Identifying Patterns of Breast Cancer Genetic Signatures using Unsupervised Machine Learning

Rifat Hamoudi[1], Meriem Bettayeb[2], Areej Alsaafin[3], Mahmood Hachim[4], Qassim Nassir[5], Ali Bou Nassif[6]
[1,4]*Sharjah Institute for Medical Research; Department of Clinical Sciences, College of Medicine*
[2]*Department of Electrical and Computer Engineering, College of Engineering*
[3]*Department of Computer Science, College of Science*
[5]*Department of Electrical Engineering, College of Engineering*
[6]*Department of Computer Engineering, College of Computing and Informatics*
University of Sharjah
Sharjah, UAE
{rhamoudi, u17105766, u15100713, u16101425, nasir, anassif}@sharjah.ac.ae

***Abstract*—Deploying machine learning to improve medical diagnosis is a promising area. The purpose of this study is to identify and analyze unique genetic signatures for breast cancer grades using publicly available gene expression microarray data. The classification of cancer types is based on unsupervised feature learning. Unsupervised clustering use matrix algebra based on similarity measures which made it suitable for analyzing gene expression. The main advantage of the proposed approach is the ability to use gene expression data from different grades of breast cancer to generate features that automatically identify and enhance the cancer diagnosis. In this paper, we tested different similarity measures in order to find the best way that identifies the sets of genes with a common function using expression microarray data.**

Keywords—gene expression, unsupervised machine learning, hierarchical clustering, adaptive filtering, breast cancer

I. INTRODUCTION

Disease in the human body result from modifications in genes at the human DNA level which transmits signals in genetics terms (i.e. genetic signals) [1]. For simplicity, we can think about the cell as a software engineering paradigm. The start is at the DNA level which is analogous to the source code, and the transcription from the DNA to RNA is analogous to compiling the code. The RNA can be treated as the abstract syntax of the source code and the protein is analogous to generating machine code to create executable software. Therefore, a fault in the DNA will cause the machine code to behave differently. This is because that the mutant DNA can cause the resulting protein to be produced differently and this different protein can lead to disease through loss of regulation on cellular pathways. In computer programming, it is difficult to troubleshoot a problem at the semantic level or in machine code [2]. However, it is easy to detect syntax errors by the compiler (RNA). Therefore, working at the RNA level (middle layer) is easier as it is possible to diagnose problems in both the DNA (source code) and protein (machine code).

As disease is a result of modifications in genes in the DNA, diseases vary according to the functions performed by the mutated gene. Furthermore, the same genetic signal can have a specific pattern in healthy people, but when this pattern changes, it can lead to disease by modifying the regulation of cellular pathways [3].

The authors would like to acknowledge OpenUAE Research and Development group at the University of Sharjah and Al-Jalila Foundation (Grant code: AJF201741) for funding this work.

This is analogous to a situation when someone feels relaxed whilst listening to music at a reasonable level of volume. The same music audio signals can be frustrating if the volume is raised to a higher level. This makes it more complex and difficult to differentiate between the role of the single gene in normal state and disease state.

In this paper, we use the RNA layer to search for the genetic changes in breast cancer using gene expression microarray images as a starting point. Each gene expression microarray image contains expression data of 25,000 different genes where the expression level is represented by the pixel intensity. This is similar to Hubble Space Telescope (HST) images in astrophysics, where Doppler shift is used to see if the stars are moving away from (redshift) or towards (blueshift) Earth [4]. However, many data points exist in the captured images, making it difficult to pinpoint the pattern that relate to the disease. In breast cancer, the expression level of 25,000 different genes need to be differentiated among the different breast cancer disease grades. Therefore, the aim of this study is to search publicly available gene expression data of breast cancer cases from early (grade 1), intermediate (grade 2), and late stage (grade 3) which tend to be aggressive and use them to identify genetic signature patterns using signal processing approach. This was carried out using unsupervised learning to check whether the signature on its own is able to separate the genetic signals. In other words, we examined if applying some engineering mathematics principles followed by unsupervised clustering is enough to separate the genetic signature to its original compartments.

Most artificial intelligence (AI) applications requires training set within the model. However, medical applications and models are non-linear (stochastic) and therefore there is no defined training set in biomedical fields [5]. This is because that the disease is changing (adaptive), making it difficult to predict what the signal would look like.

