Breast Cancer Prediction Using Machine Learning Techniques

Apoorva V[1,*], Yogish H K[2], Chayadevi M L[3]

[1,2] *Dept. of IS &E, MSRIT, Bengaluru*
[3] *Dept. of IS & E, JSSATE, Bengaluru*
[*]*Corresponding authors. Email:* <u>apoorvav133@gmail.com</u>

ABSTRACT

Breast cancer affects the majority of women worldwide, and it is the second most common cause of death among women. However, if cancer is detected early and treated properly, it is possible to be cured of the condition. Early detection of breast cancer can dramatically improve the prognosis and chances of survival by allowing patients to receive timely clinical therapy. Furthermore, precise benign tumour classification can help patients avoid unneeded treatment. This paper study uses Convolution Neural Networks for Image dataset and K-Nearest Neighbour (KNN), Decision Tree (CART), Support Vector Machine (SVM), and Naïve Bayes for numerical dataset, whose features are obtained from digitised image of breast mass, as to forecast and analyse cancer databases in order to improve accuracy. The dataset will be analysed, evaluated, and model is trained as part of the process. Finally, both image and numerical test data will be used for prediction.

Keywords: IDC (Invasive Ductal Carcinoma), FNA (Fine Needle Aspirate) , Breast cancer prediction, Classifier algorithms, CNN (Convolutional neural network).

1. INTRODUCTION

Breast-cancer is among the most serious illnesses/diseases in India, causing many deaths in the current situation. Due to changes in food and lifestyle, the number of cancer cases in women is increasing day by day. It is the second most common cause of death in women in the world. [1], This uses concepts of Deep learning (DL) and Machine learning (ML) to predict breast cancer based on the data obtained. This cancer is produced by abnormal growth of fatty and fibrous tissues, and the different phases of cancer are caused by cancer cells spreading throughout the tissue [4]. This is one of the most common cancers that affects women, but other types of cancer and those who are affected by them can be treated greatly, according to a government survey, when compared to breast cancer. The various phases of breast cancer are identified via proper treatment and detailing. If we do not provide proper therapy to our patients, it will result in their death. A number of methods for establishing an accurate diagnosis of breast cancer have been presented. Because the dataset contains a variety of distinct report attributes, machine learning may be easily applied to the dataset for prediction [22-25]. Even by using Technology which is not fully automatically designed to give the output. Hence here we propose the fully automatic classification and prediction of breast cancer based on dataset. Using deep learning technique. This learning technique is recognised as best method to predict and classify for image dataset.

Earlier methods for classifying data were used, despite their lower accuracy, because they could be used for proper categorization and prediction. Deep learning algorithms and numerical dataset machine learning techniques are used to extract features and hidden features. The convolution value is obtained from the stride function, which is used to harvest features from different sizes of images. CNN is one that gives proper output for the dataset we utilised in this [13-17].

Typical cancer screening procedures are grounded on the "gold-standard", that consists of three tests: clinical evaluation, radiological imaging, and pathology testing. [18]. This traditional technique, which is based on regression, detects the existence of cancer, whereas new ML techniques and algorithms are built on model creation. In its training and testing stages, the model is meant to forecast unknown data and offers a satisfactory predicted outcome [19]. Pre-processing, feature selection or extraction, and classification are the three major methodologies used in machine learning [20]. The feature extraction part of the machine learning method is crucial for cancer diagnosis and prediction. This process may differentiate between benign and malignant tumours [21].

