A Systematic Study on The Techniques Used for The Prediction of Breast Cancer

B. LAVANYA[1], I. DILSHAD BANU[2]
[1] *Associate Professor, Department of Computer Science, University of Madras, Chennai, India.*
[2] *Mphil (Computer Science), Department of Computer Science, University of Madras, Chennai, India.*

Abstract*— Breast cancer is regarded as the most serious problem among women. This is a severe threat to women's society, similar to climate change, urbanization, food culture, etc. The tools available in data mining provide a significant contribution to the field of medical diagnostics in terms of accurate disease prediction. The probability of saving a breast cancer patient is majorly dependent on its stage detection and initiation of treatment. To address these challenges, this paper involves an overview of several data mining approaches that are specifically applied to breast cancer prediction. This paper shows the comparison of diverse classification and clustering algorithms. Varied classification algorithms and clustering algorithms are used in this survey paper. We have compared the performances of various machine learning and deep learning techniques such as Support Vector Machine, Decision Trees, Logistic Regression, Naive Bayes, Convolutional Neural Networks (CNN) etc. We have used publicly available breast cancer datasets, for testing several approaches for autonomous tumour classification. However, the proposed Convolutional Neural Networks (CNN) model classifier has the highest classification accuracy, according to experimental results.

Indexed Terms*— Breast cancer; Accurate prediction; Treatment; Early diagnosis; Convolutional Neural Networks (CNN)

I. INTRODUCTION

Breast cancer has passed the prostate as the most common malignancy in girls in current years. The preliminary step in the improvement of breast cancer is for breast cells to develop. Healthful cells divide and spread extra slowly than those cells, resulting in a lump or mass. These cells can pass from your breast to your lymph nodes and someplace else at some stage in your body. Breast cancer is not unusual in younger girls, particularly the ones below thirty years old. It desires to be sorted right now. Systemic and nearby remedy for breast cancer is divided into corporations. Chemotherapy and hormone remedy systemic treatments, while surgical treatment and radiation are local. These sorts of remedies are usually used to attain excellent consequences.

No matter the fact that breast cancer is the second largest reason for demise among women. 97% of women live for five years or longer if they are diagnosed early. Therefore, it should be predicted at an early stage so that appropriate remedies be given. Predicting breast cancer using diverse information mining algorithms is crucial in this regard. The use of system gaining knowledge of and other techniques are described in this survey paper with compiled records from a ramification of breast cancer-related papers. According to Pei Liu et al., [1] XG improves the finest survival evaluation, that's used to forecast breast cancer improvement. The proposed EXSA method turned into constructed on XG boost in system gaining knowledge of and the CPH model in survival analysis. The model completed further throughout 5 and ten years, with a C-index of zero.83454 and AUCs of zero.83851 and 0.78155. The EXSA prognostic version has excessive discriminative functionality in predicting breast most cancers disorder progression hazard, consistent with the results of the experiments.

Also, Farhad Imani et al., [2]. Explained the RSF technique aids in the estimation and prediction of the survival function via sampling and bootstrapping into huge datasets with an ensemble model of survival trees. Breast most cancers recurrence is shown to have a 7% recurrence fee, with the number of recurrences starting from 2 to 6. And in step with Yash Mate et al., [3] The accuracy of most device learning algorithms improves whilst critical qualities are determined by the feature choice procedures including Pearson's coefficient, Chi-square test, Logistic regression, Random Forest, RFE, and light gradient boosting. To summarize, the Ada boost Classifier had a maximum

accuracy of 97 % with feature selection, but the Tree Classifier had a maximum accuracy of 96.2 % with characteristic selection by deliberating 20 key features for breast cancer prediction.

The goal of the paper "A Systematic Study on the techniques used for Breast Cancer Prediction" is to show a comparison of recent Machine Learning research papers. This research also includes a discussion on breast cancer diagnosis and prognosis concerns, as well as a wide range of approaches. The major goal was to employ non-invasive and painless data mining classification algorithms to predict and diagnose breast cancer early, even if the tumour is small. Table:1 Shows the comparison analysis of different techniques used for the prediction of Breast Cancer.

