International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Predicting Prognosis in Cancer Patients Using Machine Learning and Imaging Data

[1]Dr. Rashmi Gudur, [2]Himani Sivaraman,[3]Nilesh Gupta

Submitted:22/03/2023 **Revised:**26/05/2023 **Accepted:**12/06/2023

Abstract: Many distinct forms of cancer have contributed to cancer's label as a heterogeneous illness. In cancer research, an emphasis on early detection and prognosis of a cancer type is essential since it may improve clinical care of patients. Many research organizations in the biomedical and bioinformatics fields have studied the use of ML approaches for risk stratification of cancer patients because of its significance. Therefore, the goal of using these methods to simulate the development and management of malignant situations has been pursued. The capacity of ML algorithms to identify crucial elements in complicated datasets further highlights their significance. Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs), and Decision Trees (DTs) are only few of the methods that have been extensively used in cancer research to construct prediction models, leading to efficient and precise decision making. Although it is clear that ML approaches may enhance our knowledge of cancer development, they still need sufficient validation before they can be taken into account in routine clinical practice. In this paper, we summarize current ML methods used to simulate cancer development.

Keywords*: Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs), and Decision Trees (DTs)*

1. Introduction

The human body is made up of trillions of cells. A living organism consists of cells. Every part of our bodies, from our fingers to our noses and ears, reaches its optimal size by raising the number of cells to a predetermined maximum. Cell proliferation, also known as cell reproduction, is the process through which the body's cell count increases. Cell division is tightly controlled, which is why our nose fingers and other organs are all around the same size. Due to unchecked cell growth, an abnormal mass of cells (a tumor) has formed. Thus, cancer is characterized by the uncontrolled growth of cells. Cell growth is regulated by a plethora of chemicals. Cancer may be caused by alterations in any of these substances. Since several components change throughout tumor development, each tumor offers a rich source of information. Cancer is a global epidemic. Cancer is the second leading cause of death worldwide, behind cardiovascular disease. It has the distinction of being the most varied and complicated illness ever investigated. There is a plethora of information on cancer throughout the globe. However, a major obstacle to prognosis and therapy is presented by the disease's heterogeneity. Therefore, the tumor's categorization offers a window of opportunity for prognosis and therapy. However, it is challenging and time consuming to manually categorize the data. Furthermore, such categorizations may lead to incorrect cancer diagnosis and therapy. Classifying cancer data for diagnosis and therapy using ANNs is another use of this kind of technology. Machine learning (ML) is a growing field of study that has the potential to enhance cancer diagnosis, prognosis, and therapy. The field of ML falls within the umbrella of AI. Medical professionals may soon have access to a second opinion thanks to the ML categorization of cancer. The use of ML techniques has the potential to improve cancer prediction.

Classification using ML works especially well for genomic and proteomic measurements-based applications in biology. As cancer data sets continue to grow, ML has become an indispensable tool in the fight against the disease. Decision trees (DTs) and artificial neural networks (ANNs) have previously been used in diagnosis and cancer detection, but the ML technique is in line with the growing interest in customized and predictive medicine. All eyes were on ML because of how well it predicted cancer recurrence, mortality, and vulnerability. When it comes to cancer, the fundamental goals of diagnosis and detection are vastly different from the major goals of prediction and prognosis. Cancer prognosis and prediction focus on three primary areas: (a) Survivorship prediction for various cancer types. (b) Risk assessment for developing cancer. (c) Foretelling whether or not cancer will return.

