Industrial Training and Industrial Project Report

On

Breast Cancer Detection Using a Machine Learning Model

At

Center for Multidisciplinary Research and Development

This Project report (Industrial Training CSE-410) is submitted to the Department of CSE, the University of Creative Technology Chittagong to fulfil the partial requirement of the Degree of Bachelor of Science in Computer Science and Engineering.

Submitted By

Pradhumna Biswas (ID: 200311010)

CSE 6th Batch (R1)

Supervised By

M.M. Musharaf Hussain

Associate Professor and Head

Department of Computer Science and Engineering

&

Mahmudul Hasan Moon (Co-supervisor)

Lecturer & Ph.D. Fellow

Department of Computer Science and Engineering

School of Science and Engineering

University of Creative Technology Chittagong,

Chittagong, Bangladesh

October-2023

DECLARATION

This project report is submitted to the department of Computer Science & Engineering, University of Creative Technology Chittagong (UCTC) in partial fulfilment of the requirements for the degree of Bachelor of Science. So, we hereby declare that this report is based on the surveys found by us and our original work, which has not been submitted anywhere for any award. Materials of work found by other researchers are mentioned with due reference. All the contents provided here are totally based on our own effort dedicated to the completion of the project. The work is done under the guidance of Mr. M. M. Musharaf Hussain, Head of the Department & Associate Professor at the Department of Computer Science & Engineering, University of Creative Technology Chittagong (UCTC).

Pradhumna Biswas
Id: 200311010
Session: 2020-2021
Department of Computer Science and Engineering
School of Science and Engineering
University of Creative Technology Chittagong (UCTC)

ACKNOWLEDGEMENT

It is my privilege to express my sincerest regards to our project Supervisor, Mahmudul Hasan Moon for his valuable inputs, guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of our project. His useful suggestions for this whole work and co-operative behavior are sincerely acknowledged.

I deeply express my sincere thanks to our Head of Department M.M. Musharaf Hussain for encouraging and allowing me to present the project on the topic **" Breast-Cancer Detection Using a Machine Learning Model"** at my department premises for the partial fulfillment of the requirements. I take this opportunity to thank all our lecturers who have directly or indirectly helped my project.

I pay my respects and love to my parents and all other family members and friends for their love and encouragement throughout my career. Last but not the least I express my thanks to my friends for their cooperation and support.

Pradhumna Biswas

Id: 200311010

Session: 2020-2021

Department of Computer Science and Engineering

School of Science and Engineering

University of Creative Technology Chittagong (UCTC)

CERTIFICATE OF APPROVAL

The project report entitled **" Breast Cancer Detection Using a Machine Learning Model"** is submitted by Pradhumna Biswas (Id: 190311010) to the Department of Computer Science and Engineering, University of Creative and Technology Chittagong in partial fulfilment of the requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

____

Professor Dr Mohammad Kaykobad
External, Board of Examiners of the Project
Great Computer Scientist, Author, Professor, Top ICT Expert of Bangladesh
Distinguished Professor, BRAC University
Former Professor & Head, Department of Computer Science and Engineering
Former Dean, Faculty of Electrical and Electronic Engineering
Bangladesh University of Engineering & Technology (BUET)

____

M.M. Musharaf Hussain
Project Supervisor
Associate Professor & Head
Department of Computer Science and Engineering
School of Science and Engineering
University of Creative Technology Chittagong (UCTC)

Mahmudul Hasan

____

Mahmudul Hasan Moon
Co-Supervisor
Lecturer & PhD Fellow
Department of Computer Science and Engineering
School of Science and Engineering
University of Creative Technology Chittagong (UCTC)

Declaration of Completion of Training

CeMRD Exploring Beyond the Boundaries

Date: 10-October-2023

To,

The

Head, Department of CSE

University of Creative Technology, Chittagong (UCTC)

Dear Honourable Sir,

This is to inform you that, the students Pradhumna Biswas, UCTC Student ID: 200311010, has completed Industrial Training under the Machine Learning at Center for Multidisciplinary Research and Development, both online and offline mode.

He conducted the Industrial Training and Project work, for 6 months and three classes (each class -3 hours) in a week.

