P-ISSN : 2715-2448 | E-ISSSN : 2715-7199
Vol.3 No.2 July 2022
Buana Information Tchnology and Computer Sciences (BIT and CS)

A Time Series Based Gene Expression Profiling Algorithm for Stomach Cancer Diagnosis

Teresa Kwamboka Abuya[1]
Study Program
Computer Science
Kisii University, Kenya tkwambokaa@gmail.com

Bayu Priyatna [2]
Study Program
Information System
Universitas Buana Perjuangan Karawang
bayu.priyatna@ubpkarawang.ac.id

‹β›

Abstrak— Eksperimen biologis telah menghasilkan sejumlah besar data ekspresi gen yang memiliki nilai sangat besar untuk diagnosis, pengobatan, dan pencegahan penyakit. Namun, kelemahan yang cukup besar memang ada dalam pemanfaatan yang tepat dari data ini karena skala yang besar dan kerumitannya. Sejumlah algoritma telah dikembangkan untuk menginterpretasikan data ini dalam bentuk profil gen untuk tujuan diagnosis. Diantaranya K-means, pengelompokan hierarkis, pengelompokan berbasis kepadatan, pengelompokan subruang, dan peta yang mengatur sendiri. Sayangnya, algoritme ini mengabaikan ketergantungan berurutan di antara titik waktu yang berurutan, tidak memadai dalam penemuan pola untuk mengubah aktivitas selama interval terbatas dari kerangka waktu eksperimen, dan tidak mampu membedakan antara pola faktual dan acak. Dengan demikian, ada kebutuhan untuk algoritme pembuatan profil gen yang mengatasi kekurangan yang dibatasi waktu dalam algoritme saat ini dan karenanya memfasilitasi pembuatan profil gen yang efisien untuk mendiagnosis kanker perut secara dini. Selama bertahun-tahun, eksperimen ekspresi gen deret waktu telah banyak digunakan untuk mempelajari berbagai proses biologis seperti siklus sel, perkembangan, dan respons imun. Dalam makalah ini dikembangkan algoritma profil gen berdasarkan deret waktu untuk diagnosis awal kanker lambung. Dengan menetapkan gen ke satu set profil model yang telah ditentukan sebelumnya yang menangkap pola potensial yang berbeda, signifikansi masing-masing profil ini dapat ditetapkan. Profil signifikan ini kemudian dapat dianalisis lebih lanjut dan digabungkan untuk membentuk cluster yang kemudian dapat dimanipulasi oleh algoritma clustering. Idenya adalah untuk mengukur aktivitas gen selama rentang waktu yang singkat sehingga dapat menghasilkan gambaran universal tentang fungsi seluler. Singkatnya, ini termasuk mendeteksi pola berulang dalam data biologis. Pola-pola ini kemudian digunakan untuk mengungkapkan informasi diagnostik yang mungkin penting bagi praktisi medis. Desain penelitian eksperimental digunakan untuk mencapai tujuan penelitian. Data yang berkaitan dengan genom biologis digunakan untuk pekerjaan penelitian ini. Karena perkembangan penyakit kanker saat ini, hasil dari penelitian ini diharapkan dapat menjadi signifikan dalam diagnosis dini kanker lambung sehingga pengobatan yang tepat dapat diberikan..

Kata kunci: Data microarray, respon imun, clustering, profil signifikan, diagnosis kanker.

Abstract— Biological experiments have produced enormous amount of gene expression data that possess enormous value for the diagnosis, treatment, and prevention of diseases. However, considerable drawbacks do exist in the appropriate utilization of this data due to its massive scale and intricacy. A number of algorithms have been developed to interpret this data in form of gene profiling for diagnosis purposes. They include K-means, hierarchical clustering, density-based clustering, subspace clustering, and self-organizing maps. Unfortunately, these algorithms ignore the sequential dependency among successive time points, are inadequate in the discovery of patterns for changing activity over a restricted interval of an experiment's time frame, and are incapable of discriminating between factual and random patterns. As such, there is a need for a gene profiling algorithm that addresses the time-constrained shortcomings in the current algorithms and hence facilitating efficient profiling of genes for early stomach cancer diagnosis. Over the years, time series gene expression experiments have been widely used to study a range of biological processes such as the cell cycle, development, and immune response. In this paper a gene profiling algorithm based on time series for early stomach cancer diagnosis is developed. By assigning genes to a predefined set of model profiles that capture the potential distinct patterns, the significance of each of these profiles can be established. These significant profiles can then be analyzed further and combined to form clusters that can then be manipulated by clustering algorithms. The idea is to measure the genes' activities over a short period span so as to come up with a universal depiction of the cellular functionality. In a nutshell, this includes detecting recurring patterns in biological data. These patterns are then employed to reveal diagnostic information that may be important for the medical practitioners. An experimental research design was utilized to achieve the study objectives. Data pertaining to biological genomes was employed for this research work. Due to the upsurge of cancer in the current times, the outcomes of this research work is anticipated to be significant in the early diagnosis of stomach cancer so that appropriate medication can be administered.

Keywords: Microarray data, immune response, clustering, significant profiles, cancer diagnosis.

I. INTRODUCTION

Functional genomics is the discipline in which genes are utilized in the determination of their function whereas gene expression is an approach employed to examine the functional changes in these genes. According to [1], the expression level for a given gene across different experimental conditions are collectively referred to as the gene expression profile and the expression levels for all the genes under an experimental condition are jointly referred to as the sample expression profile.

One of the goals in microarray data analysis is the identification of genes for which the expression level is significantly changed under different experimental conditions. Another objective is to cluster the expressed genes or samples having similar expression profiles to make a meaningful biological inference from the set of genes or samples (Martin et.al., 2016).

The field of bioinformatics essentially deals with biological information processing. One of the requirements for effective bioinformatics is an extensive range of computational models that helps in representation and computation of massive biological data. As [2] point out, biological experiments and processes analysis require too much effort. Additionally, this process can prove to be very slow. This can be attributed to the ever-growing intricacy of the processes and fiery growth of biological data emerging from laboratories universally. The recent drawback, as [3] noted, is on how to convert this enormous data repository into knowledge that can facilitate understanding of biological processes and experiments pertaining to both health and diseases. According to [4] timeseries gene expression analysis allows for principled estimation of unobserved time-points, clustering, and dataset alignment. In this technique, every expression profile is modeled as a piecewise polynomial which is estimated from the observed data and every time point sways the overall smooth expression curve. Gene expression experiments carried out using time series show that unobserved timepoints can be reconstructed with 10-15% less error when compared to other profiling methods. The time series-based clustering algorithm operates directly on the continuous representations of gene expression profiles. This is particularly effective when applied to non-uniformly sampled data.

