ISSN: 0067-2904

Deep Belief Network for Predicting the Predisposition to Lung Cancer in TP53 Gene

Zahraa Naser Shahweli
Al- Nahrain University, Baghdad, Iraq

Received: 5/4/ 2019 Accepted: 18/ 8/2019

Abstract

Lung cancer, similar to other cancer types, results from genetic changes. However, it is considered as more threatening due to the spread of the smoking habit, a major risk factor of the disease. Scientists have been collecting and analyzing the biological data for a long time, in attempts to find methods to predict cancer before it occurs. Analysis of these data requires the use of artificial intelligence algorithms and neural network approaches. In this paper, one of the deep neural networks was used, that is the enhancer Deep Belief Network (DBN), which is constructed from two Restricted Boltzmann Machines (RBM). The visible nodes for the first RBM are 13 nodes and 8 nodes in each hidden layer for the two RBMs. The enhancer DBN was trained by Back Propagation Neural Network (BPNN), where the data sets were divided into 6 folds, each is split into three partitions representing the training, validation, and testing. It is worthy to note that the proposed enhancer DBN predicted lung cancer in an acceptable manner, with an average F-measure value of 0. 96 and an average Matthews Correlation Coefficient (MCC) value of 0. 47 for 6 folds.

Keywords: Deep Belief Network, Lung Cancer, TP53 gene, RBM, Neural Network.

التنبؤ بالاستعداد لسرطان الرئة في الجين القامع للاورام53 بأستخدام شبكة المعتقد العميق

زهراء ناصر شاه ولي
جامعة النهرين، بغداد، العراق

الخلاصة

سرطان الرئة كغيره من انواع السرطان الاخرى ينتج عن التغير الجيني لكنه يعتبر أكثر خطورة نتيجة انتشار التدخين. حاول العلماء والباحثين ولفترة طويلة جمع وتحليل البيانات البيولوجية كمحاولة لايجاد طرق التنبؤ بالسرطان قبل حدوثه. تحليل هذه البيانات يتطلب استخدام خوارزميات الذكاء الاصطناعي والشبكات العصبية. في هذا البحث تم استخدام إحدى انواع الشبكات العصبية العميقة وهي Enhancer DBN التي بنيت من اثنتينRBM. حيث تتكون العقد المرئية RBM من 13 عقدة بينما هناك 8 عقد لكل طبقة مخفية في كل من RBM الاثنتين. تم تدريب Enhancer DBNبأستخدام خوارزمية BPNN حيث قسمت البيانات إلى 6 أقسام وجزء كل قسم إلى 3 أجزاء هي التدريب، التحقق من الصحة والاختبار. ومن الجدير بالقول ان الخوارزمية المقترحة تنبأت بسرطان الرئة بنسبة مقبولة وصلت إلى 0.96 لمقياس F و0.47 لمعامل ارتباط ماثيو.

Email: stcs-zns16@sc.nahrainuniv.edu.iq

1. Introduction

Deep l earning is a type of computational neural network based on an enormous representation of given data from multiple layers. Deep learning is consisted of many architectures, each with its own structure [1]. One of the most notable developments in the recent decades is the wide use of deep learning in the bioinformatic field to transform colossal valuable data into invaluable knowledge [2]. One of the types of deep learning is the Deep Belief Network (DBN) which depends on the Restricted Boltzmann Machine (RBM). RBM consists of ``visible`` or input layers and ``hidden`` or output layers. The visible unit consists of visible units and multi hidden units and there is a bias in each unit [3].

Today, machine learning techniques such as neural networks, support vector machines, decision trees, and deep learning are used to improve predicting and classifying cancer types, such as breast, prostate, lung and other cancers [4]. Lung cancer is the most killing cancer in the world. It accounts for about 1 in 4 cancer deaths each year and it is more increasing among smokers than others. Nevertheless, smoking tobacco, especially in older people, causes the disabling of the TP53 tumor suppressor gene programming when there is a mutant or deleted codon. The prediction of these mutations that occur in the TP53 gene facilitates the classification, diagnosis, and development of treatment for cancer [5].

2. Related Works

Many researchers used AI algorithms for diagnosing cancer through many approaches, including image analysis or gene expression data. A previous study [6] used the data of gene expression to detect one cancer from other types of cancer. The authors proposed splitting the features regarding learning into two phases; the first phase is Principal Component Analysis (PCA-based, which is used to reduce the feature dimensionality. The second phase develops sparse encoding of data, followed by the use of unsupervised feature regarding learning with a labeled data for specific cancer type to learn a classifier.

