Exploring Cancer Risk Factors using Data Mining Techniques: A Case Study from Yemen

Abdullah Al-Hashedi [1,*]
Abdullah A. Sallam [2]
Abdulqader M. Mohsen [1]

[1] University of Science and Technology, Yemen
[2] Open University Malaysia, Yemen Centre
* Corresponding author: a.alhashedi@ust.edu

https://doi.org/10.20428/JST.23.2.1

Exploring Cancer Risk Factors using Data Mining Techniques: A Case Study from Yemen

Abstract:

Data Mining is a hot scientific research topic that has many applications in various life aspects. Healthcare and medicine are among those aspects that attracted data mining researchers who sought to solve decision-making problems. Cancer diagnosis, treatment, and prediction are procedures that have been using data mining for decades. In Yemen, some cancer risk factors seem to be different from those in other parts of the world. By mining the data available at National Cancer Control Foundation (NCCF), useful knowledge have been extracted. In this paper, decision tree classification was selected for building a model to predict cancer risk factors. As the NCCF database contained data describe some social life aspects, environmental circumstances, lifestyle, etc., mining those data can contribute in the endeavors of clearing ambiguity about cancer risk factors in Yemen. The informative attributes that were selected for model building included gender, marital status, number of family members, province, chewing Qat, chewing tobacco (Shamaa), smoking, age, relatives with cancer, and cancer class. These data was prepared for Knowledge Data Discovery process. Then, it was prepared for feeding into C4.5 learning algorithms. The results shown that smoking, chewing tobacco (Shamaa), province of residence, marital status, and age are the most important cancer risk factors. The model produced found of high performance. In addition, the rules extracted from the model tree can also be of high value for both people and healthcare sector.

Keywords: Data mining, Decision tree, C4.5 algorithm, Cancer data mining, Cancer prediction.

https://doi.org/10.20428/JST.23.2.1

استكشاف عوامل خطر السرطان باستخدام تقنيات التنقيب في البيانات: دراسة حالة من اليمن

الملخص:

التنقيب في البيانات هو موضوع بحث علمي ساخن يحتوي على العديد من التطبيقات في مختلف جوانب الحياة. الرعاية الصحية والطب من بين تلك الجوانب التي اجتذبت الباحثين في التنقيب في البيانات الذين سعوا لحل مشاكل صنع القرار. تشخيص السرطان وعلاجه والتنبؤ به هي اجراءات تستخدم التنقيب في البيانات منذ عقود. تبدو بعض عوامل اخطار السرطان في اليمن مختلفة عن تلك الموجودة في أجزاء أخرى من العالم. من خلال التنقيب في البيانات المتاحة من المؤسسة الوطنية لمكافحة السرطان (NCCF)، تم استخراج المعرفة المفيدة. في هذه الورقة، تم اختيار تصنيف شجرة القرار لبناء نموذج للتنبؤ بعوامل خطر السرطان. بما أن قاعدة بيانات NCCF تحتوي على بيانات تصف بعض جوانب الحياة الاجتماعية، والظروف البيئية، ونمط الحياة، وما إلى ذلك، يمكن أن تساهم هذه البيانات في جهود إزالة الغموض حول عوامل خطر السرطان في اليمن. وشملت الصفات المعلوماتية التي تم اختيارها لبناء النموذج: الجنس، والحالة الاجتماعية، وعدد أفراد الأسرة، والمحافظة، ومضغ القات، ومضغ التبغ (الشمعة)، والتدخين، والعمر، والأقارب المصابين بالسرطان، وفئة السرطان. تم إعداد هذه البيانات لعملية اكتشاف المعرفة من البيانات. ثم، تم إعداده لإدخاله في خوارزميات التعلم C4.5. أظهرت النتائج أن التدخين ومضغ التبغ (الشمعة) ومحافظة الإقامة والحالة الاجتماعية والعمر هي أهم عوامل أخطار الإصابة بالسرطان. حقق النموذج المقدم اداء عالي. بالإضافة إلى ذلك، يمكن للقواعد المستخرجة من شجرة النموذج أن تكون ذات قيمة عالية لكل من الأفراد وقطاع الرعاية الصحية.

الكلمات المفتاحية : التنقيب في البيانات، شجرة القرار، خوارزمية سي 4.5، تنقيب بيانات السرطان، التنبؤ بالسرطان.

1. Introduction

Cancer is a disease that leads to abnormal uncontrolled cells divide. Cancer cells can propagate to other cells, and the disease spread from a body part to another through the blood and lymph system[1] . Cancer is a leading cause of death. In 2012 alone, the cancer death toll reached 8.2 million[2] . The rise in the cancer deaths, especially in the third world and least developed countries, can be attributed to late or inaccurate diagnosis. Hence, cancer prediction will be helpful in taking proactive measure for early disease control and, consequently, increase survival rate.

According to Mona Abdu Ali, a physician at the Yemen National Oncology Centre, breast cancer in Yemen seems to be spreading due to unknown factor other than those already known worldwide. In Yemen, as far as authors' best knowledge, no medical KDD researches have been carried out so far. Moreover, most of the previous studies did not approach the same set of factors as the ones in this study. Thus, this study, which will be the first of its kind as far as the author knows, will be an attempt to get closer to the cancer risk factors that have not been considered before in order to uncover unknown reasons that are indirectly connected to cancer spread. Those factors could be environmental, social or lifestyle ones. Furthermore, the outcomes of this study are expected to serve as guidelines for further exploration of the unseen causes of cancer spread in Yemen. The rest of this paper is organized as follows: section 2 highlights the related work in this field. Section 3 provides a research methodology and demonstrates framework for building a cancer prediction model. The data analysis and findings are discussed in section 4. Section 5 demonstrate results analysis. The discussion and conclusion are provided in section 6.