II. DATA MINING TECHNIQUES USED FOR THE PREDICTION OF BREAST CANCER

MACHINE LEARNING:

Machine Learning algorithms are classified as either supervised or unsupervised. Through labelling, both input data and intended output data are provided for supervised learning algorithms. Unsupervised algorithms operate with data that is neither classed nor labelled. For example, an unsupervised algorithm could sort the unsorted data into groups based on their similarities and differences.

Many ML techniques, such as transfer learning and active learning, use semi-supervised algorithms. Active learning allows an algorithm to query the user or another source for further information, whereas transfer learning employs knowledge obtained from performing one task to help tackle a separate but related problem. In cases when labelled data is scarce, both systems are routinely employed.

Reinforcement learning is used to find solutions and strategies for complex problems based on the trial-and-error method. It also explains one of the three ways of machine learning. In comparison to other learning approaches, the agent requires no data material to be trained (the learning system). AlphaGo,a game developed by Google, is a well-known example of reinforcement learning in action.

MACHINE LEARNING TECHNIQUES:

2.1SUPPORT VECTOR MACHINE:

Support Vector Machine is one of the supervised machine learning classification techniques that is widely applied in the field of cancer identification and prognosis. SVM is functioned by choosing the critical samples from all categories. These samples are called support vectors. These classes are separated by generating a linear function that divides them broadly as possible using these support vectors.

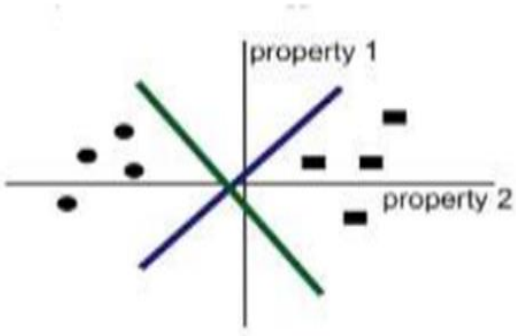

Fig.1 SVM generated hyper-planes.

2.2 RANDOM FOREST:

To ensemble a forest of trees, RF carries numerous decision trees with him. Each iteration of the RF methodology comprises selecting one random sample of size N from the data set with replacement and another random sample from the predictors without replacement. Following that, the data is partitioned. The remaining information is then removed. Depending on how many trees are required, these stages are repeated at different times. Finally, the trees that classify the observation into one of several categories are counted.

2.3 LOGISTICS REGRESSION:

LR is a supervised learning approach with a larger number of dependent variables. The binary form of this algorithm's response. Logistics regression can produce a consistent result from a set of data. A statistical model with binary variables is used in this approach.

2.4 DECISION TREE:
The classification and regression models are used to create the decision tree. The data set is broken down into smaller subsets. With this limited set of data, the most precise predictions may be made. CART, C4.5, C5.0, and conditional tree are all decision tree methods.

ADVANTAGES:
The decision tree provides the following advantages in Machine Learning: It is comprehensive in that it considers every conceivable outcome of a decision and follows each node to its conclusion Specific. Each problem, decision, and result in a Decision Tree is assigned a specific value (s).

2.5 K-NEAREST NEIBHOUR (KNN):
This algorithm is used to recognise patterns. It is an effective method for predicting breast cancer. Each class has been given equal weight to the pattern. From a huge dataset, K -Nearest Neighbour extracts similar highlighted data. We classify a large dataset based on feature similarity.

DEEP LEARNING:
A subfield of machine learning involves deep learning. It is a method of learning that is not monitored. It is possible to have data that is unstructured or un labelled. A deep neural network is described as having more than two hidden layers. The first is the input layer, and the second is the output layer. The intermediate layer is known as the hidden layer in comparison to a neural network, and it has more layers. The nodes that hold the layers are called neurons. In comparison to machine learning, deep learning is closer to its goal. In order to predict breast cancer, a convolutional neural network was used to categorise the dataset. CNN classifies the photographs in order to forecast breast cancer.

III. RELATED WORKS

Several techniques for predicting breast cancers have been published and implemented in the scientific literature.