Another investigation [7] compared three approaches (Stacked DenoisingAutoencoder, SDA;Convolutional Neural Network, CNN; and Deep Belief Networks, DBN) on the same data set, where the DBN showed the best accuracy. The data set used was collected from seven academic centers and eight medical imaging companies in the Lung Image Database Consortium (LIDC) database.

Authors of another publication [8] trained 3D Computed Tomography (3D-CT) scan images at U-net for nodule candidate detection after performing a preprocessing step on the scan images using segmentation, normalization, down sampling, and zero-centering. They finally classified the scanned images into positive or negative by the 3D CNN. The Kaggle Data Science Bowl 2017 (KDSB17) data set was used in that paper for learning.

Another published investigation [9] proposed three stages of work; the first stage was the preprocessing of the raw Computed Tomography (CT) scan image. The second stage was the pre-training of the nodule classifier for external data sets. Finally, nodules' features were fed into the lung cancer classifier and the output determined if patients have lung cancer or not.

Another article [10] used a high-level medical image representation for learning into new deep learning architectures. TheKDSB17wasthe data set used in order to be classified by the deep convolutional neural network.

3. Lung Cancer and TP53 Gene

In 1979, TP53 gene was described as an oncogene, but after 10 years, it was proved as a tumor suppressor gene [11]. Significantly, most human cancer types are related to mutations in TP53, as these mutations lead to the loss of TP53 role and the control of oncogenes and cancer cell activity [12]. The occurrence of mutations in the TP53 is strongly affected by tobacco smoking. Remarkably, exposure to tobacco whether by smoking or inhalation exposes the lungs to cancer. In all cases, this occurs as a result of mutations that cause the weakness or loss of certain genes for their function, especially theTP53 tumor suppressor gene [13]. In general, lung cancer is classified into two main types:Non – Small Cell Lung Cancer (NSCLC) which consists of several types and occurs in 50% of TP53 mutations, and Small Cell Lung Cancer (SCLC) which occurs in 70% of TP53 mutations [14]. Figure-1 contains information about the leading site of new cancer cases and deaths, depending on the date of the American cancer society 2018.

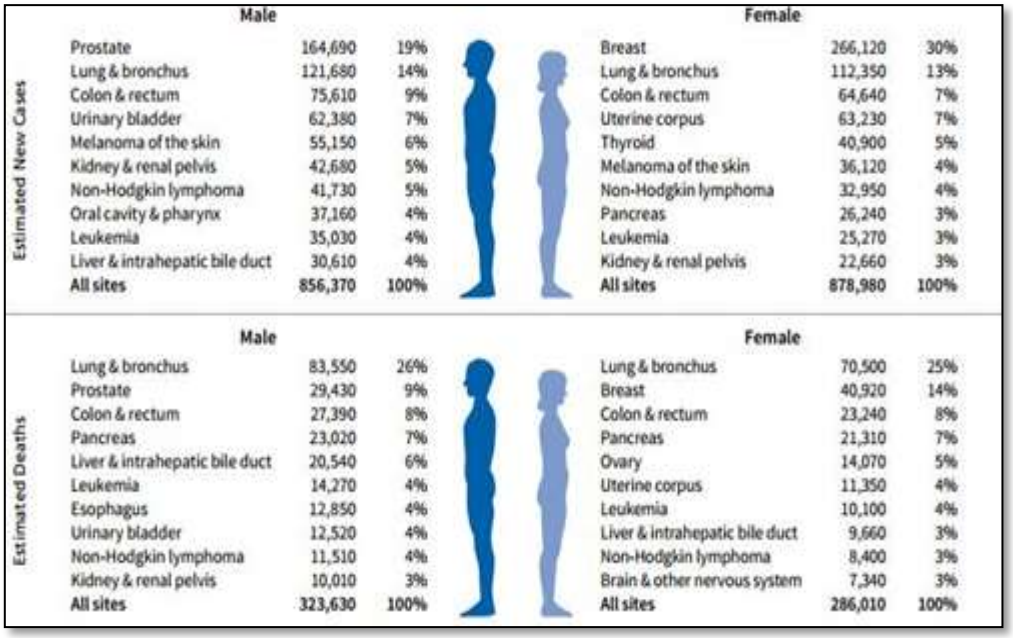

Figure 1- Leading site of new cancer cases and death- 2018 estimates [15].