2. Related Work

Previous related work falls into four categories of studies as shown in the following subsections

[1] Northern Illinois Cancer Treatment: http://www.nicancer.com/treating-cancer/what-is-cancer
[2] IARC World Cancer Report: http://www.iarc.fr/en/publications/books/wcr/

https://doi.org/10.20428/JST.23.2.1

2.1 Cancer Detection and Diagnosis

Predictive data mining is the most common type of data mining [20]. For cancer detection and diagnosis, classification techniques were mostly used in addition to clustering and association rules. Detection is usually based on mining data extracted from medical images such as X-ray, CT and RMI images and medical lab tests results. Attributes such as age and gender are considered. Here is the papers review.

[30] demonstrated how cancer registry data could be processed using data mining techniques in order to improve the statistical analysis outcomes. After data have been pre-processed for analysis, a feature selection method was applied to evaluate the contribution of each feature in predicting patient's survival. In order to evaluate the ability to predict survival of patients, the authors trained several classifiers. Finally, statistical analysis of cancer morbidity and mortality rates was performed in order to validate the initial findings. [29] used integrated ensemble learning and five data mining approaches, including support vector machine (SVM), C5.0 and extreme learning machine (ELM) in addition to others to rank the importance of risk factors and diagnose the recurrence of ovarian cancer. Experimental results elaborated that the integrated C5.0 model is a superior approach in predicting the recurrence of ovarian cancer. Furthermore, [31] sought to replace pathology report by the clinical information. DM was used in order to find correlation between clinical information and pathology report to support lung cancer diagnosis. [22] proposed a framework for constructing a data mining model for cancer early detection. The suggested framework meant to be a base for developing a diagnostic system that can provide technology and knowledge needed by users for organizing data and discovering patterns. The framework consisted of five main steps which are, 1) data collection, 2) data preparation, 3) building a model, 4) validation of the model and 5) model update and deployment. Data preparation, in turn, comprises two sub-steps, which are data integration and data preparation. Data preparation was further divided into three sub-task as follow i) handling missing values, ii) handling noisy data and iii) handling data inconsistency. [8] sought to investigate performance of different classification methods and algorithms along with demonstrating the importance of selecting the predictors (input variables) that highly improve breast cancer detection. The authors conducted a comparison among three classification algorithms: SMO (an SVM algorithm), IBK (a KNN classifier) and Best-first Decision Tree in order to determine the weight of the predictors.

In this performance examination, the authors carried out Chi-square, Info Gain and Gain Ratio tests, then calculated the average rank. This led to building high-performance classifiers. Although the three algorithms showed high comparable accuracies, SMO. The predictors employed were mostly human body factors, in addition to family history and alcohol habits.

In a contribution in proving possibility of using association rules mining in addition to classification, [24] approached oral cancer early detection and prevention. Their analysis reached eight rules with the highest confidence possible, which made those rules of high usability for physicians according to the authors. Weka data mining toolkit was used for analysis of 1025 records and 33 variables (which are all liver function parameters, Fine Needle Aspiration Cytology, Biopsy, Ultra-Sonography, CT Scan and an MRI variables). X-ray images were also used after enhancing their reliability in diagnosing lung cancer at an early stage of disease development. [1] approached this issue by classifying X-ray chest films images based on features extracted from them. The authors also demonstrated the role of Feature Reduction techniques in improving performance of a classifier, which was evident in reducing the number of false positive. Neural Networks (ANN) and Support Vector Machine (SVM) were also used in this paper to classify X-ray films into two classes: normal and abnormal (with cancer).

[19] demonstrated the predictive ability of unsupervised machine learning by using ANN for cancer detection. The authors proposed a methodology for early detection of lung cancer by using ANN Kohonen Map, which they found (from literature survey) to be of high performance. The authors integrated data from different healthcare centers. They used input variables which were all related to human body. ANN weight vector values were matched to causes and symptoms. They concluded that the above weighted vectors can be analyzed by doctors, and the cause of the disease can be found and treated accordingly. Another attempt to conduct performance comparison between three DM methods in detecting cancer was made by [9] who built accurate models for breast cancer prediction using RepTree (a fast DT learner), RBF Network (an ANN) and Simple Logistic. The variables used were all related to patient›s body. Weka toolkit was used for experiments. The Simple Logistic model was found the most accurate due to the best relevance of the variables to the algorithm. This relevance then was found by using Chi-square, Info Gain, and Gain Ratio tests, then the average was calculated. Prediction ability of Simple Logistics in cancer detection was also demonstrated.

https://doi.org/10.20428/JST.23.2.1

2.2 Cancer Prognosis

Cancer prognosis is to assess the cancer recurrence probability. Many researchers studied this issue by mining data of body parameters. However, the prognostication of breast cancer is being more approached by genetic microarray analysis, which is beyond the scope of this study.

