JOIV : Int. J. Inform. Visualization, 6(4) - December 2022 784-790

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

A Microarray Data Pre-processing Method for Cancer Classification

Tay Xin Hui [a], Shahreen Kasim [a,*], Mohd Farhan Md Fudzee [a], Zubaile Abdullah [a], Rohayanti Hassan [b], Aldo Erianda [c]

[a] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400, Johor, Malaysia*
[b] *Faculty of Computing, Universiti Teknologi Malaysia, 83100, Johor, Malaysia*
[c] *Department of Information Technology, Politeknik Negeri Padang, Sumatera Barat, Indonesia*
*Corresponding author: *shahreen@uthm.edu.my*

***Abstract*—The development of microarray technology has led to significant improvements and research in various fields. With the help of machine learning techniques and statistical methods, it is now possible to organize, analyze, and interpret large amounts of biological data to uncover significant patterns of interest. The exploitation of microarray data is of great challenge for many researchers. Raw gene expression data are usually vulnerable to missing values, noisy data, incomplete data, and inconsistent data. Hence, processing data before being applied for cancer classification is important. In order to extract the biological significance of microarray gene expression data, data pre-processing is a necessary step to obtain valuable information for further analysis and address important hypotheses. This study presents a detailed description of pre-processing data method for cancer classification. The proposed method consists of three phases: data cleaning, transformation, and filtering. The combination of GenePattern software tool and Rstudio was utilized to implement the proposed data pre-processing method. The proposed method was applied to six gene expression datasets: lung cancer dataset, stomach cancer dataset, liver cancer dataset, kidney cancer dataset, thyroid cancer dataset, and breast cancer dataset to demonstrate the feasibility of the proposed method for cancer classification. A comparison has been made to illustrate the differences between the dataset before and after data pre-processing.**

***Keywords*—Data pre-processing; microarray data; gene expression data; GenePattern.**

Manuscript received 15 Jan. 2022; revised 29 Apr. 2022; accepted 12 Oct. 2022. Date of publication 31 Dec. 2022.

I. INTRODUCTION

DNA microarray technologies allow researchers to measure thousands of genes' expression patterns in various experimental conditions. This high-throughput technology opens up the possibility of organizing, analyzing, and interpreting biological data to solve biological problems at the molecular level. In general, statistical analysis can be categorized into three studies: (a) association studies, to discover the relationships between interesting genes or biological pathways; (b) prognostic or prediction studies, to classify patients concerning clinical endpoints based on molecular markers; and (c) class discovery studies, to discover clusters based on molecular data [1]. The ability to derive biological inferences from microarray data allows researchers to identify key disease pathways and find potential therapeutic targets [2].

Microarray data or gene expression data is composed of huge tables with thousands of rows corresponding to the genes or clones present in the DNA array, and several columns, one for each experimental condition measured [3]. This massive genomic data requires practical data pre-processing techniques for their analyses. Effective computational-based methodologies highly depend on the quality of input data. There are numerous sources of systematic and random changes introduced along the various phases in assessing gene expression levels [4]. These variations in expression levels might lead to false positives under certain changing experimental conditions. Thus, applying data pre-processing techniques is important to enhance the quality of results.

Data pre-processing is a data mining technique used to transform raw data into an efficient and useful format. A basic data pre-processing method involves three steps: (a) data cleaning, to remove missing and noisy data; (b) data transformation, to transform data into an appropriate form; and (c) data reduction, to increase the storage efficiency and reduce data storage and analysis costs [5]. The unprocessed raw data are susceptible to missing, noise, outliers, and inconsistency, affecting the quality of data mining results.

Hence, data pre-processing is a mandatory procedure to undergo before the dataset can be applied to other mainstream research algorithms [6].

The structure of the paper is arranged as follows. Section 2 provides details about the use of the gene expression dataset and its information, followed by the method to pre-process the dataset. Section 3 presents the outcome of pre-processed data, and a comparison will be made to showcase the difference before and after pre-processing of the dataset. Section 4 provides a concluding summary before ending this research paper.

II. Materials and Methods

Data pre-processing involves preparing and transforming the dataset into a clean and useful format. It aims to remove irrelevant and missing data, normalize data, reduce the size of data, and extract features for data [7]. This section will explain the materials and methodology applied in this study. Gene expression dataset and the available pre-processing software tool will be introduced for further data modeling in cancer classification. The proposed data pre-processing method will be described thoroughly in this section.

A. Microarray Data

The microarray gene expression dataset is the dataset obtained from microarray technology. These data are deposited in many different databases, which can be extracted depending on the issues of researchers. Some of the common public microarray databases are National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database, The Cancer Genome Atlas (TCGA) database, the ArrayExpress database, Stanford Microarray Database (SMD) and so on. Table I shows the descriptions of the mentioned public databases.