The "gold-standard" method for detecting cancer previously consisted of three parts: clinical evaluation, radiological imaging, and pathology testing. [18]. The proposed technique indicates the presence of cancer based on regression while new algorithms are available. Model which has been designed for prediction of new data and should give good result in their training and testing phase [19]. Here there are 3 main steps preprocessing features, extraction and classification. Figure 1 shows the types of breast cancers, in this paper we consider IDC.
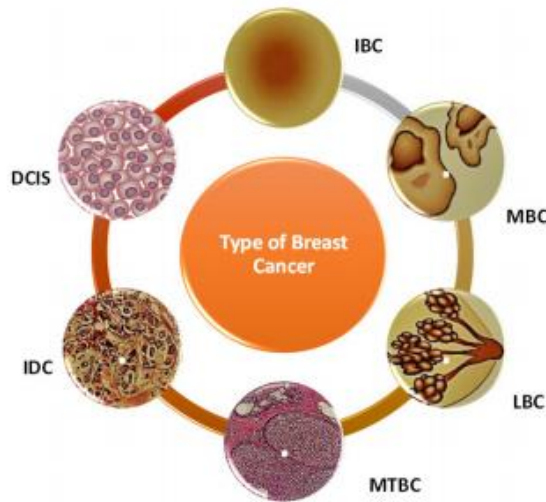

Figure 1 The various kinds of breast cancer.

2. RELATED WORKS

This section provides an orderly description of several approaches focused on chest harm examination to grasp the concerns and challenges addressed by previous research studies. [1] noted that among the different harmful developments, chest infection is possibly the most basic type of malignancy. Chest harmful development is a well-known illness kind that is a hot research topic with enormous potential [2].

Clinical consideration companies are encouraging unusual assistance for clinical experts in unique situations by utilising data science and AI estimates. Nowadays, detecting a case of chest injury is a considerable challenge, as plans range in shape, surface, and other clinical aspects. As a result, the clinical consideration sector is concentrating more on developing a practical AI estimations application [3].

Previously, a group of researchers concentrated on engaging image examination to discriminate between chest illnesses for separating the dangerous developments have infected past the chest, numerous organs and lymph nodes in the vicinity [3-5], and cell science [6-8] utilising unique but little datasets derived from computation appraisal challenges [9-12].

Radiology professionals frequently struggle with mammography mass lesion labelling, which can lead to unneeded and costly breast biopsies. The paper's implementation was evaluated using three publicly available benchmark datasets: the DDMS, INbreast, and BCDR databases for training and testing, and the MIAS dataset for testing only. The results showed that when PCNN is paired with CNN, it outperforms other approaches for the same publicly available datasets.[1]

If the mammographic breast tissue is dense, the federal law requires patient notice because increasing sensitivity us a sign of breast cancer risk and can impair sensitivity of mammography. Our goal was to get our deep learning model externally validated using radiologist breast density evaluations in a community breast imaging practise.

3. METHODOLGY

3.1 Data Set Description

WBCD repository provided the numerical dataset. Features are composed of fine needle aspirate (FNA) of a breast mass. There are 30 features that was extracted to describe characteristics of cell nuclei present in the scanned images. The dataset consists of 569 patients ,212 have an outcome of malignancy and 357 are Benign. The classes in the dataset are separated into 2 or 4 groups, with 2 corresponding to the benign case and 4 corresponding to the malignant case.

IDC dataset consists of 162 whole mount slide images,2,77,524 patches were extracted among which 1,98,738 are non-IDC and 78,786 are IDC. They are labelled 1 and 0, where 1 indicates cells with IDC characteristics and 0 indicates cell with non-IDC characteristics.

3.2 Data Pre-processing

The term "data pre-processing" refers to the process of converting unstructured data to structured data, as well as resizing and removing undesirable data from a dataset. The dataset's missing traits are replaced by the mean value. The data is then randomly selected from the dataset to ensure that the data is circulated properly.[27]

3.3 Training and Testing Phase

This phase extracts the features from the dataset, and the testing phase will deliver new data to be examined to see how well our algorithm works and behaves when it comes to prediction. As previously stated, the dataset is divided into two pieces. To avoid fitting, cross-validation is performed. Each iteration of our approach uses a ten-fold approach to portion data,

with nine-fold used for training and the remaining for testing.[28]

3.4 Proposed convolution neural network for image dataset analysis

The entire procedure is separated into three phases, the first of which is data generation, analysis, and prediction, as shown in Figure 2.