Cancer prognosis was approached by [4], who applied both DT and Association Rules for pattern recognition only on a portion of data which was more relevant to the particular type of cancer. Before pattern recognition was carried out, the authors employed clustering for determining the data relevant to lung cancer. K-means clustering used to sort data into two clusters: relevant and irrelevant. On the relevant portion, the DT and association rules mining were applied. This led to extracting significant rules, which found of the same weight, according to the authors. 400 patients› had been selected from different diagnostic centers, with 20 risk factors (variables) which were related to the patient›s body parameters and diseases in addition to food habits, drugs being taken, tobacco, alcohol, physical activity, environment, diet, radial therapy and hereditary. In order to check the possibility of building proper strategies for malignancy prediction among the ENT patients, [21] conducted a comparison study between ANN and Boosted DT. The author used twelve variables which are related to patient›s body and history of addiction (of alcohol, smoke, etc.). The authors validated the model›s performance using ten-fold cross-validation method. According to the author, for the training and validation data, Multilayer Perceptron (MLP) (an ANN algorithm) and TreeBoost showed the same specificity and sensitivity (100% for both). Zero misclassification was reached. TreeBoost and MLP model both are optimal for predicting malignancy among patients. Another comparative study on different data mining classification techniques was conducted by [17], with taking in consideration the size of the dataset this time. They applied 14 data mining classification algorithms on three different sizes datasets in order to predict the presence of three cancer types. Using Weka toolkit, they applied different classification algorithms such as Bayes, ANN, Simple Logistics, SMO, Lazy classifiers, Rules, and Trees. The variables used were those of Micro-array Gene Expressions. The study concluded that none of the classifiers used outperformed all others when applied to different-size datasets. The authors observed that the accuracies of the tools depended on the size of the dataset, and the larger the dataset is, the better the performance of the algorithm would be. Thus, the authors recommended not to stick to a particular classification method.

2.3 Cancer Survivability Prediction

This category of papers aimed to estimate the rate of survivability after cancer development. This is important for both doctors and patients. Many research papers that sought to find survival rate have been published. The papers varied from long-procedure approach to a handy user-friendly calculator.

An attempt by [23] to demonstrate prediction power of different mutations of DT. The authors sought to build a model for predicting the rate of survival among the Oral Cancer patients. They used three DT algorithms which are: single DT, forest DT and Boosted DT to develop three prediction models. The variables used were related to symptoms and medical procedures. The authors managed to produce three models of 100% accuracy and zero misclassification. All models performances were comparable. However, Boosted DT was slightly better. [7] sought to more enhance the high reliability of DT in cancer prediction. The study aimed to predict survivability rate of breast cancer patients using three data mining algorithms: Naïve Bayes, Back-propagated Neural Network, and C4.5 DT. The paper sought to find the most suitable techniques for predicting cancer survivability rate with high performance in terms of accuracy, precision and recall metrics. Using WEKA open-source toolkit, authors developed a set of tools to extract and clean up the raw SEER data. Naïve Bayes was lowest performance algorithm, while C4.5 algorithm was the best in terms of accuracy and precision, in spite of that its performance was comparable to that of ANN. The predictor employed were extracted form diagnostic in addition to race and marital status data.

[5] demonstrated how data balancing, feature selection, and ensemble voting collaborate to yield high performance classifier. Classification was used for survival prediction. SMOTE was adopted to class-balance dataset for the purpose of avoiding low classification accuracy. Then, the number of features was reduced using CFS and Information Gain. The resulting prediction probabilities from several classifiers are combined using an ensemble-voting scheme. The classification schemes used were of two types: basic classifiers (trees, functions, and statistical methods), and meta-classifiers (Ensemble voting, which is multiple classifications by using a set of algorithms whose individual predictions were combined in some way to classify new examples). The ensemble voting classifier found to be the most accurate for one-year survivability prediction.

https://doi.org/10.20428/JST.23.2.1

[2] also made an effort to estimate lung cancer survival depending on data extracted from CT-Scan images for lungs. Then, the image features were reduced using PCA. Finally, the data subset along with the reduced features were passed to ANN for survival estimation.

Also, [3] found that, among 30 different classifiers, the top five ones were all DT-based. The authors went further to seek much better performance by applying ensemble voting with five tree methods. Features relating to lung anatomy and pathology were selected to find the possibility of survival of a cancer patient after diagnosis. The paper concluded that the performance of the above prediction scheme (ensemble voting along with five DT- based methods) was found to be the best in terms of accuracy and area under the ROC curve than using one of the above five individually.

Unlike [7], [16] showed that Naïve Bayes is not that reliable. They built a breast cancer prediction system using Naïve Bayesian classifiers with maximum accuracy of 93%. The authors observed that "the Naïve Bayesian Classifiers is sensitive as it predicts the disease on the ground of probability of disease being present in the findings".

2.4 Cancer Treatment

In cancer treatment, few research papers have been produced. These papers tried to assess the impact of medicine on people with cancer. One of the few cancer treatment published papers is that of [32], who showed the significance of concurrent chronic diseases in the course of treatment. Authors created two comorbid data sets using the SEER's cancer data. One for breast and female genital cancers and another for prostate and urinal cancers. A predictive model was then built by applying several popular machine learning techniques to the data sets. The results showed that the model predictive power can be improved by having more information about comorbid conditions of patients, which in turn, can help practitioners make better diagnostic and treatment decisions.

