Journal of Theoretical and Applied Information Technology
31st October 2017. Vol.95. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PROGNOSIS CANCER PREDICTION MODEL USING DEEP BELIEF NETWORK APPROACH

[1]FAKHIRAH D. GHAISANI, [2*]ITO WASITO, [3]MOH. FATURRAHMAN, [4]RATNA MUFIDAH

[1, 2, 3, 4]Faculty of Computer Science Universitas Indonesia, Indonesia

E-mail: [1]fakhirah.dianah61@ui.ac.id, [2]ito.wasito@cs.ui.ac.id,
[3]moh.faturrahman@ui.ac.id,[4]ratna.mufidah51@ui.ac.id
[2*)] Corresponding Author

ABSTRACT

Cancer is one of main non-communicable disease. Analysis of cancer prognosis is necessary to determine the proper treatment for each patient. However, cancer data analysis is challenging because multiple risk factors may influence the prognosis of cancer, including genes and clinical condition of patients. This study aims to develop prediction model for cancer prognosis using clinical and gene expression (microarray) data. In this research, Principal Component Analysis (PCA) is applied to microarray data to reduce its dimension, then two Deep Belief Network (DBN) models for both clinical and microarray data are trained separately. Probabilities obtained from Clinical DBN model and Microarray DBN model are integrated using softmax nodes on Bayesian Network structure. Based on various experiments, the best DBN-BN integration model obtains prediction accuracy 73.3535% for overall survival prediction and 71.3434% for disease-free survival prediction.

Keywords: *Cancer, Prognosis, Principal Component Analysis, Dimensionality Reduction, Deep Belief Network, Bayesian Network, Data Integration, Data Harmonization, Microarray, Gene Expression*

1. INTRODUCTION

Cancer is one of main non-communicable diseases (NCD) together with cardiovascular disease, chronic respiratory disease, and diabetes. This disease causes approximately 8.2 million human deaths in the world each year [1].

Analysis of cancer prognosis is necessary to determine the proper treatment for each patient. However, cancer data analysis is challenging because multiple risk factors may influence cancer prognosis, including gene, clinical condition of patient, and cancer stage [2]. Previous cancer studies have successfully collected an enormous amount of cancer patient data [3]. Applying machine learning techniques, these data can be used to develop prediction model for cancer prognosis. This model can be used to predict cancer progression on patient, such as recurrence and survival of patient [2].

Most previous works in disease diagnosis have used only patient clinical data [4, 5, 6]. Meanwhile for cancer diagnosis, some studies use microarray data [3] or both clinical and microarray data [2, 7, 8], considering cancer is a genetical disease [2]. Challenge in clinical and microarray data analysis is high-dimensional data (particularly microarray data) compared to number of samples. Thus, the number of variables are much larger compared to the number of equations. Besides, data integration method is needed to combine information from clinical and microarray data which have different characteristics.

Bayesian Networks (BN) has been used for data integration in some previous researches. Gevaert et al. (2006) [8] integrate clinical and microarray data with experiment three types of integration method (full/early, partial/intermediate, and decision/late integration) using BN for prognosis of cancer. Based on experiments conducted in [8], the average AUC obtained is 0.793 for intermediate integration and 0.747 for early integration. Late integration and intermediate integration outperforms early integration because clinical and microarray data which have different characteristics are processed separately, not combined as a dataset as in early integration [8, 9]. In other work, Khademi and Nedialkov (2015) [2] use late integration method for prognosis of cancer. In [2], clinical and microarray data are trained separately, then the two models obtained (clinical model and microarray model) are integrated using

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

softmax nodes on BN structure. In that study, clinical model is constructed using BN, meanwhile microarray model is constructed using Deep Belief Networks (DBN). Experiment results in [2] using NKI dataset show that DBN and BN integration model obtains accuracy 97% for survival prediction, while Clinical BN model obtains 96% accuracy. Thus, DBN and BN integration model outperforms Clinical BN model. However, structure and parameter learning in BN is challenging, especially when using more complex data. This problem can be addressed by reducing complexity of the data. Gevaert et al. (2006) [8] applied feature selection method to microarray data to select some genes. But, this gene selection technique may lead to loss of important genetic information for cancer prognosis.

In this study, Deep Learning method, specifically DBN is utilized to develop prediction model using clinical data and microarray data. These two DBN models for both clinical and microarray data are trained separately. Then, probabilities obtained from Clinical DBN model and Microarray DBN model are integrated using softmax nodes on BN structure [2]. In previous work [2], BN is used to construct Clinical model. Meanwhile, in this research Clinical model is constructed using DBN. DBN is widely used for classification and clustering tasks, especially when complex and large-scale data is used. Some previous works also have been used DBN to predict disease prognosis [2, 4-7].

