Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6 –Issue 2,August 2023
Paper:50

Breast Cancer Detection and Classification using Random Forest and KNN on Kaggle Breast Cancer Data using Mammography

Sanjay S Tippannavar
Department of Electronics and Communication Engineering
JSS Science and Technology University,
Mysuru, India
Sanjayu2345@gmail.com

Gayathri S
Department of Electronics and Communication Engineering
JSS Science and Technology University,
Mysuru, India
sgmurthy_65@sjce.ac.in

Swetha S
Department of Electronics and Communication Engineering
JSS Science and Technology University,
Mysuru, India
shwetha2212s@gmail.com

***Abstract*—The idea behind this initiative is to eliminate the laborious process of finding breast cancer. This method is mechanised using CAD and the mammographic picture of the breast is used as the input in order to go around many of the superstations in society. For processing the user-provided database using a computer device in our system, we suggested a machine learning technique. With the help of this programme, women may avoid breast cancer instances while remaining in the comfort of their own homes. This would increase the diagnostic results' openness, removing any room for patient scepticism. The technology will allow users and computers to communicate directly. The screening picture of the mammography must be verified by an expert, but occasionally patients are unable to see them in person, causing the appointment to be delayed. The usual method of diagnosis would need at least a week to complete. As a consequence, the diagnosis will take longer, which might lead to lengthy therapy or perhaps even death. To solve these issues, we turn to CAD so that early diagnosis would make it simpler and less laborious for both patients and medical professionals to treat the condition**

Keywords*—Breast Cancer Detection, CNN, Machine Learning, KNN, Image Processing, Mammography.

I. INTRODUCTION

One of the most frequent malignancies in women is breast cancer (BC). An accurate system-based early detection of breast cancer and metastases among patients may raise patient survival to >86%. Breast cells begin to develop into malignant lumps that are carcinogenic during the beginning stages of breast cancer. Malignant tumours may be mistakenly diagnosed by doctors as benign tumours, which are non-cancerous. To accurately diagnose breast cancer, a computer-aided detection (CAD) system that employs machine learning is required. These computer-aided detection (CAD) devices may help in the early diagnosis of breast cancer. The survival rate rises when breast cancer is found early enough so that better care may be given. To increase breast cancer survival and lower the high mortality rate of breast cancer, early diagnosis of the disease is essential. Around 50% of patients will acquire distant metastases throughout their follow-up period despite early identification and the introduction of new therapies. According to WHO, India has over 1.5 million breast cancer patients annually, with 500,000 women expected to pass away from the disease in 2015 alone and approximately 9.6 million expected to do so in 2018. Given the large difference between incidence and death, early breast cancer diagnosis requires performance improvement. Thus, advancements in current methods are necessary to detect breast cancer at an early stage. There are two distinct types of breast cancer:

The cells that make up a benign tumour are not malignant. It won't intrude on nearby tissues or spread to other sections of the body (metastasize). A benign tumour is less problematic unless it is pressing on or injuring nearby blood vessels, nerves, or tissues. Lymphomas and uterine fibroids are two examples of benign tumours. Benign tumours may need to be surgically removed. They might grow to be fairly large and weigh pounds. When benign tumours are excised, they often don't come back, but if they do, they usually do so in the same place.

Malignant describes a tumor's capacity to invade surrounding tissues and the presence of malignant cells inside it. When cancer cells move to other physiological tissues after getting into the bloodstream or lymph nodes, this is referred to as metastasis. Cancer may manifest everywhere in the body, including the breast, intestines, lungs, reproductive organs, blood, and skin. Breast cancer may spread to other organs like the liver or bones if it has reached the lymph nodes. Hence, breast cancer-derived tumours may form there. A biopsy of these tumours may reveal characteristics of the primary breast cancer tumour. The researchers' primary areas of interest were studies using artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and k-nearest neighbour (k-NNs) techniques. The Wisconsin breast cancer database was also utilised. In order to identify breast cancer at an early stage, we will analyse mammography pictures using various machine learning approaches and artificial intelligence algorithms. Breast cancer contributes significantly to the world's many cancer cases, but many of these cases go undetected for a variety of reasons, one of which is women's reluctance to undergo screening, which is the first step in breast cancer detection.

