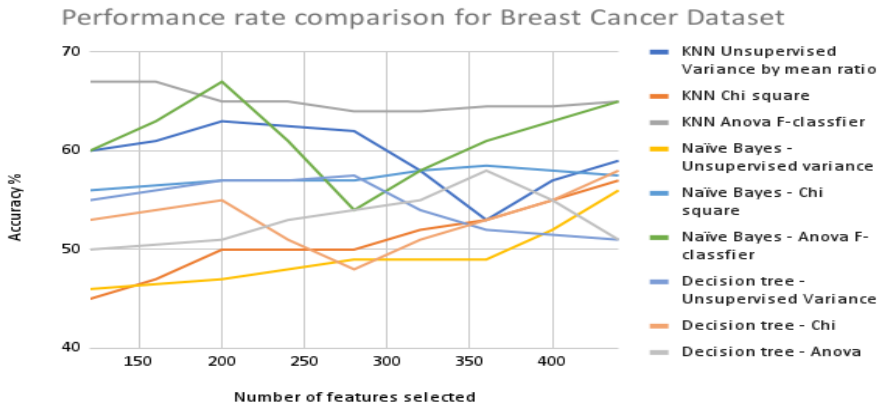
In the lung cancer (GSE19804) data set, the 2 statistics and KNN classifier for 100 chosen features set produced the maximum accuracy. This is better with only 100 selected features as compared to the greatest accuracy without feature selection. The accuracy does not improve much as the number of features grows. Although they need more characteristics (>100), the other two classifiers, Naive Bayes and Random Forest can also achieve a reasonable level of accuracy. We can also see that the Std. by Mean Ratio shows some improvement over the initial feature set (by 2%-9%).

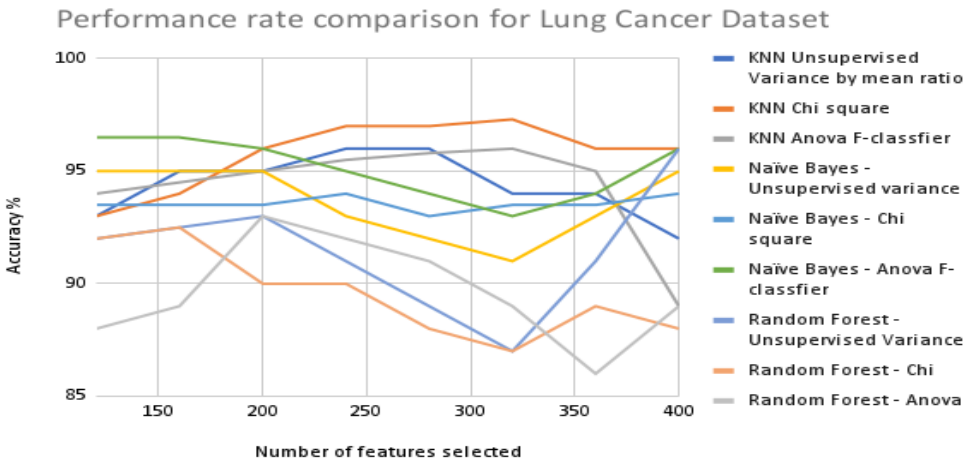
In Tumor Dataset, feature selection helps us get a good level of accuracy, although it should be noted that two out of three classifiers may achieve 100% accuracy even without feature selection. Here, we achieve greater accuracy for between 100 and 200 characteristics. The accuracy of Naive Bayes and Random Forest only diminishes as the number of features increases.

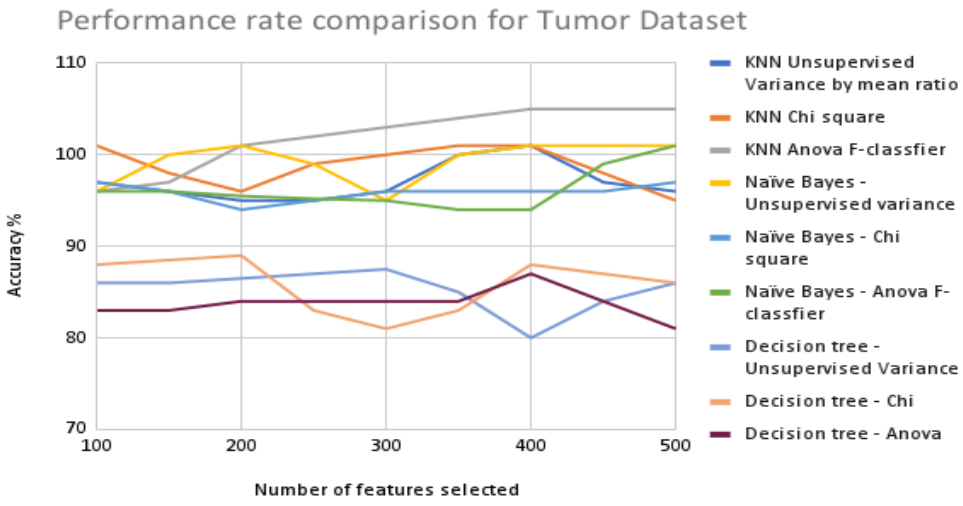
In the Leukemia Dataset, the Naive Bayes classifier provides the maximum accuracy. The naive Bayes tends to have more accuracy over each feature set for 100 features when we compare the classifier with the feature. While feature selection would undoubtedly increase accuracy, adding more features than 100 will cause the classifier to converge on a less accurate result.