TABLE I
MICROARRAY DATABASES

Microarray Databases	Descriptions
Gene Expression Omnibus - NCBI	Gene Expression Omnibus – NCBI is an international public repository that stores and distributes high-throughput gene expression and other functional genomics datasets [9].
The Cancer Genome Atlas (TCGA)	The Cancer Genome Atlas is a public free-access database that catalogs a collection of different cancers' expression data [10].
ArrayExpress	ArrayExpress is an open-source microarray database storing and providing access to high-throughput functional genomics data [11].
Stanford Microarray Database (SMD)	Stanford Microarray Database stores raw and normalized data from microarray experiments and is made available to researchers for applications [12].

Microarray data are stored in the format of large matrices of gene expression levels. The rows represent the genes that have been under different experimental conditions or samples represented by the columns [13]. Two types of profiles are exhibited in the microarray matrix structure: the gene profile and the array profile. Gene profile is the expression values of a single gene in a variety of samples or conditions [13]. In comparison, an array profile is the expression values of many genes in one sample or condition.

In this study, six datasets were obtained from the NCBI GEO database: the lung cancer dataset [27], stomach cancer dataset [28], liver cancer dataset [29], kidney cancer dataset [30], thyroid cancer dataset [31], and breast cancer dataset [32]. Table II presents the details of the selected cancer datasets.

TABLE II
GENE EXPRESSION DATASETS

Cancer	GEO ID	Platform ID	Number of Cancerous Samples	Number of Normal Samples
Lung	GSE10072	GPL96	58	49
Stomach	GSE13911	GPL570	38	31
Liver	GSE17856	GPL6480	43	44
Kidney	GSE15641	GPL96	69	23
Thyroid	GSE33630	GPL570	60	45
Breast	GSE3494	GPL96	60	176

The range of sample identification (ID) for each cancer dataset is shown in Table III.

TABLE III
SAMPLES ID OF GENE EXPRESSION DATASETS

Cancer	GEO ID	Platform ID	Range of Samples ID	Total Number of Samples
Lung	GSE10072	GPL96	GSM254625-GSM254621	107
Stomach	GSE13911	GPL570	GSM350411-GSM350479	69
Liver	GSE17856	GPL6480	GSM446165-GSM446251	87
Kidney	GSE15641	GPL96	GSM391107-GSM391198	92
Thyroid	GSE33630	GPL570	GSM831749-GSM831853	105
Breast	GSE3494	GPL96	GSM79114-GSM79615	236

B. GenePattern

GenePattern is an open-source software package that provides access to various computational methods used to analyze genomic data [14]. It aims to provide four important functionalities, which are accessibility, reproducibility, extensibility, and multiple interfaces [15]. From the perspective of accessibility, GenePattern provides access to over two hundred genomic analysis tools for researchers to develop, capture, and reproduce genomic analysis methodologies. These genomic analysis tools (referred to as "modules") in the GenePattern module repository allow for the analysis and visualization of microarray, Single Nucleotide Polymorphism (SNP), proteomic, and sequence data [14].

GenePattern ensures the reproducibility of analysis methods and results by capturing the source of the data and analytic methods [14]. It provides automated history and provenance tracking (along with methods applied and parameter settings) for users to share and reproduce a computational analysis [15]. In addition, GenePattern facilitates simple creation and integration that allows users to

import their methods and code for sharing. Multiple interfaces are made available to a broad range of users, such as web browsers, applications, and programmatic interfaces for users to analyze without any programming through a point-and-click user interface. Fig. 1 shows the web interface of GenePattern.