The main objectives of the proposed work are to;

- Dataset Collection or acquisition.
- Dataset Preprocessing.
- Training and testing the dataset using machine learning algorithm.
- To develop an accurate system for the early detection of cancer.
- To compare the system with the existing system.

According to experts, women over the age of 40 are particularly susceptible to developing this type of cancer, so in order to lower their risk, it's preferable to undergo screening. Second, a lot of individuals worry about what will happen if the test is positive. Nevertheless, early-stage cancers are simpler to treat than later-stage tumours, and the likelihood of survival is better. The literature study clearly demonstrates the widespread usage of machine learning techniques for classifying breast cancer diagnoses. Moreover, the Wisconsin breast cancer database was utilised. The researchers produced a simple and understandable catalogue of data. The Wisconsin Breast Cancer Database (WBCD) was used by the researchers to demonstrate that several algorithms have attained extremely high accuracy, but new algorithms still needed to be developed.

II. Related Work

Medical data has been studied by Medisetty Hari Krishna and others using a variety of data mining and machine learning approaches. On the Wisconsin Breast Cancer (original) datasets, they used four main algorithms: Support vector classifier, Random Forest, Gradient Boosting, Naive Bayes, Cart Model, Neural Network, and Linear Regression algorithm, comparing each algorithm's efficiency and effectiveness in terms of accuracy, precision, sensitivity, and specificity to determine which one provided the best classification accuracy. Support vector has shown its effectiveness in predicting and diagnosing breast cancer; it delivers the best results in terms of accuracy and low error rate [1].

A CV approach like k-fold cross validation should be used, according to a research by Abien Fred on the Wisconsin Diagnostic Dataset. The use of such an approach will provide a more precise measurement of model prediction performance and help in choosing the best hyper-parameters for ML algorithms [2].

Several machine learning methods, including Support Vector Machine (SVM) and Relevance Vector Machine, were evaluated in a survey conducted by Gayathri and colleagues (RVM). They discovered that numerous scientists have used neural network algorithms to forecast diseases, particularly breast cancer. The usage of Relevance Vector Machine (RVM) will probably become considerably more beneficial for identifying breast cancer if research on RVM continues [3].

Using three distinct experiments and the breast cancer dataset, Habib Dhahri and colleagues developed a machine learning method to tackle the issue of automated identification of breast cancer. The three most well-known evolutionary algorithms may attain the same performance after effective setup, they demonstrated in the first test. The second experiment examined the idea that combining approaches for features selection enhances accuracy performance. They discovered how to automatically create the machine learning supervised classifier from the third experiment. The suggested model seems to be well suitable for automating the detection of breast cancer on the one side and controlling parameter setting of machine learning algorithms on the other [4].

The paper's unique network, which was suggested by Leena Pal and others, does not have any permanent cable communication infrastructure or other network hardware. Future wireless networks known as mobile ad hoc networks (MANETs) are made up completely of mobile nodes and operate without any centralised management. These networks' nodes provide network control and routing functions as well as user and application traffic generation. DSR protocol is more useful in small cluster sizes, but as cluster sizes grow, AODV protocol exhibits dramatic changes in performance and becomes more useful, whereas DSDV assessment findings are unfavourable in contrast to the other two reactive routing protocols [6].

