**International Journal of E-Health and Medical Communications**
Volume 12 • Issue 1 • January-March 2021

# Rummage of Machine Learning Algorithms in Cancer Diagnosis

Prashant Johri, SCSE, Galgotia's University, Greater Noida, India
https://orcid.org/0000-0001-8771-5700

Vivek sen Saxena, INMANTEC Institutions, Ghaziabad, India
https://orcid.org/0000-0001-6107-5506

Avneesh Kumar, SCSE, Galgotia's University, Greater Noida, India
https://orcid.org/0000-0001-5860-3689

## ABSTRACT

With the continuous improvement of digital imaging technology and rapid increase in the use of digital medical records in last decade, artificial intelligence has provided various techniques to analyze these data. Machine learning, a subset of artificial intelligence techniques, provides the ability to learn from past and present and to predict the future on the basis of data. Various AI-enabled support systems are designed by using machine learning algorithms in order to optimize and computerize the process of clinical decision making and to bring about a massive archetype change in the healthcare sector such as timely identification, revealing and treatment of disease, as well as outcome prediction. Machine learning algorithms are implemented in the healthcare sector and helped in diagnosis of critical illness such as cancer, neurology, cardiac, and kidney disease as well as with easing in anticipation of disease progression. By applying and executing machine learning algorithms over healthcare data, one can evaluate, analyze, and generate the results that can be used not only to advance the prior health studies but also to aid in forecasting a patient's chances of developing of various diseases. The aim in this article is to present an overview of machine learning and to cover various algorithms of machine learning and their present implementation in the healthcare sector.

## KEYWORDS

AI-Enabled Support System, Artificial Intelligence, Clinical Decision Making, Computer-Aided Diagnosis, Healthcare, Machine Learning Algorithm

## 1. INTRODUCTION

Machine learning, a subfield of Artificial Intelligence, provides algorithms to learn from past experiments while performing a particular task and measuring performance. By working continuously on a task, the performance of the task gets improved and the user experience as well. A Machine learning system has a training data set working as knowledge base and rules for decision making (Blum 2007). Machine learning is the building and exploring of methods in a computer programming language and making them "learn". The program developed using machine learning algorithms accesses the data, trains the machine and tests it again for performance evaluation. The most important characteristic of machine learning is its ability to forecast. In machine learning, a model for prediction

DOI: 10.4018/IJEHMC.2021010101

is build by existing information and it is further used for predicting the data. The major aspect of learning is the features selection from the data set as all the features cannot be used in learning. The data set may have multiple fields and perspectives. The selection of features is done according to their relevance, implication and scenario (Du & Swamy 2013). The primary goal of machine learning is to produce and enhance the learning algorithms and models in order to facilitate their easy application in various disciplines such as agriculture (Patrício & Rieder 2018), banking (Erdogan 2013), cyber security (Buczak & Guven 2016), economics (Einav & Levin 2014), finance (Lin et al., 2012), insurance (Gan 2013), natural language processing (Collobert & Weston 2008), online & traditional marketing (Tripathy et al., 2006), healthcare (Crown 2015), network & telecommunication (Richter & Khoshgoftaar 2018) and others as well.

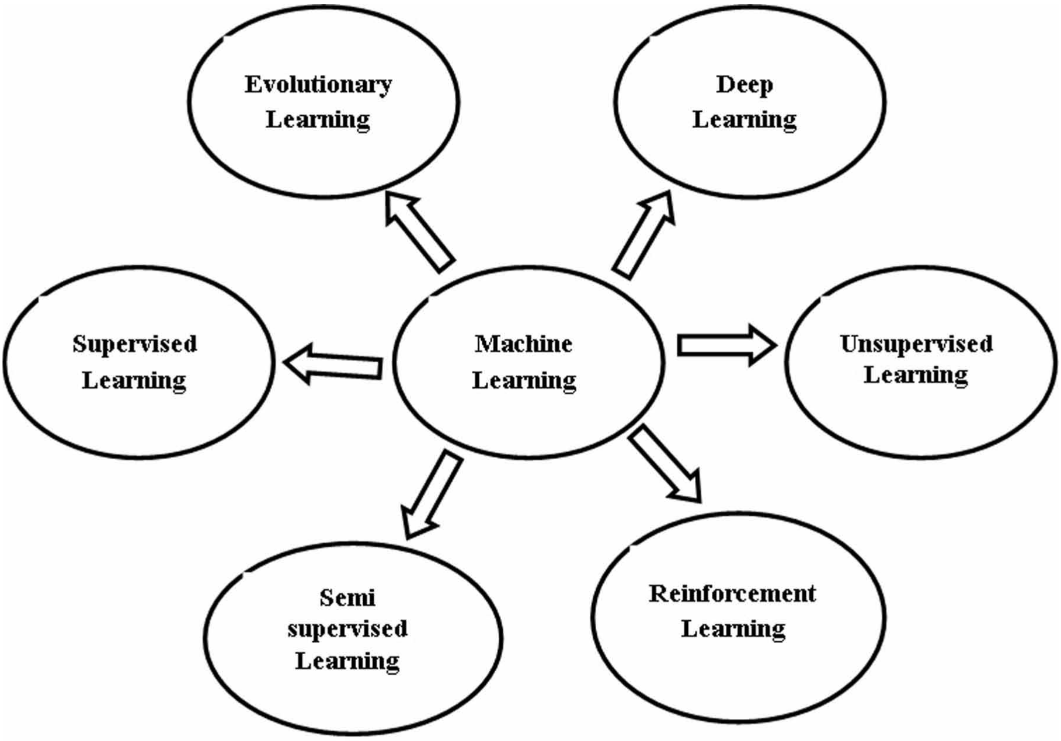
## 2. MACHINE LEARNING PARADIGMS

Machine learning algorithms are categorized according to their design, required input, produced output and applications. The majorly of machine learning algorithms are categorized as supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, evolutionary learning and deep learning as shown in Figure 1 (Fatima & Pasha 2017). Supervised Learning algorithms construct a model that uses a set of labeled data and desired output. The machine has been trained with the sampled data having essential features and preferred output. Once the machine is trained, testing is performed against the test data and the results are matched with desired results and the accuracy of the machine and algorithm gets measured. With the availability of huge amount of unlabeled data, supervised learning is not possible. Hence it is necessary to use unlabeled data to train a machine. The learning process using unlabeled data is called unsupervised learning. Unsupervised learning accomplishes the training process of an algorithm, with unlabeled data that are grouped on basis of relationship, variations and patterns within the data. Based on similarities and differences a hieratical structure is formed up to the point that similar objects are grouped together (Goyal & Kishnan 2019). Unsupervised learning algorithms are also termed as clustering algorithms as a cluster of data is formed from large amount pf unlabeled data while considering resemblance and variation (Mohssen & Eihab 2017). Clustering algorithms are further categorized as Density based methods; Connectivity based methods, Hierarchy Based Methods, Centroid Based Method or Partitioning Methods and Distribution Based Methods. Semi supervised learning uses labeled data (in small amount) and unlabeled data (in large amount) (Zhu 2005). Reinforcement learning is quite different from supervised learning and unsupervised learning. There is neither a knowledgeable expert supervisor nor an input/output pair. Reinforcement learning is a goal oriented technique of machine learning in dynamic environments. Reinforcement learning is machine learning process to produce intelligent programs or agents through learning and adopting from environmental changes. Learning can be done even though the information about the environment is not completely known. Agents get the feedback about the action performed and reward/punishment immediately. The methods for reward/punishment signal are Finite- horizon model, receding-horizon control, infinite-horizon discounted model, and average-reward model (Kaelbling et al., 1996). Evolutionary learning is a learning process where an algorithm learns from its past result and improves its performance (Zhang et al., 2011). Deep learning is the improvement over Multi layer perceptron model where the hidden layers can be increased, to a given computational level (Litjens et al., 2017). In deep learning the minimum number of hidden layers must be more than two. Categorization of machine learning algorithms is given in Figure 1.