Hierarchal clustering is a mathematical technique that is used in various engineering fields such as robotics and computer vision [6]. The advantage of clustering is that it relies purely on the data to classify the components of the model according to the data given, thus, it does not require any other parameters or complex probabilities. In addition, the clustering algorithm is based on geometry and thus it uses simple matrix algebra to find similarity measurements that define the cluster boundaries. This light-weight approach is useful in circuit design and robotics where the code can be embedded on the chip itself leading to improvements in circuit design leading to better robotics performance.

Therefore, in this study, we identified different grades from breast cancer cases by mining publicly available gene expression data. We then tested several clustering techniques to find the optimal clustering algorithm that can be applied on this data to separate the signal patterns. Therefore, the aim of the study is to test the ability of various algorithms to separate the genetic signals and test the possibility of separating signals that are too close to each other and have very similar patterns.

The novel characteristics introduced in the presented study are summarized in the following points:

- Define a roadmap for identifying novel pathways and genes that are responsible for causing different types of breast cancer.
- Perform data mining by filtering the noisy genes using semi-intelligent adaptive filters which will overcome an important optimization problem named curse of dimensionality which will be discussed later.
- Demonstrate the effects of applying unsupervised clustering to identify functional groups related to breast cancer progression.

The outline of this research paper is as follows: Section 2 introduces the conducted methodology to identify the breast cancer types. The results of performing data pre-processing and unsupervised clustering to breast cancer datasets are illustrated in section 3. Section 4 discussed the obtained results and benefits of our methodology and the conclusion is presented in section 5.

II. Materials and Methods

This section explains the conducted methodology in this research as illustrated in Figure 1. The following subsections *A* to *D* shows the details of the steps utilized in this paper.

A. Patient Materials

Fresh frozen tissues from 31 well-characterized breast cancer cells, 5 samples in the early stage which is in grade 1, 3 samples form grade 2, and 23 samples from late stage which is grade 3 were used for gene expression microarray analysis. This is reflective of the fact that early breast cancer is difficult to detect and therefore that is the reason why we got 8 samples from grade 1 and 2.

B. Gene Expression Microarray

All microarray data have been taken from Gene Expression Omnibus (GEO) Database [1] study number GSE87007 [7]. The search was filtered by several factors: biopsy, expression profiling by array (RNA), Affymetrix, GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array.

C. Data Pre-processing

Data pre-processing mainly consists of two main stages. Firstly, background correlation which removes the background signal from the probe intensity in order to get more accurate data. Secondly, normalization followed by filtering which involves the elimination of systematic variation in a microarray, so it does not affect the amount of gene expression in the probe of the array [8].

There are different techniques to perform normalization, such as Robust Multi-array Average (RMA) and quantile normalization (GCRMA), which became the default for gene expression microarray studies as it incorporated biological information on GC content within the algorithm [8]. Another normalization method is MicroArray Suite (MAS) which is similar to the band pass filter. The MAS normalization was modified to adaptive variation filter (similar to Kalman filter) by using a cutoff of 100 for the absolute noise and variation of 3 (since the minimum sample is 3 belonging to grade 2). In this study, we consider the combination result of both GCRMA and MAS.

D. Unsupervised Machine Learing Using Hierarchical Clustering

In order to properly divide or merge the clusters, similarity calculations between two clusters need to be obtained. There are different methods to get the clustering similarity such as: min, max, average linkage, distance between centroids and ward's linkage method.

In this study, we use and compare between average and ward methods. Average linkage method takes all the pairs of points and compute their similarities and calculate the average of the similarities. Ward's method of calculating the similarity between two clusters is the same as average linkage except that Ward's method calculates the sum of the square of the distances within the clusters and merges them to minimize it. From a statistical viewpoint, the process of agglomeration leads to a reduction in the variance of each resulting cluster. The motivation behind choosing these two states of art approaches is that they both do well in separating clusters if there is noise between clusters.

