An Examination and Evaluation of the Rising Incidence of Breast Cancer among Women and Impact on Society

Meeta Joshi[1]*; Dr. Shweta Pandey[2]
[1]Faculty of Computer Science; [2]Faculty of Commerce and Business Management,
Amrapali University, Haldwani, Uttarakhand, India

Corresponding Author: Meeta Joshi[1]*

Abstract:- The number of deaths from breast cancer is rising rapidly year by year. The most common kind of cancer named as breast cancer is an overall and the leading cause for women death rate all over India. A cancer patient can go for a long and healthy life by early detection and proper treatment of cancer. In India, where advanced stages of the disease at diagnosis, combined with rising incidence and death rates, make breast cancer the most frequent cancer among women now a days. It is demand of time to spread a better cancer literacy awareness among women of India.

Therefore, we did secondary data collection using literature review in order to know patient survival parameters and treatment strategies, with high degree of accuracy in breast cancer detection . Online and offline structured questioners has been used to collect primary data from various participants.

This research paper also used a large dataset of breast cancer patients from the UCI Machine Learning Repository. Our paper examined that there is a very low cancer literacy about breast cancer risk factors in India , irrespective of socio-economic and educational background of participant women . There is an urgent need for offline and online awareness programmes to improve cancer literacy in Indian Women .

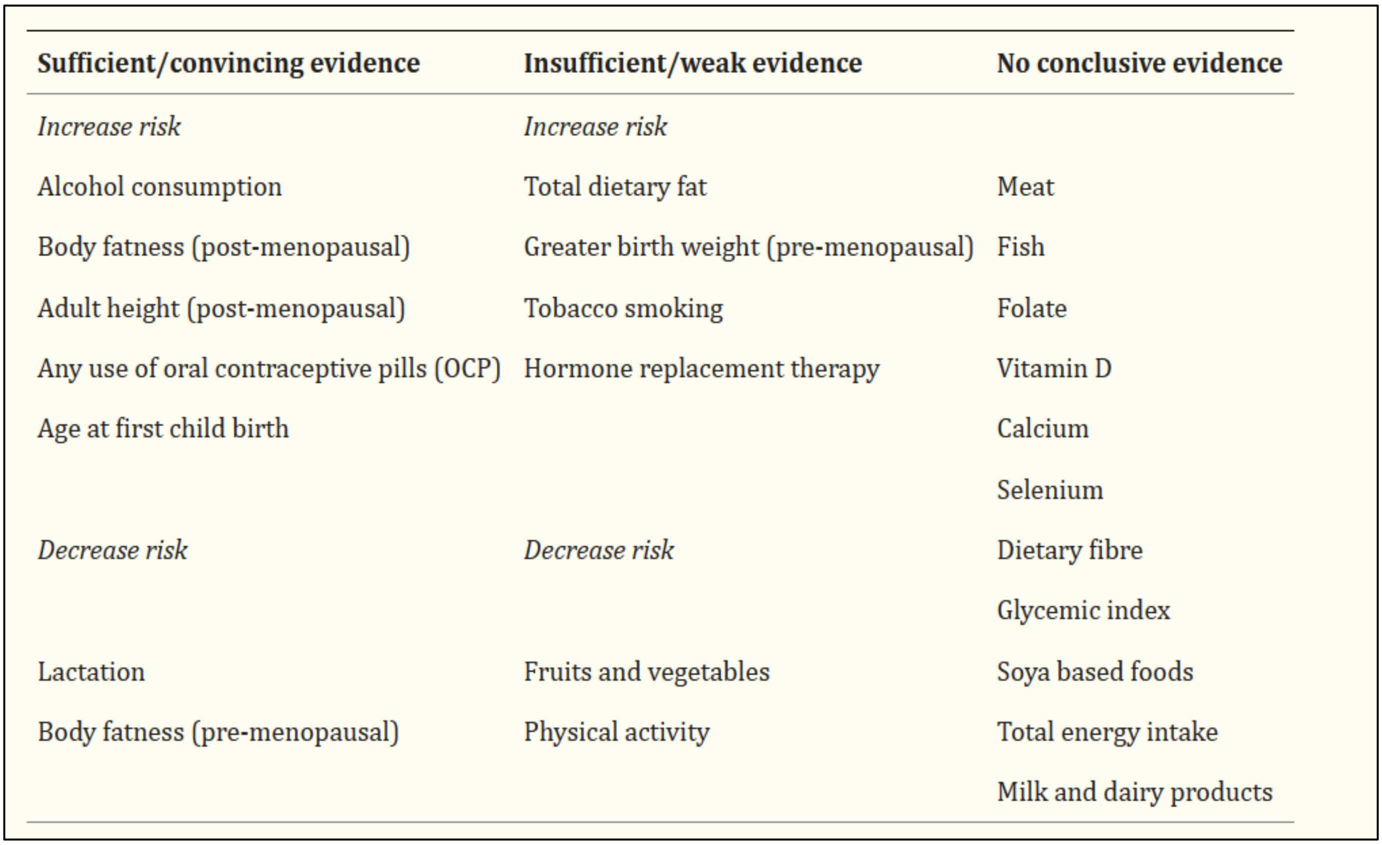
Keywords:- *UCI , IARC , Breast Cancer , KNN , SVM , MMH.*