Opinion Mining and Sentiment Analysis is a brand-new field of study that was suggested by Sakshi Koli. Sentiment analysis is the technique of identifying the feelings, opinions, and emotions that readers have towards a text that falls under the categories of blog views, article reviews, product reviews, social media buzzing, etc. Support vector machines performed best in recall, whereas native bayes performed best in accuracy. Out of four machine learning techniques, improved j48 is the best classifier. 86.6% for the enhanced J48 decision tree and% for the J48 decision tree were the different classifier accuracy results. Improved j48 emerged as the top classification strategy in a comparison of machine learning techniques [7].

The study by Adam Nover and colleagues describes how to find breast cancer utilising techniques including self-examinations, clinical breast exams, full-field digital mammography, computer-aided detection, modality employing ultrasound, and magnetic resonance imaging (MRI). a brief summary of the state of the art in early detection and screening for breast cancer. Moreover, various upcoming technologies that could complement or take the place of the present modalities were highlighted [9].

Using 1112 matched case-control couples in screened populations, Norman and Boyd suggested three nested case-control studies. Examining the relationship between the percentage of density recorded in the baseline mammography and breast cancer risk, as determined by the technique of cancer diagnosis, the length of time since screening began, and the patient's age. The National Breast Screening Study (NBSS) was a randomised study combining mammography and physical examination for breast cancer screening. 15,16 At its screening facilities, the Screening Mammography Program of British Columbia (SMPBC) solely employs mammography, whereas the Ontario Breast Screening Program (OBSP) employs both mammography and physical examination [10].

Vikas Chaurasia and Saurabh Pal presented the approach, which analyses the effectiveness of three traditional decision tree classifiers with findings that are acceptable for direct interpretation. By optimising the smallest subset with two items at each iteration, the primary concept is taken from solving a dual quadratic optimization problem. SMO has the benefit of being straightforward and analytical to adopt. Instances are categorised using K-Nearest Neighbor (KNN) classification depending on how similar they are. Each example is treated as a point in a multidimensional space, and each category is based on its closest neighbours. For closest neighbours, k might have different values. In order to decide how to categorise an unknown instance, this defines how many examples should be taken into account as neighbors [11].

In order to categorise breast cancer, Mohammad Abdula presented the article that is now being read. A simple classifier based on the Bayes theorem is known as naive Bayesian (NB). In literature, there have been many applications. This research introduced a novel NB (weighted NB) classifier and demonstrated its use in the identification of breast cancer. A weighted NB was suggested and its use in the identification of breast cancer was reported in order to address the shortcomings of the NB classifier. Based on the performed studies, the weighted NB used was able to achieve values of 99.11% sensitivity, 98.25% specificity, and 98.54% accuracy [12].

The accuracy of thermography has substantially improved as a result of developments in infrared (IR) cameras that are used to collect thermal pictures of the breast and in computational tools that are used to precisely predict heat flow inside the breast, according to Satish and colleagues. To get a high resolution picture of the breast at various crosssections during an MRI, a powerful magnetic field and pulsating radio waves are used. To provide a more accurate picture of the breast, a contrast agent is applied [14].

A novel computer-aided detection (CAD) approach is suggested in the study by Dina A. Ragab for categorising benign and malignant mass tumours in breast mammography pictures. Two segmentation techniques are used in this CAD system. Although the second strategy employs the concept of threshold and region based, the first approach includes manually selecting the area of interest (ROI). Further breast abnormalities like MCs are found using the suggested CAD system [15].

III. Methodology

One of the most frequent malignancies in women is breast cancer (BC). An accurate system-based early detection of breast cancer and metastases among patients may raise patient survival to >86%. Breast cells begin to develop into malignant lumps that are carcinogenic during the beginning stages of breast cancer. These computer-aided detection (CAD) devices may help in the early diagnosis of breast cancer