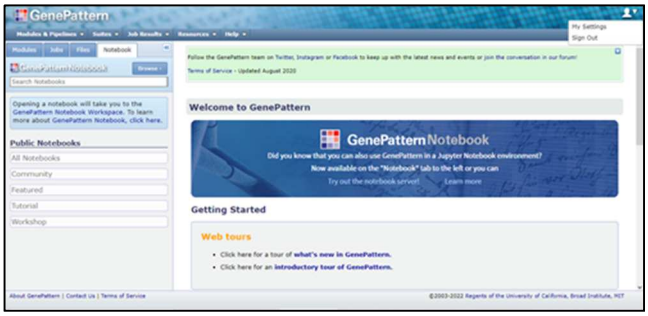

Fig. 1 The Web Interface of GenePattern

This study utilizes two GenePattern modules, the AffySTExpressionFileCreator module and the PreprocessDataset module. AffySTExpressionFileCreator module is aimed to create a Gene Cluster Text (GCT) file from a set of CEL files (Affymetrix Probe Results File) from Affymetrix ST arrays [16]. This module allows the transformation of a gene expression data file (.CEL) to a computer-readable tab-delimited text file (.GCT) to analyze matrix-compatible gene expression datasets. There are six parameter settings available in this module, which includes input file, normalize, background correct, clm file, annotate probes, and output file base [16]. The input file parameter accepts one or more Affymetrix ST CEL files for analysis. Normalized parameters allow users to normalize data using quantile normalization. Background correction aims to remove geographical biases in fluorescent intensity. clm file is a tab-delimited text file containing one scan, sample, and class per line. Annotate probes parameter provides rows annotation with the gene symbol and description. The output file base parameter sets the base name of the output file. Fig. 2 shows the screenshot interface of AffySTExpressionFileCreator module available on GenePattern.

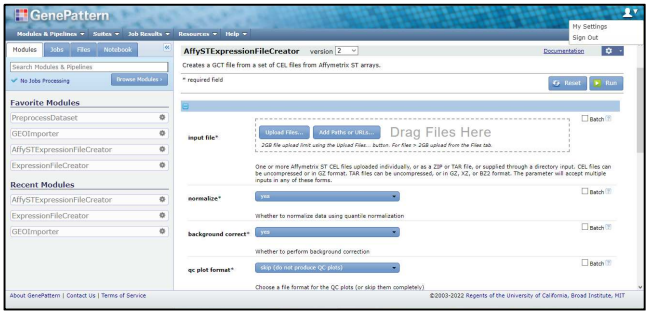

Fig. 2 The Interface of AffySTExpressionFileCreator Module

On the other hand, the PreprocessDataset module provides a variety of pre-processing operations which aim to remove platform noise and genes that have little variation so the subsequent analysis can identify interesting variations, such as the differential expression between tumor and normal tissue [17]. This module performs several pre-processing steps on GCT input file to produce filtered, pre-processed gene expression data. There are four main module parameter settings: thresholding/ceiling, variation filtering, normalization, and log2 transform. The threshold filtering parameter removes a gene whose expression profile contains insufficient values greater than a specified threshold. Floor and ceiling values can be set manually by users. The variation filtering parameter removes a gene if the variation of its expression values across the samples does not meet a minimum threshold. Row normalization or log2 transform parameters remove systematic variation of gene expression values between microarray experiments. Fig. 3 presents the screenshot interface of PreprocessDataset module available on GenePattern.

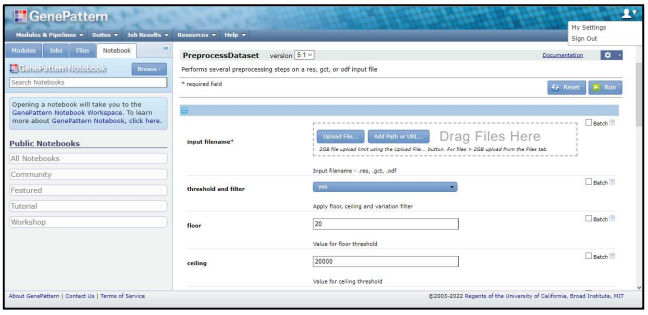

Fig. 3 The Interface of PreprocessDataset Module

C. *Pre-analysis*

Pre-analysis of gene expression data is aimed to generate a computer-readable tab-delimited text file (.GCT) for data pre-processing. Fig. 4 illustrates the pre-analysis of gene expression dataset. The pre-analysis conducts the following steps in order.

1. Download microarray gene expression dataset from microarray database
2. Create a ZIP package of CEL files for the usage of GenePattern modules
3. Run module in GenePattern
4. Output GCT files of gene expression datasets for further data pre-processing

