IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Empowering Diagnostics with Advanced Pattern Recognition and Data Mining

[1]Safla M, [2]Sreeji K B

[1]Student, [2]Assistant Professor

[1]Department of MCA,

[1]Nehru College of Engineering and Research Centre, Pampady,

Abstract : There are various combinations of database and machine knowledge approaches used to prize retired and unknown patterns from large data sets. Data mining is demanded to handle large quantities of data. Data booby- trapping deals with data diversity and correctness. Likewise, medical data mining is an extremely pivotal exploration subject, and substantial sweats have in this area in recent times since delicacy in medical data systems can lead to seriously deceptive medical treatments. Suitable mining styles should be used to examine medical data collections. Data booby - enmeshing styles have been used in the development of medical systems for complaint vaticination predicated on a set of medical data sets for the prosecution of affiliated tasks. In this, an examination of the survivability rate vaticination of bone cancer cases using data mining approaches is presented in this paper. We used SEER Public- Use Data as the source of information. The preprocessed data collection contains 151,886 records with all 16 fields from the SEER database. We experimented with three data mining styles Nave Bayes, back- propagated neural networks, and C 4.5 decision tree algorithms. These algorithms have been applied in a number of trials. The reached vaticination results are as of other approaches. Still we've established formerly that C4.5 is markedly outperforming the other two approaches.

IndexTerms **- Breast cancer survivability, Data Mining, SEER, WEKA**

I.INTRODUCTION

Mining is a knowledge chancing procedure that deals with the evaluation of data that may be hidden in enormous quantities. It's a data birth fashion of history records that's used to make concrete opinions for unborn protrusions. Data booby-trapping exemplifications include image mining, opinion mining, web mining, textbook mining, graph mining, and medical data systems. It has grown in significance in medical exploration as a tool to help reveal new patterns in medical data that might be preliminarily unknown. Croakers can view and explore conditions grounded on the prognostications the vaticination mode provides. In the United States moment, one in eight women will develop bone cancer at some time in her life.

According to the most recent data, the survival rate is 88 percent five times after opinion and 80 percent ten times after opinion. Knowledge birth from illness- related data is what enables chancing the survival rate or survivability of that complaint. Foreseer (Surveillance Epidemiology and End Results) is one of the data sources and is a one - of -a - kind, reliable, and pivotal resource for probing colorful aspects of cancer. The foreseer database summations patient - position data on cancer position, excrescence histology, stage, and cause of death. The features of a population can be studied to determine the rudiments that impact a given outgrowth. experimental exploration, similar as statistical literacy and data mining, can establish the relationship between the variables and the outgrowth, but not always the cause- and- effect relationship. numerous scientific fields, similar as drug and biotechnology, are decreasingly counting on data- driven statistical exploration. The current study uses data mining ways to prognosticate the survival rate of bone cancer cases. The experimenters anatomized SEER data and developed a pre-classification system that considers three variables Survival Time Recode (STR), Vital Status Recode (VSR), and Beget of Death (COD).

II . LITERATURE SURVEY

Substantial growth in breast cancer diagnosis has been observed in recent years, largely thanks to the introduction of data mining and machine learning techniques in the field. Deep Learning Techniques: A Review [1] points out the biggest transition in breast cancer diagnosis is associated to systems that are capable of deriving feature maps from mammograms and ultrasound images on their own using convolutional neural networks (CNNs), framing it on the grounds for more precise and earlier detection. Furthermore, as seen in Ensemble Machine Learning Techniques [2], ensemble learning unifies different classifiers like decision trees and support vector machines (SVM) to push the boundaries of predictive accuracy. These methods help in reducing overfitting and assist the model to generalize with diverse datasets, thus showing better classification accuracy in breast cancer detection. Besides the dimensionality reduction and feature selection techniques in high-dimensional datasets [3] consisting of principal component analysis (PCA) and recursive feature elimination (RFE), being the most highlighted, these dimensionality reduction and feature selection techniques select the most relevant features with fewer information from noise by measuring these qualities of the output while further increasing the efficiency of the model. The integration of radiomics and machine learning [4] further enhances breast cancer diagnosis by extracting quantitative features from medical images, providing more detailed information about tumor

behavior, which aids in both diagnosis and prognosis. However, while machine learning models have demonstrated remarkable accuracy, the need for explainable AI methods [5] remains essential in healthcare. Techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) add transparency to an AI model. This makes an AI model understandable and reliable to be used clinically. Finally, machine learning is used to successfully predict the possibility of recurrence for breast cancer by identifying high-risk patients [6]. In general, these are going to reshape breast cancer diagnosis: deep learning, ensemble learning, feature selection, radiomics, and explainable AI - towards more accurate, personalized, and interpretable tools for the clinicians.