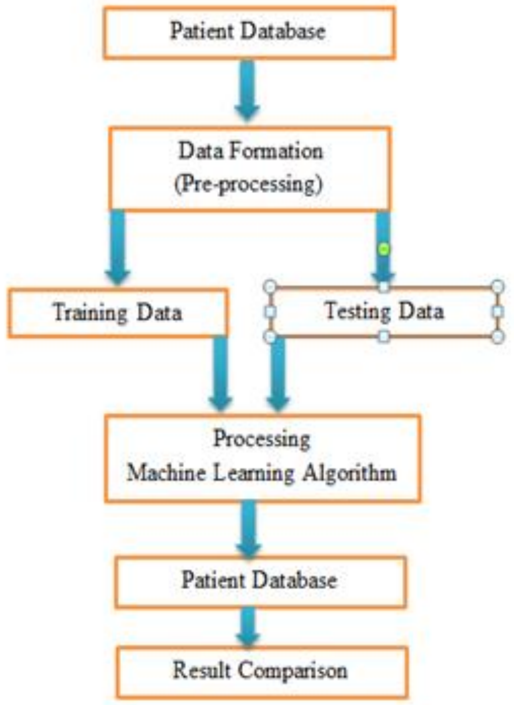


Fig. 1. Block diagram of the proposed system.

The first phase is gathering information from the retrospective cohort of patients who have received a breast cancer diagnosis. Data pre-processing is done to make a dataset better so that it can provide clean data that is good for modelling. Data cleaning entails reducing noise and irregularities to improve the data's quality. The machine learning model is trained using a training dataset. At the prediction step, the validation dataset is employed.

On the other side, feature extraction minimises the number of dimensions by converting characteristics in a high dimensional space to fewer dimensions. The two most common feature selection methods are CFS and RFE approaches. PCA and LDA are the two most used feature extraction methods (linear discriminant analysis). Algorithms for feature selection may often be divided into the following classes: Wrapper feature selection approaches combine a random feature subset to train the model. As a result, the elements of the higher error rate combination were eliminated while the lower error rate combination was maintained.

Filter Methods: The filter feature selection technique chooses features by assigning a score to each feature. With big datasets, filter feature selection might be utilised as a pre-treatment step for the wrapper approach.

Embedded Methods: Embedded feature selection methods choose the features while building the model.

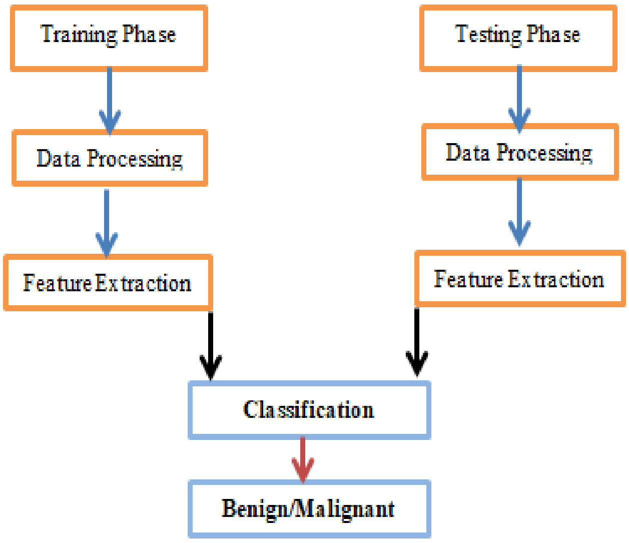


Fig. 2. Training and Testing Phase

The accuracy of prognosis prediction is improved by the use of feature selection/extraction techniques and classification-based algorithms. ANN, SVM, KNN, ELM, and MLP

