EasyChair Preprint

№ 1608

Prediction of Diabetes Disease Using Data Mining and Deep Learning Techniques

Tharak Roopesh, Asadi Srinivasulu and K.S. Kannan

October 9, 2019

PREDICTION OF DIABETES DISEASE USING DATA MINING AND DEEP LEARNING TECHNIQUES

P.T.S.S Roopesh and Dr. Asadi Srinivasulu and Dr.K.S.Kannan
tharakroopesh@gmail.com, srinu.asadi@gmail.com ,saikannan2012@gmail.com
Data Analytics Research Laboratory, Sree Vidyanikethan Engineering College

Abstract **-** Diabetes Mellitus is one of the growing fatal diseases all over the world. It leads to complications that include heart disease, stroke, and nerve disease, kidney damage. So, Medical Professionals want a reliable prediction system to diagnose Diabetes. To predict the diabetes at earlier stage, different machine learning techniques are useful for examining the data from different sources and valuable knowledge is synopsized. So mining the diabetes data in an efficient way is a crucial concern. In this project, a medical dataset has been accomplished to predict the diabetes. The R-Studio and Pypark software was employed as a statistical computing tool for diagnosing diabetes. The PIMA Indian database was acquired from UCI repository will be used for analysis. The dataset was studied and analyzed to build an effective model that predicts and diagnoses the diabetes disease earlier.

Keywords: *Diabetes, Classification, Clustering, Regression, SVM, K-NN, Neural Networks, CNN, RNN*

1. INTRODUCTION

As we know that the growth in technology helps the computers to produce huge amount of data. Additionally, such advancements and innovations in the medical database management systems generate large volumes of medical data. Healthcare industry contains very large and sensitive data. This data needs to be treated very careful to get benefitted from it.Diabetic Mellitus is a set of associated diseases in which the human body is unable to control the quantity of sugar in the blood. Diabetic Mellitus which results in high sugar levels in blood, may due to the the body not producing sufficient insulin, .The focus is to develop the prediction models by using certain machine learning algorithms. The Machine Learning is an application of artificial intelligence as it helps the computer to learn on its own. The two classification of ML are supervised and unsupervised. The Supervised learning calculation utilizes the past experience to influence expectations on new or inconspicuous information while unsupervised calculations to can draw derivations from datasets.

Machine learning algorithms are:

1.1. Supervised learning techniques:

Classification
The procedure of finding the obscure information of the class name which is utilizing recent known information is called as class mark which is intern called as classification .Popular Classification Algorithms are given below.

i. Random forest
ii. SVM
iii. K-Nearest neighbors
iv. Decision tree
v. Naïve Bayes

Regression

A supervised learning algorithm which is used to find the between the independent variables and with some depent variables is called as regression .

The popular Regression algorithms are:

i. Simple Linear Regression
ii. Multiple Linear Regression
iii. Logistic Regression
iv. Polynomial Regression
v. Linear Discriminant Analysis(LDA)

1.2. Unsupervised Learning techniques:

Clustering

The process which classifies the similar objects into groups called as clustering mechanism.
Some of the clustering techniques are:

i. K-means clustering
ii. Hierarchical clustering

1.3. R studio:

An Integrated Development Environment (IDE) for R programming language which was founded by Jjallaire is called as R Studio. The command line that R Studio uses is interpreter. R studio used for statistical computing and graphics. R Studio is having many built-in packages so it can manipulate huge dataset for analysis.

2. LITERATURE REVIEW

S. No.	Paper	Author(s)	Name of the Journal	Methods	Findings	Notes/Critique
1.	Predicting Diabetes in Medical Datasets Using Machine Learning Techniques	Uswa Ali Zia, Dr. Naeem Khan.	International Journal of Scientific & Engineering Research (IJSER).	Boot strapping resampling technique to enhance the accuracy and then applying i. Naive Bayes, ii.Decision Trees iii.k-Nearest Neighbors (k-NN)	After Bootstrapping Accuracy : i.NaiveBayes- 74.89% ii.Decision Trees- 94.44% iii. k-NN(for k=1) 93.79% 4. k-NN(for k=3) - 76.79%	i.Plan to use further more advanced classifiers such as Neural Networks. ii. It should consider some other important factors that are related to gestational diabetes, like metabolic syndrome, family history, habit of smoking, lazy routines, some dietary patterns etc.
2.	Prediction of Diabetes Using Data Mining Techniques	FikirteGirma, Woldemichael, Sumitra Menaria	International Conference on Trends in Electronics and Informatics (ICOEI)	i. Back Propagation Algorithm ii. J48 Algorithm iii. Naïve Bayes Classifier iv. Support Vector Machine.	Back Propagation Algorithm has Accuracy-83.11% Sensitivity- 86.53% Specificity-76%	i. Increment the accuracy of the algorithms.