Data from the patient, such as x-rays, mRI, and so on, has been generated utilising IoT devices. The information gathered may have come from photographs or from sensors for numerical data gathering, and it has been recorded in a database. These records can be viewed by authorised individuals at any time and from any location. Medical specialists analyse the data and use various algorithms to forecast cancer classifications.

We are utilising the CNN algorithm for analysis and prediction in this proposed model. It is vital to understand the many levels and adjustments made in each layer in order to comprehend the rest of the process. For example, as seen in Figure 2 input data should be in pixel format, which will be converted to a 3X3 image matrix in the conv 2D layer, and then to a single linear vector for classification.

Subject to appropriate channels, the CNN can obtain spatio-transient information. On the images, a planning communication is used to better comprehend the full visual information. The 4 x 4 x 3 picture is showed up in Figure 3

Multiple ConV filters are stacked on top of each other and Max pooling is performed. Using depth wise separable convolution.
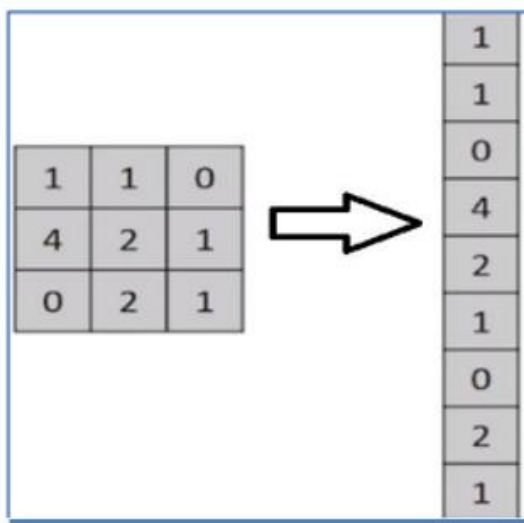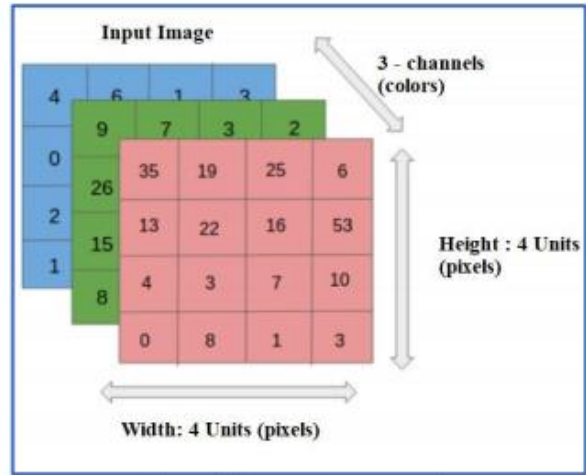

Figure 2 Data Matrix Representation

Figure 3 Image Representation

Pooled images are flattened into long vector which is fed to fully connected ANN. An artificial neural association's core design consists of a large number of interconnected neurons grouped in three layers: input, concealment, and yield. By considering enough models, this type of association usually sorts out a way to do tasks. The following is a representation of forward inducing and a single neuron:

$$Output = b_i + \sum_{j=1}^{n_x} w_{ij} x_i$$

$$(1)$$

The weight to input layer is wij, the bias value is bi, and the input value is xi. The fully connected layer uses Relu activation functions and softmax classifier.

$$Rectified\ Linear\ Unit: Activation(x) = \begin{cases} 0, & for\ x \leq 0 \\ x, & for\ x > 0 \end{cases}$$

$$(2)$$

$$Leaky\ Rectified\ Linear\ Unit: Activation(x) = \begin{cases} 0.01x, & for\ x < 0 \\ x, & for\ x \geq 0 \end{cases}$$

$$(3)$$

$$Softmax: Activation(x) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}}, where\ i = 1, 2, \ldots, j$$

$$(4)$$

After the input layer, there is a hidden layer with several sub layers, and after propagation, the loss value