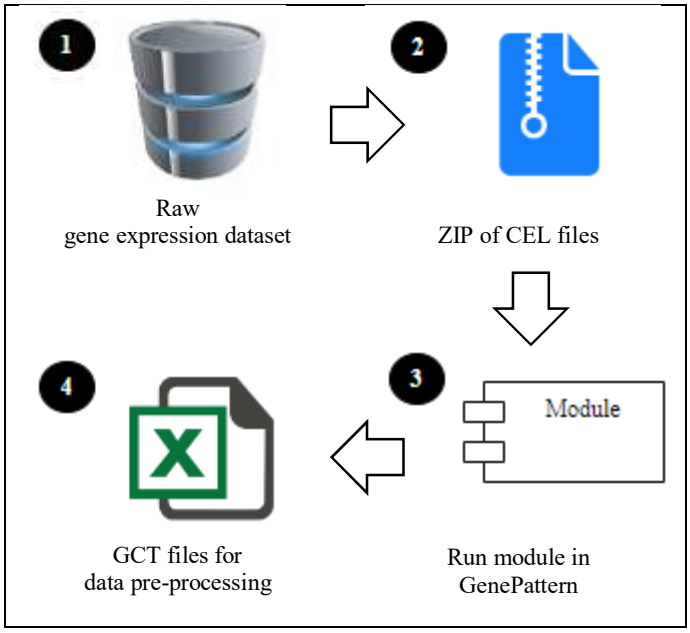

Fig. 4 Pre-analysis of Gene Expression Dataset

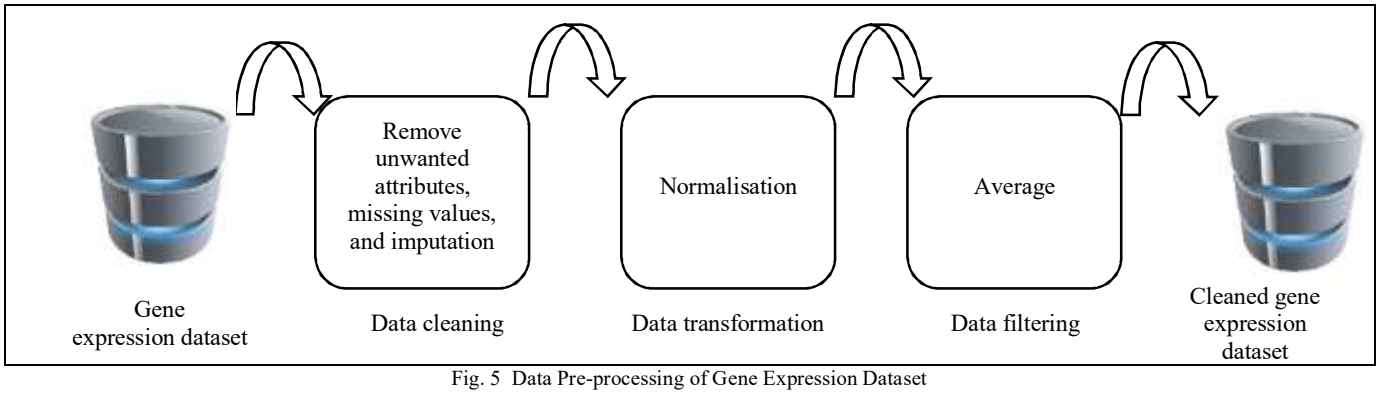

Fig. 5 Data Pre-processing of Gene Expression Dataset

This raw gene expression data file contains abundant information extracted from the cell [18]. In order to generate a GCT file for data pre-processing, a ZIP package of CEL files downloaded from the database is created for the usage of GenePattern modules in the next step. Then, the created ZIP package of CEL files was uploaded to the AffySTExpressionFileCreator module for processing. The module's normalized and background correct parameters were set to 'no' to extract the raw dataset. Other parameters were set to default behavior to obtain a matrix containing one intensity value per probe set per sample in the GCT file format [16]. The analysis module will output a GCT file of gene expression dataset that a computer can process for further data pre-processing.

D. Data Pre-processing

Microarray experiments produce huge amounts of data, and systematic pre-processing methods are required to extract meaningful expression relations [13]. The mass numbers of microarray data collected from a single experiment could be tens of thousands of data points for thousands of genes [13]. This data represents the key information for responding to crucial biological questions and hypotheses. In order to enhance the reliability of data, it is necessary to apply pre-processing techniques to extract accurate data.

After completing the pre-analysis of gene expression data, the actual data pre-processing starts. In this study, data pre-processing involves three phases: Phase 1: data cleaning, Phase 2: data transformation, and Phase 3: data filtering. Fig. 5 demonstrates the phases in data pre-processing. Data cleaning is the first step in microarray data pre-processing, which aims to correct or remove inaccurate, damaged, improperly formatted, duplicate, or incomplete data from a dataset. These dirty data will affect the mining procedure and lead to unreliable and poor output [7]. First, unwanted and empty values of attributes were removed. The unwanted attributes include patient biological information, dataset information, and dataset descriptions not applicable to cancer classification.