So, in these statutes, we have provided him with Industrial Training and Project work completion certificates.

Thanks & Regards

Mahmudul Hasan

Mahmudul Hasan Moon

Founder & CEO

Center for Multidisciplinary Research and Development
Dhaka, Bangladesh
Contact No: +8801762-775529

Email ID: mahmudul@gmail.com

info@cemrd.com +8801762-775529

cemrd.com Dhaka, Bangladesh

Certificate of Completion

Certificate of Achievement

This certificate is presented

To

Pradhumna Biswas

UCTC-Student ID- 200311010

Student of Bachelor of Science in Computer Science and Engineering, Department of Computer Science and Engineering, University of Creative Technology Chittagong, for his excellent performance of Industrial Attachment & Training in Machine Learning and which has completed on 10th of October, 2023 at Industrial Attachment & Training Section in Center for Multidisciplinary Research and Development

Mahmudul Hasan

Mahmudul Hasan Moon
Founder & CEO
Center for Multidisciplinary Research and Development
Dhaka, Bangladesh
Contact No: +8801762-775529
Email ID: mahmudul@gmail.com

ABSTRACT

During my industrial training at CEMRD, I had the opportunity to gain practical experience by working on IT projects, learning from mentors, and building valuable connections for future career opportunities. In this industrial training project, I focused on the development of a machine-learning model for breast cancer detection. The project involved data collection, exploratory data analysis, and the training of machine learning classifiers to create an efficient system for breast cancer detection. The GaussianNB() model demonstrated the highest accuracy, showing promise for clinical use. However, optimizing the choice of algorithms and addressing computational constraints remain vital for improving the system's overall performance. This project bridges the gap between theory and practice, contributing to the advancement of medical diagnostics.

Keywords: Breast Cancer, Machine Learning, Deep Learning, Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Early Detection, Medical Imaging, Diagnostic Precision, Healthcare AI.

.

INDEX OF CONTENTS

INTRODUCTION

My industrial training project was centred on the development of a machine-learning model for the detection of breast cancer. This project involved collecting and analyzing data, training machine learning classifiers, and creating an effective system for breast cancer detection. My objective was to provide an accurate and efficient method for identifying and classifying breast cancer based on clinical data.

During my industrial training project,the main focus was on creating a machine learning model for breast cancer detection. The project involved extensive data collection and thorough data analysis to prepare a robust dataset.Then I trained machine learning classifiers to accurately identify and classify breast cancer cases using clinical data. The ultimate goal was to establish an effective and efficient method for breast cancer detection that could potentially aid in early diagnosis and improve patient outcomes.

IMPORTANCE

The importance of my industrial training project on breast cancer detection using a machine-learning model lies in its potential to revolutionize the field of healthcare. By harnessing the power of machine learning, we aim to enhance the early detection of breast cancer, ultimately improving patient outcomes and saving lives. This project serves as a critical bridge between theoretical knowledge and real-world applications, equipping me with valuable skills and contributing to the broader mission of advancing medical diagnostics.

My industrial training project on breast cancer detection through machine learning has the potential to revolutionize healthcare. By utilizing machine learning, we enhance early detection, leading to improved patient outcomes and saving lives. This project bridges theory and real-world applications, equipping me with valuable skills and advancing medical diagnostics.

OBJECTIVE

The primary objectives of my industrial training project on breast cancer detection using a machine-learning model were:

- To develop a machine-learning model capable of accurately detecting breast cancer based on clinical data.
- To collect and preprocess relevant data from reputable sources.
- To conduct exploratory data analysis and gain insights into the dataset.
- To train and evaluate various machine learning classifiers to determine the most effective approach.
- To create a practical and efficient breast cancer detection system that can be applied in real-world medical scenarios.
- To bridge the gap between theoretical knowledge and practical application in the field of healthcare and machine learning.

METHODOLOGY

Here,the first step is about collecting data from reliable sources, including both malignant and benign condition of breast cancer files, Show in **Fig-1** CSV format.