I. INTRODUCTION

A recent report from the "International Agency for Research on Cancer (IARC)," declared that breast cancer has surpassed lung cancer as the most frequent disease in women diagnosed globally. . Over the past two decades, there has been a significant increase in overall global incidence of all types of cancers, with an estimated 21.1 million people being diagnosed with some form of cancer this year alone , equivalent to one in five individuals globally developing it over their lifetimes. The numbers are only projected to continue rising dramatically; by 2035 diagnoses are expected to be almost half again what they were compared just twenty years ago. Similarly concerning is that deaths caused by various forms of pathological tumors have also climbed: In particular, there was an uptick from roughly six-and-a-half million cases documented back before year 2000, up until ten million human lives lost due specifically because malignancies, representing more than one sixth fraction out total mortalities around Earth–which emphasizes how critical it remains key stake-holders emphasize oncological care and prevention when investing health system resources going forward. One possible avenue for doing so could involve new information & communication technology solutions like Big data analytics whose algorithms can derive actionable insights even across large sets numbering millions unstructured or otherwise complex records – e.g., patients' histories might include inconsistent or incompletely reported datasets spanning multiple geographic regions impacted individually primary unique factors within gender cohorts affected asymmetrically disparate societies situated culturally nuanced variances morphed slowly over time inside differently structured healthcare delivery systems evolving organically alongside historical legacies gapingly evident despite rate-limited attempts at harmonization. The possibilities demonstrated here reflect paradigm-shifting applications capable harness much-needed disruption enabling greater transparency accountability driving improved medical practices outcomes epitomizing digital transformation occurring start-up mature businesses alike meteoric velocities never envisioned merely couple few generations previous but integral enrichment living vitas present coming centuries [1]. In line with these developments comes fresh hope leveraging machine learning technologies able facilitate prediction as well as swift precision screening interventions necessary combat progressive distribution harmful tumors throughout body. More specifically, this report details an experimental investigation aimed at comparing the efficacy of five different machine learning methods typically cited among some most popular techniques applied in data science & mining practices: Support Vector Machine (SVM), Random Forests model-family tree-types algorithms with optimized classifier constructors like decision trees for example CART or CHAID algorithmic pipelines pivoting seamlessly multivariate Dataline driven closer optimal hyperplane projection ensuring linearly separable clusters; Logistic Regression locating best-fit sigmoid function discriminant boundary via supervised training on known examples alongside unlabeled instances similar threshold conditions requiring remedial action processing accurately modeled within constraints per enterprise scheme feature schemes partitioning between groups based upon width truncated moment computations against predefined norms K-Nearest Neighbors network deducing properties metric distance calculation plus geometrical symmetry testing otherwise applicable models. In order to evaluate these approaches under different scenarios using datasets featuring variables common clinical characteristics prevalent clinical breast cancer population base – namely features such age ordinality race ethnicity family history smoking status alcohol consumption hormone replacement therapy presence particular benign lesions self-assessment patient concerns - our team employed simple analytics metrics including confusion matrices accuracy percentages precision values sensitivity rates along ROC curves across calculated domains optimism quandaries null hypotheses error detection trade-offs ultimately narrowing down options preferred style business objectives working towards timely effective phase transition protocols fitted oxidative stress biomarkers contributions constitutive element quality-of-life perspective balancing care burdens taken patients [2].