Stomach Cancer (SC) is the fourth most frequently diagnosed malignancy and the second leading cause of cancer death worldwide (Yang et.al.,2018). Although the incidence of SC has declined for decades, the prognosis of SC remains very poor, especially in China. At present, the pathogenesis of SC is unclear, thereby necessitating effective biomarkers and targeted therapeutics. Traditionally, clinic pathological parameters were used in risk stratification of SC outcomes. However, a number of advanced SC patients remained stable for a couple of years, whereas some early-stage patients progressed rapidly [5]. Therefore, reliable biomarkers or stratification systems that can be used for more accurate prediction are highly essential [6].

The greatest challenge in cancer diagnosis is the identification of a subset of genes with crucial roles in diverse stages of these diseases' progression from early stages of carcinogenesis to its final stage of metastasis. As [7] explains, reliable identification of molecular determinants of clinical outcomes can facilitate the discovery of functional biomarkers predictive of therapy response or disease progression. In addition, this can provide insights into new therapeutic targets in this aggressive disease. [8] further point out that the complexity of genomic networks and the vast volume of genes present increase the challenges of understanding and interpreting the resulting mass of data. The problem is compounded by the vagueness, imprecision, and noise present in this data. According to [9], the current algorithms such as Hierarchical gene profiling algorithm, Self-Organizing Maps (SoM), Support Vector Machines (SVM) and K-means algorithm, can only detect relationships where there is sufficient variability in gene expressions and as such, functional interactions are only detectable if they induce changes in transcriptional state that persist over a reasonable timescale. To address this problem, algorithms for visualizing high-throughput single-cell datasets and identifying putative functional relationships between genes are required [10].

Due to the potential of time series to unravel biological processes that take place over short time duration, this research work employed this nonconventional data type to come up with a gene profiling algorithm that is instrumental in disease diagnosis in human beings. In this paper, a time series - based gene profiling algorithm for early stomach cancer diagnosis was developed. Early and accurate diagnosis of stomach cancer can significantly improve the design of personalized therapy and enhance the success of therapeutic interventions. Since time series has the potential of identifying significant chronological expression profiles and the genes associated with this profile, it can enable the comparison of cancer infected genes behavior across multiple conditions over short time duration. Specifically, the response of gastric epithelial cells infected with the vacAmutant strain of the pathogen Helicobacter pylori was investigated [11].

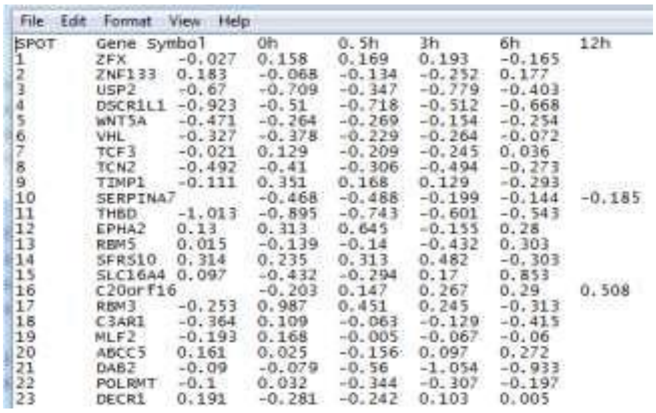
The contributions of this paper include the derivation of mathematical parameters that were shown to help in the generations of gene profiles over a limited duration of time. The rest of this paper is organized as follows. Section 2 presents the related work while section 3discusses gene profiles derivation. Section 4 gives a presentation of results and discussion while part 5 concludes the paper.

II. METHOD

This paper adopted an experimental research design to develop an algorithm that aided in the derivation of gene profiles. The approach involved the derivation of gene profiling parameters which were then employed to develop a time-series based algorithm. This algorithm was then experimented on sample genomic data described in section A below, to provide the required gene profiles visualization in the form of graphs. This visualization provided a straight forward means of establishing the sequential dependency among successive time point. In addition, the visualization facilitated the discovery of patterns for changing activity over a restricted interval of an experiment's time frame.

3.1 Data Set

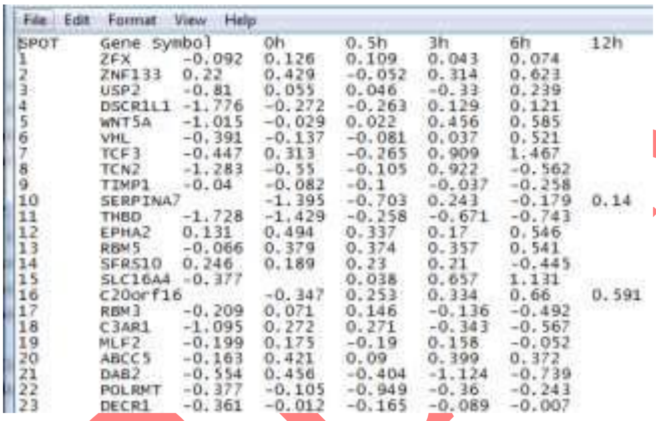
The genomics data employed in this paper were from two experiments measuring the response of gastric epithelial cells infected with the vacA-mutant strain of the pathogen Helicobacter pylori. The data is sampled at five time points 0 h, .5 h, 3 h, 6 h, and 12 h. A sample of these data is shown in Figure 1 for G27 TC1 trial 4.



```
SPOT    Gene Symbol     0h      0.5h    3h      6h      12h
1       ZFX     -0.027  0.158   0.169   0.193   -0.165
2       ZNF133  0.183   -0.068  -0.134  -0.252  0.177
3       USP2    -0.67   -0.709  -0.347  -0.779  -0.403
4       DSCR1L1 -0.923  -0.51   -0.718  -0.512  -0.668
5       WNT5A   -0.471  -0.264  -0.269  -0.154  -0.254
6       VHL     -0.327  -0.378  -0.229  -0.264  -0.072
7       TCF3    -0.021  0.129   -0.209  -0.245  0.036
8       TCN2    -0.492  -0.41   -0.306  -0.494  -0.273
9       TIMP1   -0.111  0.351   0.168   0.129   -0.293
10      SERPINA7        -0.468  -0.488  -0.199  -0.144  -0.185
11      THBD    -1.013  -0.895  -0.743  -0.601  -0.543
12      EPHA2   0.13    0.313   0.645   -0.155  0.28
13      RBM5    0.015   -0.139  -0.14   -0.432  0.303
14      SFRS10  0.314   0.235   0.313   0.482   -0.303
15      SLC16A4 0.097   -0.432  -0.294  0.17    0.853
16      C20orf16        -0.203  0.147   0.267   0.29    0.508
17      RBM3    -0.253  0.987   0.451   0.245   -0.313
18      C3AR1   -0.364  0.109   -0.063  -0.129  -0.415
19      MLF2    -0.193  0.168   -0.005  -0.067  -0.06
20      ABCC5   0.161   0.025   -0.156  0.097   0.272
21      DAB2    -0.09   -0.079  -0.56   -1.054  -0.933
22      POLRMT  -0.1    0.032   -0.344  -0.307  -0.197
23      DECR1   0.191   -0.281  -0.242  0.103   0.005
```