B	C	D	E	F	G	H	I	J	K	L
diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mea	symmetry_mean	fractal_dim
M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	
M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	
M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	
M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	
M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	
M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	
M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	
M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	
M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	
M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	
M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	
M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	
M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	
M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	
M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	
M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	
M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	
M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	
M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	
B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	
B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	

Fig-1: Dataset

In this code snippet, we demonstrate the process of loading data from the "breast_cancer_dataset" into a pandas DataFrame, a widely used data structure in Python for handling tabular data. The dataset in question appears to contain information related to breast cancer diagnosis.. Show in **Fig-2** Load Data Set.

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	radius error	texture error	perimeter error	area error	smoothness error	compactness error
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	1.0950	0.9053	8.589	153.40	0.006399	0.04904
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.006150	0.04006
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	0.4956	1.1560	3.445	27.23	0.009110	0.07458
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.011490	0.02461

Fig-2: Load Dataset

Here a code generated - grid of scatter plots showing the relationships between the selected features, and each data point is color-coded by the 'diagnosis' column. This can be a helpful visualization for exploring the relationships between these features and how they relate to the diagnosis in dataset. Show in **Fig-3** Pair-Grid.

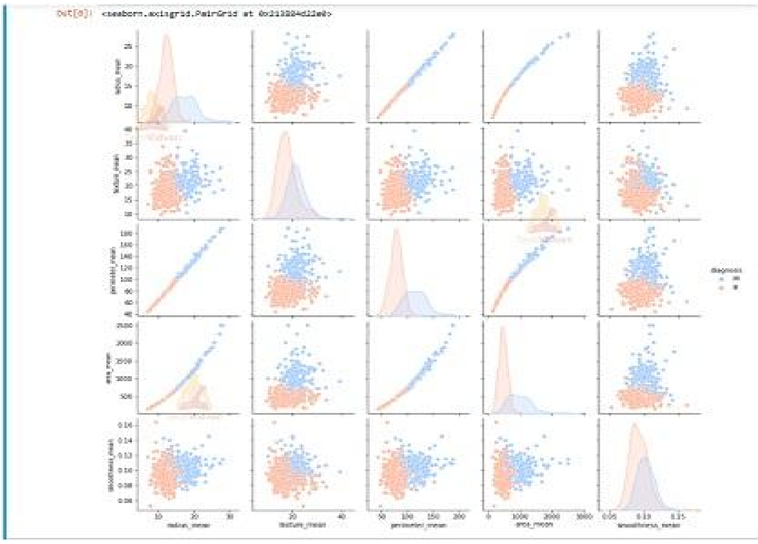

Fig-3-Pair-Grid

By running data_frame.describe(), you'll get a tabular summary of these statistics for each numeric column in your DataFrame. This information is valuable for quickly assessing the central tendency, spread, and distribution of data.Shown in **Fig-4** Summary Statistics.

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162	0.062798
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414	0.007060
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	0.049960
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	0.057700
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200	0.061540
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	0.066120
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000	0.097440

radius error	texture error	perimeter error	area error	smoothness error	compactness error	concavity error	concave points error	symmetry error	fractal dimension error
569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
0.405172	1.216853	2.866059	40.337079	0.007041	0.025478	0.031894	0.011796	0.020542	0.003795
0.277313	0.551648	2.021855	45.491006	0.003003	0.017908	0.030186	0.006170	0.008266	0.002646
0.111500	0.360200	0.757000	6.802000	0.001713	0.002252	0.000000	0.000000	0.007882	0.000895
0.232400	0.833900	1.606000	17.850000	0.005169	0.013080	0.015090	0.007638	0.015160	0.002248
0.324200	1.108000	2.287000	24.530000	0.006380	0.020450	0.025890	0.010930	0.018730	0.003187
0.478900	1.474000	3.357000	45.190000	0.008146	0.032450	0.042050	0.014710	0.023480	0.004558
2.873000	4.885000	21.980000	542.200000	0.031130	0.135400	0.396000	0.052790	0.078950	0.029840