4. Deep Belief Network (DBN) and Restricted Boltzmann Machine (RBM)

The DBN is an unsupervised machine learning algorithm that has been introduced by Professor Geoffrey Hinton. It consists of two types of neural networks: Deep Belief Network and Restricted Boltzmann Machine (RBM) [16]. RBM is a structure of multi deep neural networks; it is consisted of undirected connection between stochastic units in its network, which have pairs of visible and hidden units [17]. The visible unit represents the first layer, which is corresponding with samples that has been trained, while the hidden layer is corresponding with the feature of the input. All visible and hidden neurons are connected to the bias unit. In RBM, all neurons are connected to all neurons in other layers. On the contrary, there are no inner connections inside the same layer, and this restriction gives the RBM its name [18]. RBMs is used in the learning of DBN when the feature activation is produced in one RBM, which acts as a data for the training of the next RBM. DBN provides an efficient learning technique [19].

5. Data Set

Dr. Curt Harris and his cooperators established a database of TP53 mutations to simplify the analysis and utilize these mutations in scientific research [20]. This database was established in 1991 and since then the International Agency for Research on Cancer database has been continuously maintained [21]. The latest version of this database is R19, August 2018, which contains somatic and germ line mutations related to all types of cancer, in addition to other information about mutations. IARC database [22] is available free to all researchers on: (http://p53.iarc.fr/) website.

6. Material and Methods

The International Agency for Research on Cancer TP53 (IARC TP53) database of somatic mutation is used in this paper for learning the DBN.It consists of 3174 samples of mutant and normal cases. Out of69 IARC TP53selected features, 13 features were used to learn neural network. All selected features were converted to numerical forms to facilitate the learning for neural work.

Features with high values were normalized with the min/max normalized equation to harmonize among all features.

The DBN is constructed by training a number of RBMs and additional output layers. Then, the enhancer DBN is optimized by training the DBN by Back Propagation Neural Network(BPNN).The Enhancer DBN classifier is constructed of two steps:

First: the unsupervised way in the two RBMs.

Second: the supervised way, when the DBN is trained with the output layer by BPNN.

The proposed DBN consists of 2RBMs, each having its visible layer (v) and hidden layer (h).The nodes in the visible layer for the first RBM are 13, whereas the hidden layers for the two RBMs are 8

nodes. The weights (w) represent the interaction between the visible nodes and hidden nodes. The energy function can be expressed as:

$$E(v,h;\ \theta) = -\sum_{ij} Wij\ vij\ hj - \sum bi\ vi - \sum aj\ hj \quad 1$$

Where $\theta = \{W, a, b\}$ represents the model parameter, a_i is the bias of a visible node i, and b_i is the bias of a hidden node j. The probability distribution of the model is expressed as:

$$P_\theta\ (\mathrm{v,h}) = \frac{1}{Z(\theta)} \exp\big(-E(v,h;\theta)\big) \quad 2$$

$$Z(\theta) = \sum_{h,v} \exp(-E(v,h;\theta)) \quad 3$$

Where $Z(\theta)$ is the normalization constant.

6.1 Training the Enhancer DBN

The DBN is built with weights and bias as learned in two RBMs.The training data are loaded into the 1st RBM as a visible layer, and the data trained in its hidden layer are inserted into the 2nd RBM as a visible layer. Finally, the data trained in the2nd RBM are in a hidden layer before going to the output layer.

The RBM is trained when

1. The visible layer is set according to the training data.
2. Calculating the binary state of the hidden nodes by using the sigmoid activation function with bias *a* according to equation 4:

$$P(h_j=1 \mid v) = \mathrm{sigmoid}(\textstyle\sum_i Wij\ vi + aj) \quad 4$$

3. The state of visible nodes are calculated also by sigmoid activation function with bias *b* according to equation 5.

$$P(v_i=1 \mid h) = \mathrm{sigmoid}(\textstyle\sum_j Wij\ hj + bi) \quad 5$$

4. The weights, bias *a* and bias *b* are updated by using the gradient descent algorithm, also the gradients of weight are evaluated by contrastive divergence (CD) learning algorithm.

After constructing the enhancer DBN with two RBMs, the enhancer DBN can classify the training dataset using PBNN. The 6 fold cross validation was performed positively, where each fold is divided into three partitions (Training, validation, and testing data) as in previous work [23].

The 3174 samples are used to train the enhancer DBN that is partitioned into 6 folds. The 529 samples for each fold are divided into 60%, 20%, and 20% for training, validation, and testing,respectively.Figure 2 explains the structure of the proposed DBN, while figure 3 demonstrates the block diagram of the proposed DBN classifier.