Figure 1. Sample G27 TC1 Trial 4 Data

This Figure 1 shows TC1 gastric epithelial (AGS) cells infected with wild type H. pylori (G27) and isogenic mutants in cagA and vacA for 0, 0.5, 3, 6, and 12 hours. Figure 2.0 shows the G27 TC1 trial 5 data.



```
SPOT    Gene Symbol     0h      0.5h    3h      6h      12h
1       ZFX     -0.092  0.126   0.109   0.043   0.074
2       ZNF133  0.22    0.429   -0.052  0.314   0.623
3       USP2    -0.81   0.055   0.046   -0.33   0.239
4       DSCR1L1 -1.776  -0.272  -0.263  0.129   0.121
5       WNT5A   -1.015  -0.029  0.022   0.456   0.585
6       VHL     -0.391  -0.137  -0.081  0.037   0.521
7       TCF3    -0.447  0.313   -0.265  0.909   1.467
8       TCN2    -1.283  -0.55   -0.105  0.922   -0.562
9       TIMP1   -0.04   -0.082  -0.1    -0.037  -0.258
10      SERPINA7        -1.395  -0.703  0.243   -0.179  0.14
11      THBD    -1.728  -1.429  -0.258  -0.671  -0.743
12      EPHA2   0.131   0.494   0.337   0.17    0.546
13      RBM5    -0.066  0.379   0.374   0.357   0.541
14      SFRS10  0.246   0.189   0.23    0.21    -0.445
15      SLC16A4 -0.377          0.038   0.657   1.131
16      C20orf16        -0.347  0.253   0.334   0.66    0.591
17      RBM3    -0.209  0.071   0.146   -0.136  -0.492
18      C3AR1   -1.095  0.272   0.271   -0.343  -0.567
19      MLF2    -0.199  0.175   -0.19   0.158   -0.052
20      ABCC5   -0.163  0.421   0.09    0.399   0.372
21      DAB2    -0.554  0.456   -0.404  -1.124  -0.739
22      POLRMT  -0.377  -0.105  -0.949  -0.36   -0.243
23      DECR1   -0.361  -0.012  -0.165  -0.089  -0.007
```

Figure 2. Sample G27 TC1 Trial 5 Data

In these data samples, hybridizations of G27 (trial 4) and cag A- (trial 3) time-courses are accomplished in parallel. A technical replicate of the G27 time course (trial 5) and hybridization of vacA- (trial 3) time course is also accomplished in parallel. The cag A 6- and 12-hour time points technically replicated (trial 4) (the cag A 6-hour sample of trial 3 are lost).

3.2 Gene Profiling Modeling Process

This research dealt with the profiling, comparing and visualizing gene expression data from short time series of two experiments measuring the response of gastric epithelial cells infected with the vacA-mutant strain of the pathogen Helicobacter pylori. The gene expression profiling comprised of four major steps as shown in Figure 3. As show in this figure, the steps included the generation and normalization of expression signals, testing each probe for its differential or association with the phenotype, the application of proper statistical significance criteria to identify the gene expression profile, and the investigation of the functions and pathways of the genes in the expression profile.

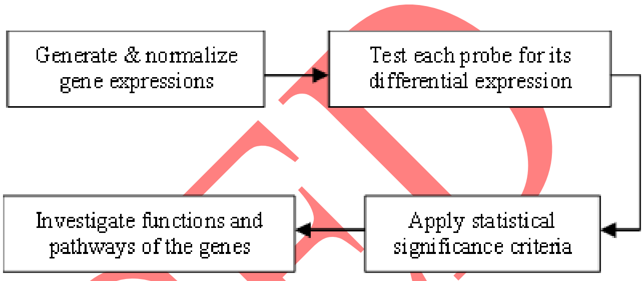

Generate & normalize gene expressions
Test each probe for its differential expression
Apply statistical significance criteria
Investigate functions and pathways of the genes

Figure 3. Gene Profiling Steps

Thereafter, a number of statistical significance criteria such as Pearson correlation, P-value, Euclidean distance, Logistic regression, Bonferroni correction, False discovery rate and Time Points Permutation were applied to help identify specific list of genes differentially expressed or associated with the phenotype.

Although mutual information (MI) measure is superior over simpler measures such as Pearson correlation as it is capable of capturing complex non-linear and non-monotonic dependencies. In addition, it can reflect the dynamics between pairs or groups of genes, computing MI involves estimating pair-wise joint probability distributions which requires density estimation or data discretization, with the accuracy of these estimates depending on sample sizes. As this measure was not deployed in this research study. Table 2 gives a summary for the deployment of the various performance metrics.

Table 2 Performance Metrics Deployment

SNO	Statistical Measure	Deployment
1.	Pearson correlation	Weighted Relation between all genes
2.	P-value	Significance of gene coexpressions
3.	Euclidean distance	Correlation distance between gene profiles
4.	Logistic regression	Estimate of cancerous probability
5.	Bonferroni correction	Adjustment to the confidence levels

6.	False discovery rate	Adjustment to the confidence levels
7.	Time Points Permutation	Optimize the number of required profiles

3.3 The Algorithm of modeling gene profiles The first step was the commencement of the algorithm while the second step in the processing activities was the input of the genomics data as shown in Figure 4.0. in the next page. In step three, validation is done against empty genomic file upload such that if this field is empty, then an error message is generated for this effect in the fourth step. During the fifth step, the validation against spot IDs not included is done such that if these IDs are not included, then they are computed in the sixth step. The value of the spot ID is initialized to 1 which are thereafter incremented by one until the value of 24192 is reached, which is equivalent to the number of genes in the file that were investigated. Whereas spot IDs were unique for each gene entry, the same gene symbol may appear multiple times in the data file corresponding to the same gene appearing on multiple spots.