Furthermore, [15] used Association Rules mining to explore patterns of Chinese medicinal formulas in treating and preventing breast cancer recurrence and metastasis in an aim to find patterns or rules in the treatment or control of breast cancer. An initial statistical analysis was carried out to categorize the herbs according to their medicinal types, dosage, natures, flavors, channel tropism, and functions. Based on the categorization, the frequencies

of occurrence were computed. The rules were discovered using the SPSS and Clementine Data Mining System. The above research work covered a wide range of DM methods, in addition to preprocessing tools that added significance to the classification or, in general, to KDD process. However, the preprocessing is the cornerstone of the success of KDD processes. In your KDD battle, preprocessing can be the borderline between sweet victory and bitter defeat. In a comparison between two DM algorithms, feature selection or degree of variable relevance to the learning method can bring one of them to success and the other to failure. For instance, Naïve Bayes model used by [16] was not as good as that of [7]. The latter›s success can be mainly attributed to the relevance of the variables and the data itself to the learning algorithm. This draws attention to the impact of the relevance of data on the classification process, which was demonstrated by [4], who used clustering model to define this relevance. Data relevance is not the whole story. Data size is also of high importance in DM process. As there are no DM method or algorithm that suits all dataset sizes, there is no particular method that fits for any DM process. A data miner should not stick to particular methods. Instead, he/she should continuously go through different methods and algorithms to choose the best for his/her data size nature [3]. However, by looking to [7], [3], [4], [8],[9] and others, DT different algorithms can, to some extents, be the most reliable tool in cancer prediction. But to add more reliability to a DT model, meta-classification methods, such as boosting and ensemble voting, can be added [3], [21], [5].

3. Theoretical Framework

The population considered in this study is the whole observations collected by National Cancer Control Foundation (NCCF) between 2009 and 2014, which are the data of the people arrived at NCCF for treatment or for referring to other medical facility. The model building steps listed below is depicted in Figure 1.

https://doi.org/10.20428/JST.23.2.1

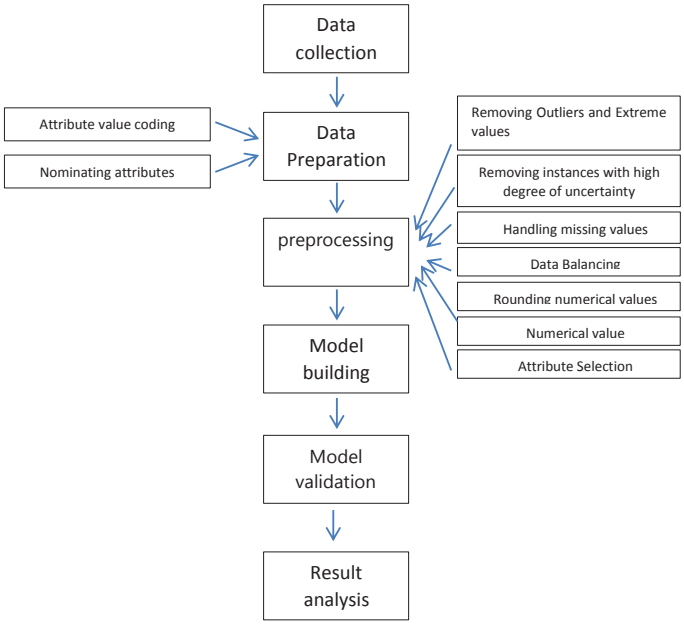

Figure 1: Propsed prediction model framework

4. Methodology

The following subsections describe in details the main steps involved in the methodology.

4.1 Data Collection

The data to be worked on in this paper were initially recorded by the National Cancer Control Foundation (NCCF). NCCF is a charitable non-profit society that helps people with (or suspected to be with) cancer medically and financially. The experiments were restricted to the data obtained from NCCF in spite of the hard work required to prepare the data for the KDD process. The dataset are for patients arrived at NCCF between 2009 and 2014 and consists of 17,000 records. The attributes of the dataset are related to marital status, geographical residence, date of diagnosis, and others.

4.2 Data Preparation

As the majority of the NCCF data tuples were with the type-of-cancer value missing, the attribute was dropped, and classification was limited to YES/NO binary classification; "Yes" class means "with cancer" while "No" means "without cancer". However, it seemed that the NCCF had been mostly being visited by those who were already diagnosed (by another healthcare facility) with cancer. Thus, the majority of the tuples were of Yes-class, which implies data imbalance and can lead to over-fitting [5]. The dataset format was converted from MS Access to MS Excel to facilitate manipulation. For several reasons, the whole attribute set could not be considered for the experiment. For instance, due to process and inaccurate data entry, the values of directorate were duplicated due to misspelling. Hence, it was time-consuming to extract the correct values from thousands of instances. For another reason, the profession (job) attribute also had to be dropped. There was no consistent definition for the attribute values; i.e. the values sometimes represented the profession and sometimes the organization for which the observed person works. For instance, while teacher, farmer, and carpenter are specific job names, public-employee, worker and self-employed do not tell the exact nature of the profession. This inconsistency would not help in linking the disease to the actual career. One more attribute was recommended to be disregarded by the NCCF data entry clerk, where it was entered without enough care. This attribute which is a degree of illness (low/medium/advanced). Similarly, the quality-of-life attribute was omitted due to that its values were selected based on personal guess, with the absence of standard measure. The most useful attribute that it had to be disregarded was type-of-cancer. This attribute was dropped for two reasons. First, the number of instances with the values of this attribute is present in only 2,400 (about 7% of) instances. Second, the problem of value coding was also faced here, so that a single cancer type was entered in five different values. Moreover, a cancer name sometimes expressed with the main type and other time with its sub-type. Finally, two attributes were combined together and made as a single attribute, since there was no reason for being separated and to simplify the learning process. The number of family male members and number of family female members were made single attribute called a number of family members. Table 1 shows the attributes that were omitted from the model due to the problems mentioned above.

https://doi.org/10.20428/JST.23.2.1

Table 1: Attribute that were omitted

No	Attribute Name	Reason for omit
1	Job	No consistent definition for the attribute values
2	Degree of illness	Entered without enough care
2	Quality of life	Its values were selected based on personal guess
4	Type of cancer	Low representation for its value.