## 3. STUDY OF MACHINE LEARNING ALGORITHMS IN HEALTHCARE SECTOR

Healthcare sector is one of the most emerging sectors, which present lifesaving to millions of people; it is also becoming one of the top revenue-earning sectors in various countries. Today in India, approximately 4.7% of total GDP is spent on the healthcare sector per year (Global Health Observatory

Figure 1. Machine learning categorization



Report - India n.d.). As a result, healthcare experts and researchers of closely related or other fields, all around the globe are working to find innovative ways to deliver quality and timely out comes in health services. Due to the demand of fast and quality health service in patient care, billing, and updated medical records, technology based expert system are being developed. Implementation of Machine learning in healthcare is widely accepted. Machine learning in healthcare assists in analyzing various data points and outcome generation while providing timely medical toolset. In this section, we discuss various applications of machine learning algorithms in identification and predication of Cancers (Breast Cancer, Lung Cancer and Skin Cancer). Cancer is a disease in which body cells grow or change abnormally. Due to this growth, lumps could develop in any part of the human body. Each specific cancer is named upon the body part it develops in such as Oral Cancer, Stomach Cancer, Melanoma, Breast Cancer and Lung Cancer. Due to large number of cancer cases reported and an increase in the use of MRI, CT scans and mammographic images the task of medical practitioner has become critical and crucial. They need to quickly identify and analyze the nodules in the images and make a prognosis. A small negligence in the task may turn into disaster for the patient. As a result computer aided diagnosis (CAD) systems are evolved for early and correct detection of cancer. CAD systems are designed for detection (CADe) and diagnosis (CADx). The CADx system produces a diagnosies based on MRI, CT as well as Mammographic images, feature extraction can be performed over those images and as result, they are automatically classified as Malignant (Cancerous) or Benign (Non-Cancerous) (Valente et al., 2016).

### 3.1 Breast Cancer

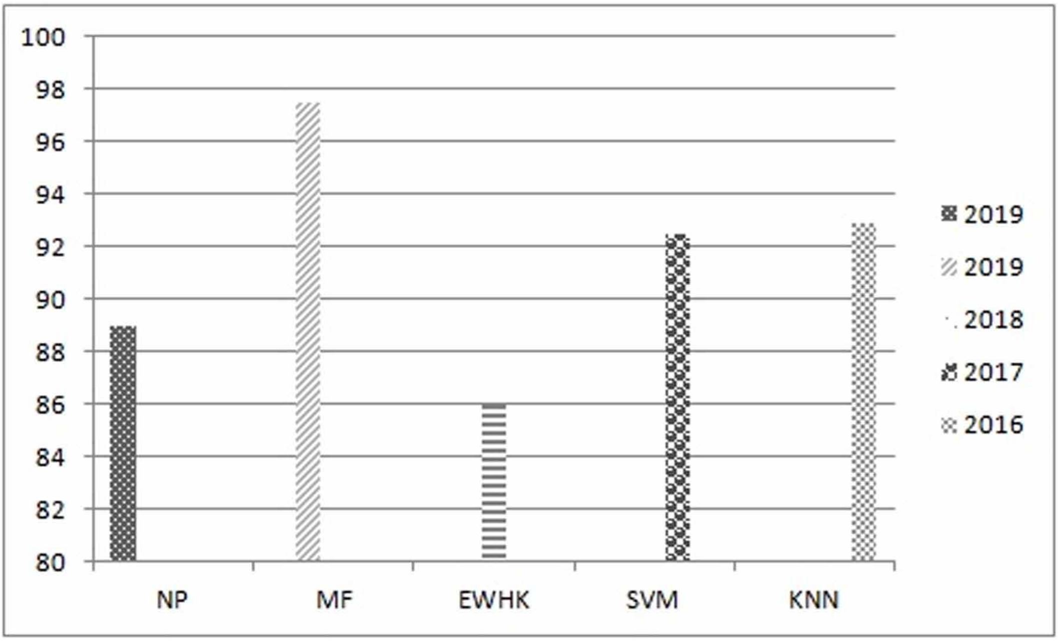
An abnormal growth of malignant breast tissue is termed breast cancer. Breast cancer is the leading cause of death in women by cancer. At the same time, it is also one of the most curable cancers