1Krishna Institute of Medical Sciences, Krishna Vishwa Vidyapeeth "Deemed To Be University" Karad Malkapur, Karad (Dist. Satara), Maharashtra, India. PIN – 415539
Emai ID: anandgudur@gmail.com"
2Asst. Professor, Department of Comp. Sc. & Info. Tech.
Graphic Era Hill University, Dehradun Uttarakhand 248002, hsivaraman@gehu.ac.in
3Assistant professor Medi-caps University, Pigdamber, Rau, Indore
Email id: nileshgupta1112@gmail.com"

Cancer susceptibility prediction is estimating the likelihood of developing a certain cancer before the illness actually manifests itself. Tobacco use risk evaluation in the healthy population, for instance. Recurrence prediction is the estimation of how likely it is that cancer will return after first disappearing. Predicting the likelihood of a return of oral cancer after initial tumor suppression is one such example. However, cancer survival refers to how long a person is expected to live after receiving a cancer diagnosis. If someone has been diagnosed with oral cancer, they may wonder how their tumor will respond to treatment, how quickly it will spread, how long they have before they die, or what other alternatives they have. The success of a prognostic prognosis depends on the quality of the diagnosis in cases of cancer survival and cancer recurrence.

2. Machine Learning

For the purpose of making accurate diagnoses of cancer, machine learning (ML) methods are employed to build classification systems. These ML techniques allow for rapid, accurate inspection of healthcare datasets, benefiting professionals and novices alike in their efforts to minimize error. One obstacle to efficient machine learning categorization is the sheer volume of available data.

After a classification model has been developed, the performance of individual classifiers is measured by their accuracy, specificity, sensitivity, and area under the curve (AUC). The percentage of correctly detected outcomes/features by the ML model is represented by its sensitivity. The percentage of True Negatives (TNs) that are correctly detected by the ML classifier is indicated by the model's specificity.

The area under the ROC curve (AUC) provides a unified metric for evaluating the efficacy of an ML model at any given classification cutoff. You may get an overview of the ML classifier's efficiency in terms of various cutoffs from this curve.

Using ML methods, several cancer prediction models have been developed to learn diagnostic features and foretell disease development. Experts in the field of health management are hard at work creating new algorithms and tailoring ML in order to better categorize patients' data and evaluate it in order to provide accurate, error-free predictions. Many parameters are used in the ML platform for this aim. The following types of practical ML tools are extensively used to evaluate large datasets of cancer patients using filtering and classification techniques:

a) Naive Bayes Classifiers (NBC)

b) Logistic Regression (LR)

c) Support Vector Machines (SVM)

d) Artificial Neural Networks (ANN)

e) Classification and Regression Trees (CART)

f) K-Nearest Neighbors (KNN)

g) Multi-Layer Perceptron (MLP)

These methods unquestionably provide the correct response and aid in extending healthy aging and decreasing mortality rates. By aiding healthcare personnel with clear guidance and rapid identification, ML technologies have the potential to save lives. ML programs are very useful for accurately classifying large datasets. This provides cancer sufferers with an immense amount of hope. If surgeons had access to reliable prediction techniques, they could track the drug's efficacy in treating breast cancer, as well as any side effects, countermeasures, or individual responses to treatments. Useful inquiry methods for uncovering latent links in huge BC data may be built with the use of ML technologies.

The research design describes in-depth methods for coordinating the study's many parts so that they work together effectively and efficiently to solve a complex issue. Methods for prediction and prognosis modeling using ensembles of machine learning classifiers were the focus of this research. In order to better the current intelligent computational tools in the detection and prognosis of breast cancer, this research compared the performance of several Machine Learning classification algorithms. We used a machine learning (ML) ensemble classification approach on many cancer datasets. The data set on cancer may include duplicate information or missing values. Consequently, we place a premium on ensuring that the database contains neither redundant nor missing data. Standardizing or normalizing the biological dataset and extracting crucial features from the input biological dataset have received more focus in the future.

3. Literature Review

The efficacy and accuracy of gene expression data in contrast with medical data in survival projections has been evaluated (Bashiri et al., 2017). Confirmation that a very large quantity of biological data is indeed accessible. There is fresh work being done in this area to accurately analyze such data using machine learning.