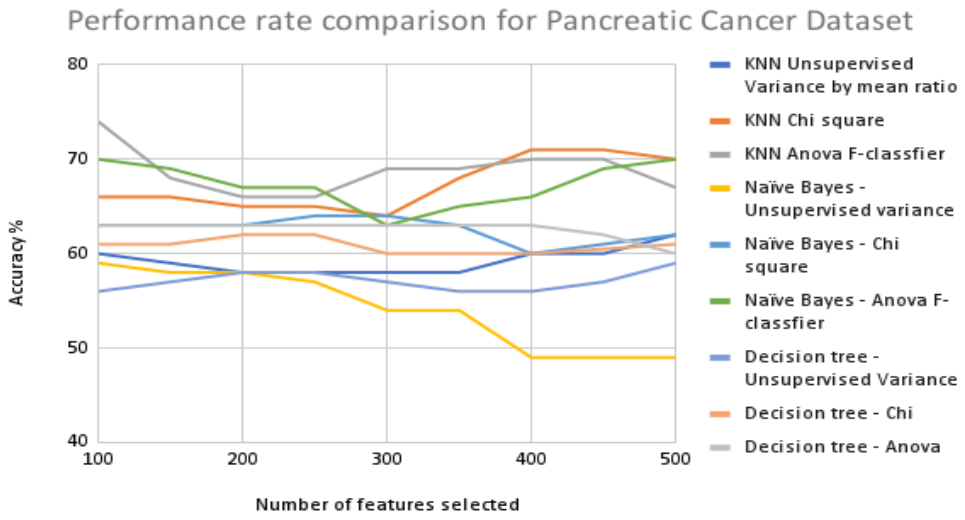
In the Pancreatic Dataset Test, the best accuracy is attained using KNN and a pair of 2 classifier-feature selection algorithms with 500 features (Going against the trend of 100 features). There are the most features that have been chosen thus far. In this instance, fewer features do not equate to more accuracy. Since certain features tend to be more crucial for categorization, we observe this variation (from 100 features) here.

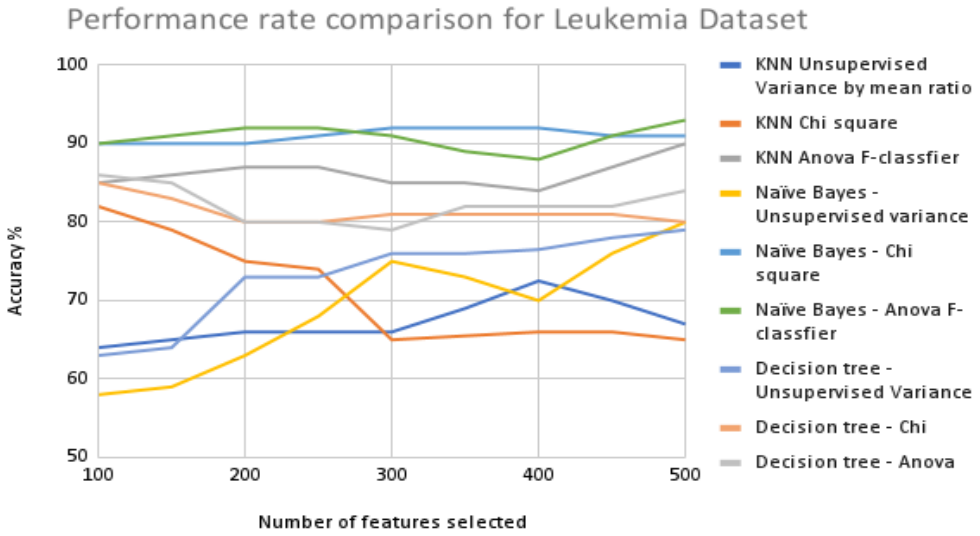
The following graphs show a performance rate comparison for each feature selection for all the data sets:-











Conclusion

In all the five above-mentioned datasets, the highest accuracy was observed while keeping the number of features to 100. KNN classifier gave the best accuracy in most data sets.

The accuracy can be further improved by selecting features based on biological significance. Important features can be narrowed down for classification if we use the larger database as the algorithm will have better training.

Individual Contribution of members:

1. Data extraction and preprocessing of datasets:-
    - Breast Cancer Dataset, Lung Cancer Dataset:- Amarjeet Kumar
    - Tumor Dataset, Pancreatic Cancer Dataset:- Nikhil Yerra
    - Leukemia Dataset: Ujwala
2. Comparative analysis and study of algorithms:-
   All three members

3. Running the classifier and checking the accuracy:-
    - K Nearest Neighbors:- Amarjeet Kumar
    - Naïve Bayes:- Ujwala
    - Random Forest:- Nikhil Yerra

4. Performance comparison and Result Analysis:-
Every member compared the performances and discussed the results with each other.

References

[1] Khan, J., Ringner, M. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks

[2] Pati,J. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks

[3] Liu, S., Yao, W. Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection

[4] Sotiriou, C., Neo, S. Breast cancer classification and prognosis based on gene expression profiles from a population-based study

[5] Xie, Y., Lynn, B. Breast cancer gene expression datasets do not reflect the disease at the population level

[6] Batra, I., Luthra, M. Gene Expression-Assisted Cancer Prediction Techniques