The seventh step was the computation of the average value for the expression values for the same gene. This was accomplished using the median before further analysis on the data was carried out. The eighth step was that option of filtering some specific genes using P-value metric. In situations where a gene was filtered, then it was excluded from further analysis. Gene filtering was accomplished for those genes that did not show a sufficient response to experimental conditions, those genes that had too many missing values, or the gene expression pattern over repeats was too inconsistent as dictated by the minimum correlation between repeats. The ninth step was the usage of additional parameters namely the maximum Pearson correlation and maximum number of candidate model profiles to dictate the selection of model profiles along with the maximum number of model profiles and maximum unit change in model profiles between time points as shown in Figure 5.0. In this algorithm, the candidate model profiles were designed to be nonconstant profiles which started at zero and increased or decreased an integral number of units that was less than or equal to the value of the maximum unit change in model profiles between time points.

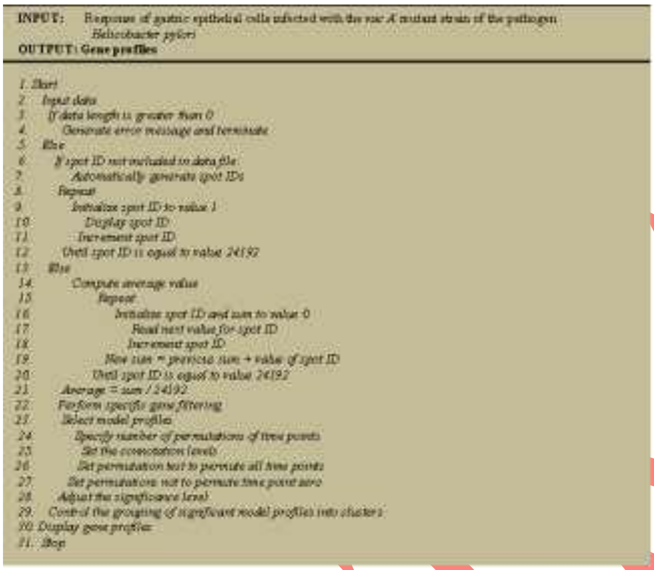
Figure.4. Gene Profiling Algorithm Pseudo-Code

When this parameter was set to zero, all permutations were used. In the eleventh step, the P-value based significance level was utilized to set the connotation level at which the number of genes assigned to a model profile as compared to the expected number of genes assigned was regarded as significant. During the twelfth step, the permutation test was set to permute all time points including time point zero when computing the expected number of genes assigned to a profile.

In this case, the developed algorithm located profiles with significantly more genes assigned than expected on condition that all the input columns had been randomly reordered. On the other hand, during the thirteenth step, the permutation test was configured not to permute at time point zero and as such, the algorithm found profiles with more genes assigned than expected on condition that all the columns except for the first column had been randomly reordered. In the developed algorithm, permuting time point zero was preferred since it was the only test that took into account the significant changes that took place between time point zero and the immediate next time point (0.5 h).

In the fourteenth step, the correction method was utilized to adjust the significance level since this algorithm was meant to test multiple profiles for significance. Two types of corrections were utilized in this algorithm. The first one was the Bonferroni correction while the second one was the conservative false discovery rate (FDR) control. In the third scenario, no correction was made for the multiple significance tests.

The logic here was that when the number of candidate model profiles exceeded the p -value of seeing t more genes in the intersection, then instead of explicitly generating al l candidate model profiles, a subset of candidate model profiles of this size was randomly selected. In the tenth step, the number of permutations per gene parameter was employed to specify the number of permutations of time points that were randomly selec ted for each gene when computing the expected number of genes assigned to each of the model profiles.



```
INPUT:   Response of gastric epithelial cells infected with the vac A mutant strain of the pathogen
         Helicobacter pylori
OUTPUT: Gene profiles

1. Start
2.   Input data
3.     If data length is greater than 0
4.       Generate error message and terminate
5.   Else
6.     If spot ID not included in data file
7.       Automatically generate spot IDs
8.       Repeat
9.         Initialize spot ID to value 1
10.          Display spot ID
11.        Increment spot ID
12.      Until spot ID is equal to value 24192
13.    Else
14.      Compute average value
15.          Repeat
16.              Initialize spot ID and sum to value 0
17.                Read next value for spot ID
18.                Increment spot ID
19.            New sum = previous sum + value of spot ID
20.          Until spot ID is equal to value 24192
21.      Average = sum / 24192
22.      Perform specific gene filtering
23.      Select model profiles
24.        Specify number of permutations of time points
25.          Set the correlation levels
26.          Set permutation test to permute all time points
27.        Set permutations not to permute time point zero
28.      Adjust the significance level
29.    Control the grouping of significant model profiles into clusters
30. Display gene profiles
31. Stop
```

Figure 5. Modeling Gene Profiling Process

In the fifteenth step, two parameters namely the minimum correlation and the minimum correlation percentile were utilized to control the grouping of significant model profiles into clusters. In so doing, these parameters served to control how similar two model profiles had to be if they were grouped together. For the case of the minimum correlation, any two model profiles assigned to the same cluster of profiles had to have a correlation above this parameter's value. On its part, the minimum correlation percentile was employed in cases there were repeat data from different time periods. It was used to specify that any two model profiles assigned to the same cluster of profiles had to have a correlation in their expression greater than the correlation of this percentile in the distribution of gene expression correlations between the repeats. The last step was the display of the gene profiles based on the Euclidean distance after which the algorithm halted in the seventeenth step. Figure 6 gives a diagrammatic representation of the gene profile derivation process. As this figure shows, the process gene profile derivation process involves the input of the genomic data containing the gene expressions to be profiled.

These data items are analyzed using parameters such as Pvalue, Pearson correlations, permutations, logistic regression and median to yield probable profiles as already discussed above. Correction methods are then employed to adjust the significance level to permit the testing of multiple profiles for significance.