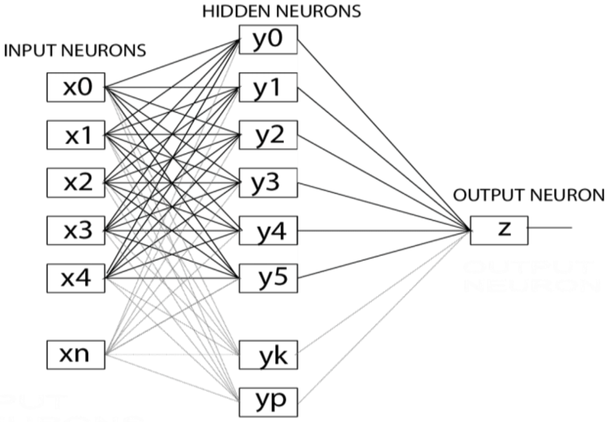
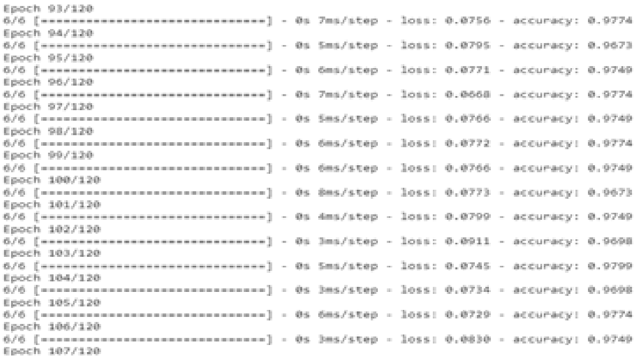


Fig. 3. SVM architecture of the proposed system

IV. Results and Discussions

In this section, the classifier's operation is reviewed, confirmed, and verified along with a comparison to earlier, more current work. In the picture in Fig. 4, the suggested work's epochs are shown.



```
Epoch 93/120
6/6 [==============================] - 0s 7ms/step - loss: 0.0756 - accuracy: 0.9774
Epoch 94/120
6/6 [==============================] - 0s 5ms/step - loss: 0.0795 - accuracy: 0.9673
Epoch 95/120
6/6 [==============================] - 0s 6ms/step - loss: 0.0771 - accuracy: 0.9749
Epoch 96/120
6/6 [==============================] - 0s 7ms/step - loss: 0.0668 - accuracy: 0.9774
Epoch 97/120
6/6 [==============================] - 0s 5ms/step - loss: 0.0766 - accuracy: 0.9749
Epoch 98/120
6/6 [==============================] - 0s 6ms/step - loss: 0.0772 - accuracy: 0.9774
Epoch 99/120
6/6 [==============================] - 0s 6ms/step - loss: 0.0766 - accuracy: 0.9749
Epoch 100/120
6/6 [==============================] - 0s 8ms/step - loss: 0.0773 - accuracy: 0.9673
Epoch 101/120
6/6 [==============================] - 0s 4ms/step - loss: 0.0799 - accuracy: 0.9749
Epoch 102/120
6/6 [==============================] - 0s 3ms/step - loss: 0.0911 - accuracy: 0.9698
Epoch 103/120
6/6 [==============================] - 0s 5ms/step - loss: 0.0745 - accuracy: 0.9799
Epoch 104/120
6/6 [==============================] - 0s 3ms/step - loss: 0.0734 - accuracy: 0.9698
Epoch 105/120
6/6 [==============================] - 0s 6ms/step - loss: 0.0729 - accuracy: 0.9774
Epoch 106/120
6/6 [==============================] - 0s 3ms/step - loss: 0.0830 - accuracy: 0.9749
Epoch 107/120
```

Fig. 4. Epochs of the proposed system

Fig. 4 represents the Epochs of the SVM which has the accuracy range of 97.5%.