Others have noted (Suryachandra et al., 2016; Maurya et al., 2019) that comparing the efficiency of two additional algorithms is a good way to improve the efficiency of ML tools. An accurate diagnostic assessment algorithm that can reliably forecast the illness and risk of cancer and aid the treating physician in his therapy is the ultimate objective of the ML algorithm in cancer. This is

shown by a level that represents the stage of the patient's cancer. The variety of data and reported situations is the primary challenge for ML algorithms.

According to research (Mattiuzzi & Lippi, 2020), one of the most crucial steps in monitoring is the timely prediction of BC and their signs. Using ML to discover patterns is a great way to reduce the amount of false positives and negatives in your evaluations. About 80% of all malignancies in affluent nations are detected in persons aged 50–55, and this is true for both men and women. Cancer may have several causes, including but not limited to excessive cigarette use, radiation exposure, obesity, drug use, alcohol use, and the introduction of faulty genes.

Proposed System

As can be seen in Figure 1., there are three distinct stages to the proposed system. Here are the specifics of each of the three phases: During Stage 1, biological datasets are being compiled. The experimental study of cancer datasets comes from a variety of reference databanks. The strategy is excellent for pre-processing or standardizing cancer datasets and extracting significant features.

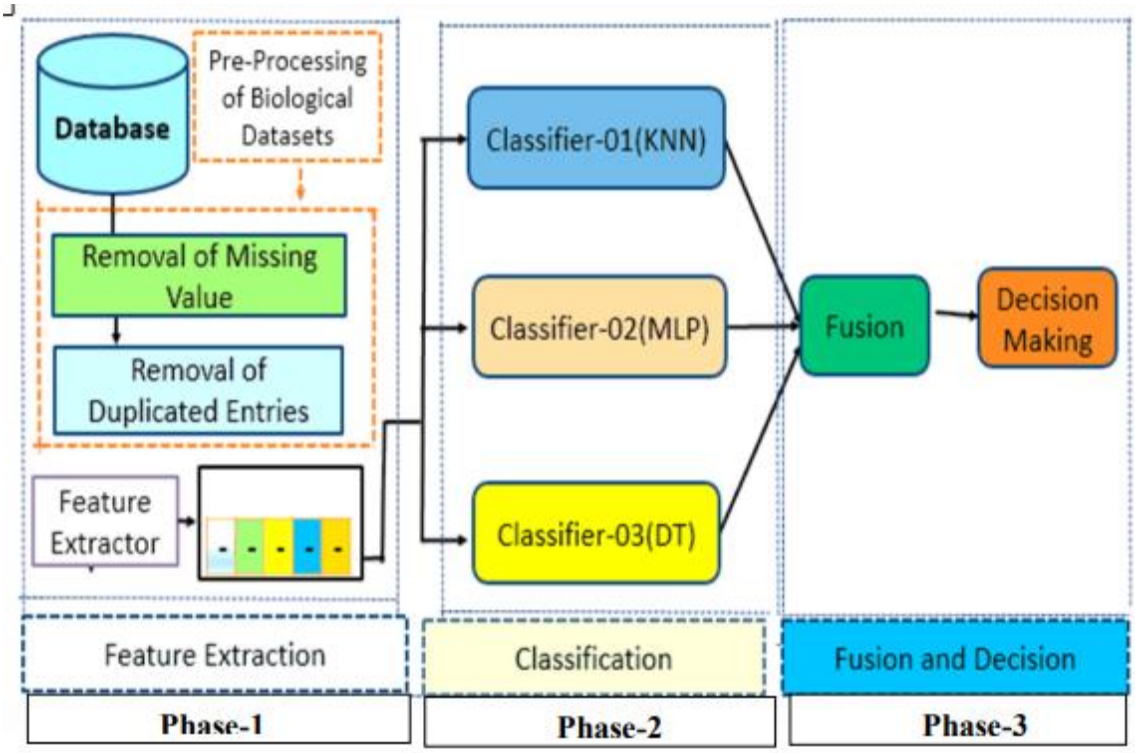

Fig 1 Proposed Methodology for Adaptive Integrated Model (AIM)

After completing the trimming phases of preprocessing, in contrast to the missing value removal method, the later dataset was constructed without any room for mistake. Feature extraction is one of the other crucial and often utilized phases in the proposed architecture (refer to Figure 1), and its primary purpose is to reduce the amount of features derived from gene expression. This helps immensely in breaking the jinx of high-level Dimensionality issues. When preparing a dataset for subsequent categorization, feature extraction may be used to weed out irrelevant information and reveal the most useful details.