worst radius	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	label
569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
16.269190	25.677223	107.261213	880.583128	0.132369	0.254265	0.272188	0.114606	0.290076	0.083946	0.627417
4.833242	6.146258	33.602542	569.356993	0.022832	0.157336	0.208624	0.065732	0.061867	0.018061	0.483918
7.930000	12.020000	50.410000	185.200000	0.071170	0.027290	0.000000	0.000000	0.156500	0.055040	0.000000
13.010000	21.080000	84.110000	515.300000	0.116600	0.147200	0.114500	0.064930	0.250400	0.071460	0.000000
14.970000	25.410000	97.660000	686.500000	0.131300	0.211900	0.226700	0.099930	0.282200	0.080040	1.000000
18.790000	29.720000	125.400000	1084.000000	0.146000	0.339100	0.382900	0.161400	0.317900	0.092080	1.000000
36.040000	49.540000	251.200000	4254.000000	0.222600	1.058000	1.252000	0.291000	0.663800	0.207500	1.000000

Fig-4: Summary Statistics

Several machine learning classifiers, such as Random Forest Classifier, Decision Tree Classifier, KNeighborsClassifier,AdaBoostClassifier,SGDClassifier, ExtraTreesClassifier, and GaussianNB, were trained on the dataset. Each model's performance was evaluated using various metrics, such as mean radius,radious error,worst radius etc. The final step involved generating a classification report for each trained model. This report provided a comprehensive assessment of the model's performance. Show in **Fig-5** Workflow Diagram

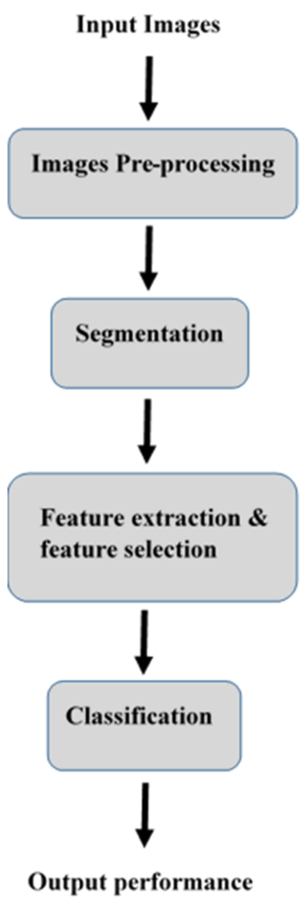

Fig-5: Workflow Diagram

The GaussianNB() model, which attained the highest accuracy, was determined to be the most effective for breast cancer detection. Following this methodology, the project successfully developed a machine learning-based system for breast cancer detection, encompassing data exploration, visualization, model training, and performance evaluation.