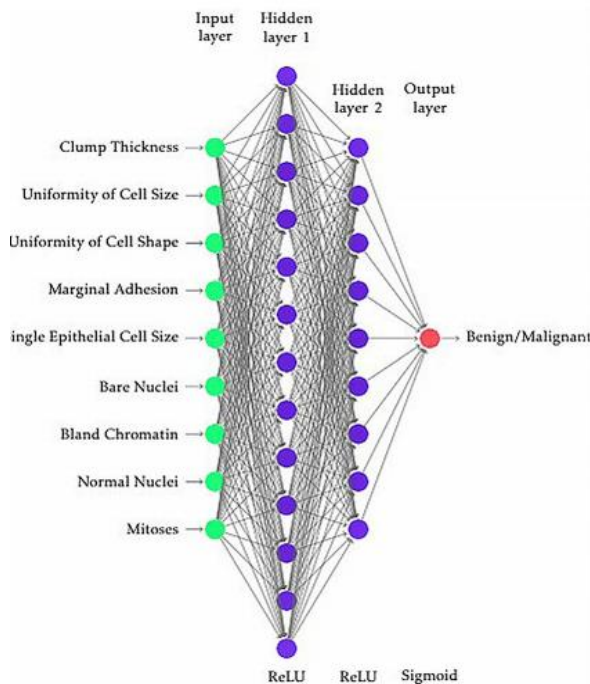is derived from the anticipated and actual values.

Figure 4 The system architecture of ANN

The proposed CNN model and suggested computation achieved the most essential precision for the dataset, which is greater than the other current algorithms. The proposed CNN architecture clearly demonstrates that it is significantly suitable for analysing the chest data, as evidenced by the aforementioned explanation.

3.5 Proposed Method for Numerical Dataset

In Numerical dataset we are using 4 different algorithms and identifying which algorithm is suitable for the dataset and gives most accurate output.

Algorithms used are SVM, KNN, Decision Tree and Naïve bayes.

3.5.1. SVM

Support vector machines are AI models that are managed. While requesting items, an assistance vector machine creates a hyperplane. The two classes are perceived by a hyperplane, which is a line on a plane. A non-probabilistic mixed direct classifier, an SVM getting ready estimation generates a model that assigns new advisers for at least one grouping or the backwards, given a social occasion of training models, each set aside as having a spot with in any case one or something in spite of two characterizations.[29]

3.5.2. K Nearest Neighbour

In the field of machine learning, KNN is a supervised machine learning algorithm. It is a classifier algorithm that is used to classify data. It is used to calculate the distances between points in the data, and subsequently votes are used to make a judgement.[30-32]

3.5.3. Decision Tree

A controlled mimicked knowledge computation is a Decision tree. A decision tree is just a tree with each handle serving as a leaf placement point or decision centre. The procedures for the Decision tree are quick and appropriate for making a decision. Inside and outside focal habitats were linked together in a decision tree.[33-35]

3.5.4. Naïve Bayes

The Nave's Bayes learning computation is influenced by social events. The restrictive likelihood hypothesis is used to establish the class of another component vector. The NB employs the arranging dataset to obtain the restrictive likelihood respect of vectors for a particular class. After dealing with each vector's likelihood restrictive assessment, the new vectors class is determined based on its chance likelihood. NB is used to depict issues that are content-related.[36-40]

4. RESULTS AND DISCUSSION

Based on the above explanation of proposed methodolgy, analysis of the numerical dataset using four different algorithms results in namely SVM, KNN, Decision tree (CART), Naïve bayes yields following results.