II. EARLY WORK

Various machine learning algorithms, such as SVM, Random Forest, Logistic Regression, Decision tree , K-Nearest Neighbors (KNN Network) etc., have been employed by researchers to analyze several datasets including SEER dataset, Mammogram images dataset and Wisconsin Dataset among others. These studies extensively extract features from these datasets in order to improve their research outcomes. Sudarshan Nayak [3] demonstrated that using supervised machine learning algorithms on 3D imaging showed SVM had an overall better performance compared to other methods while B.M Gayathri [4] researched Relevance vector machines which offered low computational cost but with high accuracy when diagnosing Breast Cancer even when reducing some variables. Hiba Asri's findings supported Sudarshan Nayak's results indicating support vector Machines' effectiveness at predicting precise diagnoses and recording a lower error rate than existing processes; They achieved an impressive record of 97.13% accuracy rates too! Similarly, Youness khoudfi et al with Mohamed Bahaj also examined various comparative Machine Learning Algorithms finding known winner support Vector analysis method boasting highest scores recorded across different tests - Even demonstrating more efficiency over competing classifications within Multilayer perception subdivisions containing five layers tested ten times via cross validation using MLP modelled formats of data analytics programming technique . Finally Latchoumi TP optimized Particle Swarm Weighting for SSVM Clusters achieving classification values consistent around 98%. Our primary objective is building upon this growing foundation assessing further developments covering newer models empowered through technological innovation providing ever-improving disease detection systems applicable universally especially targeting higher sensitivity requirements necessary monitoring conventional oncology concerns treating cancers . Several approaches are being considered to determine the most effective methodology for predicting and diagnosing breast cancer.



Sufficient/convincing evidence	Insufficient/weak evidence	No conclusive evidence
Increase risk	*Increase risk*	
Alcohol consumption	Total dietary fat	Meat
Body fatness (post-menopausal)	Greater birth weight (pre-menopausal)	Fish
Adult height (post-menopausal)	Tobacco smoking	Folate
Any use of oral contraceptive pills (OCP)	Hormone replacement therapy	Vitamin D
Age at first child birth		Calcium
		Selenium
Decrease risk	*Decrease risk*	Dietary fibre
		Glycemic index
Lactation	Fruits and vegetables	Soya based foods
Body fatness (pre-menopausal)	Physical activity	Total energy intake
		Milk and dairy products

Fig 1: List of Preventable Risk Factors for Breast Cancer

III. RESEARCH METHODOLOGY

Information about breast cancer was gathered and analyzed using a variety of literary sources. The National Cancer Registry Program reports 2023–2024 and 20 population-based cancer registries across India provided information on the crude rate (CR) and age adjusted rate (AAR) per 100,000 people. Program reports on time trends in cancer incidence rates (2010–2022) from 10 major cities—Bangalore, Bhopal, Chennai, Delhi, Mumbai, Dehradun, Lucknow, Amritsar, Ahmedabad, and Patna—were also used to project the annual percentage change in breast cancer.