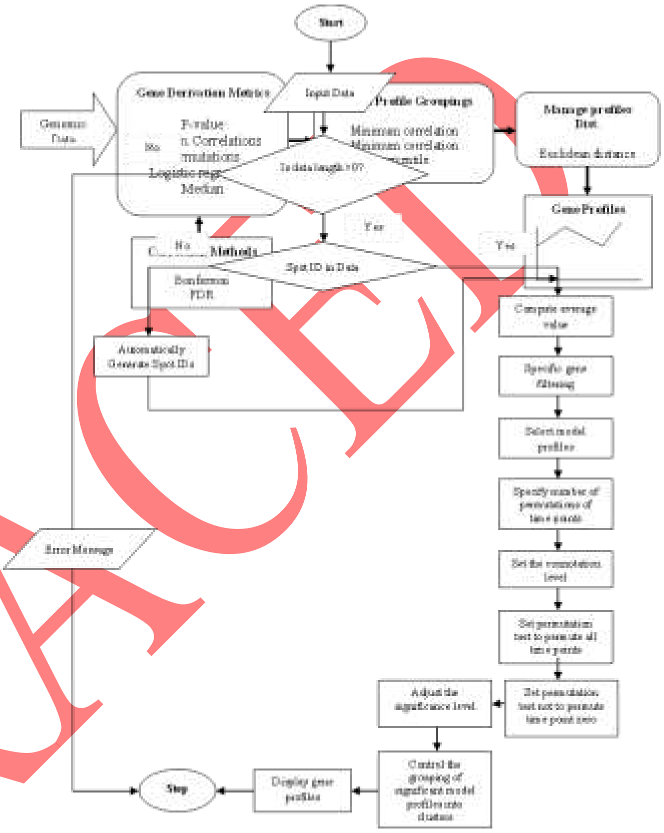
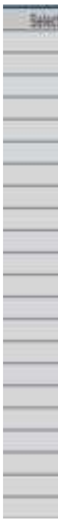


Figure 6. Schematic Gene Derivation Process

The output gene groupings are then clustered using minimum correlation and minimum correlation percentile before Euclidean distance is applied to them to distinguish the various gene profiles. The final outputs are the gene profiles in form of graphs.



III.RESULTS AND DISCUSSION

In this section a time series-based gene profiling algorithm is developed. To test the derived parameters and their gene profiling abilities, the algorithms and statistical computations were put into use to achieve some functionality as shown in Table 3. The genomics data from two experiments measuring the response of gastric epithelial cells infected with the vac A-mutant strain of the pathogen Helicobacter pylori were then fed as input to this algorithm.

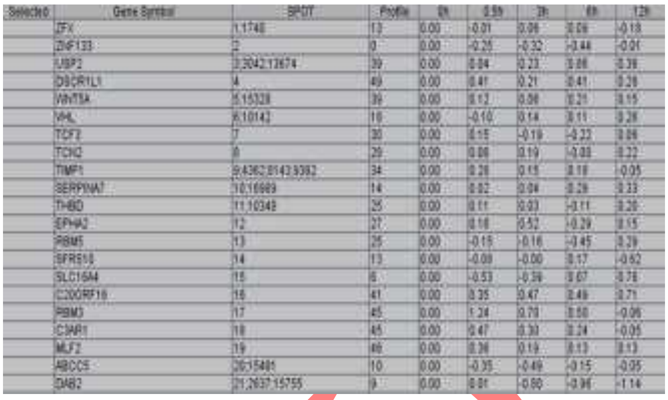
Table 3. Gene Derivation Process

Step	Parameter	Activity
1	n/a	-Commence gene derivation process
2	n/a	-Input genomic data
3	n/a	-Validation is done against empty genomic file upload
4	n/a	-Prompt genomic data input error
5	n/a	-Validation against spot IDs
6	n/a	-If not included in file compute spot IDs
7	Median	-Computation of the average value for the expression values for the same gene
8	P-value	-Filtering specific genes
9	Pearson correlation, p-value	-Model profiles selections.
10	Permutation	-Computation of the expected number of genes assigned to each of the model profiles
11	P-value, Logistic regression	-Setting the connotation level at which the number of genes are assigned to a model profile
12	Permutation	-Compute the expected number of genes assigned to a specific profile
13	Permutation	-Configure permutation test not to permute at time point zero
14	Bonferroni, FDR	-Adjust the significance level to test multiple profiles for significance
15	Minimum correlation, minimum correlation percentile	-Control grouping of significant model profiles into clusters
16	Euclidean distance	-Display generated gene profiles
17	n/a	-Halt gene derivation process

The minimum absolute expression change was any value more than -0.05. As an illustration, using the maximum number of missing values to be 2, the minimum correlation between repeats to be 0, and the minimum absolute expression change to be 0.05 yielded the information in Table 4.0 for the sample filtered genes. **Table 4.** Sample Filtered Genes

The genes that were devoid of these three characteristics were regarded as standard genes and were the ones that took part in further analysis. Table 5 gives information on the sample genes that passed the classification criteria.

Table 5. Sample Genes Passing Classification Criteria



Selected	Gene Symbol	SPOT	Profile	0h	0.5h	3h	6h	12h
	ZFX	1;1748	13	0.00	-0.01	0.06	0.06	-0.18
	ZNF133	2	0	0.00	-0.25	-0.32	-0.44	-0.01
	USP2	3;3042;13674	39	0.00	0.04	0.23	0.06	0.38
	DSCR1L1	4	49	0.00	0.41	0.21	0.41	0.28
	WNT5A	5;15328	39	0.00	0.12	0.08	0.21	0.15
	VHL	6;10142	16	0.00	-0.10	0.14	0.11	0.28
	TCF2	7	30	0.00	0.15	-0.19	-0.22	0.06
	TCN2	8	29	0.00	0.06	0.19	-0.03	0.22
	TIMP1	9;4362;8143;9382	34	0.00	0.28	0.15	0.18	-0.05
	SERPINA1	10;16989	14	0.00	0.02	0.04	0.28	0.23
	THBD	11;10348	25	0.00	0.11	0.03	-0.11	0.26
	EPHA2	12	27	0.00	0.16	0.52	-0.29	0.15
	RBM5	13	25	0.00	-0.15	-0.16	-0.45	0.29
	SFRS16	14	13	0.00	-0.08	-0.00	0.17	-0.62
	SLC16A4	15	6	0.00	-0.53	-0.39	0.07	0.78
	C20ORF16	16	41	0.00	0.35	0.47	2.48	0.71
	RBM3	17	45	0.00	1.24	0.78	0.50	-0.06
	C3AR1	18	45	0.00	0.47	0.30	0.24	-0.05
	MLF2	19	46	0.00	0.36	0.19	0.13	0.13
	ABCC5	20;15481	10	0.00	-0.35	-0.49	-0.15	-0.05
	DAB2	21;2637;15755	9	0.00	0.01	-0.80	-0.96	-1.14

Afterwards, eight parameters were utilized for the computational derivation of gene profiles from this set of data: maximum correlation, maximum number of candidate model profiles, maximum number of model profiles and maximum unit change in model profiles between time points, number of permutations per gene, significance level, and correction method as shown in Table 6 below.