3.	Diabetes Disease Prediction Using Data Mining	Dheeraj Shetty,Kishor Rit,Sohail Shaikh,Nikita Patils	International Conference on Innovations in Information, Embedded and Communication Systems(ICIIECS).	i. Naïve Bayes ii. k-NN algorithms	prediction of the disease will be done with the help of Bayesian algorithm and KNN algorithm and analyze them by taking various attributes of diabetes.	i. Increment the accuracy of the algorithms. ii. So Working on some more attributes which is used to tackle the diabetes even more.
4.	Classification of Diabetic Patients by using Efficient Prediction from Big Data using R Studio	K. Sharmila, Dr. S.A. Vetha Manickam	International Journal of Advanced Engineering Research and Science (IJAERS).	Decision tree	i. Using R, the dataset is analyzed and the correlation coefficient for two attributes is calculated. ii. Decision Tree is used to predict the type of Diabetes.	Possibility of developing efficient predictive models using the information from the analysis which is already carried out.
5.	Diagnosis of diabetes using Classification mining Techniques	AiswaryaIyer, S. Jeyalatha and Ronak Sumbaly	International Journal of Data Mining & Knowledge Management Process (IJDKP).	Decision tree Naïve Bayes.	J48 Cross validation-74.8698 % J48 Percentage Split-76.9565 % Naive Bayes-79.5652 %	i.In future the work, planned to be gathering the information from different locales over the world. ii.This work can be improved and extended for the automation of diabetes analysis.
6.	An Disease Diagnosis using Data Mining Techniques And Empirical study.	M. Deepika, Dr. K. Kalaiselvi	The 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT).	i.Artificial Neural Network ii.Decision Tree iii.Logistic Regression iv.Naïve Bayes v.SVM	Artificial Neural Network : 73.23% Logistic Regression :76.13% Decision Tree :77.87%	Efficient and Accurate classifier can be developed.

3. PROPOSED SYSTEM

We propose a classification model with boosted accuracy to predict the diabetic patient. In this model, we have employed different machine learning techniques are using like classification, regression and clustering. The major focus is to increase the accuracy by using resample technique on a benchmark well renowned PIMA diabetes dataset that was acquired from UCI machine learning repository, having eight attributes and one class label. The proposed framework is shown in Figure 1.

