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

The Application of Data Mining Techniques to the Detection of Cancer

[1]Raghavendra R, [2]Neeraj Kumari, [3]Surendra Yadav, [4]Prabha Shreeraj Nair

Submitted:18/04/2023 **Revised:**07/06/2023 **Accepted:**23/06/2023

Abstract: Cancer is one of the leading causes of mortality worldwide. In 2018, there were approximately 1,735,350 new instances of cancer identified in the United States alone, and 609,640 individuals passed away as a direct result of the disease. Cancers include skin melanoma, lung bronchus cancer, breast cancer, prostate cancer, colon and rectum cancer, bladder cancer, kidney and renal pelvis cancer, and others. Cancer has risen to prominence in the scientific community due to the wide variety of cancers and the enormous number of people it affects. There is still active research on cancer prevention and diagnostic strategies. Using data mining methods, we sought to create a reliable and workable system for cancer diagnosis. Machine learning techniques may assist professionals in creating tools that enable early cancer detection. To improve cancer diagnosis rates, this research aims to introduce a novel machine learning method called the Elephant herding optimized logistic regression (EHOLR) strategy. Histogram equalization (HE) was used for preprocessing the acquired cancer data, and linear discriminant analysis (LDA) was used to extract the data's features. Finally, cancer detection is accomplished using our recommended strategy. The effectiveness of the suggested strategy is then assessed using the performance matrix, namely accuracy, recall, and precision..

Keywords: *Cancer detection, data mining, histogram equalization (HE), linear discriminant analysis (LDA), Elephant herding optimized logistic regression (EHOLR).*

1. Introduction

Cancer analysis has made extensive use of data mining and machine learning. Furthermore, data mining and machine learning aid medical researchers in recognizing correlations among factors, allowing for the prediction of illness outcomes based on records. Breast tumor detection, diagnosis, and the avoidance of unnecessary treatment are all areas where machine learning may be put to use. It could also help in making sound judgments. The purpose of this study is to evaluate the impact of data mining and machine learning on finding and diagnosing breast cancer[1]. Cancer is a prevalent illness brought on by alterations in genes controlling cell division and proliferation. Mutation in the DNA building blocks of genes is one example of these easily detectable shifts. Compared to healthy cells, cancer cells usually have many more mutations. However, individual tumors might have a wide variety of genetic abnormalities. Still, the disease may be the indirect cause of some of these noticeable shifts. Further shifts, including the determination of clinical issues, scientific data, and the implementation of the new field of neurooncology, will occur as cancer continues to spread. As a result, better treatment options and longer life expectancy are possible because of cancer identification at an earlier stage[2].

The goal of data mining is to unearth previously unknown but highly pertinent data. It's a method that has been used effectively for making forecasts. Developing fully automated breast cancer diagnosis models has previously attempted using a number of Collecting and analyzing data with computers methodologies. Breast cancer prevention and treatment methods detection include logistic regression, decision trees, K-Nearest Neighbor, neural networks, AdaBoost algorithms, and Support Vector Machine (SVM) [3]. When it comes to female death rates, breast cancer is in second place. The American Cancer Society predicted that 272,600 women would be diagnosed with carcinoma of the breast and 62,970 women would be treated with cancer that is not invasive in 2019. The greatest strategy to improve the likelihood of treatment and survival is via early identification. Marketers, social scientists, financiers, and medical researchers have all found success using data mining as a technique for knowledge discovery. To anticipate patients' medical diagnoses, various classification algorithms have recently been used for healthcare data sets. Tumor behavior in cancerous breast patients might, for instance, be evaluated using artificial intelligence methods[4]. Estimating the amount of time cancer patients will live is crucial. Patients