Data integration mechanism can be applied to combine patient data with different characteristic attributes, such as clinical and microarray data. Beside variation in data type and characteristic, there are enormous sources of patient data that have been collected in the previous researches around the world, thus researchers need to combine data from different sources to expand the scope of study. However, there are challenges in combining some sources of patient data, such as heterogeneous data and there is no alignment of terminology used in patient data recording. Thus, to use heterogeneous data from various researches, data homogenization or harmonization is needed [10]. Spjuth et al. (2016) [10] propose data harmonization method. First, variable of interest (VOI) is determined, then list of harmonized vocabulary (HV) id created, HV mapping, and information integration.

In previous work, Khademi and Nedialkov (2015) [2] proposed a classification model using DBN and BN for prediction of cancer prognosis. Manifold learning is used in pre-processing step to reduce the dimension of microarray data. However, we think that the pre-processing step may affect the performance of classification model. Thus, in this work, experiments in clinical and microarray data pre-processing are performed, such as KNN-impute and dimensionality reduction experiments. Moreover, experiments in data harmonization are also performed.

The aim of this research is to construct classification model using DBN and BN to integrate clinical and gene expression data of cancer patient. Thus, this model can be used for prognosis of cancer. Besides, this research aims to harmonize clinical patient data.

2. METHODOLOGY

2.1 Algorithms

2.1.1 Dimensionality reduction

In this research, three dimensionality reduction techniques are used. They are Gene-Shaving, Isometric Projection (IsoProjection), and Principal Component Analysis (PCA). PCA and IsoProjection is linear and nonlinear dimensionality reduction technique, meanwhile Gene-Shaving is gene (biomarker) selection technique.

2.1.2 Classification

Classification algorithms used in this research are:

1. Deep Belief Network (DBN)

   Deep Belief Networks (DBN) is one of deep learning model. DBN is a graphical generative model which consists of Restricted Boltzmann Machine (RBM) on the top two layers and Sigmoid Belief Networks (SBN) on other layers below. The top two layers are connected indirectly and symmetrically forming associative memory, other layers below are connected top-down (directed), while the bottom layer represents data vector [11].

   Learning process in DBN starts with unsupervised pre-training on RBM until equilibrium sample is reached. There is a fast RBM training algorithm named contrastive divergence. The result of pre-training forwarded to the next layer, until states of each layers obtained. This process is a generative model of DBN. In this paper, DBN is utilized to construct prediction model with clinical and microarray data.
2. Bayesian Network (BN)

   Bayesian Networks is a probabilistic graphical model which

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

represents random variables and conditional dependencies. Its structure is Directed Acyclic Graph (DAG). Each node Xi (random variable) has Conditional Probability Distribution (CPD) P(Xi|parent(Xi)), which defines probability of certain node given its parent. The most common CPD used in BN are Table CPD, Gaussian CPD, and Softmax CPD. Table CPD is used for discrete node and discrete parent, Gaussian CPD is used for continuous node and discrete and continuous parent, and Softmax CPD can be used to represent discrete node with continuous parent [2]. In this paper, we use BN with Softmax CPD to integrate probability obtained from Clinical DBN and Microarray DBN.

2.1.3 Data integration

Utilizing DBN and BN, [2] proposed data integration method using Bayesian Network. In this research, data integration process illustrated in Figure 1.

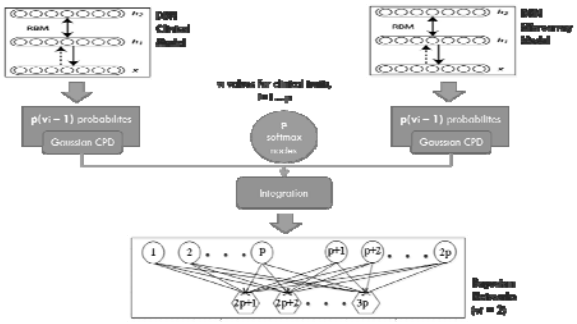

Figure 1: Data Integration Algorithm

1. Clinical and microarray data are trained separately using two different DBNs, they are Clinical DBN and Microarray DBN. Generative mechanism processed by stacked RBMs, then discriminative fine-tuning mechanism is processed by backpropagation algorithm on Multi-Layer Perceptron. DBN outputs are $p(v_i\text{-}1)$ probabilities, where p is number of attributes we want to predict and v_i is number of possible values for each attribute we want to predict. In this research, we want to predict OS_STATUS and DFS_STATUS which are binary attributes. Thus, $p = 2$, $v_1 = 2$ and $v_2 = 2$.
2. Probabilities obtained from two DBNs can be represented as Gaussian Conditional Probability Distribution (CPD) on BN.
3. Gaussian CPD from two sources (clinical and microarray model) are integrated using p softmax nodes. Softmax node will produce probability of certain attribute has value 1. These probabilities determine the prediction label.

2.2 Data Description

In this research, two breast cancer datasets are used. They are METABRIC Dataset and TCGA Dataset.