6.2 Validating and Testing the Enhancer DBN

The performance measures that are used to evaluate the work are the F-measure, MCC and accuracy, where the equation of the F-measure is:

$$F=\frac{2*PRC*SN}{(PRC+SN)} \quad 6$$

Pre (precision)is the number of true positive over the number of true positive plus the number of false positive

$$\mathrm{Pre} = \frac{TP}{TP+Fp} \quad 7$$

Sn(Sensitivity)is the number of true positive over the number of true positive plus the number of false negative

$$SN=\frac{TP}{TP+FN} \quad 8$$

While the equation of MCC depends on all basic terms in the confusion matrix:

$$\mathrm{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad 9$$

Also, the equation of accuracy extent derives the quantity of accurately recognized cases in the aggregate number of test cases:

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} 10$$

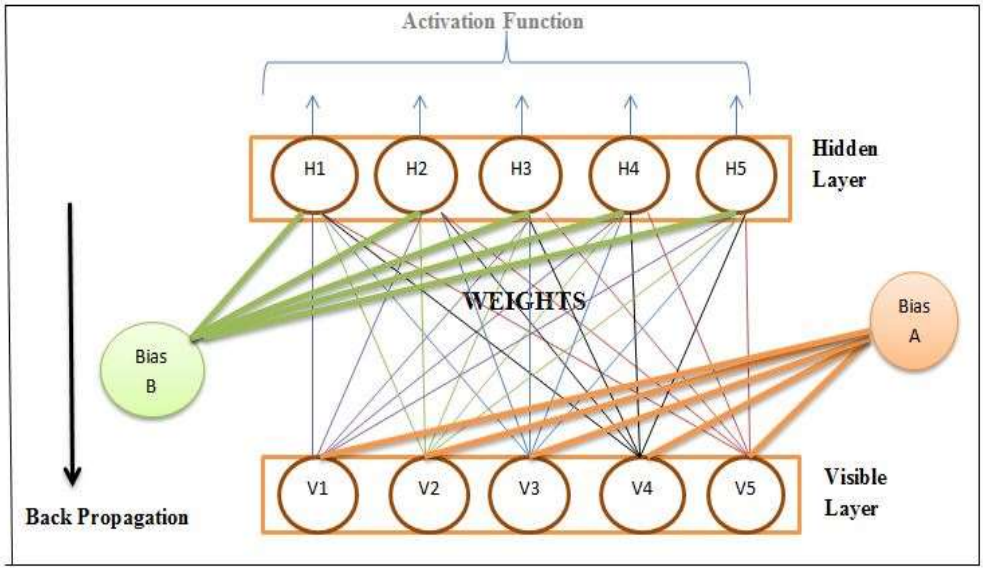

Figure 2- The structure of proposed DBN.