1Assistant Professor, Department of Computer Science and IT, Jain(Deemed-to-be University), Bangalore-27, India, Email Id: r.raghavendra@jainuniversity.ac.in
2Assistant Professor, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India, Email id: arun.k.chauhan@relianceada.com
3Professor, Department of Computer Science & Application, Vivekananda Global University, Jaipur, India, Email Id: surendra.yadav@vgu.ac.in
4Associate Professor. & Dy. HoD, Department of Information Technology and M.Tech Integrated, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India, Email id: prabha.snair@niet.co.in

diagnosed with cancer often seek out information about their outlook and possible outcomes. Planning therapy, considering possible modifications in lifestyle, and calculating potential costs are all aided by this. To enhance their ability to assist their patients, it also aids physicians in identifying potential therapy avenues, comprehending potential outcomes of various prognoses, and making more data-informed judgments. Making a medical prediction and identifying survival variables, on the other hand, are very difficult tasks. These elements may be broken down into two broad classes: time and biology. Time-dependent variables include lymph node status, tumor size, and histology stage, all of which may vary throughout treatment[5]. Survival rates for cancer patients may be correlated with a wide range of clinical characteristics using data mining methods. Due to its efficacy, data mining has become an inspiring and fascinating tool for medical professionals. It provides a wide variety of categorization and regression strategies that may be used for tumor data to construct diagnostic and prognostic models. Prognostic factors for surviving cancer have been identified using a variety of mathematical and statistical methods. Classification or regression methods are used to predict the probability of mortality based on factors that are likely to be clinical and sociological in nature [6]. Cervical cancer awareness was examined in cross-sectional research. One kind of cancer that disproportionately affects women is cervical cancer, which targets the cervix. Cervical cancer develops when cervix cells undergo a malignant transformation. Human papillomavirus (HPV) is the primary cause of cervical cancer and is spread mostly via sexual contact. The human papillomavirus (HPV) family causes infections in the reproductive system of sexually engaged people. The immune system of a woman who has been exposed to HPV can always prevent further infection, but in some women, the virus may remain dormant for decades before causing cervical cancer. According to, all women between the ages of 30 and 49 who fall within the target age range should get a screening test at least once in their lives. Cancer mortality rates have been falling because of improvements in cancer screening and early diagnosis. [7].

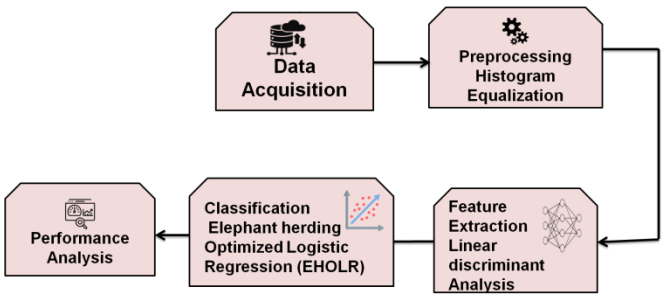

Fig 1: A block diagram showing the steps involved in detection of cancer

2. Related works

Cancer of the breast is a tumor that develops in the breast tissues. It is considered to be among the most frequent form of cancer in females and a major killer of women worldwide. This article presents an in-depth analysis of how far we've come in predicting breast cancer using deep learning, ML, and data mine. The diagnosis and prognosis of breast cancer have been the focus of a great deal of study, although the success of various approaches varies widely depending on the circumstances, the methods, and the information sets used. To determine the best suitable approach that would support the enormous dataset with outstanding precision of prediction, compare and contrast several current Machine Learning and data mining methodologies[8]. Over the last decade, data mining's prominence has grown, and a great deal of research has been conducted into its potential uses. The healthcare sector serves as the majority of the applications are launched, and these may be broken down into two categories: decision support for clinicians and policymaking. However, it remains difficult to locate relevant publications in the field of healthcare[9].Predictive analysis using data mining methods, such as decision tree algorithms, is utilized in the field of biomedicine. The lopsided dataset was retrieved from the University of California, data-set collection. The data set with an enhanced number of occurrences has been balanced using the Synthetic Minority Oversampling Technique (SMOTE). Patients' ages, the number of children they've had, whether or not they used birth control, whether or not they smoked, and when they were diagnosed with STDs are all included in the dataset.Data mining methods such as the Boosted decision tree, the decision forest, and the decision jungle algorithms predict the likelihood that a patient will develop cervical cancer based on the results of a screening test [10].Among females, breast cancer ranks highest in terms of mortality, accounting for 627,000 deaths in only one year. The very high death rate from breast cancer must be addressed, ideally by early identification leading to timely prevention. Data mining has several uses in Brest cancer prediction and may help advance cutting-edge research. In data mining classification approaches to the problem of distinguishing between the two most common types of breast cancer [11].Gastric cancer, in which malignant cells form and spread throughout the stomach, is one of the leading causes of mortality globally. Patients' survival rates may be increased if the decision-making process that leads to the selection of appropriate treatment options is enhanced. Now that hospitals collect and store so much patient data, data mining techniques may be used to enhance healthcare delivery. This CRISP-based research seeks to predict not only the death rate but also the presence of any postoperative complications in patients