Table 6. Gene Profiles Evaluation Metrics

Gene Profiling Option	Value
Maximum correlation	1
Maximum number of candidate model profiles	1,000,000
Number of permutations per gene(0 for all permutations)	0
P-value significance level	0.05
Maximum Number of model profiles	50
Maximum unit change in model profiles between time points	2
Correction method	None
Minimum Correlation	0.7

Based on the evaluation metrics of Table 6.0, the algorithm was run to yield the proposed gene profiles.

3.1 Time Series-Based Gene Profiling.

In this profiling, the maximum correlation specified the value that the maximum correlation between any two model profiles had to be below, and was therefore employed to guarantee that two very similar profiles were not selected. The maximum value for this parameter was set to 1 in order to prevent two perfectly correlated model profiles from being selected. It was observed that lowering this parameter led to the number of model profiles selected being less than the maximum number of model profiles even in situations where more candidate model profiles were available.

On the other hand, the maximum number of candidate model profiles represented non-constant profiles which commenced at 0 and increased or decreased an integral number of units that was less than or equal to the value of the maximum unit change in model profiles between time points. The number of permutations per gene parameter specified the number of permutations of time points that were randomly selected for each gene when computing the expected number of genes assigned to each of the model profiles.

When this parameter was set to 0, all permutations were used. It was also important to set permutation test to permute time point 0 or not. When computing the expected number of genes assigned to a profile, if the permutation test for time point 0 was set, the permutation test permuted all time points including time point 0. It was observed that doing this led to profiles with significantly more genes being assigned than expected if all the input columns had been randomly reordered. On the contrary, if the permutation test was not set for 0, the permutation test permuted all time points except for time point 0. In this scenario, profiles with more genes were assigned than expected if all the columns except for the first column had been randomly reordered.

Permuting time point 0 was preferred since only this test took into account significant changes that occurred between time point 0 and the immediate next time point. However in some cases based on experimental design a gene's expression value before transformation at time point 0 was expected to be known more accurately than the other time points, and because of this asymmetry, not permuting time point 0 was also be useful.

It was observed that increasing the maximum number of model profiles increased the number of candidate models as shown in Table 7.

Table 7. Maximum Number of Gene Model Profiles Viz. Significant Gene Models

Maximum Number of Model Profiles	Resulting Sig. Number of Gene Models
50	14
60	16
70	18
80	20
100	21
120	23
140	24

Based on the values in Table 7, a graph was plotted for maximum number of model profiles against the resulting significant number of gene models as shown in Figure 7.

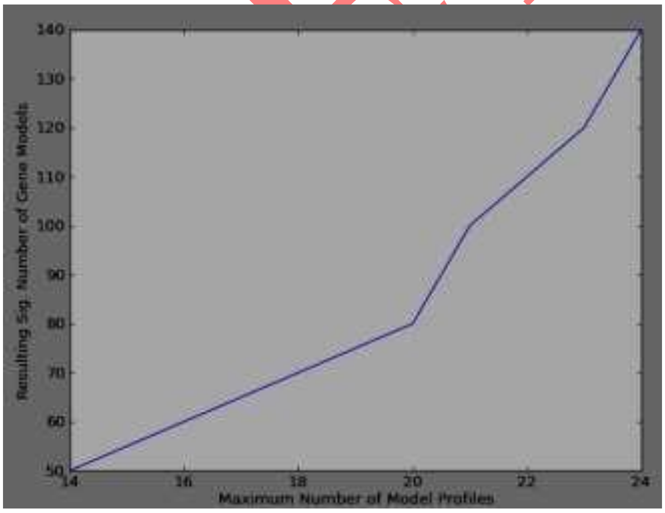


Figure 7. Maximum No. of model profiles Viz. Resulting Significant. No. of Gene Models

The graph of Figure 7.0 shows that the resulting significant number of gene models increase nearly exponentially as the maximum number of model profiles was increased. Consequently, to get fine grained gene model profiles, the maximum number of model profiles had to be increased and vice versa. Table 8.0 gives the shift in the resulting significant number of gene profiles as the maximum unit change in model profiles between time points was adjusted.

Table 8. Maximum Unit Change in Model Profiles Viz. Significant Gene Models

Maximum Unit Change in Model Profiles	Resulting Sig. Number of Gene Models
1	13
2	14
3	15
4	13
5	14
6	15
7	15
8	16
9	16
10	15

As shown in this table, generally as the maximum unit change in model profiles is increased, the resulting significant number of gene models is increased. This is due to the pronounced Euclidean distances between the gene models. Regarding maximum correlation, the value of minimum correlation was set to zero (0) and the value of maximum correlation was slowly reduced from 1 to zero. The results obtained are shown in Table 9 are observed.

Table 9. Maximum Correlations Viz. Significant Gene Models

Maximum Correlation	Resulting Sig. Number of Gene Models	Number of Genes Assigned
1	12	1005
0.9	12	1005
0.8	9	993
0.7	6	1185
0.6	4	1216
0.5	3	1243
0.4	3	1348
0.3	3	1348
0.2	3	1470
0.1	3	1470
0	1	1275

Generally, as the value of maximum correlation is reduced from one to zero, the number of resulting significant number of gene models reduced to unity (1) while the number of genes assigned to these gene models increased from 1005 to a maximum value of 1275. This implies that when the correlation value is small, gene models are basically indistinguishable hence at correlation zero, there is only one resulting significant model. On the other hand, at maximum correlation, the genes can be clearly distinguished and hence the resulting significant gene models are many.

Concerning the number of genes assigned to models, at low correlation coefficients, genes profiles are indistinguishable and hence a large number of genes are assigned to the few available models. However, as the correlation coefficients are increased, the gene profiles become increasing disparate and few genes are assigned to each of the many models now available as the rest are discriminated due to their large pvalues.