III. METHODOLOGY

Breast cancer prediction

Breast cancer is now a very common cancer among females. The age-old method for the diagnosis of breast cancer is mammography. However, radiologists vary highly in their interpretations of mammograms. Moreover, Elmore said that 90 percent of radiologists diagnosed fewer than 3% of cancers, while 10% diagnosed around 25% of cases. Another such technique for breast cancer diagnosis has been fine needle aspiration cytology, which results in a much more accurate predictive accuracy. But the average correct recognition rate has been at a mere 90%. All such studies aim at identifying the malignant type of breast cancer patients from other individuals who remain in the begin group without suffering from breast cancer. Three types of classifications occur in the case of cancer prognosis, which include:

i)The chance of gaining the disease (Risk prediction)

ii)Recurrence of the disease

iii)Chance of survival. This is in case of the AJCC-the established predictive factor in the diagnosis of breast cancer in America**.**

It is a phase approach that bases its classification upon the TNM system (T, tumor; N, node; M, metastasis), and survival is based on any form of breast cancer whereby the patient survives for at least six months after diagnosis. C4.5 is among the most used classification techniques by decision tree induction, which was combined by Abdelghani Bellaachia and Erhan Gauven through two techniques, namely Nave Bayes and Back Propagated Neural Network. They applied the above data mining techniques for the analysis of survivability rates forecasts of breast cancer patients, contained in the new version of SEER Breast Cancer Data. An analysis of predicting the survivability rate for patients suffering from breast cancer using data mining approaches is given in the paper. We used SEER Public-Use Data. Nave Bayes, back-propagated neural networks, and C4.5 decision tree algorithms are the three methods of data mining considered. Various experiments have used these algorithms. Prediction results derived are comparable with other These algorithms have been used to predict the survivorship rate of the SEER breast cancer data set. These three classifying algorithms are used to determine which one is the best that can be adapted for predicting the survivor rates of cancers. The Bayes strategy is based on Bayesian approach. The well-known Bayesian approach to decision making uses a simple, transparent, and fast classifier. It is called 'Naive' because it assumes mutually independent qualities. Actually, it is very nearly always false, although dependencies among categories could be eliminated at pre-processing stages. It had already been successful to encode and utilize probabilistic knowledge in building quite many domains within machine learning algorithms with impressive performances. In this paper, second method, uses artificial neural nets where multi-layer networks or multi-layer perceptron of back-propagation was used to be applied within this research study. The third method utilizes a decision-tree generation algorithm known as C4.5. ID3 is used to develop the C4.5. The last two approaches were proved to be better. The experiments of these three data mining techniques have been carried out using the Weka toolbox. Weka is free open source software that combines a number of data classification, regression, clustering, association rule mining, and data visualization tools in one package. It was built using the programming language Java, and is itself licensed under the GNU General Public License. A group of tools were developed for raw SEER data extraction and cleaning. A very preliminary study would expose that there exists missing values within the SEER data. Simple analysis indicates that SEER data has missing records in Extent of Disease (EOD) and Site Specific Surgery (SSS) for nearly half of cases.

Most of the missing information is found on orders submitted prior to 1988. Since we will be examining all the fields available on the SEER database, we excluded those orders from the test data set. '4' The EOD Coding System code for these types of records after 1998, the field SSS begins to be used a little differently. All the information used to appear together in one of the standard field. To fill in the blank SSS fields, a mapping technique from new SSS to old SSS is developed. The records with missing information are then removed after this stage. The EOD field contains five fields, among them is the EOD code. Among the parameters involved are the tumor size, number of positive nodes, number of nodes and number of primary that were marked with missing values by the codings '999, 99, or 9'. For clarity purposes note that any 'unknown' valued fields in Table 1.1 and 1.2 are not data but the actual 'unknown'. All the applied fields in our research appear on this table as well.

Nominal variable name	Number of distinct values
Race	19
Marital status	6
Primary site code	9
Histologic type	48
Behaviour code	2
Grade	5
Extension of tumor	23
Lymph node involvement	10
Site specific surgery code	19
Radiation	9
Stage of cancer	5

Table 1.1: Summary of Nominal Variables in Cancer Dataset

Numeric variable name	Mean	Std. Dev.	Range
Age	58	13	10-110
Tumor size	20	16	0-200
No of positive nodes	1.5	3.7	0-50
Number of nodes	15	6.8	0-95
Number of primaries	1.25	0.5	1-8

Table 1.2: Survivability Attributes

As explained in the preceding section this paper approached the pre-classification differently. Whereas here, only three fields were considered: STR, VSR, and COD. In the SEER database, the column STR has values between 0 and 180 months. Below is the step of the pre-classification procedure.