with this condition. To better anticipate outcomes, many categorization models were evaluated and compared [12].The likelihood of survival improves when cancer is diagnosed at an early stage. This work introduces an IDSS for exploiting genetic expression patterns obtained from microarrays of DNA in the early detection of cancer. This is because there are often just a few hundred examples in such databases, but thousands of genes to consider. To prevent excessive fitting, it is crucial to exclude traits (genes) that are not associated with the illness of concern. The suggested strategy prioritizes input patterns by maximizing the information gain (IG). The grey wolf optimization (GWO) technique is then used to further prune the remaining characteristics (genes). Finally, a classifier based on a support vector machine (SVM is used in the algorithm to categorize cancer types. Segmentation accuracy, the most crucial performance metric in illness diagnosis, was used to assess the effectiveness of the proposed technique across both data sets (Breast and Colon) [13].The proliferation of computing power across all industries has resulted in a deluge of data that has to be mined for insights. The goal of data mining is to uncover previously unknown information and analyze it using several criteria to conclude. One of the numerous fields where data mining has found use is medical diagnostics. Many illnesses now have a stigma of being fatal. Some of the worst include cardiovascular disease, breast cancer, and diabetes. This report examines 168 papers about the use of data mining in the diagnosis of these conditions. The research focuses on 85 articles that have gotten a lot of attention and provide careful attention to every algorithm, data mining model, and assessment technique. The objective of the research is to identify the best effective data mining techniques used for medical diagnosis. The identification of research gaps in the use of data mining in healthcare is another important finding from the study [14].Data preparation, data mining, and result expression and analysis are the primary components of data mining techniques that allow it to search for potentially useful information from enormous amounts of data. It uses databases, which are an established sort of information processing technology. Database systems are the study, management, and use of databases in the field of computer science. Database processing and analysis require an understanding of the theory and methodology of database storage, design, maintenance, and use [15].

3. Data Acquisition

The breast cancer FNA image is the source of the features used to create this dataset. The phenotypic of interest stores the prognosis (malignant vs. benign). The dataset has a total of 286 occurrences, including 85 repeats and 201 events that don't happen again. The Breast Cancer dataset had 8 records where the value for the defining condition was missing.

Due to the incomplete and unbalanced data in the two datasets utilized in this study, a significant amount of time will be spent preparing the data before running the tests to improve the classifier's efficiency. During preprocessing, your data discrepancies and values that are missing will be handled. All cases with no values are deleted to handle the absence of characteristics. Either the classifier or the training set has to be rebalanced to address the issue of uneven data. The resample filter is employed to do this artificial rebalancing.

4. Preprocessing using Histogram Equalisation (HE)

The segment filter was first used to discretize the data before the values that were missing were purged. Second, we used the resample filter to resample cases while keeping the subsample's class composition and somewhat favoring a uniform distribution among classes. The classification performance is shown to be enhanced. Finally, trials were conducted across three classifiers Naive Bayes, SMO, and J48 using cross-validation. Three phases of the data preparation approach have been implemented: discretization, resampling of instances, and elimination of missing values. Thereafter, cross-validation was used. Three classifiers were then tested using the prepared data sets.

To increase the contrast, HEs alter the brightness levels or arrangement of individual pixels. They are often classified as either global (GHE) or local (LHE). GHE-based computations use a variety of techniques to enhance images, including standard HE, Histogram Charting which converts every pixel in an image's illumination distribution into an ideal indicate histogram, and image breakdown which splits the initial image into numerous smaller ones based on intensity levels. In most cases, LHE will adjust the brightness of each pixel in an image based on the parameters of the window of movement to which it has been allocated.