CNN's APPROACH:
CNNs are Deep Learning architecture that has been successfully applied to a huge variety of pc vision problems. [4] Freezing and high-quality tuning processes had been used to increase mass-lesion categorization accuracy. The VGG16 version showed pleasant accuracy, sensitivity, specificity, AUC, and F-rating while compared with four different algorithms. Subsequently, it is said that which include the CNN into the screening mechanism and using gaining knowledge of transfer can bring about the best prediction. The accuracy becomes 98.96 %, the sensitivity was 97 % and the specificity became 99%, the precision became 99 %, the F-rating turned to 97%, and the AUC became 0.995 %.[6] The reason for this work become to extract the relevant feature before the education system and deep learning algorithms to expect the breast tumour reaction to chemotherapy. For this prediction, the pathological whole reaction (PCR) is used as a general reference.

[7] In this paper a Histopathology statistic is set to broaden the gadget and done category using SVM and CNN fashions, accomplishing 99 % accuracy for the test ratio of about 60:40.[8] The CNN model, then again, has a 98 % accuracy rate along with the ensemble model with a median accuracy of 96 % vs 94 %for models.[11] To reinforce the classifier's robustness, we use transfer getting to know with the powerful ResNet-50 CNN that was pre-trained on ImageNet. Subsequently, the received results are superior to automatic breast cancer picture evaluation which has been stated in the literature.

3.2 SVM APPROACH:
The supervised learning model (SVM) is a type of machine learning. Both regression and classification issues can be solved with SVM. Support vector classifiers or support vector regressors are the names given to them depending on the circumstance. All of the input data is shown in an n-dimensional plane for SVMs to work.[5] This paper results in experiments to suggest that the SVM Co CoA method functions brilliantly and that it has a wide variety of applications, particularly in the domains of medical sciences and radiology.

[10] The goal was to accurately diagnose Breast Cancer. It offers an Ensemble approach for improved performance, and our proposed system has an accuracy of 99.28 %, which is obtained by combining Ensemble voting for improved accuracy. Ensemble

voting improves the stability of the system. As a result, the proposed system is more resistant to unforeseen circumstances.

3.3 LOGISTIC REGRESSION APPROACH:

For binary or multiclass classification problems, logistic regression (LR) is used. The independent variable(s)(Xi) that reflect the features and dichotomous dependent variable (Y) that represents the class label are assessed using the LR. The Logistic function is.

$$P = \frac{e\ A0+ A1x1+ A2x2+ A3x3+\ldots ANxN}{1+ e\ A0+ A1x1+ A2x2+ A3x3+\ldots ANxN}$$

$\alpha + \beta = \chi$

The goal of the logistic regression is to find A0, A1, A2, A3+…. +An

By applying logistic transformation:

$$Logit(Y) = \ln (P / 1-P)$$

[14] The statistics in this research suggest that the model is 95% accurate in forecasting survival rates, implying that it could be used in clinical practice. The studies also suggest that the mid-age group has a higher survival rate than the elderly. It's possible that the immune system is to blame for these outcomes, as people in their older years have a weaker immune system than those in their middle years.

3.4 DEEP NEURAL NETWORKS APPROACH:

Finally, we looked into employing a Deep Neural Network to test their performance. This network has been used in a variety of tasks, presenting the relevant results from recent studies.[12] In a Neuroevolutionary manner, they have used NSGA III to optimise and provide hyperparameters for DNNs. The evolutionary algorithm utilised in this work (NSGAIII)but this might not be able to handle many performance criteria at the same time. [13] To compare their performance with other classification methods, we have to integrate a Deep Neural Network. We compared and discussed accuracy and ROC curve measures in particular. The data tables and graphs were used to analyse the study findings. Thus, Deep Neural Networks fared admirably used in this investigation, and also, they had better results in image-based experiments.

Table1: COMPARISON ON ANALYSIS OF DIFFERENTTECHNIQUES OF BREAST CANCER

S.NO	NAME OF THE AUTHOR	YEAR	METHODS USED	ACCURACY	DATASET USED
1	Pei Liu et al., [1]	2020	XG Boost, GBM, RSF, Cox, EXSA	0.83% 0.82% 0.81% 0.76% 0.83%	Clinical Research Centre for Breast (CRCB)
2	Farhad Imani et al., [2]	2019	Random Forest		SEER
3	Yash Mate et al., [3]	2021	Boosting(ensemble) classification AdaBoost classifier, Extra Tree classifier	 0.97% 0.96%	Wisconsin Breast Cancer (WBC)
4	Abeer Saber et al., [4]	2021	CNN (InceptionV3, ResNet50, VGG-16, VGG-19,	 0.98%	MIAS