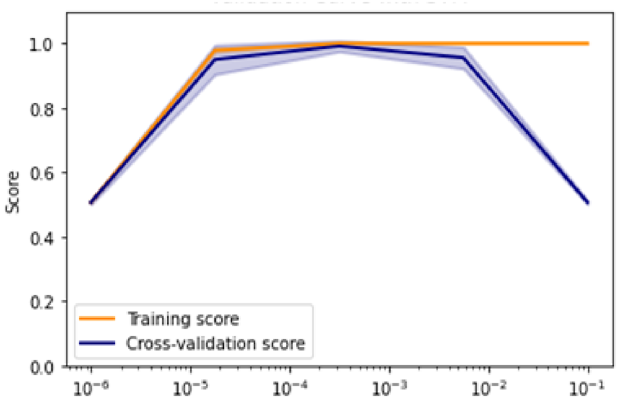


Fig. 5. Validation Curve

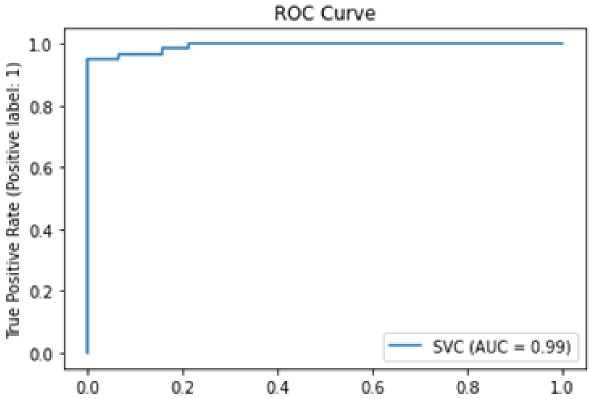


Fig. 6. ROC Curve

Fig. 6 represents the ROC curve of the classified dataset.

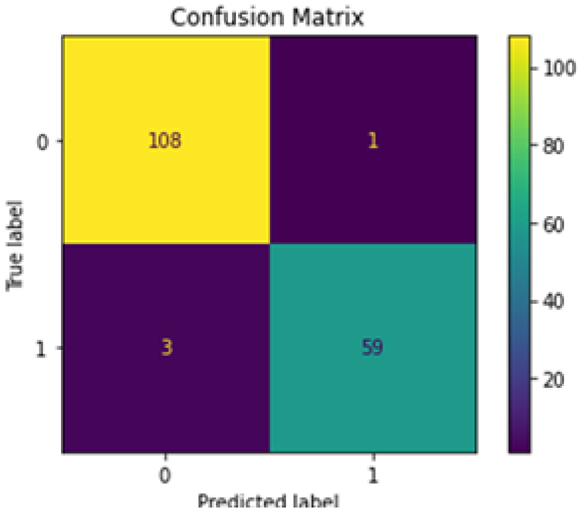


Fig. 7. Confusion Matrix of the proposed system.

Fig. 7 represents the confusion matrix which has classified datasets in which 108 data's are benign or non-cancerous and 59 datasets are malignant or cancerous datasets.

TABLE I. METRICS REPORT

Accuracy	96.49%
Precision	100%
Recall	90.47%
F1_score	95%

TABLE II. COMPARISON OF PROPOSED WORK WITH RELATED WORK

Method / algorithm	DNNS[1]	SVM[Proposed]
Dataset	Kaggle	Kaggle
Accuracy	95.61	97.66
Error rate	4.39	2.34

V. CONCLUSIONS AND FUTURE SCOPE

Scientific practitioners may benefit from reliable and accurate prediction effects with the use of a good breast cancer diagnostic model. Breast cancer is curable if it is found early, which may greatly improve the prognosis. With 97% accuracy, the suggested approach divides tumours into malignant and benign types based on traits found in cell pictures. As a result, it may be effectively used to the identification and prevention of breast cancer. Eminent tools for inference in this field may be provided by the integration of multidimensional data with various categorization, feature selection, and dimensionality reduction algorithms. The SVM classifier, which trains models to classify cancer patients based on their diagnosis, is then utilised for classification. The efficiency of the mode is shown by experimental findings. To sum up the created approach, the first phase is based on image enhancement grey level information and segments the breast tumour. Morphological characteristics are retrieved from each tumour area extract to classify the breast tumour.