The choice of distance measure in a clustering method is very important as it calculates the similarity between two elements (x,y) which will affect the clusters' shape. The choice of an appropriate metric will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. Euclidean and Manhattan are the classical techniques used for distance measures [9]. Euclidean distance is the "ordinary" straight-line distance between two points in Euclidean space as shown in Equation (1). With this distance, Euclidean space becomes a metric space. Manhattan is like Euclidean, but the distance is calculated by summing the absolute value of the difference between the dimensions as presented in Equation (2). For example, Manhattan distance implies moving straight, first along one axis and then along with the other.

$$d_{man}(x,y) = \sum_{i=1}^{m} |x_i - y_i| \quad (1)$$

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \quad (2)$$

Clustering was achieved among all the breast cancer tissues using the Euclidean and Manhattan distance measures in order to separate the different grades of breast cancer types.

[1] https://www.ncbi.nlm.nih.gov/geo/

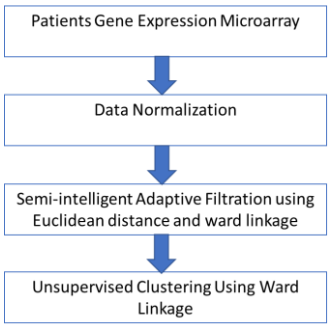

Figure 1. Data analysis methodology

III. RESULTS

This section shows the results for the application of data pre-processing and unsupervised clustering to breast cancer dataset.

The output of the gene expression microarray is illustrated in Figure 2. This raw gene expression image presents the intensity of each pixel or probe feature which represents the amount of gene in the specified sample in the microarray [10].

Image analysis is carried out to preform data pre-processing to the raw data followed by filtering the non-variant probes and, identify the differentially expressed genes in order to visualize the correlation between these data.

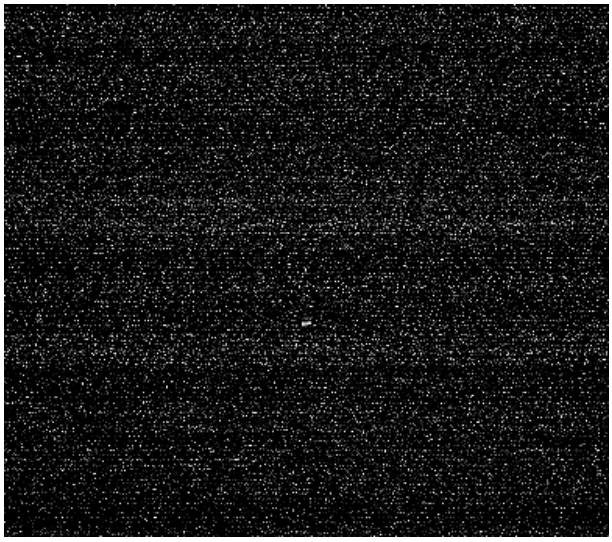

Figure 2. Gene expression microarray

The effect of GCRMA normalization algorithm is illustrated in Figure 3. As shown in the diagram, the un-normalized data show wide variations, making it difficult to compare the expression of genes across the samples. The normalized data show more consistent variations across all samples making it easier to compare genes across all the samples. The histogram of the normalized data using GCRMA is shown in Figure 4. The dotted line indicated the separation between the clean and noisy signal.