First the baseline algorithm analysis is done, after pipelining the dataset. Figure 5 shows the accuracy obtained by these classifier algorithms.

```
SVM: 0.619614 (0.082882) (run time: 0.191676)
CART: 0.907778 (0.029122) (run time: 0.044150)
NB: 0.940773 (0.033921) (run time: 0.008003)
KNN: 0.927729 (0.055250) (run time: 0.023999)
```

Figure 5 Initial output using baseline algorithm

Then we use standard scalar to for scaling and centering the data to boost accuracy, then along with scalarization the algorithm is applied.

Figure 6 depicts the accuracy obtained after using standard scalar.

```
ScaledCART: 0.918841 (0.036785) (run time: 0.055378)
ScaledSVM: 0.964879 (0.038621) (run time: 0.032821)
ScaledNB: 0.931932 (0.038625) (run time: 0.016234)
ScaledKNN: 0.958357 (0.038595) (run time: 0.031827)
```

Figure 6 Accuracy boost after standard scalar

In Figure 7, it can be observed SVM shows high accuracy. Along with accuracy mean and standard values are also calculated.
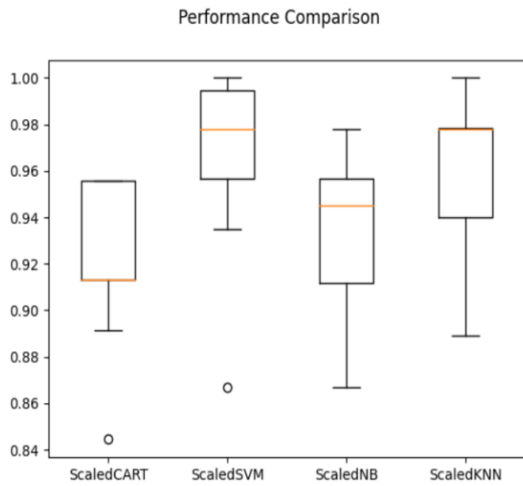

Figure 7 Plot of accuracy comparison

Thus, SVM is considered for further analysis and prediction. The train data is used to model and the test data is passed as input for prediction.

Figure 8 is providing the screenshot of accuracy obtained by using SVM.

```
The accuracy of testing data using SVM:  0.9824561403508771
The running time:  0.03998208045959473
```

Figure 8 Prediction accuracy using SVM for test data

The proposed method also talks about the prediction of breast cancer using CNN and this is applied on IDC image dataset, Figure 9 depicts the accuracy and other parameters obtained.

```
Accuracy: 0.9813647419151824
Specificity: 0.922566796121262
Sensitivity: 0.934698006851728
```

Figure 9 Results obtained using CNN on test data

These results are obtained by the model that was build using sequential API that forms a network which uses ConV filters, stacks them and apply max pooling and seperableConv2D depthwise convolutions are applied.

This model is then used to predict the results of test data.

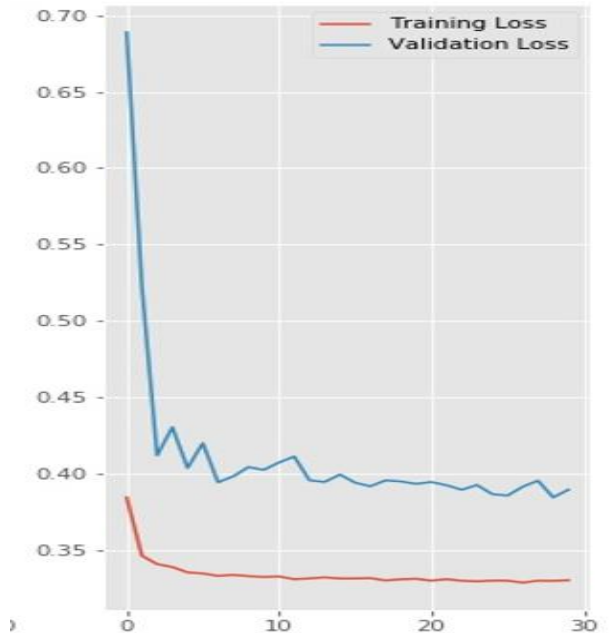Figure 10 provides plot of loss function, which shows the significant decreases.