2.2.1 METABRIC dataset

This dataset consists of clinical data (MET-clin) and gene expression data (microarray) (MET-gene) which has been collected previously by Pereira et al. This dataset consists of 1980 samples, 30 clinical attributes and 24368 gene expression attributes. This dataset is available online and can be accessed freely on cBioPortal website (http://www.cbioportal.org/study?id=brca_metabric)

Microarray data in this dataset is a 24368 x 1980 table, each cell is real value number. Each row in this data represents probe (gen) and each column represents sample. Value in cell *ij* (row *i* column *j*) represents intensity of gene expression *i* on sample *j*.

Out of 10 attributes, 2 attributes are used as clinical traits we want to predict, they are disease-free survival (cancer/no cancer) and overall survival (dead/alive). These two attributes are binary with possible values 0 or 1. Besides, 3 attributes are not used for prediction because they only have one possible value. They are SAMPLE_TYPE, CANCER_TYPE and CANCER_TYPE_DETAILED.

2.2.2 TCGA dataset

In this research, only clinical data from TCGA Dataset that are used. This dataset is collected by National Cancer Institute, National Human Genome Research Institute, National Institutes of Health. This dataset contains 1096 samples with 112 clinical attributes. This dataset is available online and can be accessed freely on cBioPortal website (www.cbioportal.org/study?id=brca_tcga). TCGA dataset will be used for data harmonization experiments.

2.3 Research Framework

In this paper, two prediction models are constructed using clinical data and gene expression data of cancer patient separately utilizing DBN. Each model (Clinical DBN and Microarray DBN) will produce probabilities for classification. Then, these two models are integrated using softmax

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

nodes on Bayesian Networks structure [2], thus probabilities for classification from two different sources of information obtained (from clinical and gene expression data). In general, framework of this research is given in Figure 2.

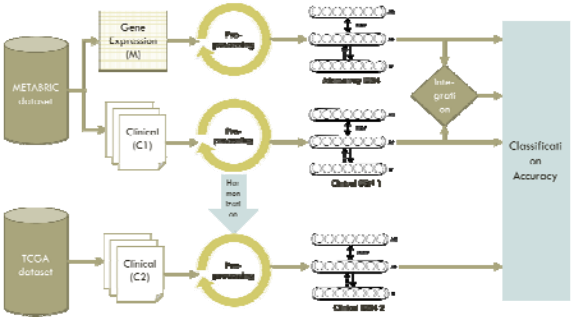

Figure 2: Research Framework

2.3.1 Data pre-processing

METABRIC dataset used in this research consists of clinical and gene expression (microarray) data which is available online in separate files. The first step of METABRIC preprocessing is patient mapping between clinical and gene expression data. Then, these clinical (MET-clin) and microarray data (MET-gene) are processed separately. Clinical data is cleaned by removing irrelevant data. Then, string values are converted to nominal (binary or categorical). After string to nominal conversion, continuous attributes are converted to discrete categorical by defining groups for each attribute. The next process in missing value imputation/estimation. The last, categorical attributes are converted to binary.

This research also aims to do experiments in data harmonization. Two clinical data will be harmonized, which are MET-clin and TCGA dataset. MET-clin data consists of 32 attributes, while TCGA has 115 clinical attributes. All attributes from both data sources are defined as Variable of Interest (VOI), while attribute names from MET-clin data used as Harmonized Vocabulary (HV). Then, attributes from both data sources are mapped.

2.3.2 Deep Belief Networks (DBN) model construction

Two DBN models are constructed, Clinical DBN and Microarray DBN. Clinical attributes from METABRIC dataset are used as input for Clinical DBN, while *d* dimension microarray data (result of dimensionality reduction) as input for Microarray DBN. Then, those two DBNs are trained separately. Experiments are conducted with variation of architectures and hyperparameters. The last layer of each DBN model has 2 nodes as many as number of clinical trait we want to predict. In this research, DeeBNet toolbox V3.2 on MATLAB is used, which is developed by [12] (http://ceit.aut.ac.ir/keyvanrad/DeeBNet%20Toolbox.html).

2.3.3 Bayesian Networks (BN) integration model construction

After the best parameter settings and models for Clinical DBN and Microarray DBN have been obtained, probability for each clinical outcome can be obtained. In this paper, 2 probabilities are obtained. These probabilities can be represented as continuous nodes with Gaussian CPD on BN. Then, these nodes are integrated using 2 softmax nodes. As results of these softmax nodes, probabilities can be used to classify patient. In this research, Bayes Net Toolbox on MATLAB is used, which is developed by [13] (https://code.google.com/archive/p/bnt/).

2.3.4 Experiments design

In general, design of experiments conducted in this research.