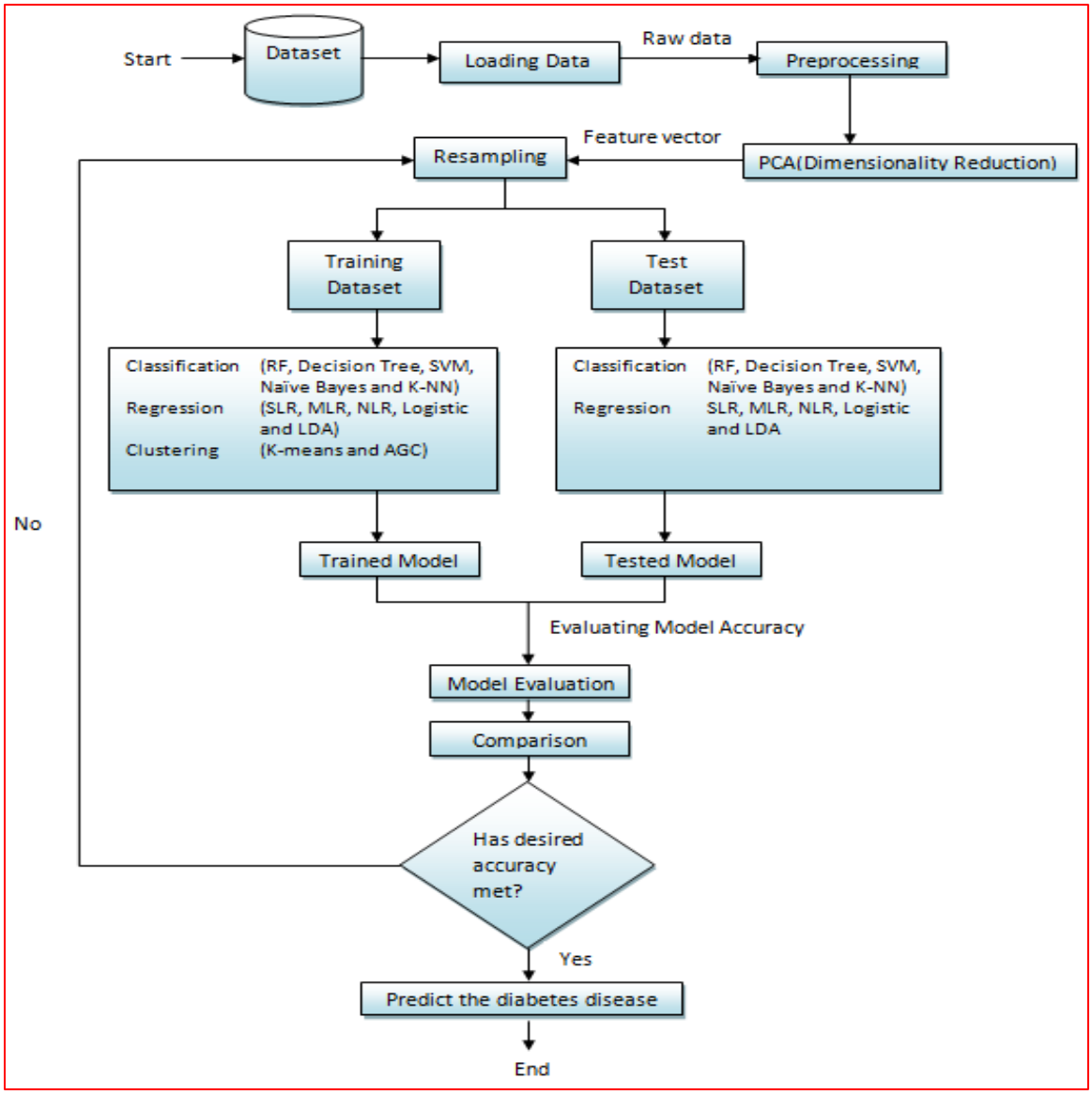

Figure 1.Proposed System for Diabetes Prediction System

The description of the each phase is mentioned below:

3.1 Data Selection:

Data selection is a process in which the most relevant data is selected from a specific domain to derive values that are informative and facilitate learning. PIMA diabetes dataset having 8 attributes that are used to predict the diabetes at earlier stage. This dataset is obtained from UCI repository.

3.2 Data Pre-processing:

Data pre-processing is a Machine Learning technique that includes changing crude information into reasonable configuration. It includes Data Cleaning, Data Integration, Data Transformation, and Data Discretization.

3.3 Feature Extraction through Principle Component Analysis:

Feature Extraction on the dataset to determine the most suitable set of attributes that can help achieve better classification. The set of attribute suggested by the PCA are termed as feature vector. Feature reduction or dimensionality reduction will benefitted us by reducing the computation and space complexity.

3.4 Resampling Filter:

The supervised Resample filter is applied to the pre-processed dataset. Re-sampling is a series of methods used to reconstruct your sample data sets, including training sets and validation sets. In this study, Boot strapping resampling technique to enhance the accuracy.

4. MACHINE LEARNING TECHNIQUES

4.1 Classification:

4.1.1 Random Forest:

An outfit learning technique which is used for regression and classification and such that other tasks which is operated by constructing an multitude of decision trees at the training time and outputting the mode of classes for the individual trees is called Random Forest .

4.1.2 Support Vector Machine (SVM):

A division of supervised learning algorithm which is the strategy used to perform the regression and classification and is used to perform the outlier detection of the data is called as Support Vector machine which is called Supervised Learning Algorithm .In SVM the information that classifies grouping will be dependent on hyper plane.