Table 2 shows the attributes remained for model building.

Table 2: The attributes nominated for use in the KDD process

Attribute Name	Attribute Type	Possible Values
GENDER	Nominal (Binary)	M: Male F: Female
MARITAL_STATUS	Nominal	S: Single M: Married D: Divorced W: Widow Y: Under-aged
No_FAMILY_MEMBERS	Numerical	The value express the number of family members
PROVINCE	Nominal	The values express the name of the province to which the patient belongs Cap: Capital San: Sanaa Ibb: Ibb Hdr: Hadramout Dmr: Dhamar Taz: Taiz Hud: Hudeidah And: Aden Rma: Rayma Haj: Hajjah Amr: Amran Shb: Shabwa Mrb: marib Dal: Al Dhale'a Mhw: Al Mahweet

Table 2: Continued

Attribute Name	Attribute Type	Possible Values
		Lhj: Lahj
		Sda: Saada
		Bay: Al Baidha
QAT	Nominal (Binary)	The values express the statues of qat habit
		T: True: Has qat; F:False: Does not have qat
SHAMMA	Nominal (Binary)	The values express the statues of shamma (tobacco chewing) habit
		T: True: Has; F: False: does not have
SMOKE	Nominal (Binary)	The values express the statues of cigarette smoking habit
		T:True: Smoker; F:False : Non-Smoker
AGE_TESTED	Nominal (Binary)	The values express the age at which a patient was diagnosed cancerous
RELATIVES_WITH_ CANCER	Nominal (Binary)	The value tells if a patient has relatives with cancer. The attribute value was changed from nominal (how many relative and how they relate) to Yes/No attribute, which will be easy to deal with and more informing.
		T: True: patient has relatives with cancer
		F: False: a patient has no relatives with cancer
CLASS	Nominal (Binary)	Class attribute; Yes: Malignant; No: Benign

4.3 Data Preprocessing

Good data preprocessing is essential to KDD success. Outlier and extreme values were discovered using Weka Interquartile Range supervised filter. From Statistics, interquartile range is a statistical dispersion measure. According to [26], interquartile range (IQR) is the difference between the upper and lower

https://doi.org/10.20428/JST.23.2.1

quartiles. A dataset can be ranked by three points that divide it into four equal parts, each represents a quarter of the data. Weka Interquartile filter adds two attributes to the existing dataset, both of Yes/No values, which tell whether an instance has an outlier or extreme attribute value. By using the "Yes" value of the added attribute, we can remove the instances with outlier/ extreme values by using another filter of Weka called RemoveWithValue. The added attributes were removed later after processing the outliers. On the top of the dataset existed several hundreds of instances with their habits attributes (SMOKING, QAT, SHAMMA) values were all FALSE. This seemed not to be normal, as the value might be omitted for being unavailable at the time of data entry. Hence, these instances were removed to avoid probable model bias or overfitting [7].

The data collected by the NCCF suffered from unbalanc, due to that NCCF center seemed to be mostly visited by people that have been diagnosed with cancer at a previous medical facility. Thus, the no-cancer (No-class) instances were extremely fewer than cancer instances (Yes-class). Therefore, it had to add synthesized balance by employing Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is an over-sampling approach that creates synthetic minority class instances that can improve the classifier accuracy for the class of minority [10]. After applying SMOTE, the number of instances changed from above 11,000 to reach over 21,000. SMOTE technique appends the synthesized data at the bottom of the dataset. When cross-fold validation is used, this can result in data folds (portions of data) that are entirely of the same class, which brings back the problem of lack of balance. Therefore, we can use Weka Class-Randomize, which is an unsupervised filter that randomly mixes the instances. Handling missing values by deleting the entire instances that contain empty attributes is the shortest way for facilitating the KDD process. However, the volume of the dataset could be reduced dramatically, especially when there are many instances with missing values. Moreover, for DT, splits gets hard at a node having an attribute with a missing value. [28]. The missing values in GENDER attribute were substituted by inferring the gender from the patient›s name while missing values in the rest of attributes were substituted with the average value in case of numerical value, and the most frequent value in case of nominal value [6]. The calculated values such as AGE and No_FAMILY_MEM-BERS values resulting from SMOTE process needed rounding. Weka NumericTransform unsupervised filter was used for this reason. Although C4.5 can handle continuous (numerical)

attributes, discretization results in better learning algorithm efficiency, higher model performance and more interpretable and easy to understand results. In addition, an effective discretization method decreases the demand for system resources, such as memory [11]. Performance of C4.5 is improved with discretization [12].

Experimentally, it was noticed that performing attribute selection before balancing the dataset results in ignoring many high-informative attributes when performing attribute selection and feature reduction, which can be justified by that unbalanced dataset is far from representing the reality. Hence, the attribute selection process was delayed to the end of preprocessing. In addition, to achieve unbiased feature selection, attribute selection carried out with the patient identity removed [25].

4.4 Model Building

The model was built as follows:

- Attribute Selection method: Wrapper with considering C4.5 (Weka WrapperSubsetEval filter + BestFirst search method)
- Learning algorithm: DT Scheme, C4.5 algorithm. This is named in Weka as weka.classifiers.trees. J48 -C 0.25 -M 2. J48 is Weka implementation for C4.5. Pruning is also used to avoid overfitting.
- PART rule algorithms used to have rule-like output by applying the Weka weka.classifiers.rules.PART decision list. PART rule algorithms uses separate-and-conquer to builds a partial C4.5 decision tree in each iteration and make the "best" leaf into a rule[3]. MinNumObj (which is the minimum number of instances per leaf) was increased to 12 to reduce rules number.