Traditional HE typically flattens and stretches the histogram of an image's spectrum to improve the image's contrast. Let's pretend i is the frequency of the grayscale value that appears in picture y. The probability that the picture will include a pixel with intensity level i is

$$p_y(i) = \frac{n_i}{n},\ 0 \le i < L \quad (1)$$

$$CDFy = \sum_{j=0}^{i} y(j) \quad (2)$$

$$CDF(i)_y = iK \quad (3)$$

$$y = CDF\ (y)_y \quad (4)$$

$$y' = y.(\max\{y\} - \min\{y\}) + \min\{y\} \quad (5)$$

5.Feature extraction in linear discriminant analysis

LDA is a transform-based technique that seeks to reduce the amount of variation both within and outside of each class. The next sections elaborate on the mathematical formulation central to LDA theory.

computed using

$$S_x = \sum_{i=1}^{L} X_i E\{(X - A_i)(X - A_i)^T\}, \quad (6)$$

where Ai is the median of class i and Xi is the proportion of instances in the training data that belong to class i. The dispersion of the predicted vectors around the global mean is defined by a between-class scatter matrix, which is calculated as

$$S_b = \sum_{i=1}^{L} X_i (A_i - A_o)(A_i - A_o)^T \quad (7)$$

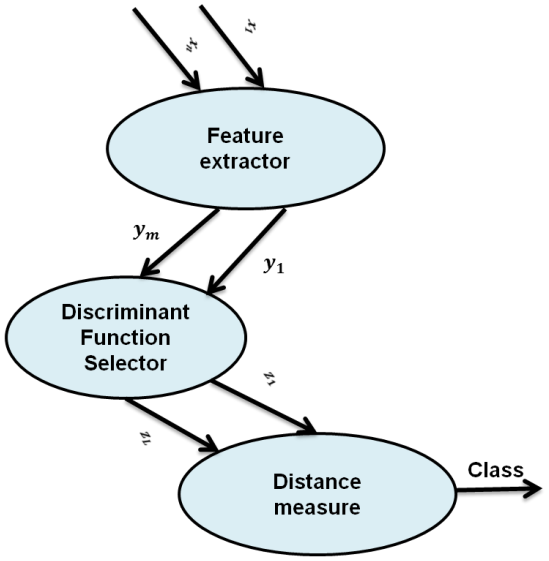

Fig 2: High-level perspective of the classification issue, outlining the typical structure of a classifier.

where A_O is the average of all classes, A_i is the average of class i, and Pi is the proportion of instances of class i in the training set.

The matrix of covariance of all samples is used in Equation 3 to get the overall or mixture scatter matrix.

$$S_b = \sum_{i=1}^{L} X_i (A - A_o)(A_i - A)^T \quad (8)$$

$$Criterion = 1n|inv\ (S_w) \times S_b| \quad (9)$$

$$Criterion = S_b - \mu(S_x - c) \quad (10)$$

$$Criterion = inv(S_x) \times S_b \quad (11)$$

To maximize the separation of classes in the transformed space, LDA employs the latter.

$$\Phi_1 = \frac{s_x^{-1}(M_i - M_j)}{|s_x^{-1}(M_i - M_j)|} \quad (12)$$

where i and j correspond to different classes.

After obtaining the transforms, both the practice and test sets are changed to the new space. Each test vector is allocated a class depending on its Euclidean distance from the centers of the classes in the modified space.