The types of the Suppourt Vector Machine Classifiers are

1.Linear

2.Non linear classifier

4.1.3 Naïve Bayes:

The algorithm performs classification tasks in the field of ML are called as **Naïve Bayes**. It can perform classification very well on the dataset even it has huge records with multi class and binary class classification problems. The application of Naive Bayes is mainly to text analysis and Natural Language Processing. It works based on conditional probability.

It can be represented as:

$$P(M|N) = \frac{P(M|N)P(M)}{P(N)}$$

Here M and N are two events and, P(M|N) is the conditional probability of M given N.P(M) is the probability of M. P(N) is the probability of N. P (N|M) is the conditional probability of N given M.

4.1.4 k-Nearest Neighbors:

The supervised classifier which is a best choice for K-NN is called as k-Nearest Neighbor. It is a best choice for the classification of k-NN kind of problems. In order to predict the target label of a test data, KNN which finds distance between nearest training data class labels and new test data point in the presence of K value? KNN uses K variable value between 0 to 10 normally.

4.2 Regression:

4.2.1 Simple Linear Regression:

The linear Regression algorithm which explains the relationship between independent and dependent variables to predict the values of the dependent variable is called as Simple Linear Regression algorithm. Simple regression uses one independent variable.

The simple linear regression model is represented as

$$\mathbf{y= (b_0 +b_1x)}$$

Here, x(independent variable) and y (dependant variable) are two factors involved in simple linear regression analysis. Also b_0 is the Y-intercept and b_1 is the Slope.

4.2.2 Multiple Linear Regressions:

It explains the relationship between two or more independent variables and a dependent variable to predict the values of the dependent variable. It uses two or more independent variables. Dependent variable has a continuous and independent variable has discrete or continuous values.

The multiple linear regression model is represented as

$$\mathbf{y= (p_0 +p_1x_1+p_2x_2+\ldots+p_nx_n)}$$

Here $x_1, x_2,..., x_n$ (independent variable) and y (dependant variable) are two factors involved in multiple linear regression analysis. Also b_0 is the y-intercept and $p_1, p_2,\ldots, p_n$ is the slope.

4.2.3 Logistic Regression:

The predictive analysis which is used for the dependent variable is categorical called as Logistical Regression. Logistical Regression explains the relationship between one dependent variable and one or more independent variables.

The various types of Logistic Regression are:

i. Multinomial Logistic Regression(many)
ii. Binary Logistic Regression(two)
iii. Ordinal Logistic Regression(1)

The categorical response has only two possible outcomes. Multinomial Logistic Regression has three or more outcomes without ordering whereas Ordinal Logistic Regression has three or more outcomes with ordering.

4.2.4 Polynomial Regression:

The form of regression analysis which explains the relationship between the independent variable and dependent variable as an nth degree polynomial is called as polynomial regression. It fits a non-linear relationship between the value of independent variable and conditional mean of dependent variable. It is represented as

$$x = a + b * y \wedge n$$

Here p is Dependent Variable, q is Independent Variable and n is Degree.

It is used to fit the data very well when the data is below and above the regression model. It minimizes the cost function and provides optimum result on the regression.

4.2.5 Linear Discriminant Analysis:

The process of using various data items and applying different functions to that set to analyze classes of objects or items separately is called Linear Discriminant Analysis. Image Recognition and Predictive analytics use this Linear Discriminant Analysis

4.3 Clustering:

4.3.1 K-means Clustering:

. The unsupervised machine learning algorithm which is used to solve clustering problems by classifying the dataset into a number of clusters k (group of similar objects),which defines the number of clusters which is assumed before classifying the dataset.

4.3.2 Hierarchical Clustering:

The type of clustering algorithm which is used to build a hierarchy of clusters is called hierarchical clustering

The two types of Hierarchical Clustering are:

Agglomerative Clustering

It is used to group objects into clusters based on their similarity. The result obtained at last is a tree representation of objects called Dendrogram.

Divisive Analysis

This is a best down methodology where all perceptions begin in one bunch, and parts are performed recursively as one moves down the pecking order. A hierarchical clustering is often represented as a dendrogram. The each cluster will be representing with centroids. Distance will be calculated by using linkage.