Step Two: Here we train and test our ML classifiers. The ML classifier was trained using a k-fold cross-validation/resampling approach on the training dataset.

In the third stage, we discuss the fusion plan. The most up-to-date literature on cancer and ML suggests that unique ranking-based algorithms have been utilized to choose and exclude appropriate classifier outputs for predicting class labels.

Proposed System Algorithm

First, a cancer dataset is constructed by collecting data from various sources.

The second phase involves cleaning the data by removing duplicates and filling up any blanks in the Cancer Dataset. a) Delete Duplicate Entries: A deletion algorithm based on Euclidean distance is utilized to do this. To address the issue of missing data in the cancer dataset, the distance-based Mean Method (k-Nearest Neighbor) is used.

Extraction of Features, Double RBF Kernels have been utilized for feature extraction using the filter-based feature ranking method.

Three classifiers (KNN, MLP, and DT) have been trained and tested to evaluate the model's performance in this stage, stage 4: Classification. To assess how well models perform on the biological dataset, a k-fold cross-validation or resampling technique was used (where k is the number of sets/groups).

Picking the Right Parameters: We used a hit-and-try approach to get the optimal values for the Machine Learning (ML) classifier parameters.

Classifier Ordering is the Sixth Step: To determine where one classifier stands in relation to another on a test dataset in terms of accuracy and variety, use the rank pointer. Review of the Model: Six different metrics have been used to quantify the AIM's effectiveness. Accuracy, Area Under the Curve, Sensitivity, Negative Predictive Value, Positive Predictive Value, and Specificity are the measures.

Biological Datasets for Experimental Analysis

Carcinoma, lymphoma, and sarcoma are all examples of cancer that begin in different tissues. Multiple heterogeneities may coexist in a single tumor. Due to the varied nature of cancer, treating it is a time-consuming endeavor. Poor diagnostics contribute to more than half of all cancer deaths in third world nations. Breast cancer is the most diverse form of the disease.

Cancer Prediction on Prognostic Dataset

Clinical research into potentially fatal diseases was evaluated using the WPBC dataset for 43 recurrent cancer patients and 143 nonrecurrent cancer patients. After feature selection, the classification findings are reviewed in detail.

Table 1: Training performance of a single ML classifier on WPBC Dataset.

Techniques	Accuracy	Sensitivity	Specificity	NPV	PPV	AUC
KNN	93.33	97.70	81.81	93.40	93.10	0.92
MLP	95.00	98.85	84.84	94.50	96.50	0.93
DT	94.82	97.86	85.84	95.60	96.72	0.94

Both Sensitivity and Specificity may be thought of as diagnostic tools. Specificity, in the context of a cancer screening test, refers to the percentage of healthy individuals in a community or subset of a population that test unfavorably for the illness. False positives (FP) occur when a screening test returns a positive result for an illness in a person who does not really have that condition. Table-1 shows the values for the KPIs that may be found in the WPBC dataset. It was hypothesized that the system performed quite well on the WPBC dataset because the predicted performance of the MLP classifier is higher than the individual performance of other classifiers in terms of accuracy (95.00%), sensitivity (98.85%), specificity (84.84%), NPV (94.50%), PPV (96.50%), and AUC (0.93%).