(Mallamala et al., 2019). Breast cancer can be further classified as Ductal Carcinoma in Situ (DCIS), Invasive Ductal Carcinoma, Triple negative breast cancer, inflammatory breast cancer, Metastatic breast cancer, Breast Cancer during pregnancy, Medullary Carcinoma, Tubular Carcinoma, and Mucinous Carcinoma (Type of Breast Cancer n.d.). The survey conducted by National Cancer Institute suggests that in the US one out of every eight women is prone to get such cancer in her lifetime (SEER Cancer Static Review 1975-2016). Breast cancer cases account for 25% of total cancer cases reported on women. 3% of women die in the early stages of the disease. Women 40 year age or older are more prone to breast cancer. To protect themselves against breast cancer, patient should closely follow their habits as to avoid getting infections, and to recognize symptoms in the early stages of the disease. It is also important to take proper medication, and consult with experts (Gbenga et al., 2017). DE Gbenga el al suggests that machine-learning algorithms are very effective and efficient in early and correct detection of breast cancer. Eight machine-learning algorithms (Radial Based Function (RBF) Network, Support Vector Machine (SVM), Simple Linear Logistic Regression Model (SL), Naïve Bayes (NB), *k*-Nearest Neighbor (*k*-NN), AdaBoost, Fuzzy Unordered Role Induction algorithm (Fuzzy), and Decision Tree) are experimented on Wisconsin Breast Cancer (WBC) dataset. The parameters taken into consideration for effectiveness and efficiency of these algorithms are accuracy, TPR, FPR, precision, F-measure for effectiveness and efficiency. SVM is reported with best accuracy (97.07%) (Gbenga et al., 2017). An automatic breast cancer classification approach is proposed using digitalized mammographic images. Feature extraction is performed by Gabor Filter and data reduction is performed using Locality Sensitive Discriminant Analysis (LSDA). Classification algorithms used to classify data are Decision tree, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), *k*-Nearest Neighbor (*k*-NN), Naïve Bayes (NB) Probabilistic Neural Network (PNN), Support Vector Machine (SVM), AdaBoost and Fuzzy (FSC). Universally accepted DDSM dataset (Digital Database for Screening Mammography) with 690 mammographic images is used for analysis. The highest accuracy reported by *k*-NN machine learning algorithm is 98.69%, with sensitivity of 99.34% and 98.26% specificity. *k*-fold cross validation (k= 5) (Raghavendra et al., 2016). In early detection of breast cancer, tumor detection via Raman Spectroscopy has an efficient role but due to low SNR of the Raman spectrum, it is difficult to diagnose breast cancer tumors with it. The authors have proposed entropy weighted local hyper plane *k*-Nearest Neighbor (EWHK) algorithm to improve the detection of breast cancer. The other approaches used and compared in this diagnosis are adaptive plane *k*-local hyper plane (AWKH) algorithm and *k*-NN. The result analysis shows that EWHK gives 92.33% accuracy as compared to 90.70% for AWKH and 89.30% specified by *k*-NN (Li et al., 2018). The BreaKHis dataset contains approximately 8,000 histopathology images of breast tumor (benign or malignant). The authors have proposed a Non-Parametric Multiple Instance Learning CAD system for breast cancer. Various MIL techniques such as APR (Axis-Parallel Rectangle algorithm), DD (Diversity Density), MI-SVM (Multiple Instance Support Vector Machine), Citation *k*-NN and MILCNN are used along with non-parametric multiple instance learning approach to study the breast tumor images. Non-parametric MIL is a modified version of the *k*-NN algorithm. It utilizes an approach on the basis of distance between *k*- neighbors (nearest). Non-parametric MIL increases strength to identify noise on different datasets. The MIL approach is also compared with various Single Instance Learning (SIL) classification algorithms such as Neural Network (1-NN), Quadratic Discriminant Analysis (QDA), Random Forest (RF), CNN and SVM. The BreaKHis dataset contacts images that are magnified 40x, 100x, 200x and 400x for better study of malignant tumor. As per the results obtained by the authors in their study presents that in comparison to MIL, Non parametric MIL gives better result in the class (40x (92.1%), 100x (89.1%) and 200x (87.2%)) an average of 89.47% and MILCNN for 400x images 83.4% (Sudharshan et al., 2018). Rakhlin A et al developed a computational model based on convolutional neural network for classifying images of breast cancer into four classes; Normal, Benign, 'Carcinoma in situ' and Invasive Carcinoma. Each class represents a specific cancer type in an image. In their research, the authors have used a different approach termed as deep convolutional feature representation (Guo et al., 2016). To normalize each microscopic image, 50 random color augmentations was performed and downsized each image with crops of 400 x 400 and 650 x 650 encoded as 20, converted into single descriptor through three norm polling (Boureau et al., 2010).

Table 1. The summary of machine learning approaches for breast cancer diagnosis

| Author | Year | Data Set/<br>No of Images | Machine Learning Algorithm/ Technique |
|---|---|---|---|
| P.J. Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, Paul Honeine (Sudharshan et al., 2018) | 2019 | BreaKHis Dataset/ 8000 with zoom 40x, 100x, 200x, 400x | APR, DD, EM-DD, Citation *k*-NN, MILCNN, Non parametric MIL |
| Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov, Alexandr A. Kalinin (Rakhlin et al., 2018) | 2019 | H&E stained breast histology microscopy image Dataset/400 | CNN Models (ResNet 50, VGG-16, Inception V3, Model Fusion) |
| Qingbo Li, Wenjie Li Jialin Zhang, Zhi Xu (Li et al., 2018) | 2018 | Peking University Third Hospital /16, QE6500 Raman spectrometer | EWHK,<br>*k*-NN (k=5)<br>AWKH |
| Dada Emmanuel Gbenga, Ngene Christopher, Daramola Comfort, Yetunde (Gbenga et al., 2017) | 2017 | WBC /699 | RBF, SVM, LR, NB, *k*-NN, AdaBoost, DT, Fuzzy |
| U. Raghavendra, U. Rajendra Acharya, Hamido Fujita, Anjan Gudigar, Jen Hong Tan, Shreesha Chokkadi (Raghavendra et al., 2016) | 2016 | DDSM /690 | DT, LDA, QDA, *k*-NN, NB, PNN,SVM, AdaBoost, Fuzzy Sugeno (FSC) |

Figure 2. Accuracy of machine learning algorithms for breast cancer diagnosis



CNNs such as ResNet 50, Inception V3 and VGG- 16 were used for features Extraction. For model fusion with ResNet 50 and Inception V3, last convolution layer (2048 channels) was replaced with GlobalAveragePooling (1D feature vector of length 2048). For model fusion of VGG 16 GlobalAveragePooling was applied on four internal convolution layers and replaced with one vector of length 1408. Dataset used for the analysis is 'Hematoxylin and Eosin stained breast histology microscopy image dataset' with 400 images of four classes. The result of proposed approach was compared with the results of classical CNN models (ResNet50, Inception V3 and VGG-16). The result shows that proposed model fusion gives the accuracy of 93.8% for 2-Class (non-carcinoma vs carcinomas) classification and for 4-class classification accuracy measured is 87.2% (Rakhlin et al., 2018). The accuracy of classification is measured over 10 fold cross validation. The Summery of machine learning approaches for breast cancer, studied in this paper is given in the Table 1. The graph plotted in Figure 2 shows annual accuracy achieved by machine learning algorithm in the diagnosis of breast cancer. The data presented in Table 1 is used to plot the graphs. To plot the graph authors have used MS Visio.

### *3.1.1 Analysis of Breast Cancer Study*

In the above literature discussed, *k*-NN offers highest accuracy of 98.69% in 2016 as shown in the Table 1. In many application areas *k*-NN shows good performance result. Features selected by authors in the paper responded well by *k*-NN algorithm. In 2018 a variant of *k*-NN termed as EWHK (Entropy Weight Local Hyper plane k-nearest-neighbor) is used to classify breast cancer. An accuracy of 92.33% is achieved which is quite low in comparison with 2016. The dataset used in both the papers were different. Overall in over study *k*-NN gives better results while comparing with other algorithms.