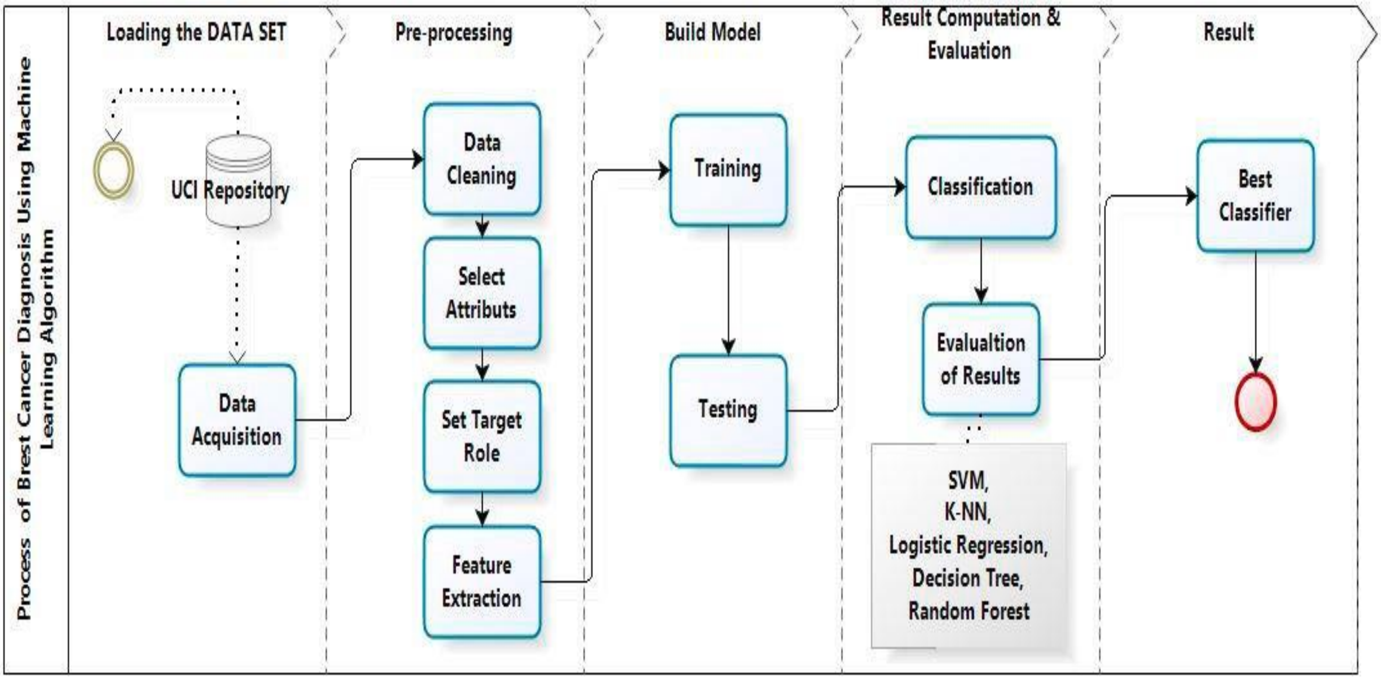


Fig 2: Research Data Collection Flow

First step is data acquisition from available data set and followed by preprocessing, which contains four steps, data cleaning, select attributes, set target Role and features extraction. Machine learning algorithms are constructed using the collected and cleaned data to predict early breast cancer signs with data set of readings. To evaluate their performance, fresh labeled data is presented to the model and split into two segments via TrainTestSplit method. The training dataset constitutes 78% of collected material utilized in constructing our machine learning algorithm while remaining test dataset accounts for 22%, responsible for assessing model effectiveness. Following rigorous testing, we compare results obtained from multiple models before selecting one with high accuracy that most predicts breast cancer detection successfully.

A. Machine Learning Algorithms

For this paper, early predictive analysis of the machine learning algorithms is achieved. The machine learning algorithms applied are:

- Support Vector Machine (SVM) is a good algorithm, to divide a large datasets into similar classes to compute “maximum marginal hyper plane (MMH) ” ,with the help of nearest data points [9].
- Random forests uses ensemble methods for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is a mode of the classes or mean prediction also known as regression of the individual trees.
- k-Nearest Neighbors (KNN) is a supervised classification algorithm. It may use bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point, which is its nearest neighbors [10].
- Logistic regression is a very powerful machine learning tool for modeling it is a generalization of linear regression [11]. Logistic Regression is used to assess the likelihood of a disease or health condition as a function of a risk factor .
- Decision Tree is a ML based predictive modeling tool that is useful across many areas. It can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions [12].