6. Classification using Elephant herding optimizing logistic regression

6.1. Logistic regression

Predicting the likelihood of an event by fitting data to a logistic function, LR is a popular quantitative decision-support tool. Using medical, socioeconomic, and other data, it predicts immediate medical results for particular patients. Indeed, LR is a multivariate approach. Two or more separate factors are used to predict a single dependent parameter. In this investigation, we used binary LR to predict one of two possible classes of outcomes. Additional information may be gathered from the LR model and should be incorporated into decision-making, even if the main output of the model is the predicted odds or probability of a binary event. If there are two possible groups, and one has a larger than 50% chance of being correct, we call that group "1." The value "0" is used for anything else.

$$P(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots.\beta_n x_n)}} \quad (13)$$

$$= \frac{1}{1+e^{-(b^T \times X)}} \quad (14)$$

6.1 Elephant herding

Swarm-based search methods like EHO are used for many optimization issues. The shepherding habits of a herd of elephants inspired this algorithm. Elephants live in groups, with a Matriarch at the head of every group. The adult male elephants separate from the herd. As a result of these mammoth actions in groups, we get two operators: the clan updating operator and the separating operator

Various forms of worldwide efficiency problems may be solved by adhering to a set of guidelines. The following are the guidelines:

1. Each family unit is led by a Matriarch who oversees a herd of elephants.
2. Elephants live in groups called clans, and within each clan, the population never fluctuates.

3. Adult male elephants gradually separate from the herd throughout many generations.

A. Clan updating operator

In clan pk, the matriarch knows where each elephant will go in the next group movement. So, the elephant's new clan pk rank may be calculated as

$$Z_{new,pk,t} = Z_{pk,t} + \delta * \left(Z_{best,_{pk}} - Z_{pk,t}\right) * r \quad (15)$$

where,Znew,pk,t = elephant t's new clan rank. pk

Zpk,t = previous elephant t's place in clan pk

Zbest, pk means the strongest elephant in the pk tribe.

In a normal distribution, r equals [0, 1].

Zpk,t is a scaling factor that is impacted by matriarch pk, and the value of falls between 0 and 1.

The healthiest member of the elephant family is characterized by:

Z_(〖new,〗_(pk,t))=γ*Z_(center,) p_k (16)

Znew,pk,t is compiled from the collective knowledge of all the elephants in

clan. pk

γ ∈ [0, 1] = determines how much Znew,pk ,t is affected by Zcenter,pk

Zcenter,pk = center of clan pk and can be obtained by the given equation for dth dimension

Z_(〖new,〗_(pk,t))=γ*Z_(center,) p_k (17)

Here 'd' shows the dth dimension i.e. 1≤d≤D where D shows the total

dimension.

Zpk,t,d is the dth of Zpk ,t elephant

θpk = number of elephants in clan pk

Pseudo-code of clan updating operator:

B. Operator separation

When the young male elephants in an elephant group mature into adulthood, they disperse and start new lives on their own. Considering that the elephant with the poorest fitness is used to apply the separation operator at every stage improves the accessibility of the EHO technique, as the resulting equation reads:

Z_(〖new,〗_(pk,t))=γ*Z_(center,) (18)

The EHO method has developed using the data provided by the clan updating operator and the separating operator, and its main features are as follows:

Z_(〖new,〗_(pk,t))=γ*Z_(center,) p_k (19)

7. Result

In order to optimize and identify cancer in the simulation of data mining, this study explores the efficacy of a novel evolutionary strategy. Data mining was regarded as the foundational strategy for reaching this goal. Then, to fine-tune the computational parameters, it was combined with an elephant herding optimization method. In this article, we undertake a comprehensive review of how to use LDA and its variants to construct classification models using microarray data. The results of the research demonstrated that the modified approaches outperformed LDA in terms of accuracy in prediction. The efficacy of the suggested approach is covered in this section. The recommended system's capacity to achieve Accuracy, Precision, Recall, F1-Score, Sensitivity, and Specificity justifies its adoption.

A.Accuracy

Accurate assessments in EHOLR require thinking about the best and worst-case scenarios. Accuracy refers to both the degree to which a prediction may be made with absolute confidence and the degree to which it can roughly anticipate the result. The detection efficacy was calculated by comparing the expected result to the actual result. Our suggested approach provides a higher rate of successful lung cancer diagnosis than other traditional approaches. Figure represents an evaluation of the accuracy.