1. Imputation simulation using KNN-impute. Before applying imputation (missing value estimation) on data, experiment is conducted to determine the best *k* (number of neighbor in KNN-impute) for each data.
2. Dimensionality reduction experiment. In this experiment, microarray data dimension is reduced by three different methods, which are Principal Component Analysis (PCA), Isometric Projection (IsoProjection) and Gene-Shaving. Then, each preprocessed data is trained using DBN and classification performance is evaluated.
3. Data integration experiment. In this experiment, classification performance of DBN-BN integration model, Clinical DBN model, and Microarray DBN model are compared.
4. Data harmonization experiment. In this experiment, classification performance of Clinical DBN trained by METABRIC data and Clinical DBN trained by METABRIC + TCGA harmonized data are evaluated.
5. Classification method experiment. This experiment aims to compare classification performance of DBN-BN integration model with other classifiers, which are Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.4 Evaluation

2.4.1 Imputation simulation

In imputation simulation, performance is evaluated and measured by imputation error, as described in Formula 1.

$$Imputation\_error = \frac{1}{n}\sum \frac{(x - x_{impute})^2}{x^2} \quad (1)$$

Where x is value in real data, while x_{impute} is value estimated by KNN-impute.

2.4.2 Classification result

Evaluation is done by applying 10-fold cross validation, then quantitative performance is measured by calculating classification accuracy of each prediction model. Classification accuracy is defined as

$$Accuracy = \frac{\#correctly\_classified\_samples}{size\_of\_data} \quad (2)$$

2.4.3 Pair-wise comparison

During experiment, if performance average ± standard deviation of method A and method B are overlapping, then this result is inconclusive, thus the best method cannot be determined. To overcome this problem, pairwise comparison method is used as in previous work by Wasito and Mirkin (2006) [14] to determine which method outperforms others.

Table 1: Pair-wise Comparison between method A and method B.

	A	B
A	-	60
B	40	-

For example, in Table 1, cell (1,2) tells percentage of experiments when method A outperforms method B, while cell (2,1) gives percentage of experiments when method B is better. If there are 50 experiments and method A outperforms method B 30 times (60%), then cell (1,2) in pairwise comparison table has value 60.

3. RESULT

3.1 Imputation Simulation

Table 2 presents average and standard deviation of imputation error. Table 2 shows that range of imputation error value for each k is overlapping, thus this result is inconclusive. Then, performance of k configuration is evaluated using pairwise comparison as given in Figure 3-5.

Experiment result shows that on MET-gene microarray data, at least 25 nearest attributes are needed to estimate missing value well. Then, for $k > 25$, imputation error raises, because non-strong-correlated attributes with missing value is also considered. It also happens on MET-clin clinical data, with best k obtained is 15, while for TCGA clinical data the best k obtained is 10.

Table 2: Imputation performance of KNN-impute with variation in k values

k	Imputation Error					
	MET-gene (n_m = 10)		MET-clin (n_{c1} = 1000)		TCGA (n_{c2} = 700)	
	Avg	Std	Avg	Std	Avg	Std
5	0.0253	0.0233	0.1384	0.0066	0.1405	0.0026
10	0.0248	0.0247	0.1153	0.0046	0.1389	0.0064
15	0.0231	0.0192	0.1148	0.0046	-	-
20	0.0234	0.0204	0.1154	0.0040	-	-
25	0.0229	0.0196	0.1146	0.0040	-	-
30	0.0232	0.0203	-	-	-	-
35	0.0233	0.0204	-	-	-	-
40	0.0241	0.0230	-	-	-	-
45	0.0240	0.0219	-	-	-	-
50	0.0238	0.0213	-	-	-	-
55	0.0241	0.0219	-	-	-	-
60	0.0243	0.0224	-	-	-	-
65	0.0242	0.0221	-	-	-	-
70	0.0243	0.0219	-	-	-	-
75	0.0243	0.0220	-	-	-	-
80	0.0244	0.0219	-	-	-	-
85	0.0245	0.0217	-	-	-	-
90	0.0246	0.0219	-	-	-	-
95	0.0248	0.0222	-	-	-	-
100	0.0249	0.0222	-	-	-	-

Experiment result also shows that imputation error for microarray data is most likely smaller than clinical data (see Figure 6). It is caused by the small number of missing values on microarray data used (n_m = 10) compared to large number of attributes (24368 attributes), thus data has a lot of information to estimate missing values. On clinical data, there are more missing values (n_{c1} = 1000 in MET-clin, and n_{c2} = 700 in TCGA data) and have fewer attributes, thus data do not have enough information to estimate missing values. This experiment also shows that standard deviation of imputation error on microarray data are most likely larger than clinical data (see Figure 7). It is because values in microarray data are more varied (real value) compared to clinical data (integer, binary value). Then, the best k values obtained for each data are used to impute real data using KNN-impute algorithm.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