## 3.2 Lung Cancer

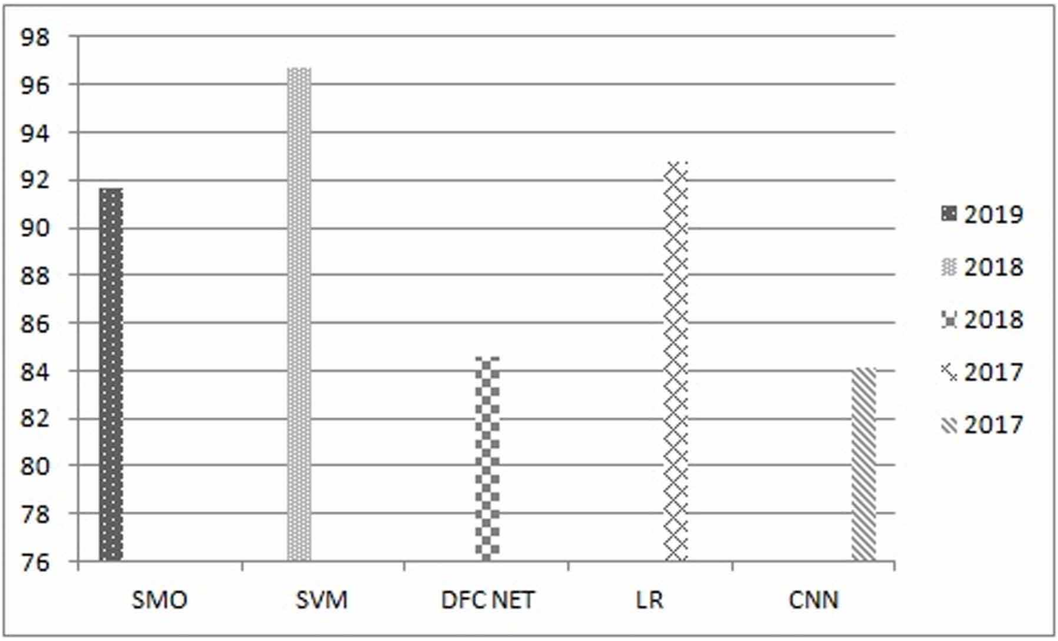
Uncontrolled growth of cell in to lungs is called Lung Cancer. Lung cancer can be categorized as Small cell lung cancers (SCLC), a cancer dominates on 10-15% of total lung cancers cases, and Non-small cell lung cancers (NSCLC) holds 85% of the total cases. NSCLS can be categorized further as Adeno carcinoma, Squamous cell carcinomas, and Large cell carcinomas. Lung, prostrate, breast, colon cancer are most common cause of death due to cancer. 46% of all the deaths due to cancer are because of these, whereas lung cancer alone holds 27% of this (Siegel et al., 2016). Early detection of lung cancer with Computed Tomography (CT) scan increases the chances of survival of patient. Various computer aided diagnosis system have been developed for early detection of Lung cancer. In this section authors have studied and analyzed few researches for detection of infected nodules. A person may get infected with cancer due to his genetic structure. Studies of Ecogenomics of cancer are more important to highlight the factors and cause of cancel. The authors had analyzed genes expression of lung cancer on KRBMD repository (Kent Ridge Bio-medical Dataset). The authors have predicted optimal subset of genes which are most probably play vital role in lung cancer. Environment as a factor associated also being considered along with human genes which may cause Cancer as well. Advanced machine learning algorithms such as Multilayer Perceptron (MLP), Random Sub Space (RSS) and Sequential minimal Optimization (SMO) were used as classifier to provide comparison on the factors accuracy, precision and recall. Out of 7129 genes 72 genes comes as most probable cause of cancer. By applying the ranking method used was Infogain. Out of that, six genes type were selected, which were associated with a specific type of cancer. The result reveals that an accuracy of 91.67% was achieved with SMO in comparison with 86.67% with MLP and 68.33% with RSS (Pati, 2019). Structural co-occurrence matrix (SCM) a feature extraction technique was applied on grayscale and Hounsfield unit images with Mean, Laplace, Gaussian and Sobel filters on the images of LIDC/IDRI dataset in order to create eight different configurations. Support Vector Machine (SVM), *k*-Nearest Neighbor (*k*-NN) and Multilayer Perceptron (MLP). Machine learning algorithms were used for classification of images as Malignant or Benign and also classify the malignancy level on the scale of five. The result of SCM was compared with GLCM, Local binary patterns, Central moments and statistical moments (Rodrigues et al., 2018). This approach results in 96.7% accuracy and F-score in image classification (Malignant or Benign) and 74.5% accuracy and 53.2% F-score in classifying Malignancy level. The authors concluded that SCM extracts the features successfully when used with mean filter with HU unit images as input and using SVM as classifier. A. Masood et al have proposed a computer assisted decision support system (DFCNet) for pulmonary lung cancer detection and classification. The CAD system is based on novel deep learning model and data collection was performed using medical body area network (MBAN). DFCNet based on fully convolutional Neural Network (FCNN), classifies the pulmonary nodules into four stages of lung cancer. The architecture of DFCNet is embraces seven convolution layer and parametric rectified liner units, seven pooling layer (max), seven normalization layer (batch) and two dense layer $\text{Conv}_{2048}$ leaky rectified linear units (LreLU), a 1000-dimentional dense layer and deconvolutional layer with activation function as softmax. The authors have used six different dataset (LIDC/IDRI, RIDER, SPIE challenge dataset, LUNAI6, LungCT diagnosis, Shanghai hospital no 6 dataset) to check the performance of the model proposed. An accuracy of 84.58% (Average) of DFCNet is achieved while comparing the results with CNN (77.6%). The common symptoms taken into consideration during study were body weight loss, breathlessness, heart rate, high body temperature, blood pressure, insomnia, and hyper-calcemia, which cause pain, fatigue and constipation (Masood et al., 2018). Yan Qiang et al proposed a deep learning based method for pulmonary nodules diagnosis using EML (Extended Learning Machine). Dual model deep supervised auto encoder (DSDAE) method, take input images, obtained from positron emission tomography (PET) and computed tomography (CT) scan in a pair. Supervised encode extracts high-level discriminative features within the nodule. For validation purpose 5-fold cross validation is used over 1600 images of 120 patients with 2810 pulmonary nodules, collected from a hospital's PET/CT centre in Shanghai, China. To increase the number of images for study new images were obtained by image augmentation (rotating and scaling of images). The authors have divided the data into five subsets; four subsets were used for training and one for testing. The process of training and testing was repeated five times. For achieving higher accuracy, multimodal information fusion (Decision Level fusion (DLF) and Feature level fusion (FLF)) techniques were used for fusion of CT and PET features. For classification, classical machine learning algorithms (Back Propagation, Neural

