ISSN 2278-3091
Volume 9, No.1.3, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse7591.32020.pdf
https://doi.org/10.30534/ijatcse/2020/7591.32020

An Analysis of Classification of Breast Cancer Dataset Using J48 Algorithm

John Heland Jasper C. Ortega[1], Michael R. Resureccion[2], Lizel Rose Q. Natividad[3], Emilsa T. Bantug[4], Ace C. Lagman[5], Shinji Robin Lopez[6]
[1]FEU Institute of Technology, Philippines, jcortega@feutech.edu.ph
[2]University of the East, Philippines, resurreccion.michael@ue.edu.ph
[3]San Beda University, Philippines, lnatividad@sanbeda.edu.ph
[4]Nueva Ecija University of Science and Technology, Philippines, emilsa.bantug@neust.edu.ph
[5]FEU Institute of Technology, Philippines, aclagman@feutech.edu.ph
[6]FEU Institute of Technology, Philippines, srlopez@feutech.edu.ph

ABSTRACT

Classifying tumors into benign and malignant take time and resources; sometimes it takes several radiologists and oncologists to diagnose if a tumor is malignant or benign, especially if features are hardly distinguishable to the human eye. To determine a way to automatically classify if a tumor is benign or malignant, the researchers developed a model using J48 decision tree algorithm to classify a tumor through analysis of cell features extracted by the X-cyt program. Based on confusion matrix analysis, the algorithm performed well by recording a 95 percent accuracy rate derived from confusion matrix analysis.

Key words: Breast Cancer, Diagnosis, Diagnostics, Oncology, Pathology, Radiology

1. INTRODUCTION

Most cancers in general are hard to diagnose, especially in its earlier stages, when left undiagnosed, the cancer, beginning from its primary origin, starts to proliferate to the immediate cells around the tumor before metastasizing to other parts of the body. Unfortunately, cancers are usually diagnosed during stage III or IV, when the patient experiences symptoms by which in the third stage, the cancer has already spread to immediate cells surrounding the tumor but not yet to distant organs. Addressing this problem leads to saving lives. With this, the researchers utilized the efficiency of machine learning algorithms particularly decision tree algorithms to extract hidden patterns that can be particularly useful in this domain.

1.1 Background of the Study

Over the past decades, breast cancer is still one of the most common cancers [1]. Diagnosing whether a tumor is malignant is of utmost importance in order to provide the right treatment. Oncologists for one will never deny the benefit of having the technology to automatically classify breast cancer types within seconds, saving their time and providing more time for the patient's treatment and recuperation. The importance of diagnosing the malignancy of cancers has led researchers to study the study of machine learning and deep learning including its application [13][2]. One of the key machine learning algorithms used in the prediction of cancer is decision tree algorithm. Decision trees classifies data into leaf nodes and internal nodes that are connected by branches, resembling an inverted tree, with the root node in the top most portion of the tree, all for the purpose of extracting useful information [3]. Although there have been numerous research validating the benefits of machine learning in the field of oncology, these papers must first undergo thorough validation before they are deployed into the healthcare industry.

1.2 Research Questions

In this section, the following research questions are formulated.

a. How to develop a model that will classify whether a tumor in the breast area is benign or malignant?
b. How effective the develop model in terms of confusion matrix analysis?
c. How effective the system as perceived by experts using ISO 9126 metrics?

1.3 Literature Reviews

Data Mining is an area under computer science that deals with the applications of machine learning algorithms in order to develop data models that can be used for prediction, cluster and association analysis [4].

Knowledge Discovery in Databases (KDD) provides a step by step mechanism in order to generate useful patterns that can be used in different domains. The main steps of this methodology involve data preprocessing, modeling,

evaluation and deployment. It encompasses the different applications of the algorithms of machine learning to extract guidelines or equations commonly used for generating predictive models [5].

Decision tree algorithm provides a powerful if then else statements generated from decision trees. These rule sets are generated based on the computation of entropy and information gain. This algorithm has been applied in multiple domains and across different domains because of its simplicity and structure. It uses a greedy search approach to determine which among the attributes has the highest information gain [6].