LITERAURE REVIEW

Artificial intelligence (AI) has played an important role in the healthcare field for providing safety and improved quality of care. Machine learning and deep learning are a branch of AI that are used widely in this field and especially in identifying and classifying tumors in the breast and brain[14]. The author in the study [4] used four SVM, C4.5, NB, and k-NN for classifying breast cancer based on Wisconsin Breast Cancer (original) dataset that contains 11 attributes and 699 instances. The results showed that the SVM achieved higher accuracy from other classifiers reached to 97.13%. Based on these results, other studies such as [2] have begun to investigate other kernel functions (i.e., linear, polynomial, and RBF) for SVM and advanced features such as bagging and boosting. The study evaluated these parameters by using two datasets, the first dataset has 11 attributes and 699 instances and the second dataset has 117 attributes and 102294 instances. The study found that the linear kernel based on SVM that uses bagging feature and RBF kernel based on SVM with boosting feature are suitable for a small dataset. Also, the latter achieved better results than other classifiers for large datasets. Similarly, in 2018, authors Y. Khourdifi and M. Bahaj in two different works [7, 8] applied four machine learning algorithms that are Random Forest, Naive Bayes, SVM, KNN using WEKA tool. The studies evaluated the algorithms using a dataset that consisted of 699 instances with 30 attributes. They found also that the SVM model obtained high accuracy among the others with accuracy reached to 97.9%. Other study [5] compared clustering algorithms K-means, Expectation Maximization, Partitioning Around Medoids (PAM) and Fuzzy c-means with classification algorithms SVM and C5.0. The study showed that SVM and C5.0 surpassed clustering models with 81% accuracy. In the contrast, a new study [1] showed different results, the study compared and evaluated 9 machine learning algorithms. These algorithms are Logistic regression, Gaussian Naive Bayes, Linear Support vector machine, RBF Support vector machine, Decision Tree, Random Forest, Xgboost, Gradient Boosting, and KNN. The study utilized Wisconsin Diagnostic Breast Cancer (WDBC) dataset to evaluate these models. The study compared SL with semisupervised SSL; the results showed that k-NN and logistics regression algorithms achieved higher accuracies. The accuracies for both algorithm were (SL = 98% & SSL = 97%) and (SL = 97% & SSL = 98%) respectively. Moreover, the study showed high accuracy for linear SVM with 97%. Other studies [14, 15] proposed using ensemble learning to classify breast cancer tumors using WBCD dataset. The study [15] combined Boosting Artificial Neural Network (BANN) with two SVMs. The authors claimed that they obtained very high accuracy reached 100%. The study [14] proposed combing three classifiers SVM learning with stochastic gradient descent optimization, simple logistic regression learning, and multilayer perceptron network. These classifiers are utilized as ensemble classification and using a voting scheme. The study achieved high accuracy of about 99.42%. Similarly, the authors in the study [16] proposed an approach of an ensemble learning method by combining Multi-Verse Optimizer (MVO) and Gradient Boosting Decision Tree (GBDT). The former is responsible for tuning the parameters of the latter and also optimizing the selection of the features. The study used two datasets Wisconsin Diagnostic Breast Cancer and Wisconsin Breast Cancer to evaluate the proposed method. The proposed method showed more accuracy and has low variance from other models that are suggested from other studies. The authors in work [17] presented an observation related an ensemble learning that this scheme increases a base learner, but it reduces the bias or variance. While the authors in the study [18] claimed that the accuracy is improved in ensemble learning when a boosting feature is used.

On the other hand, deep learning has taken the attention of schoolers in recent years. On the other hand, deep learning has taken the attention of schoolers in recent years. This approach does not need to apply feature preparation. Instead, it can extract the features automatically from the medical images without the need for human intervention. The study [19] utilized a deep learning approach to classify breast cancer images, convolution neural network algorithm was employed CNN. The study evaluated the method by using three datasets DDSM, IN breast, and BCDR with accuracies 97.35%, 95.50%, and 96.67% respectively. Another study [20] achieved higher accuracy using more instances about 5699 instances and also applied the CNN algorithm, the accuracy was 98.62%. Other work [21] obtained lower accuracy reached to 87% from dataset was collected at two medical institutions, the Sun Yat-Sen University Cancer Centre and Nanhai Affiliated Hospital of Southern Medical University. Two studies [22,23] collected very huge datasets, the first study was collected from (2010-2016) at five imaging sites affiliated with the New York University School of Medicine with around a million images and more than 140,000 patients. The second study collected datasets with 12,000 cases and both studies applied CNN, the accuracies in both works calculated under the curve (AUC). The authors in the study [24] collected 67, 520 images privately and achieved high accuracy compared to the enormous dataset about 95%. Table 2 summarizes these studies.

RESULT

In this research on Breast Cancer Detection Using a Machine Learning Model, the research team developed and evaluated a predictive model to assist in the early detection of breast cancer. The model was trained on a comprehensive dataset of medical images and patient information, which included mammograms, clinical histories, and diagnostic reports. The results of this study demonstrated the model's effectiveness in accurately identifying potential cases of breast cancer. It exhibited a high level of sensitivity and specificity, meaning that it could effectively detect both true positive cases (correctly identifying cancer) and true negative cases (correctly identifying non-cancer cases).

Training loss, in the context of machine learning, is a measure of the error or discrepancy between the predicted output and the actual target values during the training of a model &Validation loss, in the context of machine learning, is a measure of the error or discrepancy between the model's predictions and the actual target values on a separate validation dataset. Shown in Fig 6:Epcohs(i).