If STR ≥ 60 months and VSR is alive

Then classify as "survived"

Else if STR < 60 months and COD is breast cancer

Then classify as "not survived"

Else

Ignore the record

In summary, the classification can be represented as:

Survived if STR ≥ 60 and VSR = alive

Not Survived if STR < 60 and COD = breast cancer

Ignore otherwise The records that are not included in this approach above, are the patients that have a STR of fewer than 60 months, and they were alive at last return to clinic visit, or the patients that had a STR of fewer than 60 months but died out of breast cancer. Tables 2 and 3 show the pre-classification process classes and the approach used in, respectively.

Class	No of instances	Percentage
0: not survived	35,148	23.2
1: survived	116,738	76.8
Total	151,886	100

Table 2: Proposed Survivability Class Instances

Class	No of instances	Percentage
0: not survived	162,381	58.3
1: survived	116,282	41.7
Total	278,663	100

Table 3: Survivability Class Instance

A common approach after the step of pre-processing is to assess the effect of the attributes on the vaticination, or trait selection. The rates were ranked using the information gain metric because this is similar a common system, and the C 4.5 decision tree fashion uses this. When a trait provides fresh information about a class, information gain (IG) is measured as the difference in entropy (H). The information gain and entropy ahead and after observing the trait Xi for the class C are as follows:

$$H(C) = -\sum p(c)\log p(c) \ , \ c \in C$$

$$H(C|Xi) = -\sum p(x) \sum p(c|x)\log p(c|x) \ , x \in Xi, c \in C$$

$$IGi = H(C) - H(C|Xi)$$

The rated survival properties of data as obtained from the Weka toolbox are shown in Figure 1. It can be noted that Tumor Extension has a better ranking than Tumor Size.

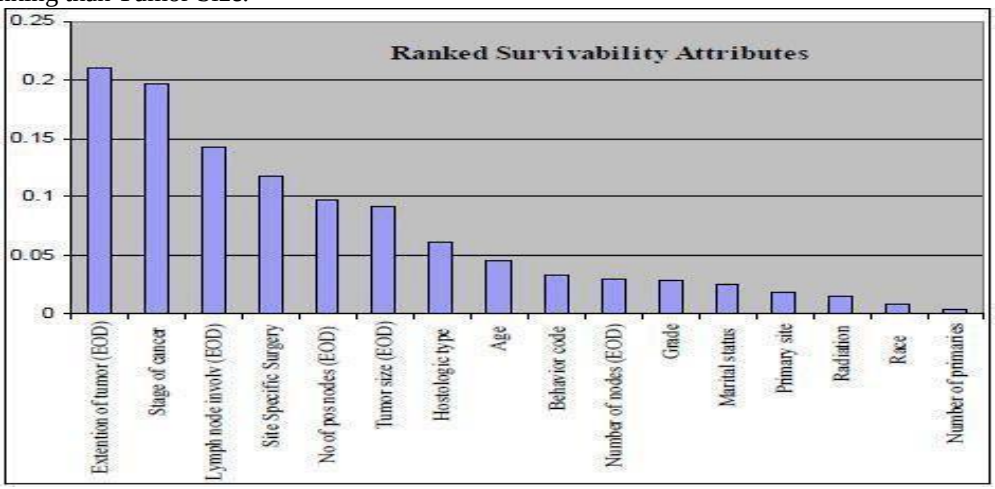

Figure1: Survivability Attributes in Order

This sets delicacy, perfection, and recall of the three strategies as the criteria for comparison. To ensure the assessment of classifier performance is exact; across-validation with ten crowds was conducted. Cross validation is the act of splitting data into k groups in its utmost introductory form. The average error rate is the estimated error rate from these k groups, and each group is projected using the bracket rule created from the remaining (k- 1) groups. The error rate can also be calculated in an unprejudiced manner. The average error rate is the estimated error rate from these k groups, and each group is projected using the bracket rule created from the remaining (k- 1) groups. The error rate can also be reckoned in an unprejudiced manner. The final classifier rule is attained from the entire set of data. We gain the measures of perfection, recall, delicacy A I and Cross Validation Accuracy (CVA) to report a classifier performance after running the classifier 10 times with 10 crowds CVA = (1/10) ∑ Ai i = 1, 2,., 10 Ai = # records rightly classified / total # records
After running a specified k-fold cross-validation, the Weka toolbox can determine all of these performance measures.

IV.RESULT ANALYSIS

The study compares the accuracy of three data mining approaches. In addition to high precision and recall metrics, the objective is to have high accuracy. Though these measures are more frequently used in the information retrieval field, we have included them because they are related to other existing metrics such as specificity and sensitivity. Results are presented in Table 4.