In comparison, empty values of attributes refer to the missing values that appeared across the rows in the gene expression dataset. Missing values occurred due to different factors, such as the corruption of the image, insufficient resolution, dust or scratches on the slide, and the robotic methods used to create the arrays [13]. Then, rows with incomplete attributes or noise data values were imputed with mean values to resolve inconsistencies in data. Noise data is a random error that is generated due to faulty data collection, or data entry errors Mean imputation method was implemented to fill the missing data elements in the gene expression dataset without reducing the sample size [20]. It creates a complete gene expression data matrix for further data analysis using classification algorithms. However, data rearrangement was run through before proceeding to the next phase. Fig. 6 shows the details of phase 1 in microarray data pre-processing.

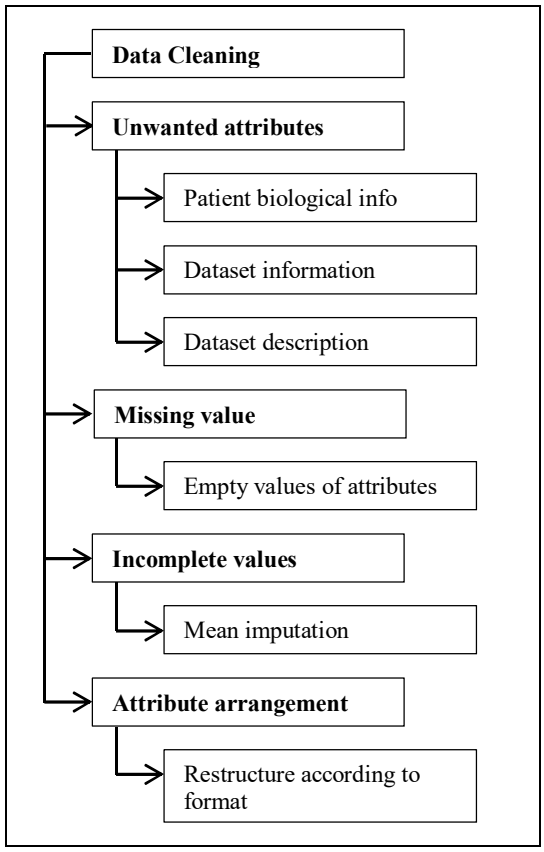

Fig. 6 Phase 1 of Data Pre-processing

In the data transformation phase, the PreprocessDataset module in GenPattern was applied to normalize gene expression data. This step aims to tune the data into a proper format suitable for analysis and other downstream processes. Fig. 7 depicts the details of phase 2 in microarray data pre-processing.

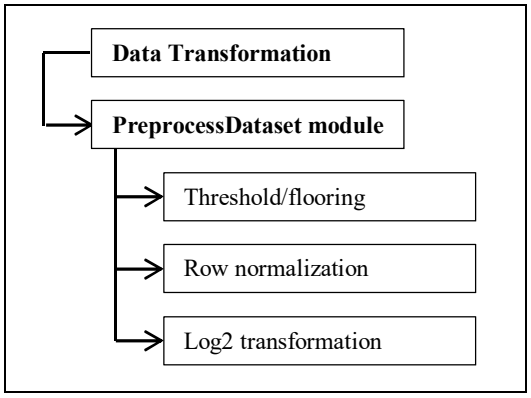

Fig. 7 Phase 2 of Data Pre-processing

The cleaned gene expression data was inputted into PreprocessDataset module for pre-processing. All the parameter settings were set to default except the row normalization and log2 transform is enabled to normalize the gene's expression values across all samples. This module undergoes a series of data transformations intended to aid in comparing gene expression data gathered across a series of hybridizations [21]. These include applying intensity thresholds or flooring to eliminate poorly detected probes and improve signal-to-noise sensitivity. Log transformation normalizes the distribution of probes across the experiment's intensity range. Row normalization scales the data into a specific range between -1.0 to 1.0 or 0.0 to 1.0. The thresholding, scaling, and log transforming data reduce variance between samples and are useful for data mining techniques like cancer classification [7].

The last phase in data pre-processing is data filtering. This step aims to reduce the huge dataset volume concerning maintaining the original dataset's integrity. Fig. 8 presents the details of phase 3 in microarray data pre-processing.

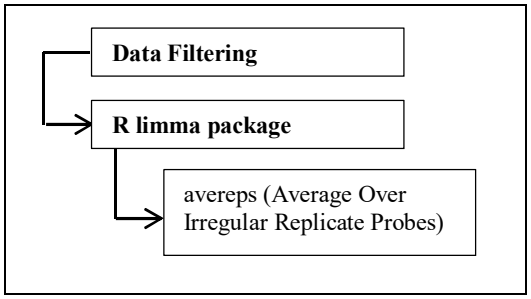

Fig. 8 Phase 3 of Data Pre-processing

Data filtering was conducted in Rstudio using R programming language [22]. The Limma package is one of the R packages build up by the R programming language for data analysis, linear models, and differential expression of microarray data [23]. Limma package was downloaded and imported in Rstudio for data pre-processing. "avereps" (Average Over Irregular Replicate Probes) function in Limma package was utilized for data reduction. It works by condensing the microarray data object so that values for within-array replicate probes are replaced with their average [23]. This method preserves highly relevant attributes and discards redundant features to reduce the size of the gene expression dataset.