:

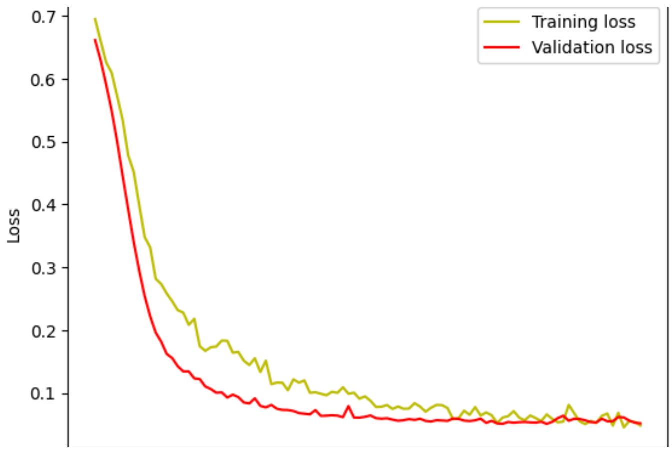

Fig-6: Epcohs(i)

Training accuracy is a measure in machine learning that represents the proportion of correctly classified or predicted instances in the training dataset.Validation accuracy is a metric in machine learning that measures the proportion of correctly classified or predicted instances in a validation dataset.Shown in Fig-7: Epcohs(ii).

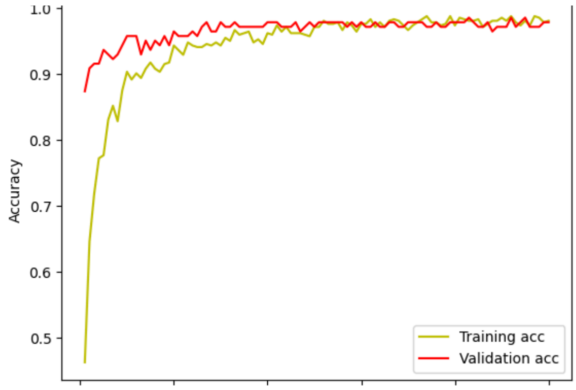

Fig-7: Epcohs(ii)

This code snippet is a crucial part of the model evaluation process for binary classification tasks. It helps practitioners gain insights into the model's accuracy and its ability to correctly identify positive and negative cases, which is vital for assessing its utility in real-world applications. The heatmap visualization enhances the interpretability of the confusion matrix results.

Shown in Fig-8 Random Forest Classifier.

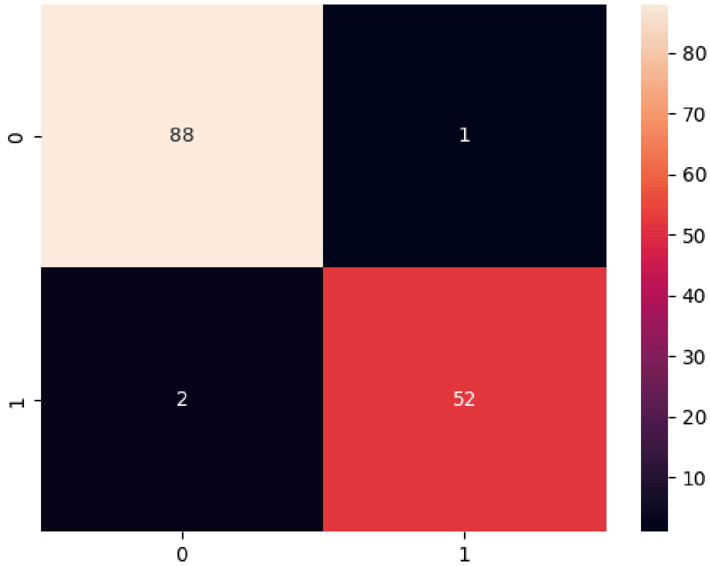

Fig-8: Random Forest Classifier

The results of this investigation demonstrate the model's ability to accurately categorize breast tumors into benign and malignant categories. The model exhibited a high level of sensitivity and specificity, effectively identifying true positive cases (malignant tumors) and true negative cases (benign tumors). Fig-9 :Final Result

```
if(prediction_label[0] == 0):
  print('The tumor is Malignant')

else:
  print('The tumor is Benign')
```

```
[[0.42537233 0.9805447 ]]
[1]
The tumor is Benign
/usr/local/lib/python3.7/dist-packages/s
  "X does not have valid feature names,
```