- More study in this area may be done to boost the ability of classification systems to make predictions using various factors.
- The selected strategy may be transformed into a projected pragmatic model, enabling clinicians to quickly consult in order to diagnose and recognise breast cancer.
- These systems may help a patient with breast cancer choose the best course of therapy.
- Patients may get a variety of therapies depending on the stage of their breast cancer; data mining and machine learning can be extremely helpful in selecting the course of therapy to be taken by extracting information from these appropriate databases

REFERENCES

[1] Anji Reddy Vaka, Badal Soni, Sudheer Reddy K., Breast cancer detection by leveraging Machine Learning, ICT Express, Volume 6, Issue 4, 2020

[2] Jun Deng, PhD Professor Department of Therapeutic Radiology Yale University School of Medicine November 4, 2017, Ohio River Valley Chapter Fall Symposium, Indianapolis, IN International Journal of Advances in Science Engineering and Technology.

[3] Nassif, A. B., Talib, M. A., Nasir, Q., Afadar, Y., & Elgendy, O. (2022). Breast cancer detection using artificial intelligence techniques: A systematic literature review. Artificial Intelligence in Medicine, 102276.

[4] Allugunti, V. R. (2022). Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. International Journal of Engineering in Computer Science, 4(1), 49-56.

[5] International Journal for Research in Applied Science& Engineering Technology ISSN: 0193-4120 Page No. 6667- 6670Volume 82 Page Number: 6667 - 6670 Publication Issue: January-February 2020. UG Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, Tamilnadu.

[6] Volume 6, Issue 6, June 2016 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper.

[7] Sanjay S Tippannavar ,Surya M S, Yadav, Pilimgole & Salman, Mohammed & M, Ajay. (2022). Hand Gesture based Speech Recognition system for Hard of Hearing People. International Journal for Research in Applied Science and Engineering Technology. 10. 545-551. 10.22214/ijraset.2022.47301.

[8] Melekoodappattu, J. G., Dhas, A. S., Kandathil, B. K., & Adarsh, K. S. (2022). Breast cancer detection in mammogram: Combining modified CNN and texture feature based approach. Journal of Ambient Intelligence and Humanized Computing, 1-10.

[9] Trang, N. T. H., Long, K. Q., An, P. L., & Dang, T. N. (2023). Development of an Artificial Intelligence-Based Breast Cancer Detection Model by Combining Mammograms and Medical Health Records. Diagnostics, 13(3), 346.

[10] Birchha, V., & Nigam, B. (2023). Performance Analysis of Averaged Perceptron Machine Learning Classifier for Breast Cancer Detection. Procedia Computer Science, 218, 2181-2190.

[11] Sechopoulos, I., Teuwen, J., & Mann, R. (2021, July). Artificial intelligence for breast cancer detection in mammography and digital

breast tomosynthesis: State of the art. In Seminars in Cancer Biology (Vol. 72, pp. 214-225). Academic Press.

[12] Allugunti, V. R. (2022). Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. International Journal of Engineering in Computer Science, 4(1), 49-56.

[13] S. S. Tippannavar, Y. S D and P. K M, "SDR – Self Driving Car Implemented using Reinforcement Learning & Behavioural Cloning," 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC), Mysore, India, 2023, pp. 1-7, doi: 10.1109/ICRTEC56977.2023.10111893.

[14] S. S. Tippannavar, V. Mishra, Y. S D, R. R. Gowda, S. H R and A. M, "Smart Transformer - An Analysis of Recent Technologies for Monitoring Transformer," 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC), Mysore, India, 2023, pp. 1-11, doi: 10.1109/ICRTEC56977.2023.10111875.

[15] S. S. Tippannavar, K. M. Puneeth, S. D. Yashwanth, M. P. Madhu Sudan, B. N. Chandrashekar Murthy and M. S. Vinay Prasad, "SR2 - Search and Rescue Robot for saving dangered civilians at Hazardous areas," 2022 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, 2022, pp. 21-26, doi: 10.1109/CENTCON56610.2022.10051203.