After completing the three phases in data pre-processing, the cleaned dataset is now prepared to be used in both the evaluation method and the classifiers. Data pre-processing is essential to build models with this cleaned dataset effectively. This process eliminates inconsistencies or duplicates in data and increases the efficiency and reliability of data for mining procedures.

III. Result and Discussion

This study used six datasets to perform the proposed data pre-processing method. Table IV shows the number of genes after data pre-processing.

TABLE IV
GENE EXPRESSION DATASETS AFTER PRE-PROCESSING

Cancer	GEO ID	Number of Genes		Number of Removed Genes
		Raw Dataset	Cleaned Dataset	
Lung	GSE10072	22283	12986	9297
Stomach	GSE13911	54675	12419	42256
Liver	GSE17856	25075	13802	11273
Kidney	GSE15641	22283	11593	10690
Thyroid	GSE33630	54675	12986	41689
Breast	GSE3494	22283	12986	9297
Total Number of Removed Genes				124502

Based on the results in Table IV, the proposed data pre-processing method removed a total of 124502 genes across six datasets. For the lung and breast cancer dataset, the raw dataset contains 22283 genes, and the cleaned dataset left 12986 genes after pre-processing. A total of 9297 genes were removed for further data modeling. The stomach cancer dataset originally consisted of 54675 genes, and the number was reduced to 12419 genes with a total of 42256 genes removed. The liver cancer dataset contains 25075 genes before pre-processing, and the number of genes decreased to 13802 with a total of 11273 genes removed. In addition, the proposed data pre-processing method eliminated a total of 10690 genes for the kidney cancer dataset. The number of genes reduced from 22283 genes in a raw dataset to 11593 genes in the cleaned dataset. For the thyroid cancer dataset, 41689 numbers of genes were extracted from the raw gene expression dataset, resulting in the alteration of figures from 54675 to 12986 numbers of genes.

In order to depict the differences between the original raw gene expression dataset and the pre-processed breast cancer dataset, GSE3494 was used as an example to compare and visualize the attribute variation. Fig. 9 illustrates the raw CEL file for the breast cancer dataset, and fig. 10 demonstrates the cleaned excel file for the breast cancer dataset after pre-processing. By visualizing the two formats of datasets, the differences in the content presentation can be observed directly to prove the feasibility of the proposed data pre-processing method.

Based on Fig. 9, the raw breast cancer dataset contains rows of unwanted information irrelevant to data analysis and modeling. This unwanted information includes dataset information, a number of attributes contained, the dataset header and footer and so on. On the other hand, Fig. 10 shows the cleaned dataset with rows represented by gene identification and columns represented by samples. The data

displays consistent gene expression levels across samples in the breast cancer dataset.

GridCornerLR=4460 4549				
GridCornerLL=185 4503				
Axis-invertX=0				
AxisInvertY=0				
swapXY=0				
DatHeader=[0..45519] Sw_100B08_A:CLS=4733 RWS=4733 XIN=				
Algorithm=Percentile				
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:				
[INTENSITY]				
NumberCells=506944				
CellHeade	Y	MEAN	STDV	NPIXELS
0	0	251	127.5	16
1	0	10808.5	1561.2	16
2	0	252.5	138.2	16
3	0	11345	1759.9	16
4	0	91.5	13.4	16
5	0	227	136.6	16
6	0	10072.5	1802.6	16
7	0	210.3	141.9	16
8	0	10750.5	2466.8	16
9	0	233.3	142.3	16
10	0	10708.3	2339.2	16
11	0	272	137.9	16
12	0	10588.5	2190.4	16