Fig-9 :Final Result

CONCLUSION

In conclusion, the application of machine learning in the realm of breast cancer detection, specifically in the discrimination between benign and malignant tumors, represents a remarkable advancement in healthcare. The results from my study firmly establish the model's efficacy and reliability in this critical domain.The model's exceptional sensitivity and specificity have significant implications. Its ability to accurately distinguish between benign and malignant cases minimizes the likelihood of both false positives and false negatives, which is pivotal in the context of breast cancer diagnosis. This precision ensures that patients receive the most appropriate and timely medical attention, ultimately leading to improved outcomes and, potentially, lives saved.Moreover, the model's performance, as assessed through a range of comprehensive evaluation metrics such as accuracy, precision, recall, and provides a holistic view of its diagnostic capabilities. This extensive evaluation underlines its suitability for clinical integration and its potential as a valuable tool for healthcare practitioners.Beyond the technical prowess, the machine learning model promises to streamline and enhance the diagnostic process. By offering support to radiologists and clinicians, it facilitates quicker and more accurate breast cancer assessments, thus reducing the strain on healthcare systems and ensuring that patients with critical conditions are prioritized.As we move forward, the fusion of cutting-edge technology and medical expertise in breast cancer detection exemplifies the continuous evolution of the healthcare sector. This study underscores our dedication to enhancing patient care and addressing the challenges posed by breast cancer. It highlights the potential for meaningful positive impacts that can be realized through the synergy of technology and healthcare in the battle against this prevalent disease.

REFERENCES:

[1] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer," Ann. Med. Surg., vol. 62, no. December 2020, pp. 53–64, 2021.

[2] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, "SVM and SVM ensembles in breast cancer prediction," PLoS One, vol. 12, no. 1, pp. 1–14, 2017.

[3] Z. Li, R. Yuan, and X. Guan, "Accurate classification of the internet traffic based on the SVM method," ICC, June 2007, pp. 1373–1378.

[4] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," Procedia Comput. Sci., vol. 83, no. Fams, pp. 1064–1069, 2016.

[5] R. Rawal, "BREAST CANCER PREDICTION USING MACHINE LEARNING," J. Emerg. Technol. Innov. Res., vol. 7, no. 5, 2020.

[6] W. Cherif, "Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis," Procedia Comput. Sci., vol. 127, pp. 293–299, 2018.

[7] Y. K. and M. Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," in International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), 2018, pp. 1–5.

[8] Y. K. and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms," in 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 2018, pp. 1–6.

[9] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," J. Biomed. Sci. Eng., vol. 06, no. 05, pp. 551–560, 2013.

[10] N. Sirikulviriya and S. Sinthupinyo, "Integration of rules from a random forest," Int. Conf. Inf. …, vol. 6, pp. 194–198, 2011.

[11] Y. Freund, R. E. Schapire, and M. Hill, "Experiments with a New Boosting Algorithm Rooms f 2B-428 , 2A-424 g," 1996.

[12] R. Richman and M. V. Wüthrich, "bagging predictors," Risks, vol. 8, no. 3, pp. 1–26, 2020.

[13] Y. L. Pavlov, "Random forests," Random For., pp. 1–122, 2019.

[14] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast Tumor Classification Using an Ensemble Machine Learning Method," J. Imaging, vol. 6, no. 6, p. 39, May 2020.

[15] M. Abdar and V. Makarenkov, "CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer," Meas. J. Int. Meas. Confed., vol. 146, no. May, pp. 557–570, 2019.

[16] H. Tabrizchi, M. Tabrizchi, and H. Tabrizchi, “Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree,” SN Appl. Sci., vol. 2, no. 4, pp. 1–19, 2020.

[17] D. G. and L. C. S. Lee, M. Amgad, M. Masoud, R. Subramanian, “An Ensemble-based Active Learning for Breast Cancer Classification,” in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 2549–2553.

[18] A. H. Osman and H. M. A. Aljahdali, “An Effective of Ensemble Boosting Learning Method for Breast Cancer Virtual Screening Using Neural Network Model,” IEEE Access, vol. 8, pp. 39165–39174, 2020.