			Inception-V2 Res Net results)		
5	Devender Kaushik et al., [5]	2018	SVM-Co CoA	0.78%	Haberman's Survival
6	Yassine Amkrane et al., [6]	2020			QIN Breast DCE-MRI
7	Anju Yadav et al., [7]	2021	SVM, CNN	0.99%	Histopathological
8	Asrar Algarni et al., [8]	2021	CNN, Decision Tree, Logistic Regression, SVM	0.98% 0.93% 0.95% 0.95%	Wisconsin Carcinoma dataset (WBCD).
9	Shuai Liu et al., [9]	2021	Binary classification	1.00%	TCGA-BRCA
10	Sunanda Das et al., [10]	2019	Naïve bayes, RBF, J48, PNN, PCA+PNN, SVM, KNN, SLSQP+ Ensemble, Proposed Ensemble	0.97% 0.96% 0.93% 0.97% 0.97% 0.98% 0.97% 0.97% 0.99%	Wisconsin Breast Cancer
11	Qasem Abu Al-Haija et al., [11]	2020	ResNet-50, Convolutional Neural Network (CNN), Deep learning, Transfer Learning		Break His
12	BEIBIT ABDIKENOV et al., [12]	2019	Entity embedding, Deep learning networks, Evolutionary algorithms, Fuzzy inferencing.		SEER, WBCD,
13	Fabiano Teixeira et al., [13]	2019	DNN, D Tree, Perceptron, RF 100, RF 50, SVM	0.91% 0.87% 0.82% 0.94% 0.94% 0.88%	University of Wisconsin Hospital

14	Shailesh Kumar Verma et al., [14]	2020	Logistic regression	0.95%	Mammographic image from SEER and NCBI
15	Sidharth S Prakash et al., [15]	2020	D Tree, SVM, Ada Boost, Random Forest, Neural Network	0.94% 0.96% 0.96% 0.99% 0.99%	The Wisconsin breast cancer dataset

IV. COMPARATIVE ANALYSIS

This study compares several machine learning and deep learning methods for breast cancer prediction based on their accuracy.[1] In the field of survival analysis, we create a prognostic model for predicting disease progression after the initial diagnosis of breast cancer and compare it to the original XG boost approach, the classical GBM, RSF, and Cox methods. On the independent test set, our proposed EXSA approach outperforms the competition. In addition, we use the prognostic model to analyse risk scores of illness progression and to demonstrate risk grouping and a continuous function between risk score and rate of disease progression in clinical practice.[2] The impact of several parameters such as age, marital status, and histological grade on the recurrence of breast cancer is investigated using an ensemble random survival forest (RSF) technique. By sampling and bootstrapping into huge data with an ensemble model of survival trees, the RSF approach can assist estimate and predicting the survival function. To define breast cancer recurrences and provide a comprehensive assessment of the problem, we use the SEER breast cancer dataset, which includes over one and a half million items. The number of recurrences varies between 2 and 6. of breast cancer, according to the findings. Furthermore, age, surgical status, tumour stage, and histological grade are among the most important characteristics that influence breast cancer recurrence.[3] Most machine learning algorithms enhance their performance when they use feature selection techniques like Pearson's coefficient, chi-square test, logistic regression, random forest, RFE, and light gradient boosting to uncover relevant qualities. To summarise, the Ada boost Classifier scored 97% accuracy without feature selection, whereas the Extra Tree Classifier achieved 96.2% accuracy with feature selection via considering 20 significant features for breast cancer prediction.