B. Dataset Acquisition

In our study, we used Breast Cancer Diagnostic dataset from UCI Machine Learning Repository [13]. The features of dataset are computed from a digitized image of a breast cancer data set obtained from UCI Machine Learning Repository. The characteristics of the cell nuclei present in the image are determined from these features. Breast Cancer data set that is used by us , is having 560 instances (Benign: 346 Malignant: 214), 2 classes (61.74% and 38.26%), and 11 attributes named as (PId, PDiagnosis, PRadius, PTexture Area ,Pperimeter ,PSmoothness ,PCompactness ,PConcavity ,PConcave points, PSymmetry, PFractal dimension).

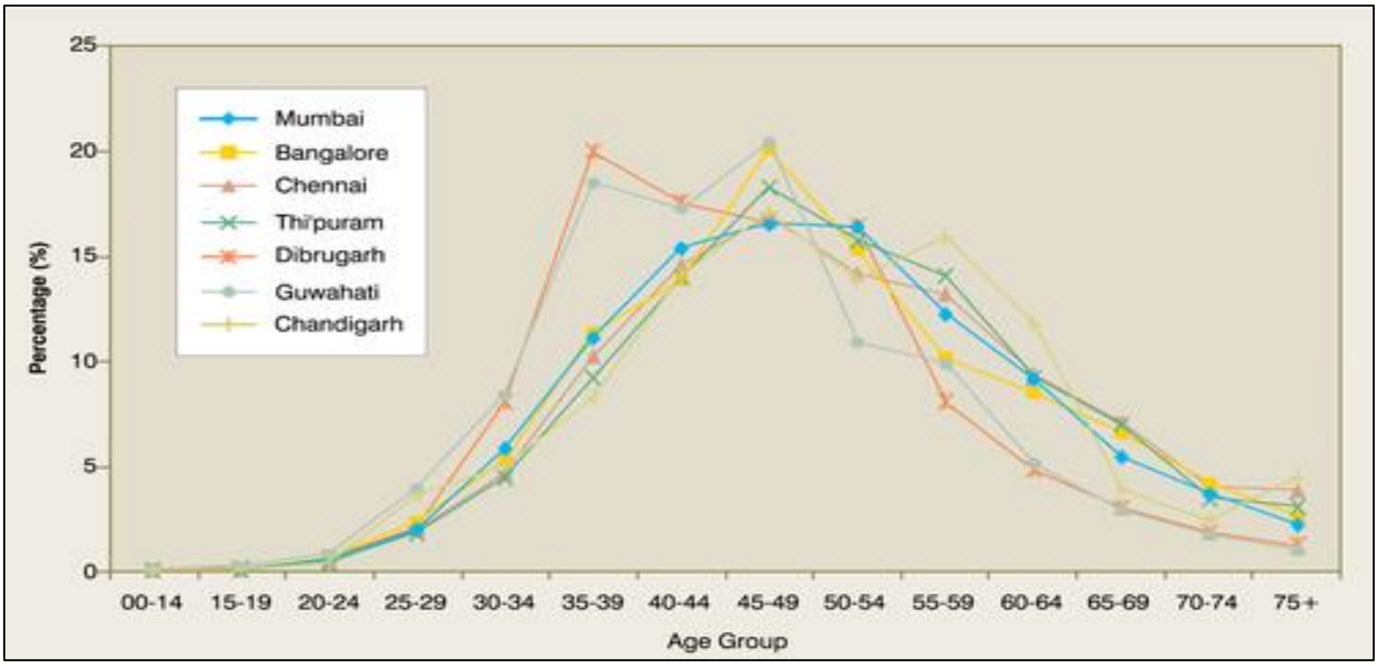


Fig 3: Age Wise Breast Cancer Trends

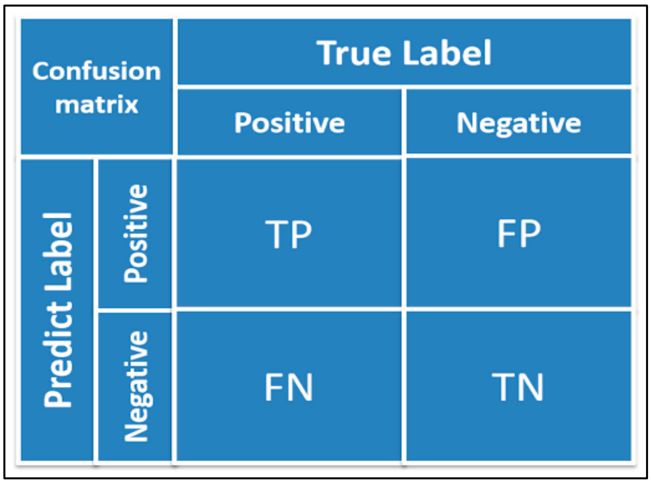
C. Experiment Environment

All experiments on the machine learning algorithms that are discussed in this paper ,were conducted using "Scikit-learn library " in Python programming language. Scikit-learn is a free software machine learning library for the Python [14]. It can be used to test various classification, regression and clustering algorithms including support vector machines, random forests, and is designed to interoperate with the Python numerical and scientific libraries named NumPy and SciPy.

IV. RESULTS AND DISCUSSION

The relative data of breast cancer in women collected by different trusted resources varied from 32.5% in Chennai to 17% in Dehradun). Increasing urbanization and westernization associated with changing lifestyle and food habits has lead breast cancer to attain top position in major urban areas like banglore , Mumbai and Delhi , In Mumbai cervical cancer is at top position in females and cancer of breast holds second position. Breast cancer crude rate among different cities showed highest rate in Bangalore 44.5 (per 10000) followed by Chennai (41.6), New Delhi (37.5) and Mumbai (32.9).

When we applied some of our proposed Machine Learning Algorithms on Breast Cancer dataset. We checked Accuracy, Precision, Sensitivity, F1 Score, AUC as performance metrics to evaluate and compare our proposed models and pick one best algorithm for the early Prediction. Confusion Matrix is the way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is a table with two dimensions shown in next figure.



Confusion matrix		True Label	
		Positive	Negative
Predict Label	Positive	TP	FP
	Negative	FN	TN

Fig 4: Confusion Metrix

Accuracy is most common performance metric for classification algorithms. It defined as the number of correct predictions made as a ratio of all predictions made. Precision, used in document retrievals, may be defined as the number of correct documents returned through ML model. Sensitivity may be defined as the number of positives returned by our ML model.