Network, Support Vector Machine, Naïve Bayes, Logistic Regression and ELM) were used on the features of CT, PET, F-DLF, and F-FLF. The effectiveness of a classification algorithm is measured on accuracy, sensitivity and specificity. Logistic regression achieved an accuracy of 92.8% on F-FLF, whereas ELM achieved 88.75% and stood second (Qiang et al., 2017). Deep learning is one of the most powerful and popular method in Medical Image diagnosis (Litjens et al., 2017). The authors have designed CNN, DNN and SAE three deep neural networks for lung cancer detection. The CNN model has two convolution layers (24 x 24 x 32, 8 x 8 x 64) and max pool layers (12 x 12 x 64, 4 x 4 x 64), two fully connected layer (512 x 1, 2 x 1) and one softmax layer (2 x 1) as activation function. The DNN model has one Input layer (28 x 28 x 1) four fully connected layer (784 x 1, 512 x 1, 256 1, 64 x 1) and one softmax layer (2 x 1) as activation function. SAE is a multilayer auto encoder neural network. It is an unsupervised learning algorithm with 3 layers (input, hidden and output). Input layer is 28 x 28 x 1; hidden layer has 3 fully connected layer (784 x 1, 256 x 1, 64 x 1) and output layer as softmax (2 x 1) activation function to classify the nodules as malignant or benign. LIDC/IDRI dataset with 5024 sample images is used for analysis. The result confirms CNN gives accuracy of 84.15% in comparison with DNN (82.37%) and SAE (82.59%) (Song et al., 2017). Overview of the applicability of the machine learning algorithms for Lung cancer diagnosis, studied in this paper, is presented in the Table 2. The graph plotted in Figure 3 shows the annual accuracy achieved by the use of machine learning algorithm for the purpose of diagnosing lung cancer. The data presented in Table 2 is used to plot the graphs. The authors have used MS Visio to produce the graph.

Table 2. The summary of machine learning approaches for lung cancer diagnosis

| Author | Year | Data Set/ No of Images | Machine Learning Algorithm/ Technique |
|---|---|---|---|
| Jayadeep Pati (Pati, 2019) | 2019 | KRBMD repository | MLP<br>RSS<br>SMO |
| Murillo b. Rodrigues, Raul victor m. Da nóbrega, Shara shami a. Alves, Pedro pedrosa rebouças filho, João batista f. Duarte, Arun k. Sangaiah, Victor hugo c. De Albuquerque (Rodrigues et al., 2018) | 2018 | LIDC/IDRI dataset | SVM<br>*k*-NN<br>MLP |
| Anum Masood, Bin Sheng, Ping Lic, Xuhong Hou, Xiaoer Wei, Jing Qin, Dagan Feng (Masood et al., 2018) | 2018 | LIDC/IDRI, LIDER, SPIE challenge, LUNAI6, LungCT diagnosis, Shanghai Hospital No 6 | DFCNet<br>CNN |
| Yan Qiang, Lei Ge, Xin Zhao, Xiaolong Zhang, Xiaoxian Tang (Qiang et al., 2017) | 2017 | Shanxi hospital's PET/CT centre China/1600 | BP, SVM, NB, ELM, LR |
| QingZang Song, Lei Zhao, XingKe Luo, XueChenDou (Song et al., 2017) | 2017 | LIDC/IDRI /5024 | CNN, DNN, SAE |

Figure 3. Accuracy of machine learning algorithms for lung cancer diagnosis



### 3.2.1 Analysis of Lung Cancer Study

SVM provides better results in detection of Lung cancer with an accuracy of 96.7% as reported in 2018. LIDC/IDRI data set was used for experiments and SVM was implemented for features extraction. In 2019, a variant of SVM named SMO was also implemented for the lung cancer nodules detection.

The accuracy achieved with this approach is 91.67%, which was lower than that of SVM, as reported in 2018. However, the datasets for both studies were different. Over-all in those studies SVM results as the best algorithm for lung cancer detection.