3.2 Prediction Power of the Developed Algorithm In the developed algorithm, sequences of gene expressions were listed in order of occurrence, starting at time point 0h to 12h. The aim was to collect and investigate precedent observations of gene expressions at various time points in order to come up with ideal models to express the intrinsic structure of the underlying genomic data. Based on these models, it was possible to predict future gene expressions. To put this into perspective, profile ID 17 was considered whose gene expressions are shown in Figure 8.

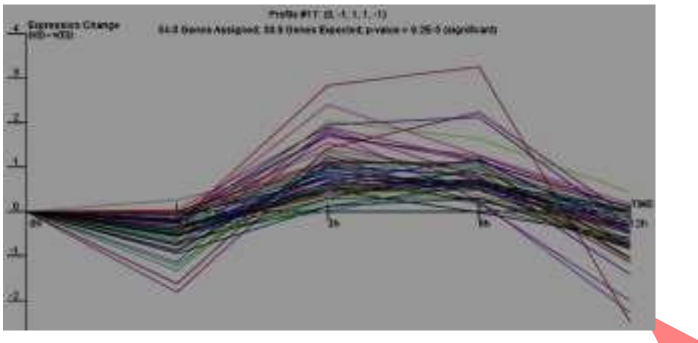


Figure 8. Gene Expressions for Profile ID 17

A total of 54 genes were assigned to this model profile whose individual expressions are shown in Figure 8. By sketching a line of best fit through these gene expressions and performing some extrapolations, the future expressions beyond the 12h time point can be obtained as shown in Figure 9 below.

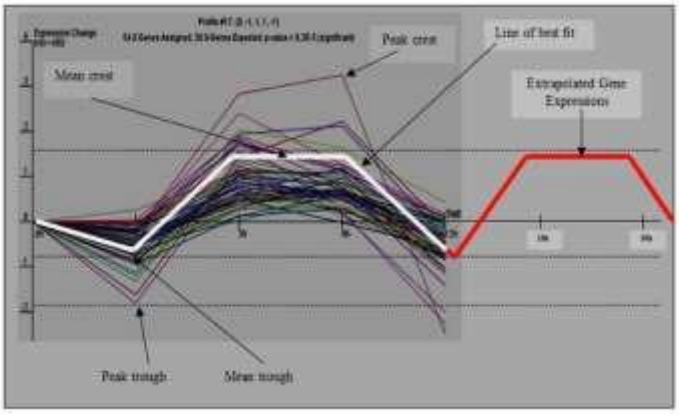


Figure 9. Gene Profile Prediction

The white thick line through the gene expressions is the line of best fit while the thick red line represents the extrapolated gene expressions for the 54 genes assigned to profile ID 17 for the future 18h and 30h time points. Suppose that the stomach cancer patient gene expressions are as shown in Figure 10 below.

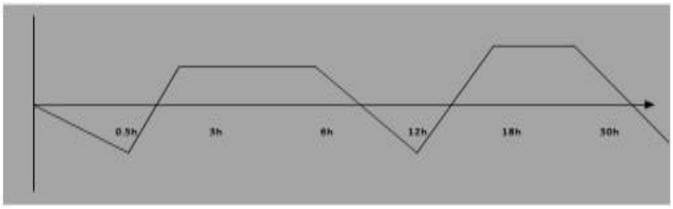


Figure 10. Stomach Cancer Patient Gene Expressions over 30h Duration

Comparing the hypothesized gene expressions over the 30h duration and the gene models in Figure 11.0 below, then considering the first few gene expressions, model profile IDs 13, 14,15,16,17 and 18 are candidates' models that the stomach cancer patient gene expressions can fit in. However, taking into account the preceding time points eliminates model profiles 14(experiences near exponential growth followed by plateau), 15(experiences linear growth followed by plateau), 16 (portrays linear growth, linear decay and plateau), and 18 (presents linear growth followed by plateau). This leaves profile ID 13 and 17 as the most probable model profiles. By drawing a horizontal line through these two profiles as shown in Figure 11, it is possible to discern which of them perfectly fits the patient gene expressions.

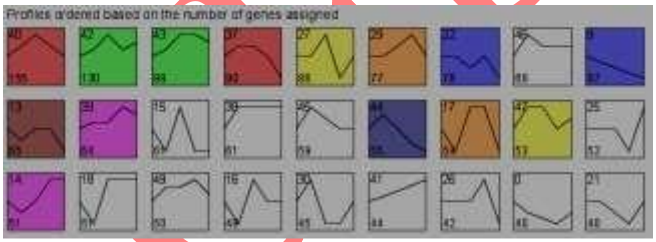


Figure 11. Gene Model Fitting

Based on this line and considering time-points at which troughs and crests appear, it is clear that model profile ID 17 perfectly fits the patient gene expressions for a duration of 30h time points. As such, it can be implied that the developed algorithm led to accurate diagnosis of stomach cancer patients within 12h time points since the commencement of the cancerous gene expressions. In the next section, this algorithm is validated against some well-known gene profiling algorithms.

3.3 Validation of the Developed Algorithm In this section, the time series-based algorithm that was developed is validated against other gene profiling algorithms such as Hierarchical gene profiling algorithm, Support Vector Machine, Self-organizing maps, and K-means algorithm. In Hierarchical gene profiling algorithm, genes with related expression patterns are grouped together and connected by a series of branches to form a dendrogram. Unfortunately, this algorithm considers each gene as an individual cluster and genes that are similar to each other form nested clusters based on the pair-wise distances.

On the other hand, the time series-based algorithm developed in this research study considered a group of genes with similar expressions as profile clusters. For instance, in Figure 6.6, a total of 155 genes were represented by a single model profile with ID 40 and 90 genes were represented by model profile ID 37. These two model profiles formed a cluster with a total of 245 genes. As such, the developed algorithm is operationally faster during gene profiling compared to Hierarchical gene profiling algorithm, rendering it ideal for large genomic data set. The genomic data that was utilized in this research consisted of 24192 gene symbols observed under 5 time points, making the total gene expressions 120960, a very big data set for the rather slow Hierarchical gene profiling algorithm.

To effectively apply the Support Vector Machine gene profiling algorithm, it requires training using the same members of each model profile that have to be identified. This training takes time and hence compared to the developed algorithm, it is slow and hence inefficient for large data sets such as the 120960 gene expressions that were under investigation in this research.