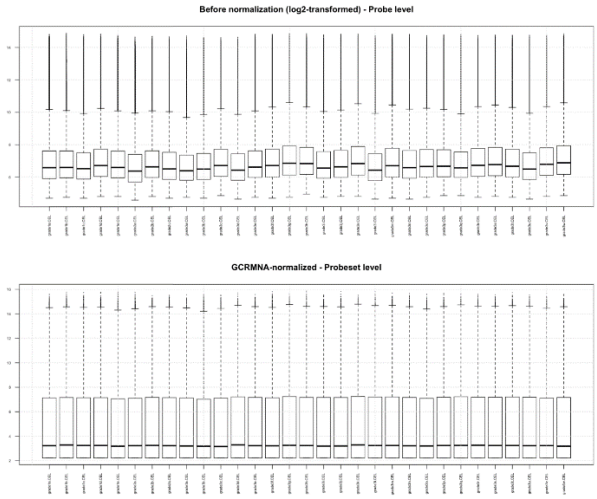

Figure 3. Normalization of gene expression microarray data using GCRMA

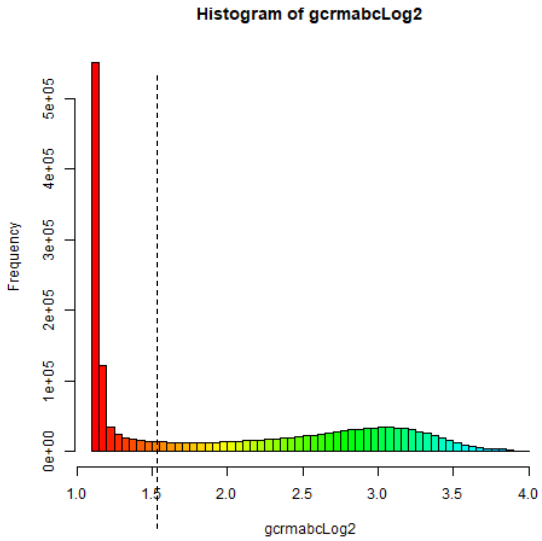

Figure 4. Histogram of filtered data using GCRMA

We started with 54675*31=1,694,925 features from the raw data. However, after filtration the number of features is reduced significantly to 9282*31=287,742 features as shown in the flowchart in Figure 5.

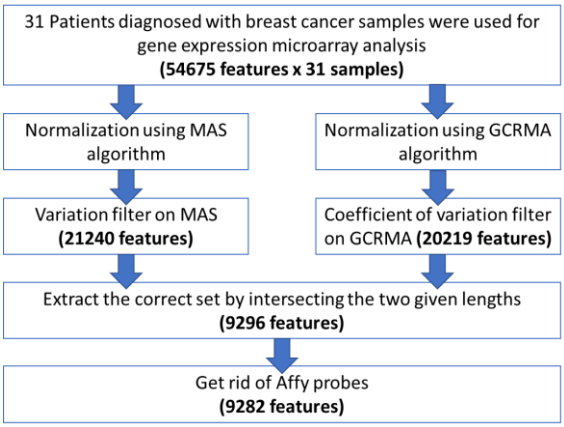

Figure 5. Feature reduction

After getting the clustering trees using average linkage and ward linkage including both types: Euclidean and

Manhattan, statistical evaluation parameters to measure the performance of each clustering type can be calculated. Furthermore, comparisons between the different clustering types can therefore be conducted.

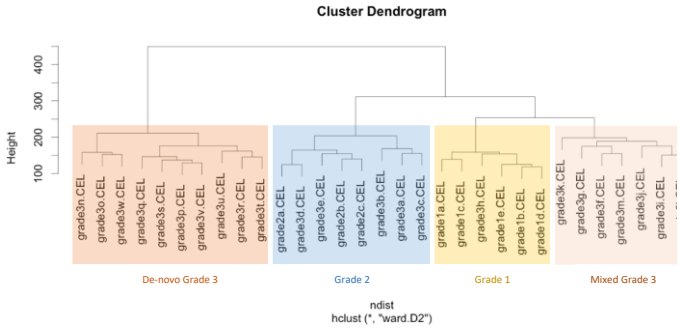

Figure 6. Cluster tree of Ward linkage and Euclidean distance of normalized (GCRMA) and filtered data

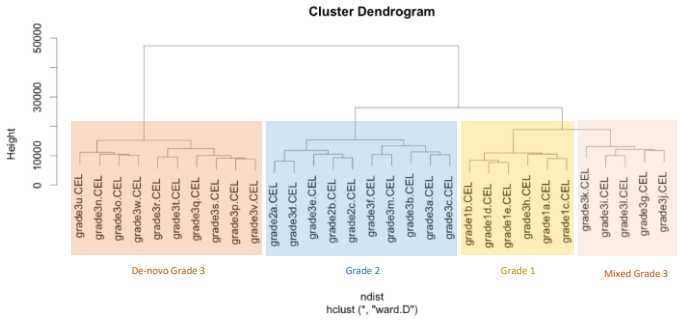

Figure 7. Cluster tree of Ward linkage and Manhattan distance of normalized (GCRMA) and filtered data