4.5 Model Validation

For examining the model validity, the dataset was split into two portions, one for learning (training) and the other is for testing. To avoid overfitting, ten-fold cross-validation was used. Ten-fold validation is to divide the dataset into ten equal portions. Then, the first fold is used for testing and the remaining nine folds are used for training. In the next steps, the second fold is used for testing and the remaining folds are used for training. This is repeated ten times; in each time one fold is used for testing and the remaining folds are for training. The confusion matrix is a very useful tool for gauging model performance.

[3] http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/PART.html

https://doi.org/10.20428/JST.23.2.1

Confusion matrix shows how successfully a model has recognized instances of different classes [14]. For our case, we have a two-class classification process. Thus, the confusion matrix is a 2×2 table as shown in Figure 2. Another performance measure is area under ROC curve, which is the area under a two-dimensional curve of which sensitivity is represented on Y-axis and 1-specificity is represented on X-axis.

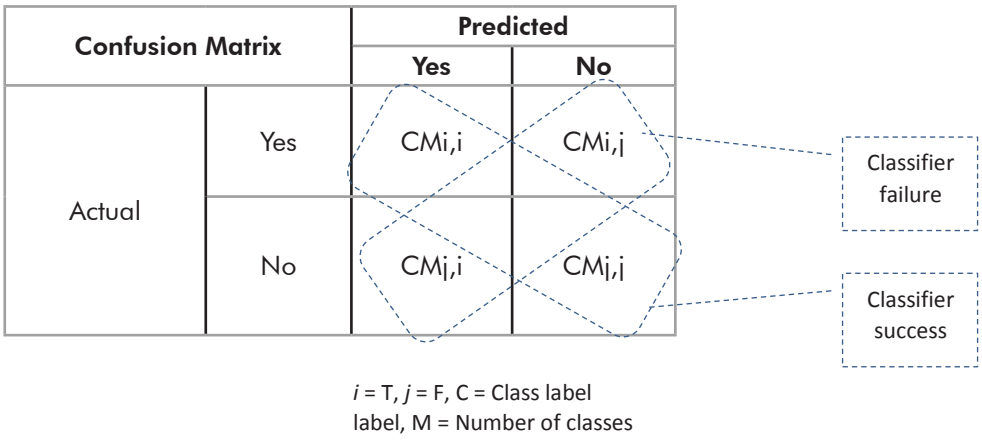

Figure 2: Confusion matrix for a binary model

For implementing the experimental work for this study, Weka open source toolkit was used. "Weka workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools" [28].

5. Data analysis and Results

The following subsections describe the main steps undertaken for data analysis and obtaining the results.

5.1 Attribute Selection

As mentioned previously, performing attribute selection before the dataset is class-balanced leads to disregarding important attributes. This can be explained by the fact that class-unbalanced dataset is far from representing the reality, and consequently, the representative attributes fail to keep their weighs. Both wrapper and filter method are performed for comparison and to try to gain an in-depth view. However, wrapper method, with DT considered as the next step learning algorithm, selected all the attributes, except for RELATIVES_WITH_CANCER as shown in Table 3.

Table 3: Outcomes of wrapper method attribute selection

Method	Weka Implementation	Attributes
Wrapper: With C4.5 considered as the learning algorithms	WrapperSubsetEval filter with j48 classifier as a parameter + BestFirst search method	GENDER MARITAL_STATUS No_FAMILY_MEMBERS PROVINCE QAT, SHAMMA,SMOKE AGE_TESTED

5.2 Model Performance

The Weka implementation of the C4.5 algorithm is called J48. We run J48 with keeping its default parameters unchanged. By default, minNumbObj (minimum number of leaves per node) value = 2. Setting minNumObj to a higher value reduces the size of the tree, and consequently decreases the rules, which enhances the tree interpretability. However, we kept the small value for maximising model performance with using ten-fold cross-validation.

5.3 Classification Rate

With using ten-fold cross-validation, more than 96% classification rate was achieved. Figure 3, shows the comparative graph, and Table 4 displays the numerical results.

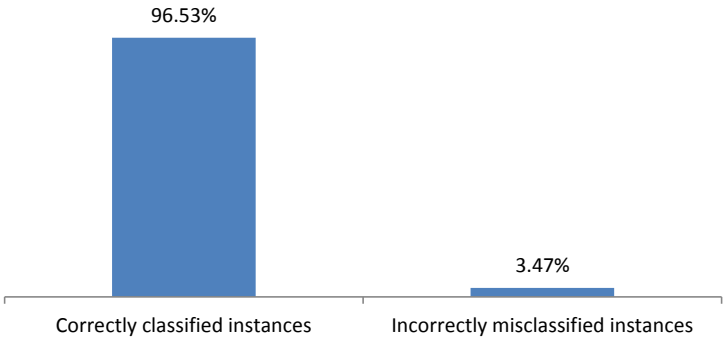

Figure 3: The model classification rate

Table 4: Model classification statistics

Total Number of Instances	21331	
Number of correctly classified instances	20626	96.53 %
Number of correctly misclassified instances	705	3.47 %

https://doi.org/10.20428/JST.23.2.1

5.4 Confusion matrix

The model built in this study achieved high performance in terms of the number of correctly classified instances as shown in Table 5. This aspect of success can be attributed to manual selection of variables from the NCCF database and then the proper method used in feature selection.