5. RESULTS AND ANALYSIS

Indian diabetes dataset named PIMA were used for analysis for this study. It consists of eight independent attributes and one independent class attribute. The study was implemented by R programming language using R Studio. Machine learning algorithms like classification (Decision Tree, Naïve Bayes, k-NN and Random Forest), regression (linear, multiple, logistic, LDA) and clustering (k-means, hierarchical agglomerative) are used to predict the diabetics disease in early stages. Measure Performance model by using accuracy.

Table1: Predictive analysis of machine learning algorithms

S.No	Algorithm	Accuracy
1	Random forest	83%
2.	Decision tree	77%
3.	SVM	92%
4.	Naïve Bayes	86%

5.	K-NN	91%
6.	Simple linear regression	98%
7.	Logistic regression	88%
8.	LDA	88%
9.	k-Means	81%
10.	Hierarchical agglomerative	74%

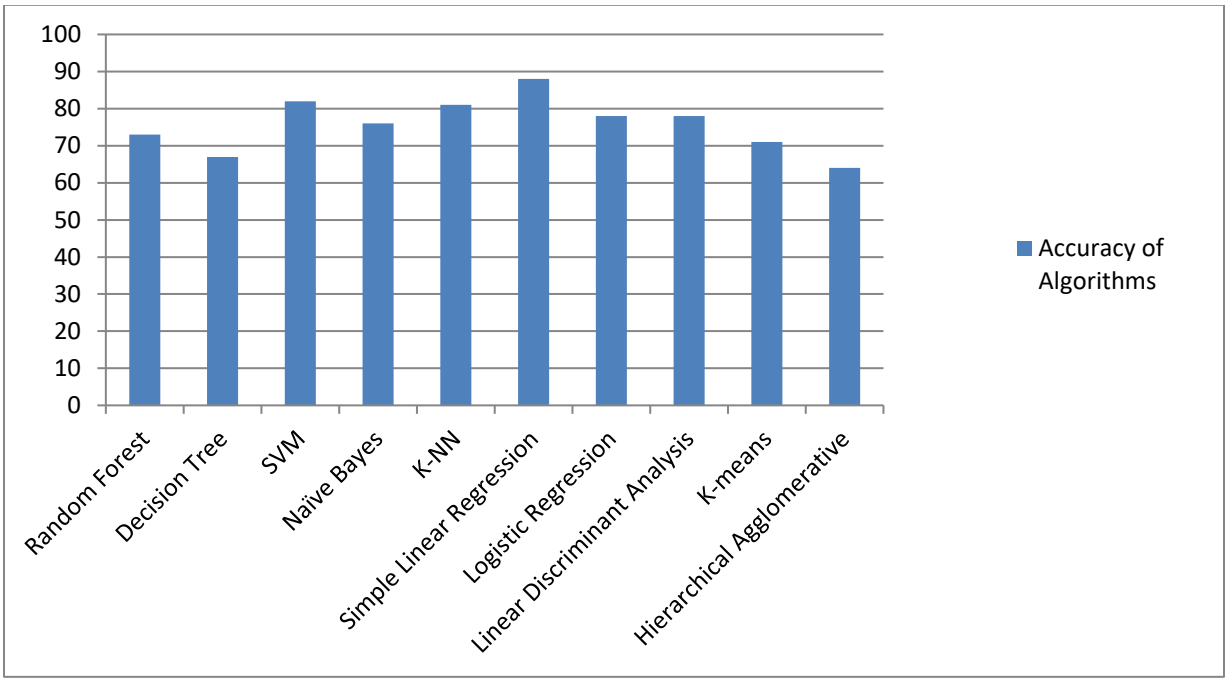

Figure 2. Comparison of accuracy of various algorithms

6. CONCLUSION AND FUTURE WORK

Deep Learning and Data mining plays an important role in various fields such as Artificial Intelligence (AI) and Machine Learning (ML), Database Systems and more. The core objective is to enhance the accuracy of predictive model**.** This PIMA dataset will increases the accuracy of almost all algorithms but the SVM and linear regression leads over others. In future many advanced deep learning techniques will be used to increasing the accuracy of the algorithms.

REFERENCES

[1] Usma Niyaz et al, “Advances in Deep Learning Techniques for Medical Image Analysis” 5th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC-2018), 978-1-5386-6026-3/18/$31©2018 IEEE, 20-22 Dec, 2018, Solan, India.