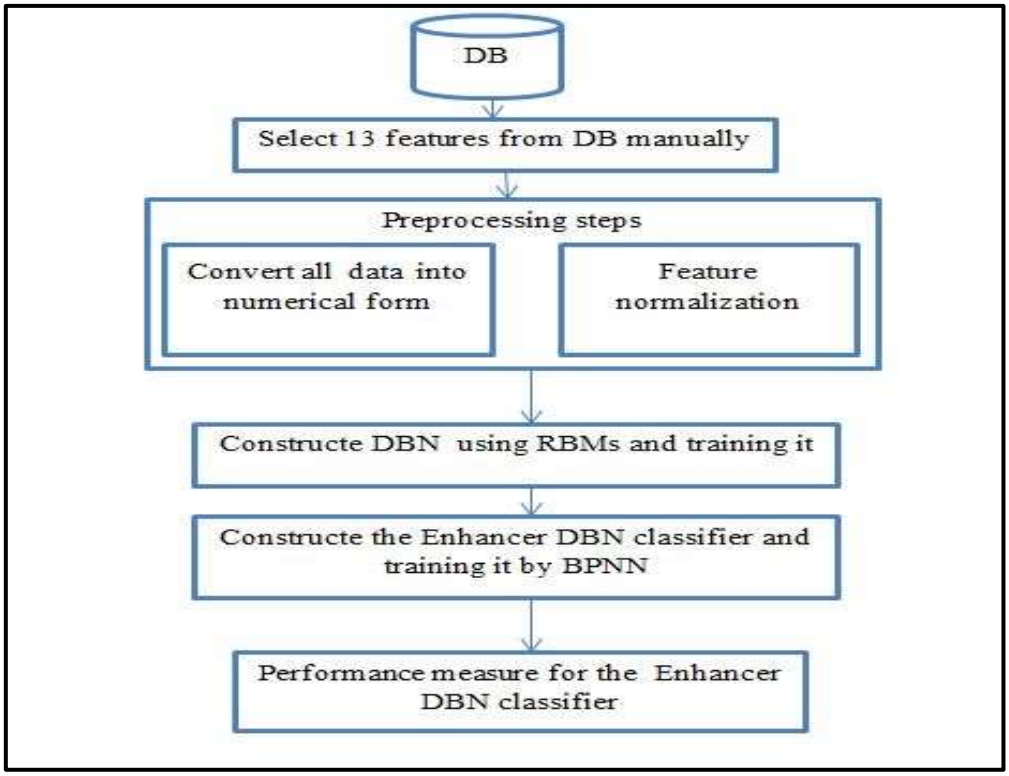

Figure 3- The block diagram of the proposed DBN classifier.

7. Results and Discussion

The F – measure or F 1 score has been used in this paper to evaluate the results of the enhancer DBN. Moreover, MCC was measured due to the symmetric significance of both cancer and non – cancerous cases. Also, it is used when a dataset is unbalanced. Table-1 contains the values of F-measure and MCC for each fold.

Table 1-The performance measures for 6 folds of the proposed method.

Performance measure	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6
F- measure	0.97	0.97	0.98	0.97	0.98	0.97
MCC	0.39	0.46	0.54	0.51	0.42	0.53

The accuracy and error values for the enhancer DBNwere0.96and0.03, respectively.
The average of the F – measure equals 0.97 while the average of the accuracy equals 0.96 for 6 folds. This proves that the proposed system has the ability to classify both cases(cancer and non – cancerous) because the accuracy measured the correct cases that the classifier can identify, but the F – measure depends on the precision and recall, so that it measured all cases, whether correct or not correct.
In comparison, with F – measure the average of MCC is 0.47, due to the few cases of normal cases comparing with the many cases of patients in the database.
When comparing the proposed DBN with a previously published work [10], we found that the effectiveness of the proposed DBN in this paper was higher (Table-2).

Table 2-Acomparison between the proposed DBN and a previously published work [10].

	Preprocessing phase	**Data set used**	**Network used**	**F- measure**
Reference [10]	No	KDSB17	CNN	0.95
Proposed DBN	Yes	IARC TP53	DBN with RBM	0.96

8. Conclusion

Due to the great increase of biological data generated from high throughput studies, the neural network approach and deep learning were widely used to conclude valuable information from these data. This information is helpful in predicting the diseases in advance, as well as in classifying and analyzing the reasons for these diseases.

In this paper, an enhancer DBN classifier was proposed, that is related to the unsupervised phase using 2 RBMs and a supervised phase when the DBN is trained by BPNN. The RBM contains 13 nodes for the visible layer and 8 nodes for each hidden layer. The BPNN training had 6 folds with cross validation, where each fold was divided into training, validation, and testing samples representing60%, 20%, and 20%, respectively.

The proposed enhancer DBN identified lung cancer with an accuracy that reached to 96% and an F-measure that reached to 97% as an average for 6 folds. However, the used data were unbalanced, containing 2672 cases of lung cancer and 502 normal cases. Dividing the data into 6 folds, then dividing each fold into 3 parts, facilitated the learning for BPNN and, therefore, the MCC reachs to a value of 0.47 as an average for 6 folds.

Future applications can use DBN to classify all types of cancer, depending on the mutations in the IARC database, and then infer the types of mutations that cause more than one cancer.

References

1. Deng, L. and Yu, D. **2014.** Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing, now the essence of knowledge. DOI: 10.1561/2000000039.
2. Min, S., Lee, B. and Yoon, S. 2017. Deep learning in bioinformatics. *Briefings in bioinformatics.* 18(5), pp: 851-869.
   URL: https://bit.ly/2UcmFye.
3. Li, C., Ding, Z., Yi, J., Lv, Y. and Zhang, G. **2018.** Deep belief network based hybrid model for building energy consumption prediction. *Energies.* 11(1),pp: 242.
   URL: https:doi.org/10.3390/en11010242.
4. Kourou, K.,Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis. **2015**. *Machine learning applications in cancer prognosis and prediction.* 13, pp: 8-17.
   URL: https://doi.org/10.1016/j.csbj.2014.11.005.
5. Campling, B. G. and El-Deiry, W. S. **2003.** Clinical implications of p53 mutations in lung cancer. *Lung Cancer*. Springer.
   URL:https://doi.org/10.1385/MB:24:2:141.
6. Fakoor, R. Ladhak, F. Nazi, A. and Huber, M. **2013.** Using deep learning to enhance cancer diagnosis and classification. Proceedings of the 30th International Conference on Machine Learning. WHEALTH workshop. Atlanta, Georgia, USA.
   URL: https://bit.ly/2K5LI15.