k	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
5	-	46	43	47	41	43	45	44	45	45	45	46	45	44	46	45	46	46	46	47
10	54	-	45	45	44	46	48	51	50	51	54	55	52	53	52	52	54	54	54	54
15	57	55	-	49	49	47	53	54	58	55	61	60	61	62	62	62	60	62	63	63
20	53	55	51	-	42	48	52	54	57	61	60	60	58	58	55	57	53	56	58	58
25	59	56	51	58	-	56	52	57	60	61	62	66	59	60	60	62	61	62	63	63
30	57	54	53	52	44	-	53	61	62	56	59	62	60	62	59	60	58	62	64	61
35	55	52	47	48	48	47	-	60	60	57	62	62	61	64	63	64	64	65	66	65
40	56	49	46	46	43	39	40	-	56	54	58	60	57	62	59	58	60	58	60	61
45	55	50	42	43	40	38	40	44	-	54	55	58	51	53	56	59	57	60	60	60
50	55	49	45	39	39	44	43	46	46	-	68	59	56	57	56	58	63	64	67	66
55	55	46	39	40	38	41	38	42	45	32	-	55	55	54	52	58	59	62	62	63
60	54	45	40	40	34	38	38	40	42	41	45	-	50	52	53	52	60	60	58	61
65	55	48	39	42	41	40	39	43	49	44	45	50	-	55	50	55	59	59	61	62
70	56	47	38	42	40	38	36	38	47	43	46	48	45	-	48	56	58	59	60	62
75	54	48	38	45	40	41	37	41	44	44	48	47	50	52	-	56	58	60	61	65
80	55	48	38	43	38	40	36	42	41	42	42	48	45	44	44	-	58	58	60	65
85	54	46	40	47	39	42	36	40	43	37	41	40	41	42	42	42	-	53	63	64
90	54	46	38	44	38	38	35	42	40	36	38	40	41	41	40	42	47	-	64	60
95	54	46	37	42	37	36	34	40	40	33	38	42	39	40	39	40	37	36	-	55
100	53	46	37	42	37	39	35	39	40	34	37	39	38	38	35	35	36	40	45	-

Figure 3: Pair-wise Comparison of Imputation Error of MET-gene Data with variation in k values (%)

k	5	10	15	20
5	-	0	0	0
10	100		16	51
15	100	54	-	62
20	100	49	38	

Figure 4: Pair-wise Comparison of Imputation Error of MET-clin Data with variation in k values (%)

k	5	10
5	-	40
10	60	-

Figure 5: Pair-wise Comparison of Imputation Error of TCGA Data with variation in k values (%)

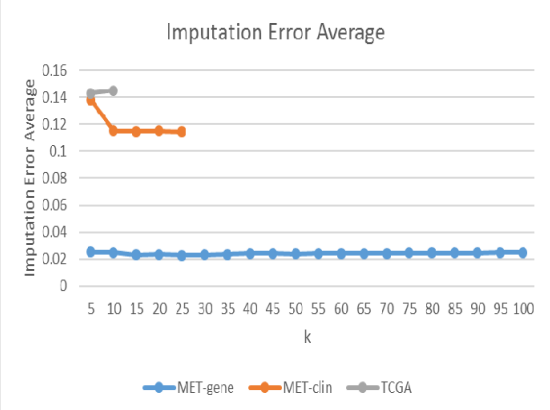

Figure 6: Chart of imputation error average

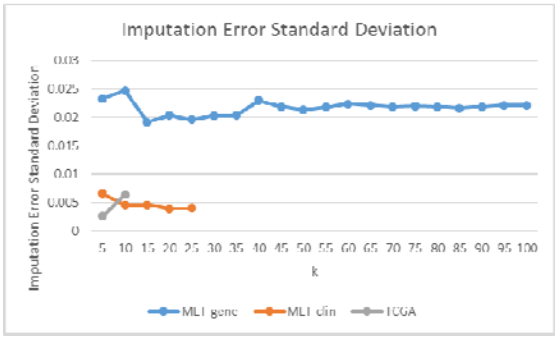

Figure 7: Chart of imputation error standard deviation

3.2 Dimensionality Reduction Experiment

In this experiment, classification performance of Microarray DBNs constructed by various preprocessed data (as result of 3 different preprocessing technique) are compared. Table 3 presents average and standard deviation of classification accuracy. However, Table 3 shows that range of classification accuracy (average ± standard deviation) for each preprocessing method is overlapping, thus this result is inconclusive. Figure 8 and Figure 9 show pairwise comparison result of accuracy for Overall Survival (OS) and Disease-Free Survival (DFS) predictions. Majority concept is applied on two pairwise comparison results to determine the best preprocessing method to predict both clinical traits (OS and DFS).

Experiment result shows that PCA and Gene-Shaving outperform IsoProjection for OS and DFS prediction. PCA is a linear dimensionality reduction technique, while IsoProjection is non-linear dimensionality reduction technique. PCA

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

outperforms IsoProjection in this experiment shows that reducing data to linear field can reduce complexity of data without losing a lot of information. In OS prediction, the best method is PCA, while in DFS prediction the best performance is obtained by Gene-Shaving. In PCA, dimension of reduced data 125 (PCA-125) is better than 250, while in Gene-Shaving, GS-250 outperforms GS-125.