[2] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-Based Learning Applied to Document Recognition,” Proceedings of IEEE, vol. 86, pp. 2278-2324.

[3] Leslie N. Smith, “Cyclical Learning Rates for Training Neural Networks,” 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 464–472, 2017.

[4] [5] J.Wang, L.Perez, “The Effectiveness of Data Augmentation in Image Classification using Deep Learning,” Computing Research Repository (CoRR) - arXiv, 2017.

[5] N. Srivastava, G. Hinton, A. Krizhevsky and I. Sutskever and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research (JMLR) vol.15, pp. 19291958, 2014.

[6] M. S. Al-tarawneh, “Lung Cancer Detection Using Image Processing Techniques,” Leonardo Electronic Journal of Practices and Technologies (LEJPT) no. 20, pp. 147–158, 2012.

[7] Abhishek S. Sambyal, Asha T., “Knowledge Abstraction from Textural Features of Brain MRI Images for Diagnosing Brain Tumor using Statistical Techniques and Associative Classification,” 2016 International Conference on Systems in Medicine and Biology, IIT Kharagpur.

[8] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” NIPS’12 Proceedings of the 25th International Conference on Neural Information Processing Systems, vol. 1, pp. 1097–1105.

[9] K. Simonyan and A. Zisserman, “Very Deep Convolutional Network for Large-Scale Image Recognition,” International Conference on Learning Representations (ICLR) 2015.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, “Going deeper with convolutions,”2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–9, 2015.

[11] K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770-778, 2016.

[12] O. Ronneberger, P. Fischer, and T. Brox,“U-Net: Convolutional Networks for Biomedical Image Segmentation,” MICCAI 2015, pp. 234?241, 2015.

[13] F. Milletari, N. Navab, and S. A. Ahmadi,“V-Net: Fully Convolutional Networks for Volumetric Medical Image Segmentation,” 2016 Fourth International Conference on 3D Vision, pp.565-571, 2016.

[14] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang Z. Eaton- Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat D. C. Barratt, S. Ourselin, M. Jorge Cardoso and T. Vercauteren, “NiftyNet: a deep-learning platform for medical imaging,” Computer Methods and Programs in Biomedicine vol. 158, pp. 113 - 122, 2018.

[15] S. Jgou, M. Drozdzal, D. Vzquez, A. Romero, and Y. Bengio,“The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation,” arXiv:1611.09326v2 [cs.CV],2016.

[16] W. Chen, Y. Zhang, J. He, Y. Qiao, Y. Chen, H. Shi, X. Tang, “W-net: Bridged U-net for 2D Medical Image Segmentation,” arXiv:1807.04459v1 [cs.CV] 12 Jul 2018.

[17] G. Yang, H. Jing, “Multiple Convolutional Neural Network for Feature Extraction,” International Conference on Intelligent Computing (ICIC) 2015.

[18] J. Ding, A. Li, Z. Hu, and L. Wang, “Accurate Pulmonary Nodule Detection in Computed Tomography Images Using Deep Convolutional Neural Networks,” Computing Research Repository (CoRR) - arXiv, 2017.

[19] Q. Song, L. Zhao, X. Luo, and X. Dou, “Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images,” Journal of Healthcare Engineering, vol. 2017, 2017.

[21] H. Chougrad, H. Zouaki, O. Alheyane “Convolutional Neural Networks for Breast Cancer Screening: Transfer Learning with Exponential Decay,”- arXiv, 2017.

[22] A. Esteva, B. Kuprel, R. A. Novoa, J.Ko, S. M. Swetter, H. M. Blau& S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” Nature 542, pp. 115–118, 2017.

[23] E. Sert, S. Ertekin, U. Halici, “Ensemble of Convolutional Neural Networks for Classification of Breast Microcalcification from Mammograms,” 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 689–692, 2017.

[24] N. C. F. Codella, Q. B. Nguyen, S. Pankanti, D. Gutman, B. Helba, A. Halpern, J. R. Smith, “Deep learning ensembles for melanoma recognition in dermoscopy images,” Computing Research Repository (CoRR), vol. abs/1610.04662, 2016.

[25] K. J. Geras, S.Wolfson, Y. Shen, S. Gene Kim, L. Moy, K. Cho, “High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks,” Computing Research Repository (CoRR) - arXiv, 2017.