Many uses decision tree as a technique for data mining for classification because of its ability to assess multi-level characteristics of a group that can lead to accurate result. This methodology classifies a population into branch-like segments that look like an inverted tree with a three primary node types being the root node, internal nodes, and leaf nodes. The algorithm does not require parameters and can efficiently deal with large, complex datasets without imposing a complicated parameter structure. Furthermore, voluminous data can be further divided into training and test datasets. Using the training dataset to build a decision tree model and a test dataset to decide on optimal efficiency of the model [7].

One of the leading causes of death worldwide is breast cancer, especially for women, worldwide [8]. As such, metrics were developed to help diagnose breast cancer accurately. [11]

Study has shown that machine learning can be used to detect many variations of cancer [12]. Over the years, detecting and classifying breast cancer using data mining and machine learning technique have been developed. This can be categorized into three section: preprocessing, extraction, and classification [9]. By preprocessing mammography films, stakeholders can visually examine not only the exact location of the cancer but its peripheral areas and identify intensity distribution. This process makes the interpretation and analysis easier [10].

2. METHODOLOGY

The KDD methodology is presented in the figure below.

.

The modified version of the KDD consists of the following steps. These steps are very important to extract quality models that can be used to convert data into a meaningful form of information.

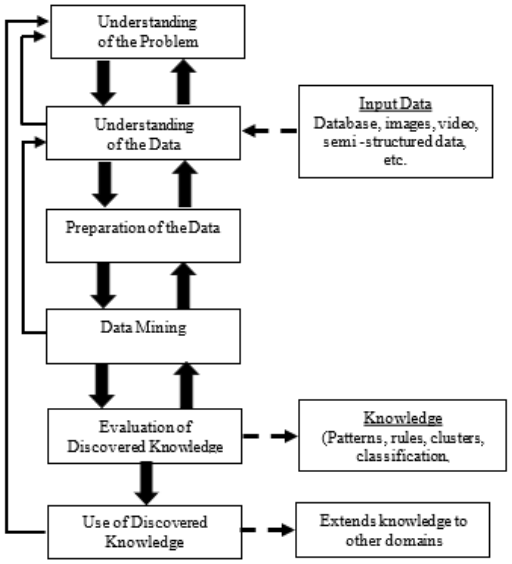

Figure 1: The Six (6) Step KDP model (*Pal & Jain, 2005*)

2.1 Data Description

The dataset used in this paper was created by Dr. William H. Wolberg in a university hospital at Wisconsin, USA. It was contributed by Olvi Mangasarian on July 15, 1992. A computer program called Xcyt will be used and digitally scan the sample fluid from patients with solid breast masses to analyze cytological features. Each attribute is evaluated on a scale of 1 to 10, with 1 being the closest to benign and 10 the closest to malignant.

Table 1: Summary of Dataset

Table Head	Summary of Dataset	
	Features	***Values***
	Sample number	Unique
	Uniformity of cell Shape	1-10
	Uniformity of Cell Size	1-10
	Clump thickness	1-10
	Bare nuclei	1-10
	Features Cell size	1-10
	Normal nucleoli	1-10
	Clump cohesiveness	1-10
	Nuclear chromatin	1-10
	Mitoses	1-10
	Class nominal	2 - Benign, 4 - Malignant
	Summary	Values
	Class distribution benign	456 (65.5%)
	Malignant	241 (34.5%)
	Number of Missing Values	16
	Number of Instances	699

The data can be considered 'noise-free' and has 16 observed missing values from the Bare Nuclei column, each coming from a different instance. Table 3 is a summary of the dataset used in this paper. The data consist of 16 missing cells belonged to a single column that is bare nuclei column. To clean the data, the researcher replaced the data by averaging the values of bare nuclei column and putting them in place of the missing values which is called imputation. Imputation is a technique to replace missing values. The result is 3.46, since the values of the data set are only composed of integers; the researcher floored the value to 3.

2.2 Modelling and Evaluation

The J48 is a reimplementation of C4.5 which is a good algorithm when dealing with imbalanced data if some of its attributes are configured correctly. Among the decision tree algorithms, J48 was the fastest to simulate compared to other decision tree algorithms.

Entropy

Entropy ***H(S)*** is a measure of the amount of uncertainty in the (data) set ***S*** (i.e. entropy characterizes the (data) set ***S***).

$$H(S) = \sum_{x \in X} -p(x)\log_2 p(x)$$

Where,

- ***S*** – The current (data) set for which entropy is being calculated (changes every iteration of the ID3 algorithm)
- ***X*** – Set of classes in S
- ***p(x)*** – The proportion of the number of elements in class x to the number of elements in set S.