Figure 10 Plot of training and validation loss

To facilitate the ease of interface the GUI is developed using Flask framework to connect the front end to the back-end model to process and provide prediction.

Medical practitioners can enter input values manually using patient records and on submission the record is classified as malignant or benign. Also, image can be uploaded which then will be process by the model built and the prediction is made.

Figure 11 Image data prediction: Malignant

BREAST CANCER PREDICTION

111.png

Predicted Label: Benign

Back to Home

Figure 12 Image data prediction: Malignant

Figure 11 and 12 shows the screenshots obtained for the image prediction. These images are uploaded as input in the webpage.
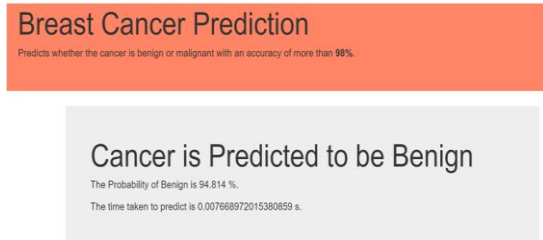

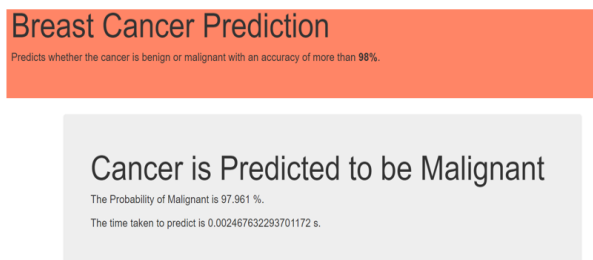**Figure 13** Numerical data record prediction: Benign

Figure 14 Numerical data record prediction: Malignant

Table 1: Comparison of Accuracies obtained by different authors

Author	Dataset	Method	Accuracy
1.Proposed methodology	Numerical dataset	SVM CART NB KNN	96.48% 91.88% 93.19% 95.8%
	IDC dataset	CNN	98.13%
2.Wang *et al.* [3]	Electronic health records	Logistic regression	96.4 %
3. Akbugday [2]	Breast Cancer Wisconsin dataset	KNN SVM	96.85% 96.85%

4. Keles, M. Kaya, [26]	Wisconsin Diagnostic Breast Cancer dataset	SVM vs KNN, decision trees and Naives bayes	up to 96.91%
5. Meteb M. Altaf [1]	DDMSdataset, INbreast dataset	PCNN	97%
6. Brian N Dontchos *et al.* [7]	Image dataset	Deep learning model	94.9%

The above-mentioned Table 1 shows the results of existing systems and the proposed system. It can be observed that there is a significant increase in accuracy obtained from the algorithms the is chosen.

5. CONCLUSION

The primary purpose of this study is to create and execute a novel computation for interpreting and orchestrating chest disease data obtained from mammography and pathology results of patient scans that is obtained and UCI repositories' cloud. Python programming, a convolutional neural association model is utilised to accomplish this, and the results are confirmed. According to the findings, the suggested CNN outperforms estimations in recognising and requesting breast cancer for image datasets. And SVM has proven to outperforms CART, NB and KNN in analysis and prediction of cancer with numerical dataset.

AUTHORS' CONTRIBUTIONS

Both authors have contributed equally to the work.

ACKNOWLEDGMENTS

REFERENCES

[1] Meteb M. Altaf- A hybrid deep learning model for breast cancer diagnosis based on transfer learning and pulse-coupled neural networks

[2] B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.

[3] Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016).