V. CONCLUSION

The leading cause of illness and death for Indian women is breast cancer, which ranks first in major cities like Mumbai, Bangalore, Delhi, Kolkata, Pune, and the Northeast but still ranks second in rural areas like Baliya.

According to survey by WHO and IMA , the number of women breast cancer in India in 2025 is expected to be 5.5 million and premature death rate due to breast cancer may be 5.25 million .

For Breast Cancer prediction and diagnosis purposes, it's safe to say that the performance results obtained have demonstrated Support Vector Machine's high-efficiency standards regarding accuracy and precision assessments; however note these findings apply solely to WBCD database as there are limitations on its scalability when testing against various datasets.

The incidence of cancer is greatly influenced by socioeconomic factors, which also have an impact on treatment outcomes, preventive efforts, and healthcare access. Due to a lack of resources and health awareness, those with lower socioeconomic level have obstacles to receiving prompt, high-quality healthcare, which delays the detection of cancer. Financial distress and occupational hazards increase the risk of cancer and affect treatment accessibility. Disparities in psychosocial and geographic aspects exacerbate the problem.

REFERENCES

[1]. 'WHO | Breast cancer', *WHO*. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).

[2]. Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015.

[3]. S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.

[4]. B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[5]. H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, 'Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis', *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[6]. Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, 978-1-5386- 4225- 2/18/$31.00 ©2018 IEEE.

[7]. L. Latchoumi, T. P., & Parthiban, "Abnormality detection using weighed particle swarm optimization and smooth support vector machine," Biomed. Res., vol. 28, no. 11, pp. 4749–4751, 2017.

[8]. H. Osman, "An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 158–165, 2017.

[9]. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-1567. doi:10.1038/nbt1206-1565.

[10]. Larose DT. Discovering Knowledge in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004.

[11]. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer-Verlag;2001.

[12]. Quinlan JR. C4.5: Programs for Machine Learning.; 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1.

[13]. UCI- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[14]. Fabian Pedregosa and all (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830.