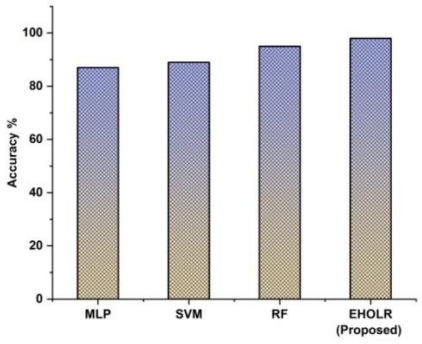

Fig 3: Evaluation of the proposed method's accuracy in comparison to the existing method

Table 1: Accuracy Comparisons

Accuracy	**%**
MLP	87
SVM	89

RF	95
EHOLR (Proposed)	98

B. Precision

The percentage of detection that concentrates on important components of cancer is referred to as precision. Accuracy in diagnosing cancer is measured by a metric called precision. That accuracy is the measure of quality is one possible interpretation. The average likelihood of correct detection is referred to as precision. Therefore, the current recommended approach is more precise than its predecessors. The accuracy of the suggested procedure is shown in Figure

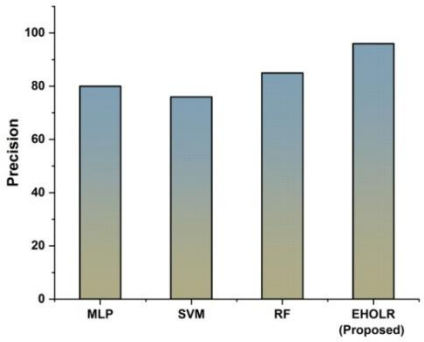

Fig 4: Comparison of the suggested approach's precision to that of the standard method

Table 2: Precision Comparison

Precision	%
MLP	80
SVM	76
RF	85
EHOLR (Proposed)	96

C. Recall

When evaluating treatment methods, recollection is one of the factors considered. The recall rate is the proportion of cancer cases correctly diagnosed via the use of CT scans. It is common practice to refer to the genuine affirmative rate as the recall. Since this is the case, the suggested approach is superior to the methods now in use. The design's high recall rate is seen in Figure 5

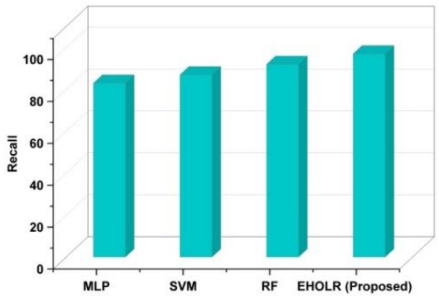

Fig 5: Recall for the proposed and existing method

Table 3: Recall Comparisons

Recall	%
MLP	83
SVM	87
RF	92
EHOLR (Proposed)	97

D. F1-Measure

Calculating the harmonic mean of the accuracy and recall scores yields the F1 score. To calculate F1, we use a simple average weighted of Precision and Recall. The result is equally affected by these two factors. In comparison to previously used approaches, the one we propose yields a high f1 score when it comes to identifying tumors. The F1-Scores for the proposed approach and the baseline method are shown in Figure 9**.**

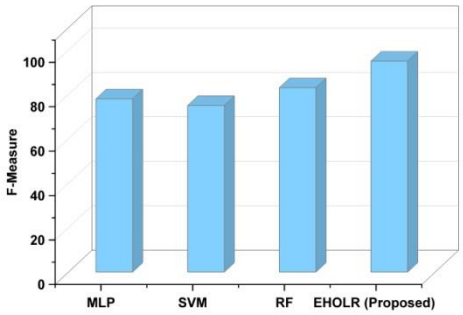

Fig 6 : F1 score for the proposed and existing method

Table 4 : F1 measure Comparisons

F-Measure	%
MLP	78
SVM	75

RF	83
EHOLR (Proposed)	95

8. Conclusion

This study outlines a few recent data mining investigations that have been conducted on cancer. Data mining techniques may be utilized efficiently to extract pertinent information from the enormous volumes of data that healthcare services produce. These experiments demonstrated that using many algorithms for data collection produces better results than applying just one method. Effective implementation of the cancer diagnosis and prognosis system results from careful algorithm combination selection and precise data set application. The necessary dataset is split into two halves; the larger portion is utilized for algorithm learning, while the smaller portion is used for method verification. The majority of research compared several classification methods on a dataset to determine if a particular patient had benign or aggressive breast cancer. A model for predicting breast cancer survival, or the breast cancer risk factor model, has been used in several investigations. It is necessary to thoroughly examine other malignancies as well. While many researchers used feature selection approaches to pinpoint the dataset's essential properties, a review of pre-processing methods showed that most researchers opted against using interpolation techniques to deal with the dataset's missing data**.**

References

[1] Eltalhi, S. and Kutrani, H., 2019. Breast cancer diagnosis and prediction using machine learning and data mining techniques: a review. IOSR Journal of Dental and Medical Sciences, 18(4), pp.85-94.