Table 3: Classification Performance of Microarray DBN with Variation in Preprocessing Method

Pre-Processing	Dim	Accuracy (%)			
		OS		DFS	
		Avg	Std	Avg	Std
GS	130	62.0808	0.4796	66.6869	0.2995
	254	62.4545	0.3985	65.9394	0.5963
IsoProj	125	58.4949	0.5160	64.7677	60.3848
	250	57.9646	0.5178	63.5253	0.3530
PCA	125	64.3182	0.4531	66.4798	0.4815
	250	64.2576	0.5287	66.3535	0.5728

	GS-125	GS-250	IscP-125	IsoP-250	PCA-125	PCA-250
GS-125	-	30	100	100	0	0
GS-250	70	-	100	100	0	0
IscP-125	0	0	-	80	0	0
IscP-250	0	0	20	-	0	0
PCA-125	100	100	100	100	-	40
PCA-250	100	100	100	100	40	-

Figure 8: Pair-wise Comparison of Microarray DBN Classification Accuracy with Variation in Preprocessing Method in Overall Survival Prediction (%)

	GS-125	GS-250	IscP-125	IsoP-250	PCA-125	PCA-250
GS-125	-	90	100	100	60	60
GS-250	10	-	100	100	30	30
IscP-125	0	0	-	100	0	10
IscP-250	0	0	0	-	0	0
PCA-125	40	70	100	100	-	50
PCA-250	40	70	90	100	50	-

Figure 9: Pair-wise Comparison of Microarray DBN Classification Accuracy with Variation in Preprocessing Method in Disease-Free Survival Prediction (%)

3.3 Data Integration Experiment

This experiment is performed to compare classification performance of Clinical DBN, Microarray DBN, and DBN-BN integration model, which is given in Table 4. Because of overlapping, classification performance also evaluated using pairwise comparison as presented in Figure 10.

Based on evaluation using pairwise comparison technique on Figure 10, Clinical DBN and DBN-BN outperform Microarray DBN. In Overall Survival (OS) prediction, DBN-BN outperforms Clinical DBN, while in Disease-Free Survival (DFS), Clinical DBN competes and slightly better than DBN-BN.

Table 4: Classification Performance of Clinical DBN, Microarray DBN, and DBN-BN Integration Model

Classifier	Accuracy (%)			
	OS		DFS	
	Avg	Std	Avg	Std
Clinical DBN	71.9596	0.9990	74.9495	0.7615
Microarray DBN	64.7020	0.6165	66.6768	0.4258
DBN-BN	73.3535	1.0414	71.3434	6.4988

	C	M	CM
C	-	100	0
M	0	-	0
CM	100	100	

	C	M	CM
C	-	100	50
M	0	-	15
CM	50	85	

Figure 10: Pair-wise Comparison of Clinical DBN, Microarray DBN, and DBN-BN Integration Model Accuracy (%)

One of DBN-BN weakness method can be found in the high standard deviation accuracy (Table 4), especially in DFS prediction. In some experiments, DBN-BN outperforms other DBNs, but in some other experiments DBN-BN has worse performance. Besides, performance of integration model can be influenced by performance of Clinical DBN and Microarray DBN which are parts of it, whether in OS or DFS prediction.

3.4 Data Harmonization Experiment

This experiment is performed to compare classification performance of Clinical DBN constructed by MET-clin data and Clinical DBN constructed by harmonized MET-clin + TCGA data., which is given in Table 5. Because of overlapping, classification performance also evaluated using pairwise comparison as presented in Figure 11.

Table 5: Classification Performance of Clinical DBN Constructed by MET-clin Data and Clinical DBN Constructed by Harmonized MET-clin + TCGA Data

Clinical DBN	Accuracy (%)			
	OS		DFS	
	Avg	Std	Avg	Std
MET-clin	71.4798	0.8918	75.0354	0.8796
MET-clin + TCGA	67.2631	0.5768	75.8748	0.3473

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 11: Pair-wise Comparison of Clinical DBN Constructed by MET-clin Data and Clinical DBN Constructed by Harmonized MET-clin + TCGA Data Accuracy (%)

MET-clin data consists of 1980 samples with 32 clinical attributes, while MET-clin + TCGA harmonized data consists of 3076 samples with 19 clinical attributes. The harmonized data has more samples because of combining two different datasets. However, not all attributes of two datasets can be harmonized, thus the harmonized data has fewer attributes than MET-clin data.

Experiment result shows that DBN constructed by MET-clin is better in Overall Survival (OS) prediction. It can be caused by some attributes that cannot be harmonized have strong correlation with OS. On the contrary, DBN constructed by harmonized data outperforms DBN constructed by MET-clin in DFS prediction. It means that 19 harmonized attributes are strongly correlated with DFS of a patient. This experiment also shows that clinical data harmonization mechanism can be used to combine some data sources, to be trained with DBN and can be used to predict prognosis of cancer.