Fig. 9 Visualization of CEL File for Breast Cancer Dataset

	GSM79114	GSM79115	GSM79116	GSM79118	GSM79119
hsa:780	8.972524	9.331111275	9.5361809	9.071774017	9.4108437
hsa:5982	6.0392565	6.60767885	6.96521125	6.63279034	6.50706725
hsa:3310	6.5069365	7.247038	6.58770955	6.804933155	6.561864
hsa:7849	5.45765025	5.435932	5.195440813	4.815377784	4.820245212
hsa:2978	5.5150135	5.47243435	4.5967861	4.535134435	5.1622289
hsa:7318	7.591369	6.3436442	7.5554179	7.68984344	7.3549348
hsa:7067	4.6533062	4.82416644	6.2428216	4.655342336	4.8824148
hsa:11099	4.8056355	4.55500415	4.52033905	4.30819094	4.65889335
hsa:6352	6.8020745	6.4429004	7.48613555	7.95628281	6.3469897
hsa:1571	4.54859725	3.760421825	3.9451424	4.495772315	4.384185475
hsa:2049	7.329829	8.53473295	6.45850435	7.14603085	7.6578807
hsa:2101	6.219669	5.9303862	6.58897665	6.49576232	6.2650185
hsa:1548	7.915119333	5.883876267	5.9637129	7.987507053	5.9113722
hsa:2621	8.574576	7.5476777	8.3037326	8.381251365	8.53752985
hsa:4323	7.6858085	7.2755046	6.789180775	7.125331688	7.9183506
hsa:8717	5.981552	5.489985467	6.1733356	6.341214963	6.4582592
hsa:2342	6.4192565	5.92204325	6.0529437	6.0574556	6.30876405
hsa:5337	4.811081	4.823439825	4.6228353	4.900894465	4.554106075
hsa:44126	7.3690825	7.91875515	7.36965795	7.39681234	7.49960455
hsa:572	6.593612	6.57549745	6.7014995	6.881980275	6.94459265
hsa:10594	8.49001	8.2623439	8.1942423	8.3616676	8.5866744
hsa:826	9.119078	8.7461525	9.33256	9.05692462	9.1938377
hsa:11224	9.704547	10.7084716	11.1748881	9.99877466	10.2526752
hsa:6158	8.948249	8.88662365	8.7332443	8.816824815	8.9556461

Fig. 10 Visualization of Excel File After Pre-processing

IV. Conclusion

The emergence of microarray technology provides solutions to crucial biological problems at the molecular level. It serves various purposes in research and clinical studies to help extract intrinsic patterns or knowledge, which may be useful for uncovering the causes of critical diseases [24, 25, 26]. However, the huge amount of microarray data is one of the unsolvable matters for researchers. Real-world microarray data are incomplete, noisy, and missing for data mining. Hence, data pre-processing is a mandatory process to facilitate the use of this powerful technology. Integrating three data pre-processing steps provides a solution to minimize the obstacles faced by analysts.

This study proposed a feasible data pre-processing method covering three phases: (1) Data cleaning, (2) Data transformation, and (3) Data filtering. Data cleaning is aimed at removing noisy, missing, and unnecessary data. Data transformation transforms data into an appropriate format and reduces data volume for efficient and effective data mining. The proposed method was applied to six cancer datasets and recorded a decrease in the number of genes after pre-processing. The differences in the number of genes between the original dataset and the cleaned dataset proved the feasibility of the proposed data pre-processing method to generate high-quality data.

Acknowledgment

Universiti Tun Hussein Onn Malaysia funds this paper. The authors appreciate the Malaysia Ministry of Higher Education (MoHE). This research was funded under REGG FASA 1/2021 (VOT NO. H888). This work also was supported/funded by Universiti Teknologi Malaysia under UTM Fundamental Research Grant (UTMFR): Q.J130000.3851.21H94.

References

[1] Owzar, K., Barry, W. T., Jung, S. H., Sohn, I., & George, S. L. (2008). Statistical challenges in pre-processing in microarray experiments in cancer. Clinical Cancer Research, 14(19), 5959-5966.

[2] Bharti, S., Krishnan, N., Veyssi, A., Momeni, M., & Raj, S. (2022). sMAP: An interactive microarray data analysis tool for early-stage researchers. bioRxiv.

[3] Herrero, J., Díaz-Uriarte, R., & Dopazo, J. (2003). Gene expression data pre-processing. Bioinformatics, 19(5), 655-656.

[4] García de la Nava, J., van Hijum, S., & Trelles, O. (2003). Pre P: gene expression data pre-processing. Bioinformatics, 19(17), 2328-2329.

[5] Deepak Jain. (2021, June 29). Data Preprocessing in Data Mining. Retrieved November 1, 2022, from https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

[6] Revathy N, Amalraj D. Accurate Cancer Classification Using Expressions of Very Few Genes. International Journal of Computer Applications. 2011;14(4):19-22.

[7] Alasadi, S. A., & Bhaya, W. S. (2017). Review of data pre-processing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.

[8] Wikipedia contributors. (2018, June 4). Microarray databases. In Wikipedia, The Free Encyclopedia. Retrieved 01:06, November 1, 2022, from https://en.wikipedia.org/w/index.php?title=Microarray_databases&oldid=844388880.