The model also performed greatly in terms of True Positives and True Negatives. True positives (TP) and true negatives (TN) provide information on where the classifier has succeeded, while false positives (FP) and false negatives (FN) reveal how unreliable the model is.

Table 5: Model confusion matrix and other statistics

Confusion Matrix		Predicted		
		Yes	**No**	**Sum**
Actual	Yes	10548	590	11138
	No	148	10053	10201
	Sum	10696	10643	
	Accuracy	98 %	95 %	
	Overall classifier accuracy		97 %	

Positive Predictive Accuracy and Negative Predictive Accuracy are other performance measures. Positive Predictive Accuracy, which demonstrates the true positive rate (true alarm rate), is the proportion of a number of positive instances that have been predicted as positive (TP) to the total number of positives (true positives plus false positives).

- Postive Prdictive Accuracy: TP/(TP+FP)

Similarly, negative predictive accuracy is the true negative rate, i.e.

- Negative Prdictive Accuracy: TN/(TN+FN)

The model showed high positive and negative predictive accuracy which were equal to 0.984 and 0.949 consecutively. This indicates very low error rates and consequently high model reliability. The Overall Predictive Accuracy is the resultant of both positive predictive accuracy and negative predictive accuracy. As the model accuracy demonstrates how large the portion of data that has been correctly classified, it also indicates high model reliability. Accuracy is also a function of both sensitivity and specificity [20]. Sensitivity is also called True Positive Rate (TPR) as it is a measure of the success of positive (Yes-class) prediction. Sensitivity (also called recall) is given by:

- Sensitivty: TP/(TP+FN)

The other component of accuracy is specificity, which is a measure of the success of negative class (No-class) prediction [14]. Specificity is given by:

- Specificity: TN/(TN+FP)

In our model, sensitivity and specificity were about equal to 0.947% and 0.985% consecutively. As we can see, the closeness in value of sensitivity and specificity to each other shows the success of the data balancing process. As the negative-class instances are synthesized, and the true negative rate is close in values to true positive rate, and the positive instances are real-world data, this means successful data SMOTE process. This, in turn, indicates, as previously stated, high model accuracy. Overall accuracy can be directly calculated by

- Overall Accuracy: (TP+TN)/(TP+P+TN+N)

Which is ~ 97% in our case. This consequently indicates less than 3% error rate, which is a minor one. Table 6 lists all the above measurements and more. High sensitivity means low false alarms while high specificity means low false positive.

Table 6: List of model performance measures

Measure	Value
Sensitivity (TPR)	0.947
Specificity (TNR)	0.985
Precision (PPV)	0.984
NPV	0.949
Fall-out (FPR)	0.015
FNR	0.053
False discovery rate (FDR)	0.016
Accuracy	0.966
Mathew correlation coefficient	0.933
Informedness	0.932
Markedness	0.933

5.5 Area under rock curve (AUROC)

The weighted average of AUROC for both classes is almost equal to 0.979 as illustrated in figure 4, which is very close to the best-case value (which is equal 1). This demonstrates high accuracy in addition to a well-balanced dataset, which again proves the success of data balancing process.

https://doi.org/10.20428/JST.23.2.1

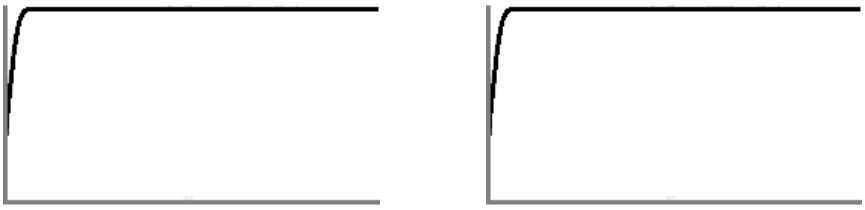

Figure 4: Area under ROC curve for both classes

5.6 Other measures

In addition to NCCF data, there was actually other data added to the dataset we worked on, which were the instances added by SMOTE. We need to measure the agreement between the real-life data and SMOTE synthesize data to find out about the data balancing process accuracy. Kappa coefficient is a measure agreement different sets of observations about the same issue [27]. Kappa coefficient ranges between 1 (perfect agreement) and 0 (no agreement), and our model of this study showed a nearing-perfect agreement as the Kappa statistics is 0.934 as shown in Table 7.

Table 7: List of other performance measures

Kappa statistic	0.934
Mean absolute error	0.056
Root mean squared error	0.173
Relative absolute error	11.161 %
Root relative squared error	34.627 %

5.7 Results analysis

The number of leaves and rules of a decision tree has an impact on the interpretability of the tree [18]. In spite that Weka J48 classifier used in this study spawned a high performance model, the tree interpretability was low and its understanding was time-consuming due to a big tree and a large number of rules. Therefore, it was required to increase the minimum number of instances per leaf (minNumObj) to 12, instead of the default value (which is 2). In addition, the continuous attributes were discretized, and finally, instead of using Weka J48, WEKA PART rule was used in order to convert the decision tree to rules. Weka PART, similarly to J48, is built on C4.5. [28]. This led to a little decrease in performance, as can be seen, form comparing Table 8 with Table 7.However, this is tolerable as it adds a great deal of simplification.