When H(S) = 0, the set S is perfectly classified (i.e. all elements in S are of the same class).

Information gain

Information gain ***IG(A)*** is the measure of the difference in entropy from before to after the set S is split on an attribute. To compute the information, gain the formula was used.

$$IG(A, S) = H(S) - \sum_{t \in T} p(t)H(t)$$

Where,

- ***H(S)*** – Entropy of set S
- ***T*** – The subsets created from splitting set S by attribute A such that
- ***P(t)*** – The proportion of the number of elements in t to the number of elements in set S
- ***H(t)*** – Entropy of subset t

To evaluate the model, classification table was used in the study. This pertains to determine the right predictions over to the entire number of data occurrences.

3. RESULTS AND DISCUSSIONS/EVALUATION

This section presents the model extracted using J48 algorithm. The figure below presents the tree structure, which consist of 14 numbers of leaves, and 27-size structure. To avoid over fitting and under fitting, the researchers used cross validation technique.

Cross validation is a statistical technique that partitions data into subset, trains it and use the other for evaluating models' performance. To reduce variability we perform multiple rounds of cross-validation with different subsets from the same data.

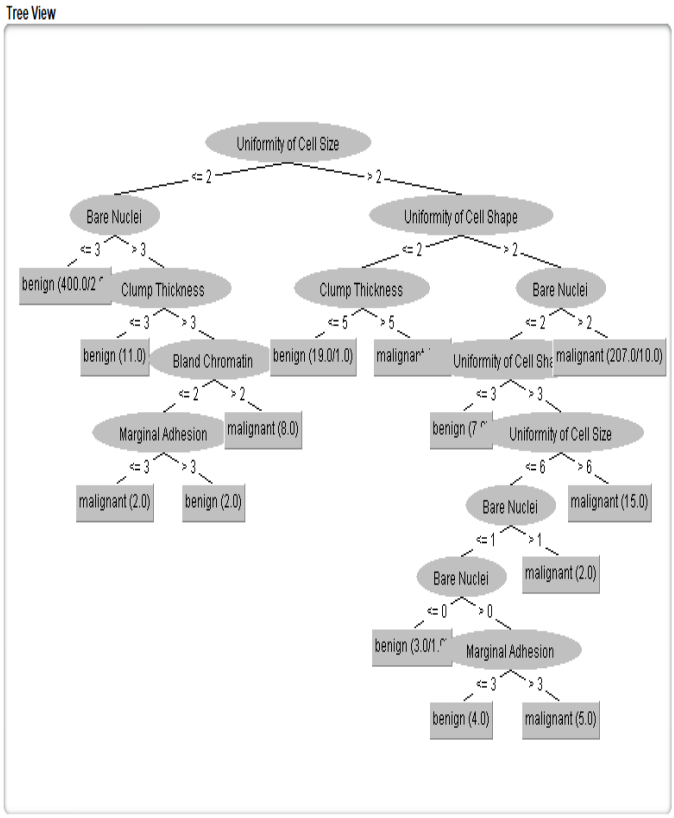

Figure 2: Tree Structure of the Predictive Model

The figure 2 shows that the uniformity of cell size and bare nuclei refer as the most significant attribute using gain ratio in predicting whether a tumor is malignant or benign.

Sample Derived Rule Sets

If Uniformity of cell size <=2, and Bare Nuclei <=3
then result is Benign
If Uniformity of cell size <=2 and Bare Nuclei >3 and Clump Thickness <=3
then result is Benign
If Uniformity of cell size <=2 and Bare Nuclei >3 and Clump Thickness > 3 and Bland Chromatin >2,
then result is Malignant
If Uniformity of cell size <=2 and Bare Nuclei >3 and Clump Thickness > 3 and Bland Chromatin <=2 and marginal adhesion <=3
then result is Malignant

If Uniformity of cell size <=2 and Bare Nuclei >3 and Clump Thickness > 3 and Bland Chromatin <=2, and marginal adhesion >3

then result is Benign

If Uniformity of cell size >2 and uniformity of cell shape <=2 and clump thickness <=5

then result is malignant

If Uniformity of cell size >2 and uniformity of cell shape <=2 and clump thickness > 5

then result is benign

3.1 Data Model Performance

To determine performance of the model, confusion matrix was used. The matrix presents the correctly classification over to the total number of instances. The model predicts a 95% accuracy result. This means that the model is very reliable and can be implemented in the system. The extracted rule sets can predict whether a tumor is malignant or benign. The table below indicates the summary result of the model.