### 3.3 Skin Cancer

Uncontrolled growth of skin cells is known as skin cancer. The major causes of skin cancer to occur skin cancer are exposed to UV rays of sunlight, tanning beds, damaged skin cell by unrepaired DNA and genetic defects. Due to these causes cell grows multiplicatively and turned into malignant tumors. Skin cancer is categorized as Actinic Keratoses (AK), Basal cell carcinoma (BCC), Squamous cell carcinoma (SCC), Melanoma (Dorj et al., 2018). Melanoma is one of the main causes of skin cancer which causes 75% of the total death from skin cancer (Kavitha et al., 2017). In this section author had studied and analyzed few researches on Melanoma skin cancer. A demographic image has various features such as color, texture and shape. Out of this texture is considered as one of the most crucial feature of the image. Texture is further categorized as Local texture feature (LTF) and Global texture feature (GTF). The global feature refers energy, entropy, homogeneity, correlation, contrast, dissimilarity and probability. To compute GTF, Grey level co-occurrence matrix (GLCM) is used with various orientations angle ($\theta = 0, 45, 90, 135$). Speeded up robust feature (SURF) a texture feature descriptor used to find LTF. For classification of dermoscopic images into melanoma and non melanoma, SVM and *k*-NN classifier algorithms were used. The performances of Machine learning algorithms were calculated on the parameters such as sensitivity, specificity, accuracy and precision. The authors have used Otsu's adaptive thresholding method for segmentation of original dermoscopic image. Total images used were 250 (150 for training & 100 for testing). The classification results shows that using SURF(86%) with SVM and *k*-NN (84%) gives better results over GLCM with SVM(77%) and *k*-NN(72%) . Local texture feature extraction method (SURF using SVM) outperform for the proposed method (Kavitha et al., 2017). J. Premladha et al proposed a CAD system for melanoma detection and prediction. 992 images from Mednode and PH2 datasets (PH2 Dataset n.d.) were used for the purpose of the study. Contrast Limited Adaptive Histogram Equalization technique (CLAHE) and Median filter were used for enhancement of images. For the validation purpose a 10-fold cross validation is performed with chi square distribution of 95% confidentiality limit. The CAD system works in five stages, starting with Image acquisition, Image segmentation, Border extraction, Feature extraction, and end with classification. In image segmentation, the ROI (Region of Interest) of an image was segmented from a complete dermographic image for which NOS (Normalized Otsu's Segmentation) algorithm was implemented. Border extraction was performed using GVF (Gradient Vector flow) and statistical region merging technique. Feature extraction is a process of finding out important features of an image for classification. The authors had used GLCM (Grey level co-occurrence matrix) and Geometrical based feature for finding the statistical and shape features of the image. Classification algorithms such as DLNN (Deep learning neural network), SVM (Support vector machine), Hybrid-Adaboost (Adaboost + SVM) and ANFIS (Adaptive Neuro Fuzzy Inference System), Adaboost (Real, Gentle & Modest) were implemented and effectiveness was checked over the parameter such as accuracy, sensitivity and specificity. The result shows that DLNN and Hybrid Adaboost (Adaboost + SVM) give better accuracy 92.89% and 91.73% respectively (Premaladha & Ravichandran 2016). The authors had followed ABCD (Asymmetrical Shape, Border, Color and Diameter of Melanoma) rule and develop a Decision Support System (DSS) that help in decision making by medical practitioner. In this research dermographic images of PH2 dataset (PH2 Dataset n.d.) had been studied through four machine learning algorithms, ANN (Artificial Neural Network), SVM (Support Vector Machine), *k*-NN(*k*-Nearest Neighbor) and DT(Decision Tree) to group the skin lesion as normal, abnormal and melanoma. The features studied to diagnose the melanoma were shape, border, color and diameter. Dataset was divided into ten different subparts by *k*-fold cross validation (k= 10) and k-1 subsets were used for training and testing was performed on remaining one subset i.e. total k times. The highest accuracy was achieved by ANN (92.50%) in comparison with DT (90.0%), SVM (89.5%) and *k*-NN (82.0%) (Ozkan & Koklum 2017). M. Nasir et al proposed a computerized method for classifying skin lesion as melanoma or benign. DullRazor (Lee et al., 1997) technique is used to process the images (removing hair from the images). In order to enhance the contrast and make the lesion clearer, a contrast stretching techniques which utilizes color and texture information is used. For image segmentation (border detection), a uniform distribution technique and active contour based methods were used. The results of uniform distribution and active contour based methods (Chan & Vese 2001) are fused together in order to get better images of the lesion. The additive law of probability was used to perform the fusion of the two results mentioned above. For feature fusion, serial feature fusion with maximum entropy technique was used on three features color, SFTA (Texture) and HOG (Shape). To validate the results of the process proposed in this paper, the authors had compared the result of ten machine learning algorithms. The results of proposed linear SVM (L-SVM) was compared with DT (Decision Tree), BT (Bagged Tree), SDA (Subspace discriminant analysis), w-*k*-NN (Weighted *k*-NN), Fine-*k*NN(f-*k*NN), subspace-*k*NN(s-*k*-NN), LDA(Linear discriminant analysis), QDA(Quadratic discriminant analysis), C-SVM(Cubic SVM) and Q-SVM (Quadratic SVM). All the results were validated on k-fold cross validation (k=10). Experiment revels that L-SVM (SVM with entropy based method) provides best result with accuracy 97.5% (Nasir et al., 2018). Tri long palm et al propose a Deep Convolutional Neural Network (CNN) study with two contributions first a model combining deep CNN and data augmentation to improve classification of skin cancer lesion, second, measuring performance of variance machine learning classification algorithms on augmented data images. The authors have proposed a system for melanoma classification with three steps augmentation, feature extraction and classification. In data augmentation deformations are applied to the data and try to overcome the problem of over fitting of labeled data. Three different data augmentation techniques were implemented by the authors. The techniques named as *Geometric augmentation*: cropping, horizontal and vertical flip of original image data. *Color augmentation*: normalizing the color of images collected from different sources or devices. *Data wrapping*: to mitigate over fitting of melanoma. Feature extraction is performed via deep convolutional network method inception V4 (Szegedy et al., 2017). InceptionV4 is an improved version of inceptionV3 (Szegedy et al., 2016) and developed on GoogleNet platform. After data augmentation and feature extraction, NN (Neural network), SVM (support Vector Machine) and RF (Random forest) machine learning algorithms were implemented for classification. Total 6762 images of three datasets 2017 ISBI challenge training and testing dataset (The Four Grand Challenges Chosen for ISBI 2017), ISIC challenge dataset (The International Skin Imaging Collaboration (ISIC) n.d.) and PH2 dataset (PH2 Dataset n.d.) with three sort of data augmentation NODAUG (no data augmentation), DAUG-50 (augment 50 samples of each image), DAUG-100 (augment 100 samples of each image). The result of DAUG-100, DAUG-50, NO-DAUG with classification algorithms NN, SVM,RF were compared with top three algorithms of ISBI 2017 challenge on the features AUC, AP, SEN, SPC, ACC, and PPV. The result shows that NN gives better accuracy (89.0%) with DAUG-100 images while comparing with Top#3 algorithm of ISBI challenge 2017 (87.2%). The results of SVM-DAUG-50 (88.8%) and RF-DAUG-50(88.5%) were better than NN-DAUG-50 (88.2%) but over all higher accuracy (89%) was achieved by NN-DAUG-100 (Pham et al., 2018). Summary of machine learning algorithms for skin cancer, studied in this paper is presented in Table 3. The graph plotted in Figure 4 gives annual accuracy achieved for machine learning algorithms used to diagnose skin cancer. The data shows in Table 3 are used to plot the graphs in Figure 4. The authors have used MS Visio to plot the graph.

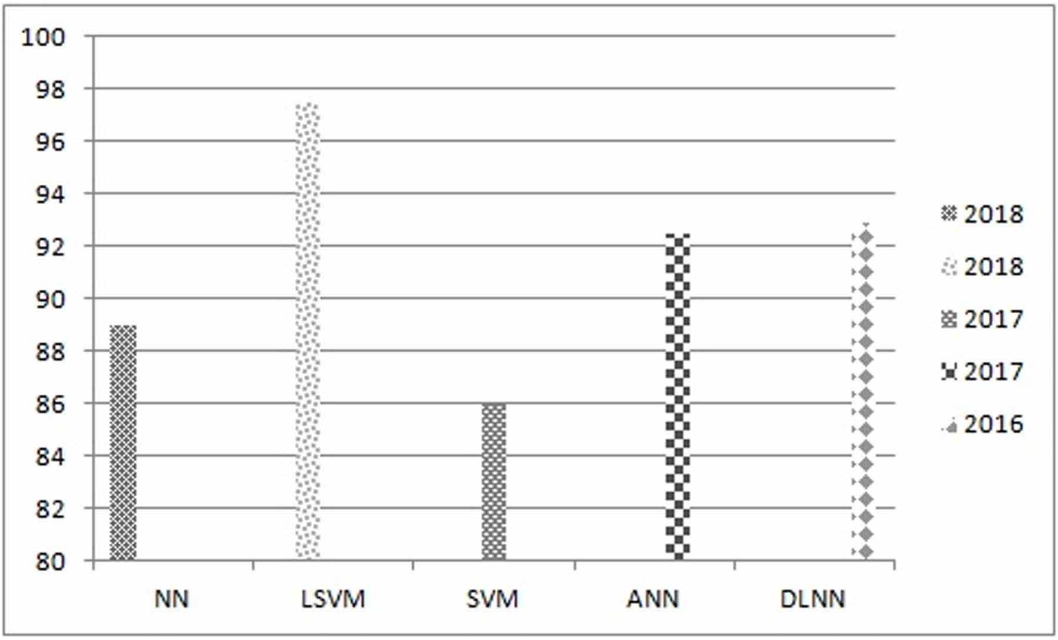
### 3.3.1 Analysis of Skin Cancer study

In the study of skin cancer diagnosis, best accuracy (97.5%) was achieved by a variant of SVM i.e. Linear SVM (L-SVM) as published in 2018. In 2017, diagnosing accuracy of 86% had been reported by SVM-SURF, which was low in comparison to the results of SVM as published in 2018. NN, ANN and DLNN were also used in 2018 and 2017 respectively. NN provide an accuracy of 89% while an