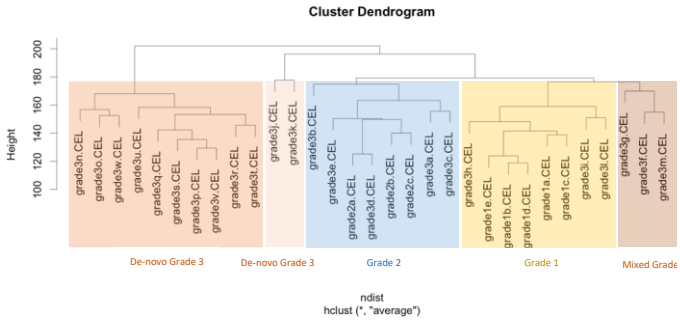

Figure 8. Cluster tree of Average linkage and Euclidean distance of normalized (GCRMA) and filtered data

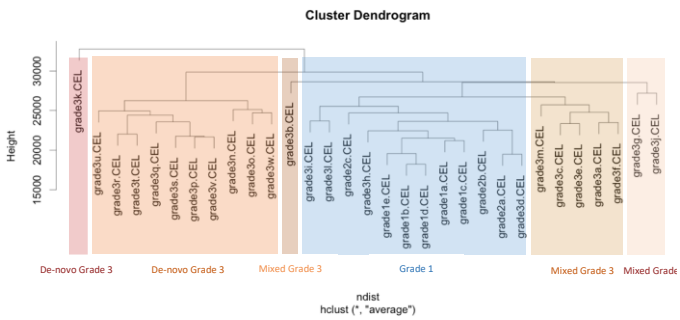

Figure 9. Cluster tree of Average linkage and Manhattan distance of normalized (GCRMA) and filtered data

Figure 6 – 9 show the cluster trees of normalized (GCRMA) and filtered data using Ward linkage and Average linkage with both Euclidean and Manhattan distance. The number of correct classified (true positive) cases, according to their actual grades, using the different clustering methods is shown in Table 1. As illustrated in the table, Ward-Euclidean outperforms the other three methods in terms of true positives.

Table 1. Statistical measures of different clustering methods

Clustering Method	True Positive (%)
Ward-Euclidean	80%
Ward-Manhattan	74%
Average-Euclidean	74%
Average-Manhattan	77%

IV. DISCUSSION

In this paper, we identified a way to filter out the noise but keep the true signal which will solve a critical problem in optimization. Filtering is not only useful for pattern recognition, eliminating the noisy areas will solve an important issue in data science which is referred to as “curse of dimensionality” or “combinatorial explosion”. Typically, the datasets in machine learning have a high dimensionality while actually the true dimensionality is often much lower than that. The curse of high dimensionality is that as dimensions grows, Euclidean distances become less meaningful and uniform distributions become exponentially harder to sample. Furthermore, many parameters become polynomialy more difficult to estimate as the number of dimensions increase and most importantly data becomes more difficult to visualize. Our algorithm can partially solve some of optimization problem by eliminating genes or data points that are invariant across all the samples using the combination of two normalization methods (MAS and GCRMA) and two effective filtering methods (Coefficient of Variation filter on GCRMA and Variation filter on MAS). Using these techniques results in a significant reduction in the number of features. As a result, the filtering algorithm will be useful in customizing the resources, reducing the time of execution and increasing the efficiency O(n).

None of the clustering algorithms are perfect but best after filtering. After filtering we get better accuracy but still not perfect and the reason behind that is the gene signals are closed to each other. However, we managed to separate them and find the subset. The heat map of the best clustering algorithm which was obtained using ward Euclidean similarity measure is illustrated in Figure 10. Interestingly, the cluster seems to identify a potentially new grade 3 group which has overlapping genetic signature with groups 1 and 2.

In addition, Figure 10 shows that the clustering identified functional groups related to breast cancer progression including cell division, oncogenic pathways such Ras pathway, immune response pathways as well as cell signaling pathways such as phosphorylation, Wnt signaling and transcription regulation. This shows that our analysis is correct as we identify Ras pathway which has been reported in the literature many times over the last 40 years. Interesting finding is the immune response pathways which links to the latest in cancer therapy termed immunotherapy and the finding that IGFBP6 which is Insulin like growth factor is differentially expressed across the grades.