Table 2: Summary Result of the Model

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.956    0.068    0.964      0.956   0.960      0.885    0.951     0.954     benign
               0.932    0.044    0.917      0.932   0.925      0.885    0.951     0.915     malignant
Weighted Avg.  0.948    0.060    0.948      0.948   0.948      0.885    0.951     0.941

=== Confusion Matrix ===

   a   b   <-- classified as
 432  20 |   a = benign
  16 221 |   b = malignant
```

To efficiently test the extracted rule sets derived from the classifiers, the researchers provided a simulation before an actual decision support system can be developed.

Table 3: Simulation of the Model

Clump Thic	Uniformity	Uniformity	Marginal Ac	Single Epith	Bare Nuclei	Bland Chroi	Normal Nuc	Mitoses	Class	Result
1	1	1	1	2	1	1	1	8	2	Correct
1	1	1	3	2	1	1	1	1	2	Correct
5	10	10	5	4	5	4	4	1	4	Correct
3	1	1	1	2	1	1	1	1	2	Correct
3	1	1	1	2	1	2	1	2	2	Correct
3	1	1	1	3	2	1	1	1	2	Correct
2	1	1	1	2	1	1	1	1	2	Correct
5	10	10	3	7	3	8	10	2	4	Correct
4	8	6	4	3	4	10	6	1	4	Correct
4	8	8	5	4	5	10	4	1	4	Correct

To determine the effectivity of the model, the rules were embedded in a decision support system. The decision support system pertains to a live platform using web application that predicts whether a tumor is malignant based on certain inputs.

The decision support system provides three main features. The first feature is an access of the administrator to input necessary values in all the parameters/attributes used. The main feature of the system provides intelligent results, which uses the extracted rule sets generated by J48 algorithm to determine whether tumor is malignant or benign. The DSS also has audit logs, report generation and data visualization of all the reported instances predicted by the application.

To determine whether the system is ready for deployment, The system was evaluated by experts using purposive sampling technique. Purposive sampling is a non-probabilistic sampling technique. However, the researcher should determine sets of parameters/filters to qualify as respondents. The parameters used are as follows (i) information technology faculty members (ii) teachings data analytics (ii).

The system is evaluated through its' functionality, usability and reliability using the following presentation of ratings.

Table 4: Representations of Ratings

RATING	INTERPRETATION
4.51 – 5.00	Excellent
3.51 – 4.50	Very Good
2.51 – 3.50	Good
1.51 – 2.50	Fair
1.00 – 1.50	Poor

Table 5: Respondents Overall Rating in Terms of Functionality

Characteristics	Statement	Mean Response	Interpretation
FUNCTIONALITY	1. The system has available functions required for its execution	4.66	Excellent
	2. The system does what was proposed correctly	4.33	Very Good
	3. The system is precise in executing its functions	4.33	Very Good
	4. The system is precise in its results	4	Very good
	OVERALL	**4.33**	**Very good**

Table 5 shows the results for the overall functionality of the system. Based from results gathered, the respondents agreed the system has the desired functions that aids its execution that got a mean response of 4.66; for the statement that the system does its functions correctly it got a mean of 4.33 the same with precise execution of functions. Lastly, the precision of results got a mean of 4.

Table 6: Respondents Overall Rating in Terms of Usability

Characteristics	Statement	Mean Response	Interpretation
USABILITY	1. The system can be easily understood.	4.33	Very Good
	2. The system's user interface looks good.	4.33	Very Good
	3. The system can be learned easily.	4.33	Very Good
	4. The user uses the system with minimal effort	4.33	Very Good
	OVERALL	**4.33**	**Very Good**

Table 6 shows the results for the overall usability of the system. Based from the results gathered, the mean for the system usability and its ability to be understood easily is 4.33, the user interface has a mean of 4.33. To summarize it all, they all got the same mean, which is 4.33 with an interpretation of Very Good.

Table 7: Respondents Overall Rating in Terms of Reliability

Characteristics	Statement	Mean Response	Interpretation
RELIABILITY	1. The software provide consistency in terms of its uses and functionality.	4.33	Very Good
	2. The software reacts appropriately when failure occurs	4	Very Good
	3. The software provides notification in terms of invalid inputs.	4.33	Very Good
	4. The software is capable of providing logs and audit report.	4.33	Very Good
	OVERALL	**4.24**	**Very Good**