Table 3. The summary of machine learning approaches for skin cancer diagnosis

| Author | Year | Data Set/ No of Images | Machine Learning Algorithm/ Technique |
|---|---|---|---|
| Tri-Cong Pham, Chi-Mai Luong, Muriel Visani, Dung Hoang (Pham et al., 2018) | 2018 | ISBI 2017 Challenge Dataset (training & testing) /6762, ISIC/ $PH_2$/200 | SVM<br>RF<br>NN |
| Mohd Nasir, Attique Khan, Mohd Sharif, Ikram Ullah Lali, Tanzila Saba, Tassawar Iqbal (Nasir et al., 2018) | 2018 | $PH_2$/200 | L-SVM, DT, BT, SDA, QDA w-*k*NN, f-*k*NN, s-*k*NN,LDA, C-SVM,Q-SVM |
| JC Kavitha, Suruliandi A, Nagarajan D (Kavitha et al., 2017) | 2017 | **Not mentioned /250** | SVM-SURF<br>*k*-NN-SURF<br>SVM-GLCM<br>*k*-NN-GLCM |
| Ilker Ali Ozkan, Murat Koklu (Ozkan & Koklum 2017) | 2017 | $PH_2$/200 | ANN, SVM<br>*k*-NN, DT |
| J. Premaladha K. S. Ravichandran (Premaladha & Ravichandran 2016) | 2016 | (Mednode & $PH_2$) /922 | DLNN, SVM, Hybrid-Adaboost, ANFIS, Adaboost (Real, Gentle, Modest) |

Figure 4. Accuracy of machine learning algorithms for skin cancer diagnosis



accuracy of 92.5% and 92.89% had been achieved with the use of ANN and DLNN. The study shows that SVM gives better results for skin melanoma detection.

## 4. CONCLUSION

Machine learning plays a vital role in many areas such as image processing, data mining, natural language processing and healthcare. Out of this healthcare is an emerging area for machine learning application and implementation. In this paper, we presented our study of different types of cancers (Breast Cancer, Lung Cancer & Skin Cancer). In addition, the machine learning algorithms were used to study their applicability in the diagnosing other cancers (colon, prostate, throat, bone, mouth), heart, liver, kidney, diabetes and other. When we compare the results of the three cancers studies that we have presented in this paper, *k*-NN provides the highest overall accuracy of 98.69% for breast cancer where as the use of SVM (or its variants) results in better accuracy in diagnosing lung cancer (96.7%) and skin cancer (97.5%).

## REFERENCES

Blum, A. (2007). *Machine learning theory*. Carnegie Melon University, School of Computer Science. cs.cmu.edu

Boureau, Y. L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. *Proceedings of the 27th International Conference on Machine Learning (ICML-10),* 111–118.

Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys and Tutorials*, *18*(2), 1153–1176. doi:10.1109/COMST.2015.2494502

Chan, T. F., & Vese, L. A. (2001). Active contours without edges. *IEEE Transactions on Image Processing*, *10*(2), 266–277. doi:10.1109/83.902291 PMID:18249617

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*. doi:10.1145/1390156.1390177

Crown, W. H. (2015). Potential Application of Machine Learning in Health Outcomes Research and Some Statistical Cautions. *Value in Health*, *18*(2), 137–140. doi:10.1016/j.jval.2014.12.005 PMID:25773546

Dorj, U.-O., Lee, K.-K., Choi, J.-Y., & Lee, M. (2018). The skin cancer classification using deep convolutional neural network. *Multimedia Tools and Applications*, *77*(8), 9909–9924. doi:10.1007/s11042-018-5714-1

Du, K.-L., & Swamy, M. N. S. (2013). Fundamentals of Machine Learning. *Neural Networks and Statistical Learning*, 15–65. doi:10.1007/978-1-4471-5571-3_2

Einav, L., & Levin, J. (2014). The Data Revolution and Economic Analysis. *Innovation Policy and the Economy*, *14*, 1–24. doi:10.1086/674019

Erdogan, B. E. (2013). Prediction of bankruptcy using support vector machines: An application to bank bankruptcy. *Journal of Statistical Computation and Simulation*, *83*(8), 1543–1555. doi:10.1080/00949655.2012.666550

Fatima, M., & Pasha, M. (2017). Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1–16. doi:10.4236/jilsa.2017.91001

Gan, G. (2013). Application of data clustering and machine learning in variable annuity valuation. *Insurance, Mathematics & Economics*, *53*(3), 795–801. doi:10.1016/j.insmatheco.2013.09.021

Gbenga, D. E., Christopher, N., & Yetunde, D. C. (2017). Performance Comparison of Machine Learning Techniques for Breast Cancer Detection. *Nova Journal of Engineering and Applied Sciences*, *6*(1), 1–8. doi:10.20286/nova-jeas-060105

Global Health Observatory Report -India. (n.d.). https://www.who.int/countries/ind/en/

Goyal, J., & Kishan, B. (2019). Progress on Machine Learning Techniques for Software Fault Prediction. *International Journal of Advanced Trends in Computer Science and Engineering, 8*(2).

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48. doi:10.1016/j.neucom.2015.09.116

Kaelbling, Littman, & Moore. (1996). Reinforcement Learning: A Survey. *JAIR, 4*.

Kavitha, J. C., Suruliandi, A., & Nagarajan, D. (2017). Melanoma Detection in Dermoscopic Images using Global and Local Feature Extraction. *International Journal of Multimedia and Ubiquitous Engineering*, *12*(5), 19–28. doi:10.14257/ijmue.2017.12.5.02

Lee, T., Ng, V., Gallagher, R., Coldman, A., & McLean, D. (1997). Dullrazor ®: A software approach to hair removal from images. *Computers in Biology and Medicine*, *27*(6), 533–543. doi:10.1016/S0010-4825(97)00020-6 PMID:9437554

Li, Q., Li, W., Zhang, J., & Xu, Z. (2018). An improved k-nearest neighbour method to diagnose breast cancer. *Analyst (London)*, *143*(12), 2807–2811. doi:10.1039/C8AN00189H PMID:29863729

Lin, W.-Y., Hu, Y.-H., & Tsai, C.-F. (2012). Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, *42*(4), 421–436. doi:10.1109/TSMCC.2011.2170420

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. doi:10.1016/j.media.2017.07.005 PMID:28778026

Masood, A., Sheng, B., Li, P., Hou, X., Wei, X., Qin, J., & Feng, D. (2018). Computer-Assisted Decision Support System in Pulmonary Cancer detection and stage classification on CT images. *Journal of Biomedical Informatics*, *79*, 117–128. doi:10.1016/j.jbi.2018.01.005 PMID:29366586

Mohssen, M. Z., & Eihab, B.M.B (2017). *Machine Learning: Algorithms and Applications*. Taylor & Francis Group.

Nallamala, S.H., Mishra, P., & Koneru, S.V. (2019). Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems. *International Journal of Advanced Trends in Computer Science and Engineering, 8*(2).