[4] Nazeri, K., Aminpour, A. &Ebrahimi, M. Two-stage convolutional neural network for breast cancer histology image classification. In International Conference Image Analysis and Recognition, 717–726 (Springer, 2018).

[5] Golatkar, A., Anand, D. &Sethi, A. Classification of breast cancer histology using deep learning. In International Conference Image Analysis and Recognition, 837–844 (Springer, 2018).

[6] Albarqouni, S. et al. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE transactions on medical imaging 35, 1313–1321 (2016).

[7] B. N. Dontchos, A. Yala, R. Barzilay, J. Xiang, C. D. Lehman, External validation of a deep learning model for predicting mammographic breast density in routine clinical practice, Acad. Radiol., 28 (2020), 475-480.

[8] Rao, S. Mitos-rcnn: A novel approach to mitotic figure detection in breast cancer histopathology images using region based convolutional neural networks. arXiv preprint arXiv:1807.01788 (2018).

[9]. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 318, 2199–2210 (2017).

[10] Bándi, P. et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE Transactions on Med. Imaging (2018).

[11] Litjens, G. et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. GigaScience 7, giy065 (2018).

[12] Aresta, G. et al. Bach: Grand challenge on breast cancer histology images. arXiv preprint arXiv:1808.04277 (2018).

[13] Gouda I Salama, M Abdelhalim, and MagdyAbd-elghanyZeid. 2012. Breast cancer diagnosis on three different datasets using multiclassifiers. Breast Cancer (WDBC) 32, 569 (2012), 2.

[14] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/] (1992).

[15] Abien Fred Agarap. 2017. A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data. arXiv preprint arXiv:1709.03082 (2017).

[16] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, HolgerSchwenk, and Yoshua Bengio, (2014), "Learning phrase representations using RNN encoder-decoder for statistical machine translation", arXiv preprint arXiv:1406.1078 (2014).

[17] C. Cortes and V. Vapnik, (1995), "Support-vector Networks.Machine Learning 20.3", (1995), 273–297. https://doi.org/10.1007/BF00994018.

[18] Gönen, M.; Alpaydın, E. Multiple kernel learning algorithms. J. Mach. Learn. Res. 2011, 12, 2211–2268.

[19] Ferroni, P.; Zanzotto, F.M.; Scarpato, N.; Riondino, S.; Nanni, U.; Roselli, M.; Guadagni, F. Risk assessment for venous thromboembolism in chemotherapy treated ambulatory cancer patients: A precision medicine approach. Med. Dec. Mak. 2017, 37, 234–242.

[20] Ferroni, P.; Roselli, M.; Zanzotto, F.M.; Guadagni, F. Artificial Intelligence for cancer-associated thrombosis risk assessment. Lancet Haematol.2018, 5, e391.

[21] Cristianini, N.; Shawe-Taylor, J. An Introduction to Support Vector Machines and other kernel-based learning methods.Ai Magazine 2000, 22, 190.

[22] Matyas, J. Random optimization. Automat. Rem. Control 1965, 26, 246–253.

[23] Jain A, Levy D. 2016. Breast mass classification using deep convolutional neural networks. In: 30th conference on neural information processing systems (NIPS 2016). Barcelona, Spain. 1_6.

[24] Jiang F. 2017. Breast mass lesion classification in mammograms by transfer learning. In: ICBCB '17. Hong Kong, 59_62 DOI 10.1145/3035012.3035022.

[25] Ragab DA, Sharkas M, Marshall S, Ren J. 2019. Breast cancer detection using deep convolutional neural networks and support vector machines.PeerJ 7:e6201 http://doi.org/10.7717/peerj.6201.

[26] Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

[27] Z. Guo, L. Tang, T. Guo, K. Yu, M. Alazab, A. Shalaginov, "Deep Graph Neural Network-based Spammer Detection Under the Perspective of Heterogeneous Cyberspace", Future Generation Computer Systems, https://doi.org/10.1016/j.future.2020.11.028.