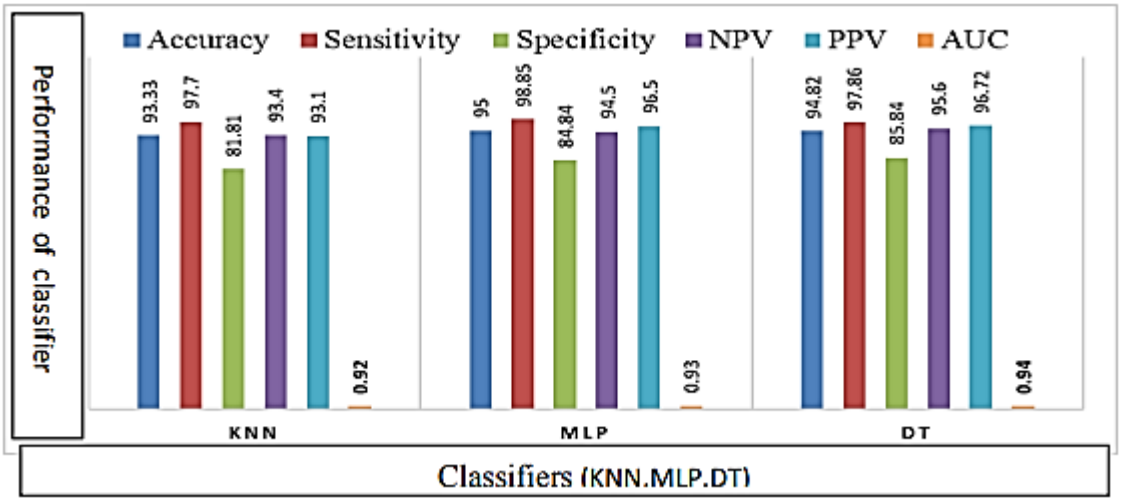

Fig 2: Training performance of classifiers on WPBC dataset

ROC Curve

Diagnostic performance may be shown using ROC curves, which plot the proportion of correct diagnoses (TPs) versus the number of incorrect diagnoses (FPs) for a given class. At the point when the AUC and the accuracy of predicting a class are equal, the graphical plot is curved. Without considering the cost of errors or the distribution of classes, the ROC curve illustrates the performance of a machine learning classifier. The area under the curve (ROC) is also known as the sensitivity versus (1)specificity graph. The ROC curve seen in Figures 3 and 4 is comprised of 1-specificity and sensitivity.