Nasir, M., Attique Khan, M., Sharif, M., Lali, I. U., Saba, T., & Iqbal, T. (2018). An improved strategy for skin lesion detection and classification using uniform segmentation and feature selection based approach. *Microscopy Research and Technique*, *81*(6), 528–543. doi:10.1002/jemt.23009 PMID:29464868

Ozkan, I.A., & Koklu, M. (2017). Skin Lesion Classification using Machine Learning Algorithms. *International Journal of Intelligent Systems and Applications in Engineering*. DOI: 10.18201/ijisae.2017534420\

PH2 Dataset. (n.d.). https://www.fc.up.pt/addi/ph2%20database.html

Pati, J. (2019). Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques. *An Eco-Genomics Approach IEEE Access Volume*, *7*. Advance online publication. doi:10.1109/ACCESS.2018.2886604

Patrício, D. I., & Rieder, R. (2018). Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Computers and Electronics in Agriculture*, *153*, 69–81. doi:10.1016/j.compag.2018.08.001

Pham, T.-C., Luong, C.-M., Visani, M., & Hoang, V.-D. (2018). Deep CNN and Data Augmentation for Skin Lesion Classification. *Lecture Notes in Computer Science*, *10752*, 573–582. doi:10.1007/978-3-319-75420-8_54

Premaladha, J., & Ravichandran, K. S. (2016). Novel Approaches for Diagnosing Melanoma Skin Lesions Through Supervised and Deep Learning Algorithms. *Journal of Medical Systems*, *40*(4), 96. Advance online publication. doi:10.1007/s10916-016-0460-2 PMID:26872778

Qiang, Y., Ge, L., Zhao, X., Zhang, X., & Tang, X. (2017). Pulmonary nodule diagnosis using dual-modal supervised autoencoder based on extreme learning machine. *Expert Systems: International Journal of Knowledge Engineering and Neural Networks*, *34*(6), e12224. doi:10.1111/exsy.12224

Raghavendra, U., Rajendra Acharya, U., Fujita, H., Gudigar, A., Tan, J. H., & Chokkadi, S. (2016). Application of Gabor wavelet and Locality Sensitive Discriminant Analysis for automated identification of breast cancer using digitized mammogram images. *Applied Soft Computing*, *46*, 151–161. doi:10.1016/j.asoc.2016.04.036

Rakhlin, A., Shvets, A., Iglovikov, V., & Kalinin, A. A. (2018). Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis. *Image Analysis and Recognition*, 737–744. doi:10.1007/978-3-319-93000-8_83

Richter, A. N., & Khoshgoftaar, T. M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial Intelligence in Medicine*, *90*, 1–14. Advance online publication. doi:10.1016/j.artmed.2018.06.002 PMID:30017512

Rodrigues, M. B., Da Nobrega, R. V. M., Alves, S. S. A., Filho, P. P. R., Duarte, J. B. F., Sangaiah, A. K., & De Albuquerque, V. H. C. (2018). Health of Things Algorithms for Malignancy Level Classification of Lung Nodules. *IEEE Access : Practical Innovations, Open Solutions*, *6*, 18592–18601. doi:10.1109/ACCESS.2018.2817614

SEER Cancer Static Review 1975-2016. (2017). National Cancer institute. https: //seer.cancer.gov/csr/1975 2016/

Siegel, Miller, & Jemal. (2016). *Cancer statistics*. Cancer J. Clinic.

Song, Q., Zhao, L., Luo, X., & Dou, X. (2017). Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images. *Journal of Healthcare Engineering*, *2017*, 1–7. doi:10.1155/2017/8314740 PMID:29065651

Sudharshan, P. J., Petitjean, C., Spanhol, F., Oliveira, L., Heutte, L., & Honeine, P. (2018). Multiple Instance Learning for Histopathological Breast Cancer Image Classification. *Expert Systems with Applications*. Advance online publication. doi:10.1016/j.eswa.2018.09.049

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. Artificial Intelligence, 4278–4284.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition* (Vol. 2016, pp. 2818–2826). CVPR. doi:10.1109/CVPR.2016.308

The Four grand challenges chosen for ISBI. (2017). https://biomedicalimaging.org/2017/challenges/

The International Skin Imaging Collaboration (ISIC). (n.d.). https://isic-archive.com/

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117–126. doi:10.1016/j.eswa.2016.03.028

Type of Breast Cancer. (n.d.). https://www.nationalbreastcancer.org/types-of-breast-cancer

Valente, Cortez, Neto, Soares, de Albuquerque, & Tavares. (2016). Automatic 3D pulmonary nodule detection in CT images: A survey. *Computer Methods Programs Biomed., 124*, 91-107.

Xiaojin, Z. (2005). *Semi Supervised Learning literature Survey*. Technical Report # 1530.

Zhang, J., Zhan, Z., Lin, Y., Chen, N., Gong, Y., Zhong, J., & Shi, Y. (2011). Evolutionary Computation Meets Machine Learning: A Survey. *IEEE Computational Intelligence Magazine*, *6*(4), 68–75. doi:10.1109/MCI.2011.942584

*Prashant Johri (PhD) is working as Professor in School of Computing Science & Engineering, Galgotias University, Greater Noida, India. He completed his M.C.A. from Aligarh Muslim University and Ph.D. in Computer Science from Jiwaji University, Gwalior, India. He has also worked as a Professor and Director (M.C.A.), Galgotias Institute of Management and Technology, (G.I.M.T.) and worked as a Professor and Director (M.C.A.), Noida Institute of Engineering and Technology, (N.I.E.T.) Gr.Noida .He has served as Chair in many conferences and affiliated as member of program committee of many conferences in India and Aborad. He has supervised 2 PhD students and many M.Tech. students for their thesis. He published Number of research papers in National and International Journals and Conferences. He also contributed numerous book chapters to the several books published with publishers of high international repute. Apart from scholarly contribution towards scientific community, he organized several Conferences/Workshops/Seminars at the national and international levels. He voluntarily served as reviewer for various International Journals, conferences, and workshops. He published Edited book in Springer . His research interest includes data retrieval and predictive analytics, information security, privacy protection, big data open platforms, Software Reliability cloud computing, Mobile cloud, Machine learning, AR & VR, Soft computing, Fuzzy systems, Healthcare, Agriculture, Pattern recognition, Bio-inspired phenomena, and advanced optimization model & computation.. He is actively publishing in these areas.*

*Vivek Sen Saxena is B-Tech and M-Tech. Currently Working as Assistant Professor at INMANTEC Institutions, Ghaziabad. He is pursuing PhD from School of Computing Science & Engineering, Galgotias University. His research area is Machine Learning implementation in healthcare industry, Big Data Analytics, Artificial Intelligence.*

*Avneesh Kumar (PhD) is working as Associate Professor in School of Computing Science & Engineering, Galgotias University, Greater Noida, India. He completed his M.C.A. from Uttar Pradesh Technical University Lucknow and PhD in Computer Science from Jiwaji University, Gwalior, India. He has also worked as an Assistant Professor at INMANTEC Institutions, Ghaziabad. His research area is Artificial Intelligence & Machine Learning, Big Data Analytics, Software Reliability, etc. He served as Chair in many conferences and affiliated as member of program committee of many conferences in India and Aboard.*