[28] Y. Sun, J. Liu, K. Yu, M. Alazab, K. Lin, "PMRSS: Privacy-preserving Medical Record Searching Scheme for Intelligent Diagnosis in IoT Healthcare", IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3070544.

[29] K. Yu, L. Tan, L. Lin, X. Cheng, Z. Yi and T. Sato, "Deep-Learning-Empowered Breast Cancer Auxiliary Diagnosis for 5GB Remote E-Health," IEEE Wireless Communications, vol. 28, no. 3, pp. 54-61, June 2021, doi: 10.1109/MWC.001.2000374.

[30] K. Yu, L. Tan, S. Mumtaz, S. Al-Rubaye, A. Al-Dulaimi, A. K. Bashir, F. A. Khan, "Securing Critical Infrastructures: Deep Learning-based Threat Detection in the IIoT", IEEE Communications Magazine, 2021.

[31] K. Yu, Z. Guo, Y. Shen, W. Wang, J. C. Lin, T. Sato, "Secure Artificial Intelligence of Things for Implicit Group Recommendations", IEEE Internet of Things Journal, 2021, doi: 10.1109/JIOT.2021.3079574.

[32] H. Li, K. Yu, B. Liu, C. Feng, Z. Qin and G. Srivastava, "An Efficient Ciphertext-Policy Weighted Attribute-Based Encryption for the Internet of Health Things," IEEE Journal of Biomedical and Health Informatics, 2021, doi: 10.1109/JBHI.2021.3075995.

[33] Puttamadappa, C., and B. D. Parameshachari. "Demand side management of small scale loads in a smart grid using glow-worm swarm optimization technique." Microprocessors and Microsystems 71 (2019): 102886.

[34] Rajendran, Ganesh B., Uma M. Kumarasamy, Chiara Zarro, Parameshachari B. Divakarachari, and Silvia L. Ullo. "Land-use and land-cover classification using a human group-based particle swarm optimization algorithm with an LSTM Classifier on hybrid pre-processing remote-sensing images." Remote Sensing 12, no. 24 (2020): 4135.

[35] Subramani, Prabu, Ganesh Babu Rajendran, Jewel Sengupta, Rocío Pérez de Prado, and Parameshachari Bidare Divakarachari. "A block bi-diagonalization-based pre-coding for indoor multiple-input-multiple-output-visible light communication system." Energies 13, no. 13 (2020): 3466.

[36] Hu, Liwen, Ngoc-Tu Nguyen, Wenjin Tao, Ming C. Leu, Xiaoqing Frank Liu, Md Rakib Shahriar, and SM Nahian Al Sunny. "Modeling of cloud-based digital twins for smart manufacturing with MT connect." Procedia manufacturing 26 (2018): 1193-1203

[37] Seyhan, Kübra, Tu N. Nguyen, Sedat Akleylek, Korhan Cengiz, and SK Hafizul Islam. "Bi-GISIS KE: Modified key exchange protocol with reusable keys for IoT security." Journal of Information Security and Applications 58 (2021): 102788.

[38] Pham, Dung V., Giang L. Nguyen, Tu N. Nguyen, Canh V. Pham, and Anh V. Nguyen. "Multi-topic misinformation blocking with budget constraint on online social networks." IEEE Access 8 (2020): 78879-78889.

[39] Arun, M., E. Baraneetharan, A. Kanchana, and S. Prabu. "Detection and monitoring of the asymptotic COVID-19 patients using IoT devices and sensors." International Journal of Pervasive Computing and Communications (2020).

[40] Kumar, M. Keerthi, B. D. Parameshachari, S. Prabu, and Silvia liberata Ullo. "Comparative Analysis to Identify Efficient Technique for Interfacing BCI System." In IOP Conference Series: Materials Science and Engineering, vol. 925, no. 1, p. 012062. IOP Publishing, 2020.