When compared to four other models,[4] the VGG16 model had the best accuracy, sensitivity, specificity, AUC, and F-score. Finally, it can be stated that by incorporating the CNN with learning transfer into the screening mechanism, a significant improvement over other existing techniques can be accomplished. The accuracy was 98.96 %, the sensitivity was 97.83 %, the specificity was 99.13 %, the precision was 97.35 %, the F-score was 97.66 %, and the AUC was 0.995 %. [5] To predict the survival of breast cancer patients, these methods use statistical or machine learning algorithms. The findings of this experiment show that the SVM Co CoA technique performs efficiently, and also demonstrate that this technique has a wide range of applications, particularly in the fields of medical sciences.[6] This model is based on a convolution and deconvolution layer (Encoder-Decoder) framework with different widths and convolution filters (Encoder-Decoder). As a result, if we feed an image into the UNet network, we get an image containing pixels that have a chance of belonging to the tumour or not. Finally, a threshold is used to segment the tumour region. The goal of this study was to identify important features from radiomics before training machines and deep learning models to predict breast tumour response to chemotherapy. For this prognosis, the pathological complete response (pCR) is used as a standard reference.[7] We used a histopathology data set to develop the system and performed classification using SVM and CNN models, reaching 99 % accuracy for the train test ratio of 60:40. In addition, the test image is classed as cancerous or noncancerous. If malignant segmentation was identified on the test image, it was performed. For segmentation, GA and K Mean were tested, with GA outperforming K-means. When cancer cells are fragmented, they are eliminated.[8] We tested four classification algorithms: decision tree, support vector machine, CNN, and logistic regression. We used k-fold cross-

validation to validate the prediction accuracy and aid in fine-tuning the model hyper-parameters. The four algorithms did a good job at classifying benign and malignant tumours. The CNN model, on the other hand, has a 98 % accuracy rate. The ensemble model was likewise more accurate than the basic models, with an average accuracy of 96 % versus 94 % for the foundation models.

The first issue is whether a patient will live longer than 5 years, and the second issue is how long a patient will live after that.[9] For the first task, the best binary classification model obtains Accu=1.0000 utilising 40 methylomic features. For the question of how long a patient would live, the best regression model with 79 methylomic characteristics gets the regression performance MAE=31.62 days. More independent validation samples may be required to fine-tune the proposed models. We were unable to identify an independent validation dataset for the proposed prediction models because of the scarcity of additional datasets with similar numbers of samples and methylome profiling technology. With more datasets available, the study's generality will be confirmed.[10] The purpose of this research was to correctly diagnose breast cancer early. Hence, we suggest an Ensemble technique for better performance, and our proposed system has achieved about 99.28 % accuracy, which is obtained by integrating Ensemble voting for better accuracy.[11] To improve the robustness of the classifier, we used the transfer learning of the powerful ResNet-50 CNN pre-trained on ImageNet. The Break His dataset is used in the model, with 75% of the images being used for training and 25% being used for testing. The proposed work presents a comprehensive paradigm for medical image processing/classification from the input layer to the output layer. Finally, to our knowledge, the obtained results outperform the automated analysis of breast cancer images described in this article.

In a neuroevolutionary manner, [12] suggests using NSGA III to optimise and provide hyperparameters for DNNs. The evolutionary algorithm applied in this work (NSGAIII) may not be capable of simultaneously addressing various performance criteria. The approach converged early in some tests, resulting in a false or pseudo-Pareto optimal front. The DNN classifier has a great performance in accuracy level (92%),[13] indicating better results than other traditional models. Random forest 50 and 100 presented the best outcome for the ROC curve metric, which is considered an excellent prediction when compared to other previous techniques.[14] The 95 % predictive accuracy of the model build shows that the machine is 95 % accurate in forecasting the survival rate of Breast Cancer patients. The study also revealed that patient age is a significant factor in influencing survival rates. It was shown that persons in the mid-age group of 22 to 35 had the best survival rate, while those in the older age group of 60+ have the lowest probability of surviving. [15] The neural network model generated the benign class an F1 score of 98 and the malignant class an F1 score of 99. This computer-assisted diagnosis approach is designed to supplement rather than replace the ability of professional doctors and medical practitioners in the diagnosis process. Table 2. Represent the top ten techniques which have maximum accuracy.