Classification Technique	Accuracy (%)	Class	Precision	Recall
Naïve Bayes	84.5	0 1	0.70 0.88	0.57 0.93
Artificial Neural Net	86.5	0 1	0.83 0.87	0.52 0.97
C4.5	86.7	0 1	0.80 0.88	0.56 0.96

Table 4: Combined Results (our study)

Classification Technique	Accuracy (%)	Class	Precision	Recall
C4.5	81.3 1	0 0.76	0.86 0.81	0.81

Table 5: Results for C4.5 (data set as in Table 3)

As shown in Table 4, neural networks and decision trees are similar in performance. Table 5 shows the experimental results grounded on the same dataset as our approach and the pre-classification approach used in.

The results obviously shown in that the categorization rate (81%) is much lower than our approach's (87%) bracket rate. The calculation times for the styles Nave Bayes, Neural Network, and C4.5 for the system were 1 nanosecond, 12 hours, and 1 hour, independently on an AMD Athlon 644000+ system. The results attained in this paper are different from those of Delen et al. Due to our use of a fresher database (2000 as opposed to 2002), as well as a different pre-classification (109,659 and 93,273 vs. 35,148 and 116,738) and different toolkits (artificial grade tools as opposed to Weka).

V.CONCLUSION AND FUTURE WORK

Difficulties, algorithms, and techniques for the breast cancer survivability prediction problem within the SEER database were defined, addressed, and resolved in this study. Aside from the Survival Time Recode (STR), the Vital Status Recode (VSR), and the Cause of Death are taken into consideration in our strategy. The outcome of the experiment shows that our strategy out performs the other. This study clearly demonstrates that preliminary results for the use of data mining approaches to the survivability prediction problem in medical databases are promising. Our study excludes records with incomplete data; future work will incorporate missing data in the EOD field from previous to 1988 EOD fields. Because the size of the data set will grow significantly, this may improve performance.

In future, the predictor can be used to design a web – grounded operation to accept the Predictor variable and an Automated system opinions. Tree grounded vaticination can be enforced in remote areas similar as pastoral regions or country sides, to mimic like mortal individual moxie for cast of complaint. The Bayesian network is farthest proved to be one of the extensively used styles in medical vaticination. Particular it has been effectively used for the prognostic and opinion of Breast cancer. In future we will design and apply such system for web grounded operations.

VI.REFERENCES

1. A review on deep learning techniques in breast cancer diagnosis. Authors: A. R. M. M. Faris, J. M. B. S. Almalki, N. M. Z. Aljawarneh, M. A. Alzubi, and A. M. L. AlBakri. *Journal of Biomedical Informatics*, 2020.
2. Ensemble machine learning techniques for breast cancer detection. Authors: M. S. N. Al-Garadi, M. A. Al-Qurashi, A. M. Al-Hadhrami, M. A. K. M. Al-Kadi, and M. A. M. R. Al-Ali. *Computers in Biology and Medicine*, 2021.
3. Dimensionality reduction and feature selection for breast cancer prediction: A comparative study. Authors: A. P. B. Elakkiya, N. R. K. Karthik, M. S. S. Vennila, and M. R. S. K. Sathish. *Expert Systems with Applications*, 2020.
4. Integration of radiomics and machine learning models in breast cancer diagnosis. Authors: S. R. Sundararajan, R. R. Sundararajan, S. A. Naveen, and P. R. Ganesan. *IEEE Access*, 2021.
5. Explainable AI methods in breast cancer diagnosis: A survey. Authors: S. P. Subramanian, R. K. R. Haritha, R. S. Rajan, and M. G. S. Madhuri. *Artificial Intelligence in Medicine*, 2022.
6. Predicting breast cancer recurrence using machine learning algorithms. Authors: M. K. S. R. S. Kannan, P. G. R. Samy, V. P. Rajendran, and P. V. N. Rajan. *Cancer Informatics*, 2020.
7. Comparison of machine learning algorithms for breast cancer prognosis and classification.Authors: M. S. R. Chandra, A. P. A. B. Rao, and R. B. G. Venkatesh. *Journal of Cancer Research and Clinical Oncology*, 2021.
8. A hybrid machine learning model for early detection of breast cancer. Authors: H. S. T. Anwar, P. R. M. Singh, and M. R. A. Khan. *Computers, Materials & Continua*, 2021.
9. Deep learning models for automated breast cancer diagnosis from mammography and ultrasound images. Authors: N. K. R. Shah, D. K. G. Kumar, and S. M. R. Agarwal. *Neural Computing and Applications*, 2021.
10. Artificial intelligence in personalized breast cancer treatment: Recent advances and challenges.Authors: J. A. T. Shah, R. K. M. Bansal, and P. K. K. Tiwari. *IEEE Transactions on Biomedical Engineering*, 2022.