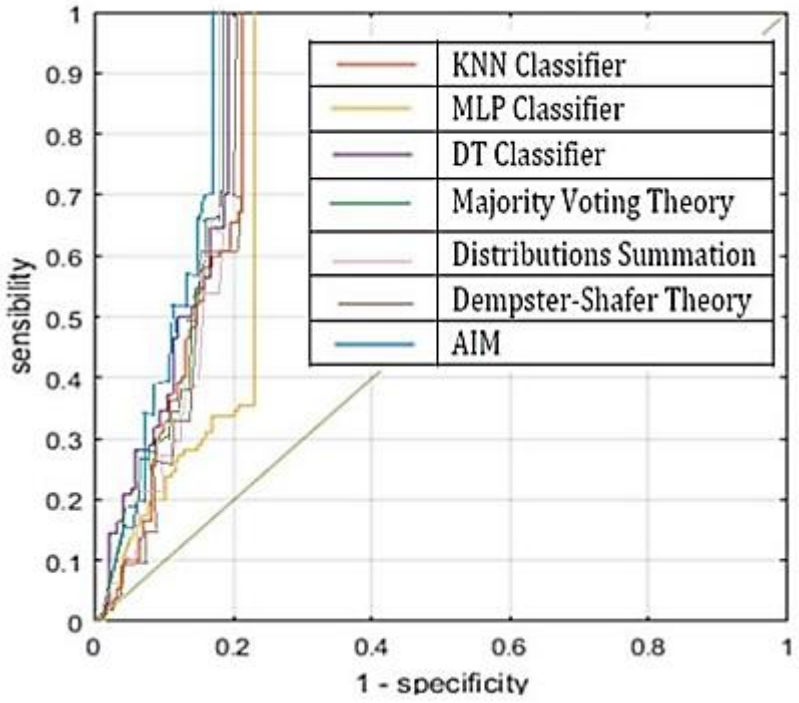

Fig 4: Training data ROC for WPBC dataset

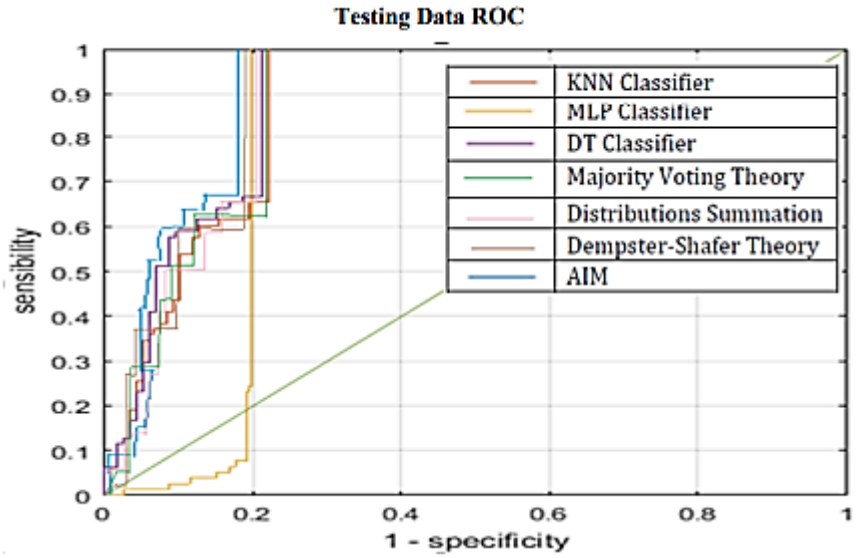

Fig 5: Testing data ROC for WPBC dataset

The ROC curve for 1-specificity and sensitivity is shown in Figures 4 and 5. The ROC curve, which favors sensitivity and increases as accuracy increases, demonstrates the reliability of the suggested AIM.

4. Conclusion

The logic behind a model will assist users evaluate the reliability of its predictions. The accurate predictions provided by ML models may be useful to human decision-makers in a variety of settings. However, such

forecasts lose value if the person cannot present evidence for why they should be believed. The suggested AIM method for cancer prediction and prognosis generates justifications for the cancer prediction domain. Task-based experiments reveal that the proposed strategy increases computational biologists' trust in their predictions and satisfaction with the reasoning. The estimated precision in the proposed AIM has been compared with the results of state-of-the-art fusion approaches like MVT, DS, and DST on a number of different biological datasets.

References

[1] Bashiri, A., Ghazisaeedi, M., Safdari, R., Shahmoradi, L., & Ehtesham, H. (2017). Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. Iranian journal of public health, 46(2), 165.

[2] Maurya, R. K., Yadav, S. K., & Tewari, P. (2019, December). Epidemiology of Breast Cancer (BC) and Its Early Identification via Evolving Machine Learning Classification Tools (MLCT)–A Study. In International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making (pp. 109-119). Springer, Cham.

[3] Suryachandra, P., & Reddy, P. V. S. (2016, August). Comparison of machine learning algorithms for breast cancer. In 2016 International Conference on Inventive Computation Technologies (ICICT) (Vol. 3, pp. 1-6). IEEE.

[4] Mattiuzzi, C., & Lippi, G. (2020). Cancer statistics: a comparison between world health organization (WHO) and global burden of disease (GBD). European journal of public health, 30(5), 1026-1027

[5] Balakumar, A., & Senthil, S. (2020). Machine Learning Is the Future for Lung Cancer Prognosis and Prediction. In Applications of Deep Learning and Big IoT on Personalized Healthcare Services (pp. 176-196). IGI Global.