Table 7 shows the overall rating of the respondents. The survey revealed that the overall for the functionality area of the system is 4.33. It is the same with the usability area of the system, which is also 4.33. Lastly, the reliability area got an overall of 4.24 which results in an overall rating of 4.3 for the entire assessment of the respondents. Based from the results of the system, we can conclude that the system functions properly as it is expected to. The overall result of the system is a 4.3 which has an equivalent interpretation of very good based from the scale given. All characteristics of the system which are the: Functionality, usability and reliability got high interpretation scores which are above average.

4. CONCLUSION

The procedures and steps in KDD methodology are very effective in extracting useful pattern from the . cell features extracted by the X-cyt program. The most crucial parts of KDD refer to preprocessing and modeling. Pre-processing involves different techniques to improve the quality of dataset while modeling checks algorithms' performance and develop predictive and cluster models used to profile and predict future instances. The J48 algorithm has been one of the most effective machine learning algorithms for predictive modeling. It is a classifier embodied by a flowchart like tree construction that has been extensively utilized to embody association models, due to its graspable nature that hold to mind the human reasoning

The confusion matrix presents the correctly classification over to the total number of instances. The model predicts a 91.5% accuracy result. This means that the model is very reliable and can be implemented in the system. The extracted rule sets can predict whether a tumor is malignant or benign.

5. RECOMMENDATION AND FUTURE WORKS

The researchers aim to test the dataset using other machine learning algorithms. This can provide an experimental search, which algorithm best fit in for prediction. The proponents would like to extend the study in the future where the system will be available through mobile. The software design should be improved and other features that can be beneficial for the users should be added.

REFERENCES

1. R. L. Siegel, K. D. Miller, and A. Jemal. **Cancer statistics, 2016**, *CA: A Cancer Journal for Clinicians*, vol. 66, no. 1, pp. 7–30, 2016. https://doi.org/10.3322/caac.21332
2. K. Korouh, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. Fotiadis, **Machine learning applications in cancer prognosis and prediction**, *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.
3. Y.-yun Song and Y. Lu. **Decision tree methods: applications for classification and prediction**, *Shanghai Arch Psychiatry*, pp. 130–135, Apr. 2015
4. C. Clifton. **Encyclopædia Britannica: Definition of Data Mining**. 2010. Retrieved 2010-12-09.
5. U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. **From Data Mining to Knowledge Discovery in Databases**. (PDF).1996. Retrieved 17 December 2008.
6. J. Han, M. Kamber and J. Pie. **Data Mining Concepts and Techniques.** 2006
7. P. Sharma. **Comparative Analysis of Various Decision Tree Classification Algorithms using WEKA**. *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, no. 2, pp. 684–690, 2015.
8. W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, **Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis**, *Designs*, vol. 2, no. 2, p. 13, Sep. 2018. https://doi.org/10.3390/designs2020013
9. J. Ferlay, I. Soerjomataram, R. Dikshit et al. **Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012**. *In International Journal of Cancer*, vol. 136, no. 5, pp. 359–389, 2014.
10. A. J. Cruz and D. S. Wishart, **Applications of machine learning in cancer prediction and prognosis**, *Cancer Informatics*, vol. 2, pp. 59–77, 2006. https://doi.org/10.1177/117693510600200030
11. S.H. Nallamala, P. Mishra and S.V. Kaneru. **Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems**. *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, No.2, pp 259-264. https://doi.org/10.30534/ijatcse/2019/26822019
12. N.Malik, V.B.Bharat, S.P.Tiwari and J. Singla, **Study of Detection of Various Types of Cancers by Using Deep Learning: A Survey**. *International Journal of Advanced*

Trends in Computer Science and Engineering, vol. 8, No.4, pp 1228 – 1233
https://doi.org/10.30534/ijatcse/2019/31842019

13. A.M.Alqudah, H.Alquraan, I.A.Qasmieh, et.al. **Brain Tumor Classification Using Deep Learning Technique - A Comparison between Cropped, Uncropped, and Segmented Lesion Images with Different Sizes**, *In International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, No.6, pp 3684-3691
https://doi.org/10.30534/ijatcse/2019/155862019