7. Sun, W., Zheng, B. and Qian, W. **2016.** Computer aided lung cancer diagnosis with deep learning algorithms. SPIE Medical Imaging: computer-aided diagnosis. San Diego,California,United States.
DOI:10.1117/12.2216307.

8. Chon, A., Balachandar, N. and Lu, P. **2017**. Deep convolutional neural networks for lung cancer detection. Tech. rep., Stanford University.
URL: https://bit.ly/2OFnY2D.

9. Gyu-tae,P., Sung Jun, S., Myung-ki,L and Joon-Jun, K.**2018.** Lung Cancer Diagnosis Using Deep Convolutional Neural Network. Graduate School of Convergence Science and Technology, Seoul National University.
URL: https://bit.ly/2CPPu8K.

10. Serj, M.F., Lavi, B., Hoff, G. and Valls, D.P. **2018**. A Deep Convolutional Neural Network for Lung Cancer Diagnostic. arXiv.org>cs>cs.CV.
URL: https://arxiv.org/abs/1804.08170.

11. Mogi, A. and Kuwano, H. **2011.** TP53 mutations in nonsmall cell lung cancer. *Journal of Biomedicine and Biotechnology*.
DOI: 10.1155/2011/583929.

12. Muller, P. AJ and Vousden, K. H. **2013**. p53 mutations in cancer. *Nature cell biology*, 15(1), pp: 2-8.
DOI: 10.1038/ncb2641.

13. Gibbons, D. L. Byers L. and Kurie J.M. **2014.** Smoking, p53 mutation, and lung cancer. *Molecular of cancer research*. 12 (1), pp: 3-13.
DOI: 10.1158/1541-7786.MCR-13-0539.

14. Toyooka, S.,Tsuda, T. and Gazdar, A.F. **2003.** The TP53 gene, tobacco exposure, and lung cancer. *Human mutation*, 21(3), pp: 229-239.
URL: https://doi.org/10.1002/humu.10177.

15. American cancer society. 2019. Cancer facts and figures **2019.**
URL: https://bit.ly/2Yh8u9o.

16. Hebbo, H. and Kim, J.W. **2013.** Classification With Deep Belief Networks. CDB publication.
URL: https://bit.ly/2FUyRL3.

17. Montúfar G. 2016. Restricted Boltzmann Machines: Introduction and Review.*Information Geometry and its Applications*,IGAIA IV 2016. Springer Proceedings in Mathematics & Statistics, Springer.
URL: https://doi.org/10.1007/978-3-319-97798-0_4.

18. Fischer, A. and Igel, C. **2014.** Training restricted Boltzmann machines: An introduction. *Pattern Recognition*, 47(1). pp: 25-39.
URL: https://doi.org/10.1016/j.patcog.2013.05.025.

19. Sarikaya, R., Hinton, G. E. and Deoras, A. **2014.** Application of deep belief networks for natural language understanding, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22(4). pp: 778-784.
DOI: 10.1109/TASLP.2014.2303296.

20. Hollstein, M,Bartsch, H. and Wesch, H. **1997.** p53 gene mutation analysis in tumors of patients exposed to alpha-particles. *Carcinogenesis*. 18(3). pp: 511-516.
DOI: 10.1093/carcin/18.3.511.

21. Olivier, M., Eeles R., Hollstein, M., Khan, M.A., Harris, C.C. and Hainaut, P. **2002.** The IARC TP53 database: new online mutation analysis and recommendations to users. *Human mutation*. 19(6). 607-614.
URL: https://doi.org/10.1002/humu.10081.

22. Bouaoun L., Sonkin D., Ardin M., Hollstein M., Byrnes G., Zavadil, J. and Olivier M.TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data.Hum Mutat. 2016 Sep;37(9):865-76.

23. Shahweli, Z.N., Dhannoon, B. N. and Ramadhan, R.S. **2017.** In Silico Molecular Classification of Breast and Prostate Cancers using Back Propagation Neural Network.*Cancer biology*, 7(3), pp: 1-7. DOI:10.7537/marscbj070317.01.