[2] AbdElNabi, M.L.R., Wajeeh Jasim, M., El-Bakry, H.M., Hamed N. Taha, M. and Khalifa, N.E.M., 2020. Breast and colon cancer classification from gene expression profiles using data mining techniques. Symmetry, 12(3), p.408.

[3] Jatain, R. ., & Jailia, M. . (2023). Automatic Human Face Detection and Recognition Based On Facial Features Using Deep Learning Approach. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2s), 268–277. https://doi.org/10.17762/ijritcc.v11i2s.6146

[4] Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P.D. and Gururajan, R., 2020. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters, 132, pp.123-131.

[5] Mohammed, S.A., Darrab, S., Noaman, S.A. and Saake, G., 2020. Analysis of breast cancer detection using different machine learning techniques. In Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5 (pp. 108-117). Springer Singapore.

[6] Simsek, S., Kursuncu, U., Kibis, E., AnisAbdellatif, M. and Dag, A., 2020. A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival. Expert Systems with Applications, 139, p.112863.

[7] Kaur, I., Doja, M.N. and Ahmad, T., 2022. Data mining and machine learning in cancer survival research: an overview and future recommendations. Journal of Biomedical Informatics, p.104026

[8] Razali, N., Mostafa, S.A., Mustapha, A., Abd Wahab, M.H. and Ibrahim, N.A., 2020, April. Risk factors of cervical cancer using classification in data mining. In Journal of Physics: Conference Series (Vol. 1529, No. 2, p. 022102). IOP Publishing.

[9] Fatima, N., Liu, L., Hong, S. and Ahmed, H., 2020. Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access, 8, pp.150360-150376.

[10] Sohail, M.N., Jiadong, R., Uba, M.M. and Irshad, M., 2019. A comprehensive looks at data mining techniques contributing to medical data growth: a survey of researcher reviews. Recent Developments in Intelligent Computing, Communication and Devices: Proceedings of ICCD 2017, pp.21-26.

[11] Alam, T.M., Khan, M.M.A., Iqbal, M.A., Abdul, W. and Mushtaq, M., 2019. Cervical cancer prediction through different screening methods using data mining. IJACSA) International Journal of Advanced Computer Science and Applications, 10(2).

[12] Kumar, V., Mishra, B.K., Mazzara, M., Thanh, D.N. and Verma, A., 2020. Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications. In Advances in Data Science and Management: Proceedings of ICDSM 2019 (pp. 435-442). Springer Singapore.

[13] Singh, M. ., Angurala, D. M. ., & Bala, D. M. . (2020). Bone Tumour detection Using Feature Extraction with Classification by Deep Learning Techniques. Research Journal of Computer Systems and Engineering, 1(1), 23–27. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/21

[14] Neto, C., Brito, M., Lopes, V., Peixoto, H., Abelha, A. and Machado, J., 2019. Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients. Entropy, 21(12), p.1163.

[15] AbdElNabi, M.L.R., Wajeeh Jasim, M., El-Bakry, H.M., Hamed N. Taha, M. and Khalifa, N.E.M., 2020. Breast and colon cancer classification from

gene expression profiles using data mining techniques. Symmetry, 12(3), p.408

[16] Ghorbani, R. and Ghousi, R., 2019. Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. International Journal of Data and Network Science, 3(2), pp.47-70.

[17] Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., Zheng, S., Xu, A. and Lyu, J., 2020. Brief introduction of medical database and data mining technology in big data era. Journal of Evidence-Based Medicine, 13(1), pp.57-69.