[9] Clough, E., & Barrett, T. (2016). The gene expression omnibus database. In Statistical genomics (pp. 93-110). Humana Press, New York, NY.

[10] Tomczak, K., Czerwinska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, Współczesna Onkologia, vol. 19, no. 1A, pp. A68-A77.

[11] Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., ... & Brazma, A. (2007). ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic acids research, 35(suppl_1), D747-D750.

[12] Sarkans, U., Parkinson, H., Lara, G. G., Oezcimen, A., Sharma, A., Abeygunawardena, N., ... & Brazma, A. (2005). The ArrayExpress

gene expression database: a software engineering and implementation perspective. Bioinformatics, 21(8), 1495-1501.
[13] Rafii, F., & Rossi, B. D. (2015). Data pre-processing and reducing for microarray data exploration and analysis. International Journal of Computer Applications, 132(16), 20-26.
[14] Kuehn, H., Liberzon, A., Reich, M., & Mesirov, J. P. (2008). Using GenePattern for gene expression analysis. Current protocols in bioinformatics, 22(1), 7-12.
[15] Wikipedia contributors. (2021, December 23). GenePattern. In Wikipedia, The Free Encyclopedia. Retrieved 03:17, November 1, 2022, from https://en.wikipedia.org/w/index.php?title=GenePattern&oldid=1061704802.
[16] David Eby, Broad Institute. (n.d.). AffySTExpressionFileCreator (v1) BETA. Retrieved November 1, 2022, from https://www.genepattern.org/modules/docs/AffySTExpressionFileCreator/1.
[17] Joshua Gould, Broad Institute. (n.d.). PreprocessDataset (v5). Retrieved November 1, 2022, from https://genepattern.org/modules/docs/PreprocessDataset/5?print=yes.
[18] Seah, C. S., Kasim, S., Fudzee, M. F., Mohamad, M. S., Saedudin, R. R., Hassan, R., ... & Atan, R. (2018). An effective pre-processing phase for gene expression classification. Indonesian Journal of Electrical Engineering and Computer Science, 11(3), 1223.
[19] Sadhvi Anunaya. (2022, June 20). Data Preprocessing in Data Mining -A Hands On Guide. Retrieved November 2, 2022 from https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/.
[20] Peterson, P. L., Baker, E., & McGaw, B. (2010). International encyclopedia of education. Elsevier Ltd.
[21] Normalization supplement: commentary on the impact of different normalization methodologies on variance distributions at a global and pathway level. Retrieved November 2, 2022 from https://doi.org/10.1371/journal.pgen.1002207.s003.
[22] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
[23] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Research, 43(7), e47. doi: 10.1093/nar/gkv007.
[24] Donatin, E., & Drancourt, M. (2012). DNA microarrays for the diagnosis of infectious diseases. Médecine et maladies infectieuses, 42(10), 453-459.
[25] Tzouvelekis, A., Patlakas, G., & Bouros, D. (2004). Application of microarray technology in pulmonary diseases. Respiratory research, 5(1), 1-18.
[26] Yoo, S. M., Choi, J. H., Lee, S. Y., & Yoo, N. C. (2009). Applications of DNA microarray in disease diagnostics. Journal of microbiology and biotechnology, 19(7), 635-646.
[27] Landi MT, Dracheva T, Rotunno M, Figueroa JD et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS One 2008 Feb 20;3(2):e1651. PMID: 18297132.
[28] D'Errico M, de Rinaldis E, Blasi MF, Viti V et al. Genome-wide expression profile of sporadic gastric cancers with microsatellite instability. Eur J Cancer 2009 Feb;45(3):461-9. PMID: 19081245.
[29] Tsuchiya M, Parker JS, Kono H, Matsuda M et al. Gene expression in nontumoral liver tissue and recurrence-free survival in hepatitis C virus-positive hepatocellular carcinoma. Mol Cancer 2010 Apr 9;9:74. PMID: 20380719.
[30] Jones J, Otu H, Spentzos D, Kolia S et al. Gene signatures of progression and metastasis in renal cell cancer. Clin Cancer Res 2005 Aug 15;11(16):5730-9. PMID: 16115910.
[31] Tomás G, Tarabichi M, Gacquer D, Hébrant A et al. A general method to derive robust organ-specific gene expression-based differentiation indices: application to thyroid cancer diagnostic. Oncogene 2012 Oct 11;31(41):4490-8. PMID: 22266856.
[32] Miller LD, Smeds J, George J, Vega VB et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 2005 Sep 20;102(38):13550-5. PMID: 16141321