3.5 Classification Method Experiment

This experiment is performed to compare classification performance of DBN-BN integration model and other classifiers (SVM and k-NN). Classification performance of each classifier is given in Table 6, while pairwise comparison of accuracy is presented in Figure 12. Beside classification accuracy, execution time of each classifier is measured. Execution time consists of training time and testing time. Using 10-fold cross validation mechanism, 1782 samples used as training data while other 198 samples as testing data. The execution time detail of DBN-BN integration model is given in Table 7 and execution time comparison for each classifier is given in Table 8.

Based on pairwise comparison result on Figure 12, classification performance of DBN-BN integration model outperforms SVM and k-NN in Overall Survival (OS) and Disease-Free Survival (DFS) predictions. One of this DBN-BN integration model superiority is late-integration mechanism, which processes clinical and microarray data separately using DBN, then integrated by BN. In this experiment, inputs of SVM and k-NN are clinical and reduced microarray data that are combined as a dataset (early integration). Thus, there is no separated-processes for two different data in SVM and k-NN.

Table 6: Classification Performance of DBN-BN Integration Model, Support Vector Machine (SVM), and k-Nearest Neighbor (k-NN)

Classifier	Accuracy (%)			
	OS		DFS	
	Avg	Std	Avg	Std
DBN-BN	73.3535	1.0414	71.3434	6.4988
SVM	69.6263	0.9932	72.1212	0.7911
k-NN	57.6970	0.3301	59.6465	0.5548

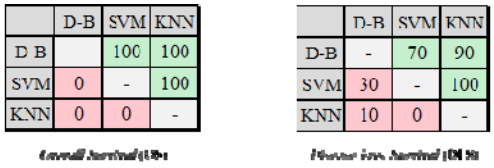

Figure 12: Pair-wise Comparison of DBN-BN, SVM, and k-NN Accuracy (%)

Table 7: Execution Time of DBN-BN Integration Model

Execution		Execution Time (s)	
		Avg	Std
Training	Clinical DBN Reconstruction	96.49947	2.52457
	Clinical DBN Fine-tuning	2.13798	0.11059
	Microarray DBN Reconstruction	46.3234	1.41308
	Microarray DBN Fine-tuning	2.12184	0.09704
	BN	0.12624	1.50231
Testing	Clinical DBN	0.01181	0.00254
	Microarray DBN	0.01074	0.00106
	BN	1.50231	0.28003

Table 8: Execution Time of DBN-BN Integration Model, SVM and k-NN

Classifier	Execution Time (s)	
	Training	Testing
DBN-BN	147.20893	1.52486
SVM	87.34850	0.0387
k-NN	0.14537	0.65228

3.6 Discussion

In previous work, Khademi and Nedialkov (2015) [2] proposed a classification model using DBN and BN for prediction of cancer prognosis. Manifold learning is used in pre-processing step to reduce the dimension of microarray data. However, we think that the pre-processing step may affect the performance of classification model. Thus, in this work, experiments in clinical and microarray data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

pre-processing are performed, such as KNN-impute and dimensionality reduction experiments. Moreover, experiments in data harmonization are also performed.

Experiment results show that classification accuracy increases when PCA is applied on microarray data on pre-processing step, outperforms manifold learning (Isometric Projection). It shows that reducing data to linear field can reduce complexity of data without losing a lot of information.

4. CONCLUSION AND FUTURE WORK

This research performs classification and data integration utilizing Deep Belief Network and Bayesian Network methods to predict cancer prognosis in patients, such as Overall Survival (OS) and Disease-Free Survival (DFS). Clinical and microarray gene expression data are used in this study. There are three main steps in this research, they are data preprocessing, Deep Belief Network (DBN) model construction, and Bayesian Network (BN) integration model construction. In microarray data preprocessing, data imputation method is applied to estimate missing values using KNN-impute, then dimensionality reduction technique is performed using Principal Component Analysis (PCA). In clinical data preprocessing, data is converted to numeric categorical, then missing values imputation is applied using KNN-impute and finally data is converted to binary. Moreover, data harmonization is performed to combine two sources of clinical data. In the second step, two DBN models are constructed, which are Clinical DBN and Microarray DBN. Various DBN architecture and hyperparameter experiments are conducted to obtain the best configuration for both DBNs. Outputs of Clinical DBN and Microarray DBN are combined in the third step, which is BN integration model construction. Conclusions of this study are:

1. The best dimensionality reduction obtained in this research is Principal Component Analysis. Based on classification accuracies, performance of Gene-Shaving competes with PCA, while Isometric Projection is good at predicting unbalanced attribute (disease-free survival). However, it cannot predict balanced attribute (overall survival).
2. Classification performance of proposed DBN-BN method competes with Clinical DBN and better than Microarray DBN, with average accuracy 73.3535% for overall survival prediction and 71.3434% for disease-free survival prediction.
3. Classification performance of proposed DBN-BN outperforms Support Vector Machine (SVM) and k-Nearest-Neighbor (k-NN).
4. The weakness of DBN-BN integration method is longer execution time (compared to SVM and k-NN).
5. Clinical data which resulted by harmonization of MET-clin and TCGA data can be used to predict cancer prognosis. Data harmonization can increase classification performance because it gains number of samples. However, the challenge in data harmonization is availability of clinical attribute/information in some sources of data.