[6] Kurian, A. W., Bernhisel, R., Larson, K., Caswell-Jin, J. L., Shadyab, A. H., OchsBalcom, H., & Stefanick, M. L. (2020). Prevalence of Pathogenic Variants in Cancer Susceptibility Genes Among Women With Postmenopausal Breast Cancer. Jama, 323(10), 995-997.

[7] Maes, F., Robben, D., Vandermeulen, D., & Suetens, P. (2019). The Role of Medical Image Computing and Machine Learning in Healthcare. In Artificial Intelligence in Medical Imaging (pp. 9-23). Springer, Cham.

[8] Maes, F., Robben, D., Vandermeulen, D., & Suetens, P. (2019). The Role of Medical Image Computing and Machine Learning in Healthcare. In Artificial Intelligence in Medical Imaging (pp. 9-23). Springer, Cham.

[9] Mangai, U. G., Samanta, S., Das, S., & Chowdhury, P. R.(2010). A survey of fusion and feature fusion strategies for pattern classification. IETE Technical Review,27(4).

[10] Mani, S., Chen, Y., Li, X., Arlinghaus, L., Chakravarthy, A. B., Abramson, V.,& Yankeelov, T. E. (2013). Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy. Journal of the American Medical Informatics Association, 20(4), 688-695.

[11] Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC research notes, 4(1), 1-14.

[12] Mathur, M. R., Singh, A., Dhillon, P. K., Dey, S., Sullivan, R., Jain, K. K& Rajaraman, P. (2014). Strategies for cancer prevention in India—Catching the 'low hanging fruits'. Journal of Cancer Policy, 2(4), 105-106

[13] Mccarthy, J. F., Marx, K. A., Hoffman, P. E., Gee, A. G., O'neil, P., Ujwal, M. L., & Hotchkiss, J. (2004). Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. Annals of the New York Academy of Sciences, 1020(1), 239-262.

[14] Mihaylov, I., Nisheva, M., & Vassilev, D. (2019). Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies. Information, 10(3), 93

[15] Moslemi, A., Mahjub, H., Saidijam, M., Poorolajal, J., & Soltanian, A. R. (2016). Bayesian survival analysis of high-dimensional microarray data for mantle cell lymphoma patients. Asian Pac J Cancer Prev, 17(1), 95-100.

[16] Ahire, P. G. ., & Patil, P. D. . (2023). Context-Aware Clustering and the Optimized Whale Optimization Algorithm: An Effective Predictive Model for the Smart Grid. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 62–76. https://doi.org/10.17762/ijritcc.v11i1.5987

[17] Paul, D., Su, R., Romain, M., Sébastien, V., Pierre, V., & Isabelle, G. (2017). Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Computerized Medical Imaging and Graphics, 60, 42-49

[18] Pearlman, R., Frankel, W. L., Swanson, B., Zhao, W., Yilmaz, A., Miller, K., ... & Goldberg, R. M. (2017). Prevalence and spectrum of germline cancer susceptibility gene mutations among patients with early-onset colorectal cancer. JAMA oncology, 3(4), 464- 471.

[19] Priya, A. M., Kannammal, K. E., Kavya, S. P., Iswarya, R., & Nivetha, M. (2019). IDENTIFY BREAST CANCER USING MACHINE LEARNING ALGORITHM. Journal Homepage: http://www. ijesm. co. in, 8(4).

[20] Sharma, A., Kulshrestha, S., & Daniel, S. (2017, December). Machine Learning approaches for breast cancer diagnosis and prognosis. In Soft Computing and its Engineering Applications (icSoftComp), 2017 International Conference on (pp. 1-5). IEEE.

[21] Prof. Sharayu Waghmare. (2012). Vedic Multiplier Implementation for High Speed Factorial Computation. International Journal of New Practices in Management and Engineering, 1(04), 01 - 06. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/8

[22] Liu, S., Xu, C., Zhang, Y., Liu, J., Yu, B., Liu, X., & Dehmer, M. (2018). Feature selection of gene expression data for cancer classification using double RBF-kernels. BMC bioinformatics, 19(1), 1-14