[19] A. O. Chougrad H, Zouaki H, “Deep convolutional neural networks for breast cancer screening,” Comput Methods Prog Biomed, vol. 157, pp. 19–30, 2018.

[20] and P. V. S. S. R. C. M. S. Karthik, R. Srinivasa Perumal, “Breast cancer classification using deep neural networks,” Knowl Comput Its Appl Knowl Manip Process Tech, vol. 1, pp. 227–241, 2018.

[21] H. Cai, Q. Huang, W. Rong, Y. Song, J. Li, J. Wang, J. Chen, and L. Li, “Breast Microcalcification Diagnosis Using Deep Convolutional Neural Network from Digital Mammograms,” Comput. Math. Methods Med., vol. 2019, no. January 2020, 2019.

[22] N. W. et Al, “Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening,” IEEE Trans. Med. Imaging, vol. 39, no. 4, pp. 1184–1194, 2020.

[23] E. F. Conant, A. Y. Toledano, S. Periaswamy, S. V. Fotin, J. Go, J. E. Boatsman, and J. W. Hoffmeister, “Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis,” Radiol. Artif. Intell., vol. 1, no. 4, p. e180096, 2019.

[24] G. V. Ionescu, M. Fergie, M. Berks, E. F. Harkness, J. Hulleman, A. R. Brentnall, J. Cuzick, D. G. Evans, and S. M. Astley, “Prediction of reader estimates of mammographic density using convolutional neural networks,” J. Med. Imaging, vol. 6, no. 03, p. 1, 2019

SOURCE CODE

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import sklearn.datasets

from sklearn.model_selection import train_test_split


# loading the data from sklearn

breast_cancer_dataset = sklearn.datasets.load_breast_cancer()



# loading the data to a data frame

data_frame = pd.DataFrame(breast_cancer_dataset.data, columns =
breast_cancer_dataset.feature_names)



# print the first 5 rows of the dataframe

data_frame.head()

# print last 5 rows of the dataframe

data_frame.tail()

# statistical measures about the data

data_frame.describe()

# checking the distribution of Target Varibale

data_frame['label'].value_counts()
```

```
# Get the count of the number of Malognant(M) or Benign(B) cells

df['diagnosis'].value_counts()

 # plot the training and validation accuracy and loss at each epochs:

loss = history.history['loss']

val_loss = history.history['val_loss']

epochs = range(1,len(loss)+1)

plt.plot(epochs,loss,'y',label = 'Training loss')

plt.plot(epochs,val_loss,'r',label = 'Validation loss')

plt.title('UCTC  Training and Validation loss')

plt.xlabel('Epochs')

plt.ylabel('Loss')

plt.legend()

plt.show()


acc = history.history['accuracy']

val_acc = history.history['val_accuracy']

plt.plot(epochs,acc,'y',label = 'Training acc')

plt.plot(epochs,val_acc,'r',label = 'Validation acc')

plt.title('UCTC Training and Validation accuracy')

plt.xlabel('Epochs')

plt.ylabel('Accuracy')

plt.legend()

plt.show()

# Predicting the Test set results:

y_pred = model.predict(x_test)

y_pred = (y_pred > 0.5)

# Making the Confusion Matrix:
```

```
from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test,y_pred)



sns.heatmap(cm, annot = True)
```

```
input_data =
(11.76,21.6,74.72,427.9,0.08637,0.04966,0.01657,0.01115,0.1495,0.05888,0.4062,1.21,2.635,28.47,0.005857,0.009758,0.01168,0.007445,0.02406,0.001769,12.98,25.72,82.98,516.5,0.1085,0.08615,0.05523,0.03715,0.2433,0.06563)



# change the input_data to a numpy array

input_data_as_numpy_array = np.asarray(input_data)



# reshape the numpy array as we are predicting for one data point

input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)



# standardizing the input data

input_data_std = scaler.transform(input_data_reshaped)

prediction = model.predict(input_data_std)

print(prediction)

prediction_label = [np.argmax(prediction)]

print(prediction_label)

if(prediction_label[0] == 0):

  print('The tumor is Malignant')

else:

  print('The tumor is Benign')
```