There are some limitations of this study. First, breast cancer patient dataset is used in this study, thus the models may not be generalized to other cancer types. Second, this study only uses two public datasets which consist of 3076 samples. Third, this study only considers clinical and gene expression attributes of cancer patients.

For future research, various microarray data preprocessing methods may be applied. Other research direction is to use more varied patient data (not only clinical and gene expression data), more patient records. In the future research, this method also can be applied to other cancer types, diseases, or in non-medical fields.

5. ACKNOWLEDGMENT

This research was funded by PITTA grant from Universitas Indonesia Number 407/UN2.R3.1/HKP.05.00/2017.

REFERENCES:

[1] WHO. “Fact Sheet: Noncommunicable Diseases”, http://www.who.int/mediacentre/factsheets/fs355/en/, 17/10/2016.

[2] Khademi, M., Nedialkov N.S. “Probabilistic Graphical Models and Deep Belief Networks for Prognosis of Breast Cancer”, *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2015.

[3] Fakoor, R., Ladhak, F., Nazi, A., Huber, M. “Using Deep Learning to Enhance Cancer Diagnosis and Classification”, *Proceedings of*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the 30th International Conference on Machine Learning, JMLR: WCP, Vol. 28, Atlanta, Georgia, USA, 2013.

[4] Abdel-Zaher, A. M., Eldeib, A. M. "Breast Cancer Classification Using Deep Belief Networks", *Expert Systems with Applications*, Vol. 46, 2016, pp. 139-144.

[5] Donoghue, J. O., Roantree, M., Boxtel, M. V. "A Configurable Deep Network for High-Dimensional Clinical Trial Data", *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2015.

[6] Li, H., Li, X., Ramanathan, M., Zhang, A. "Identifying Informative Risk Factors and Predicting Bone Disease Progression via Deep Belief Networks", *Methods*, Vol. 69, No. 3, 2014, pp. 257-265.

[7] Liang, M., Li, Z., Chen, T., Zeng, J. "Integrative Data Analysis of Multiplatform Cancer Data with a Multimodal Deep Learning Approach", *IEEE/ACM TRANSACTIONS on Computational Biology and Bioinformatics*, Vol. 12, No. 4, 2014, pp. 928-937/

[8] Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., Moor, B. D. "Predicting the Prognosis of Breast Cancer by Integrating Clinical and Microarray Data with Bayesian Networks", *Bioinformatics Oxford Journals*, Vol. 22, No. 14, 2006, pp. 184-190.

[9] Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., Beyene, J. "Data Integration in Genetics And Genomics: Methods and Challenges", 2009.

[10] Spjuth, O., Krestyaninova, M., Hastings, J., Shen, H.-Y., Heikkinen, J., Waldenberger, M., Langhammer, A., Ladenvall, C., Esko, T., Ã˘ Eke Persson, M., Heggland, J., Dietrich, J., Ose, S., Gieger, C., Ried, J. S., Peters, A., Fortier, I., de Geus, E. J., Klovins, J., Zaharenko, L., Willemsen, G., Hottenga, J.-J., Litton, J.-E., Karvanen, J., Boomsma, D. I., Groop, L., Rung, J., Palmgren, J., Pedersen, N. L., McCarthy, M. I., van Duijn, C. M., Hveem, K., Metspalu, A., Ripatti, S., Prokopenko, I., Harris, J. R. "Harmonising and Linking Biomedical and Clinical Data Across Disparate Data Archives to Enable Integrative Cross-Biobank Research", *European Journal of Human Genetics*, Vol. 24, 2016, pp. 521–528.

[11] Hinton, G. E., Osindero, S., Teh, Y.-W. "A Fast Learning Algorithm for Deep Belief Nets", *Neural Computation*, Vol. 18, No.7, 2006, pp: 1527–1554.

[12] Keyvanrad, M. A., Homayounpour, M. M. A., "Brief Survey on Deep Belief Networks and Introducing a New Object Oriented MATLAB Toolbox (Deebnet)", *CoRR*, abs/1408.3264, 2014.

[13] Murphy, K. P., "The Bayes Net Toolbox for Matlab", *Computing Science Statistics*, Vol. 33, No. 2, 2001, pp: 1024-1034.

[14] Wasito, I., Mirkin, B., "Nearest Neighbours in Least-Squares Data Imputation Algorithms with Different Missing Patterns", *Computational Statistic Data Analysis*, Vol. 50, 2006, pp: 926-949.