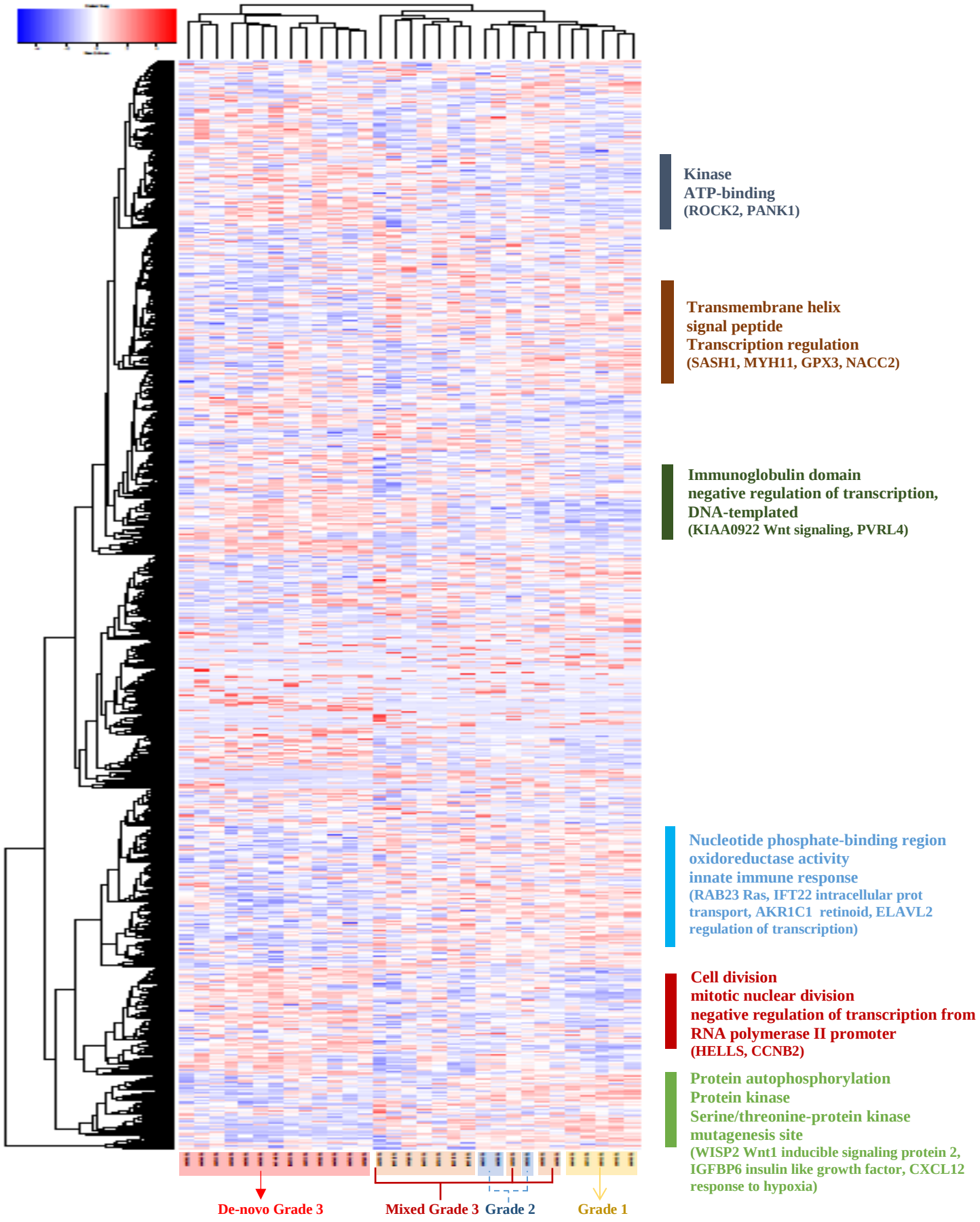

Figure 10. Heatmap of the breast cancer cluster

Such technique can be applied to other areas in artificial intelligence such as computer vision by identifying the coordinates of the image and use this clustering approach to filter out the noise to be able to recognize irregular shape.

Breast cancer is mostly detected in late stage or grade 3 as patients would already feel with the symptoms. However, grade 1 is difficult to detect at an early stage of breast cancer and is generally detected in patients by chance through screening which explains the low numbers for grade 1 and the reason why we cannot have a training set for most medical diseases. Furthermore, this explains the reason of having 74% of patient's images from grade 3 type, 16% grade 1, 10% grade 2. Having said that, the number we got in the study is reflective of the breast cancer subtypes in the population.