Although Self-organizing maps algorithm has been employed to group 1,036 genes into 24 categories, this algorithm is slow in training, hard to train against slowly evolving data and are not so intuitive since neurons close on the map (topological proximity) may be far away in feature space. Additionally, these maps do not behave so gently when using categorical data, or mixed data. Comparing the 1036 genes that Selforganizing maps algorithm profiled into 24 categories with the 24192 genes that were profiled using the developed time series-based algorithm, it is clear that the proposed algorithm is efficient.

Regarding SVM algorithm, this algorithm has been used for cancer classification with microarray data where it served as a powerful classifier together with four effective feature reduction methods namely principal components analysis (PCA), class-separability measure, Fisher ratio and t-test to the problem of cancer classification based on gene expression data. Although it very high classification accuracies, it requires feature reduction methods which renders it structurally complex compared to the time series-based algorithm implemented in this research.

On its part, the K-means algorithm operates on a series of microarray experiments measuring the expression of a set of genes at regular time intervals in a common cell line. It requires that data be normalized to permit for comparisons across these microarrays. The output produced is in form of clusters of genes which vary in similar ways over time and hence it is possible to infer that genes which vary in the same way may be co-regulated and or participate in the same pathway. Unfortunately, the numbers of clusters need to be specified which may be unknown in some instances, and figuring out the right number of clusters that represent the true number of clusters in the population is quite subjective. As such, the profiles obtained using K-means can vary greatly depending on the location of the observations that are randomly chosen as initial centroids. However, the developed time series-based algorithm employs statistical metrics such as p-value, Pearson correlation, logistics regression and Euclidean distance whose significance levels are well known.

The K-means clustering algorithm assumes that the underlying clusters in the population are spherical, distinct, and are of approximately equal size and hence tends to identify clusters with these characteristics. Therefore, this algorithm is incapable of yielding good results when clusters are elongated or not equal in size like the genomic data used in this research where some gene expressions were negative, zero and others positive. The K- algorithm is also sensitive to initial conditions, implying that different initial conditions produce varying result of gene profiles. It is also possible for a very far data from the centroid to pull the centroid away from the real one as shown in Figure 12. below. Here, 5 genes are assigned to cluster ID 55 and it is clear from the gene.

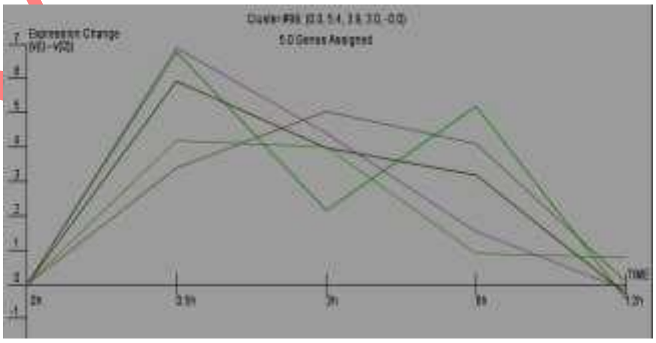


Figure 12. K-Means Based Profiling

Expressions that the profiling is not such accurate especially after the 0.5h time point. Whereas 4 gene expressions have negative gradients, one of them has a positive gradient. During the 3h time-point, some gene profiles are at the rough, others are at the crest, plateau while others are still on their descent. The same is observed during the 6h time point. These varying result of gene profiles give contradicting depiction of gene activities and hence may lead to inaccurate stomach cancer diagnosis.

IV.CONCLUSION

The aim of this paper was to develop a gene profiling algorithm based on time series to help in early stomach cancer diagnosis. Based on a number of derived gene profiling parameters, an algorithm was developed that was then experimented on sample genomic data. The results of this paper included a number of gene profiles that obtained from the underlying pathogen Helicobacter pylori data. The significance of this research lies on the fact that it helped generate gene profiles using very short time points. This feature is very critical in early stomach cancer diagnosis as it facilitates necessary preventive measures that curtail the cancerous cells advancement to other fatal phases. Since this research was purely based on stomach cancer, future work in this area lies on the implementation of this algorithm for other types of cancer or diseases.

REFERENCES

[1] Sanchita & Ashok S. (2015). Future Challenges in Application of Algorithms and Tools for Clustering of Gene Expression Data. Biotechnology Division, CSIRCentral Institute of Medicinal and Aromatic Plants. Lucknow 22601. (pp. 515-531)5 India.

[2] Brohée S., Barriot R., & Moreau Y. (2015).Biological knowledge bases using Wikis: combining the flexibility of Wikis with the structure of databases. Bioinformatics, Oxford Journals.

[3] Wong K. (2016).Computational Biology and Bioinformatics: Gene Regulation. CRC Press.

[4] Ziv B., Georg G., David K., & Tommi S. (2015). A New Approach to Analyzing Gene Expression Time Series Data. Whitehead Institute for Biomedical Research.

[5] Wang, H., Wang, X., Xu, L. et al. (2020).High expression levels of pyrimidine metabolic rate–limiting enzymes are adverse prognostic factors in lung adenocarcinoma: a study based on The Cancer Genome Atlas and Gene Expression Omnibus datasets. Purinergic Signalling 16, 347–366 (2020). https://doi.org/10.1007/s11302-020-09711-4.

[6] Wenhui, Y., Zhiyong L.., Yuan, Li., Jianbing, M., Mudan, Yang., Jun, X.(2019).Immune signature profiling identified prognostic factors for gastric cancer. Chinese journal of cancer research. Https// doi: 10.21147/j.issn.1000-9604.2019.03.08.[7] Abolfazl R., Fatemeh A., Salendra S., and Vinay V. (2016).NetworkBased Enriched Gene Subnetwork Identification: A Game-Theoretic Approach. Biomed Eng Comput Biol. Vol. 7, Issue 2, pp. 1–14.

[8] Oyelade J., Itunuoluwa I., Funke O., Olufemi A., Efosa U., Faridah A., Moses A., and Ezekiel A. (2016). Clustering Algorithms: Their Application to Gene Expression Data. Bioinformatics and Biology Insights,10, 237–253.

[9] Thalia E.C., Michael P.H., and Ann C. (2017).Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. Cell Systems, 5, 251–267.

[10] Gwang H., Peter S., Sung J., & Joo H. (2016). Screening and surveillance for gastric cancer in the United States: Is it needed? American Society for Gastrointestinal Endoscopy. Volume 84, No. 1, pp. 18-28.

[11] Siregar, Amril Mutoi, et al. "Perbandingan Algoritme Klasifikasi Untuk Prediksi Cuaca." *Jurnal Accounting Information System (AIMS)* 3.1 (2020): 15-24.