Table 2. top ten techniques with maximum accuracy

TECHNIQUES	ACCURACY
EXSA	83%
RF 100, RF 50	94%
LOGISTIC REGRESSION	95%
ADA BOOST	97%
VGG-16	98%
CNN	98%
SVM	99%
ENSEMBLE	99%
NEURAL NETWORK	99%
BINARY CLASSIFICATION	100%

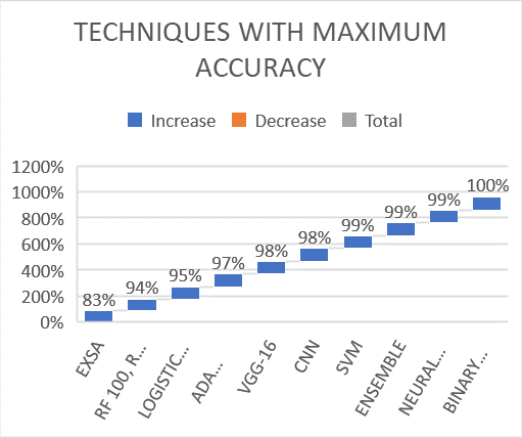

Fig 2. Techniques with maximum accuracy.

V. CONCLUSION AND FUTURE SCOPE

Hence, this study looks at a variety of methodologies as well as a review of breast cancer diagnosis and prognosis issues. Many academics examine the behaviour and performance of data mining algorithms using experimental results. The most accurate algorithm for forecasting breast cancer is simply based on the algorithm's accuracy. To analyse medical data, numerous machine learning and deep learning algorithms are available, but the issue is to create an accurate and efficient model for medical applications. This survey report uses publicly available breast cancer datasets to test several classification techniques and clustering algorithms. According to experimental data, the proposed Convolutional Neural Networks (CNN) model classifier has the best classification accuracy. In a future study, additional clinical data on breast cancer and more follow-up data will be collected and can be intended to combine Neural Networks also with other machine learning and deep learning techniques to improve its accuracy.

REFERENCES

[1] Pei Liu et al.,"Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer," in IEEE Transactions on Biomedical Engineering, vol. 68, no. 1, pp. 148-160, Jan. 2021

[2] Farhad Imani et al., "Random Forest Modeling for Survival Analysis of Cancer Recurrences," 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), 2019, pp. 399-404

[3] Yash Mate et al., "Hybrid Feature Selection and Bayesian Optimization with Machine Learning for Breast Cancer Prediction," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 612-619

[4] Abeer Saberet al., "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," in IEEE Access, vol. 9, pp. 71194-71209, 2021

[5] Devender Kaushik et al., "Post-Surgical Survival Forecasting of Breast Cancer Patient: A Novel Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 37-41

[6] YassineAmkrane et al.,"Towards Breast Cancer Response Prediction using Artificial Intelligence and Radiomics," 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), 2020, pp. 1-5

[7] AnjuYadav etal.,"Automated Detection and Classification of Breast Cancer Tumour Cells using Machine Learning and Deep Learning on Histopathological Images," 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1-6,

[8] AsrarAlgarniet al., "Convolutional Neural Networks for Breast Tumor Classification using Structured Features," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif),2021, pp. 1-5,

[9] Shuai Liu et al., "Survival Time Prediction of Breast Cancer Patients Using Feature Selection Algorithm Crystall," in IEEE Access, vol. 9, pp. 24433-24445, 2021

[10] Sunanda Das et al., "Prediction of Breast Cancer Using Ensemble Learning," 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), 2019, pp. 804-808,

[11] Qasem Abu Al-Haija et al.,"Breast Cancer Diagnosis in Histopathological Images Using ResNet-50 Convolutional Neural Network," 2020 IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS), 2020, pp. 1-7

[12] B. Abdikenov, et al., "Analytics of Heterogeneous Breast Cancer Data Using Neuroevolution," in IEEE Access, vol. 7, pp. 18050-18060, 2019

[13] Fabiano Teixeira et al., "An Analysis of Machine Learning Classifiers in Breast Cancer Diagnosis," 2019 XLV Latin American Computing Conference (CLEI), 2019, pp. 1-10

[14] Shailesh Kumar Verma et al., "Breast Cancer Survival Rate Prediction in Mammograms Using Machine Learning," 2020 2nd International Conference on Advances in Computing,

Communication Control and Networking (ICACCCN), 2020, pp. 169-171

[15] Sidharth S Prakash et al.,"Breast Cancer Malignancy Prediction Using Deep Learning Neural Networks," 2020 Second International Conference on Inventive Research in Computing Applications (ICIR), 2020, pp. 88-92