Artificial Intelligence and Machine Learning fails to converge due to the curse of dimensionality [11]. With big data, the proposed algorithm allows to minimize the search time for optimal solution without losing real information. This study showed how to use signal processing and unsupervised clustering approaches to identify the clusters that stratify the progression of breast cancer by minimizing the noise content. The noise in this case are the invariant genes that do not vary across all the samples (31 samples) so it is not involved in disease progression.

V. CONCLUSION

Use signal processing and unsupervised machine learning approaches we identified breast cancer signature patterns from publicly available gene expression data images using preprocessing procedures followed by unsupervised machine learning methodology. Pre-processing consisted of normalizing the genes signals in order to eliminate machine error variation using GCRMA and MAS algorithms. This was followed by filtering the noise using adaptive semi-intelligent filter. The filtered set of probes were subjected to unsupervised clustering of different grades of breast cancer using different similarity measure algorithms. Ward Euclidean gave the best clustering results and accuracy. The study identified novel group which overlap between grade 3 (aggressive breast cancer) and grade 1 (early breast cancer). In addition it identified several interesting novel pathways including immune response and protein phosphorylation pathways as well as novel genes such as insulin like growth factor (IGFBP6), WISP2, SASH1 and MYH11.

Furthermore, the study provided potential ways to overcome an important optimization problem of traditional approaches with feature dimensionality. It performs this by eliminating genes that do not affect breast cancer progression and are thus classified as noise and eliminated. Deploying this technique to breast cancer data, our methodology provided a good way to improve the accuracy in breast cancer classification problems, in addition to providing a more general and scalable tactic to analyze complex gene expression data across various cancer types.

Future improvement includes applying Gene Set Enrichment Analysis (GSEA) to the unsupervised clustering algorithm in order to separate analyses of the genes (signals) driving the enrichment in a more efficient way. Thus, it will identify deeper differences in various molecular pathways and biological processes of the breast cancer types. The proposed methodology of this study provides several potential applications as it can be applied to identify the gene patterns in different diseases other than breast cancer. For example, it can also be applied to identify features that cause diabetic effects such as nephropathy and retinopathy complications.

VI. REFERENCES

[1] S. P. Jackson and J. Bartek, "The DNA-damage response in human biology and disease," *Nature*, vol. 461, no. 7267, pp. 1071–1078, 2009.

[2] D. A. Schmidt, "Programming language semantics," in *Computer Science Handbook*, Second Edi., ACM, Encyclopedia of Computer Science, 2003, pp. 1463–1466.

[3] N. Presneau and R. Hamoudi, "Sequencing technologies," in *Molecular Diagnostics*, vol. 11, 2018, pp. 31–45.

[4] J. J. O'Connor and E. F. Robertson, "Christian Andreas Doppler," *MacTutor History of Mathematics archive. University of St Andrews*, 1998. [Online]. Available: http://www-history.mcs.st-andrews.ac.uk/Biographies/Doppler.html.

[5] D. D. Miller and E. W. Brown, "Artificial Intelligence in Medical Practice: The Question to the Answer?," *Am. J. Med.*, vol. 131, no. 2, pp. 129–133, 2018.

[6] C. Johnson, Stephen, "Hierarchical Clustering Schemes," *Psychom. Springer*, vol. 32, no. 3, pp. 241–254, 1967.

[7] E. Raspé *et al.*, "CDK4 phosphorylation status and a linked gene expression profile predict sensitivity to palbociclib," *EMBO Mol. Med.*, vol. 9, no. 8, pp. 1052--1066, 2017.

[8] R. Hamoudi and A. Warford, "Molecular Analysis and Interpreting Molecular Data," in *Molecular Diagnostics*, 2018, pp. 136–168.

[9] R. A. Hamoudi *et al.*, "Differential expression of NF-B target genes in MALT lymphoma with and without chromosome translocation: Insights into molecular mechanism," *Leukemia*, vol. 24, no. 8, pp. 1487–1497, 2010.

[10] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in *Proceedings of the international conference on machine learning*, 2013.

[11] A. Rosenfeld *et al.*, "MIAT: A novel attribute selection approach to better predict upper gastrointestinal cancer," in *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, 2015.