Table 8: PART rule model performance measures

Class	Positive Prediction	Negative prediction	Precision	Recall	ROC Area
Yes	0.931	0.031	0.970	0.931	0.975
No	0.969	0.069	0.928	0.969	0.975
Weighted avg.	0.949	0.049	0.950	0.949	0.975

5.8 Rule Analysis

J48 displays each terminal node (class) with two figures separated by a slash in the form of (xxxx/xx). The part before the slash represents the number of instances that support the rule (or instance weight), and the second is the number of misclassified instances [28]. Instance weight is analogous to rule support in Association Rules mining. Thus, we considered both numbers in weighing the rules. The instance weight of a group of rules the average of the individual weights and the error rate is the average of the error rates of them. The minimum instance weight considered for this rule analysis is 1, and rules with weight instance less than 1 will be dropped.

Weka PART learning produced 117 rules for which we only considered the most interesting ones. Table 9 elaborate the rules related to cancer risk factors.

Table 9: Most interesting rules

Index	Rule	Instance weight .%	Mislass .%	Class
R1	IF SMOKE: True: :> Yes (2257.0/13.0)	10.62%	0.58%	Yes
R17	IF SHAMMA: True: Yes	1.26%	0.75%	Yes
R2	PROVINCE = Haj: Yes (632.0/4.0)	2.97%	0.63%	Yes
R5	PROVINCE = Bay: Yes (305.0)	1.44%	0.00%	Yes
R7	PROVINCE = Hdr: Yes (290.0/3.0)	1.36%	1.03%	Yes
R9	PROVINCE = Mhw: Yes (262.0/2.0)	1.23%	0.76%	Yes
R15	PROVINCE = Sda: Yes (230.0/1.0)	1.08%	0.43%	Yes
R25	PROVINCE = Lhj: Yes (159.0/3.0)	0.75%	1.89%	Yes
R31	PROVINCE = And: Yes (120.0)	0.56%	0.00%	Yes
R22	PROVINCE: Taz AND MARITAL_STATUS : M: Yes (672.0/4.0)	3.16%	0.60%	Yes
R18	MARITAL_STATUS: W: Yes (264.0/6.0)	1.24%	2.27%	Yes
R40	MARITAL_STATUS = S: Yes (472.0/24.0)	2.22%	5.08%	Yes

https://doi.org/10.20428/JST.23.2.1

Table 9: Continued

Index	Rule	Instance weight .%	Mislass .%	Class
R16	AGE_TESTED = '(18.5-22.5]': Yes (297.0/22.0)	1.40%	7.41%	Yes
R8	AGE_TESTED: '(63.5-69.5]': Yes (282.0)	1.33%	0.00%	Yes
R10	AGE_TESTED = '(72.5-74.5]': Yes (261.0/15.0)	1.23%	5.75%	Yes

The link between cancer risk and the considered factors is as follows.

5.8.1 Smoking

Smoking was that the most impacting factor in cancer development. The relevant extracted rules were of weight of %10.62 (the highest among the others) with only 0.58% classification error.

Finding 1: Evident link between tobacco smoking and cancer risk.

5.8.2 Chewed Tobacco (Shamma)

Chewed tobacco or Shamma, showed much less contribution in caner development than tobacco smoking. However, although the instance weight is only 1.26%, it solely shows notable contribution that cannot be omitted, especially with less the 1% misclassified instances.

Finding 2: Slight contribution of chewed tobacco in cancer risk.

5.8.3 Chewing Qat

Chewing qat appeared to be neutral in cancer development. Only two rules with 0.53% instance weight indicated a low contribution in cancer development. Besides, 17 rules of 16.47% total instance weight showed that chewing qat did not lead to cancer development.

Finding 3: No evident relation between chewing qat and cancer development.

5.8.4 Province of Residence

Hajja people showed high vulnerability to cancer risk in comparison to its population, as about of 3% of instances showed that living in Hajja alone led to cancer development

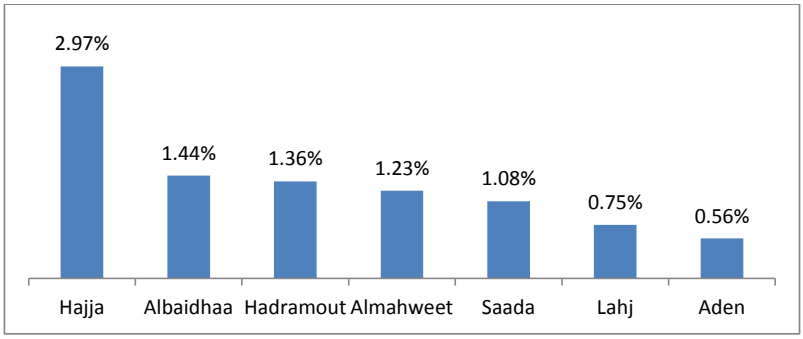

Figure 5: Cancer risk in the seven highly vulnerable province

Finding 4: There is a link between living in a particular province and cancer risk. The link is more evident in Hajja province.

5.8.5 Marital Status

The marital status had notable contribution in cancer risk in Taiz province. However, generally, there is no link between being married and cancer development. Though, unlike being married, widowhood showed relatively low connection to cancer development. Finally, being single showed some connection with cancer in provinces other than Capital City and Ibb.

Finding 5: Widowhood appeared to have a remarkable connection with cancer development.

5.8.6 Age

The rules shows unclear connection between age and vulnerability to cancer.

Finding 6: The contribution of age is not clear. Nevertheless, there is some connection between age advance and cancer risk.

5.8.7 Other Interesting Rules

Finally, Table 10 shows some other rules that uncover some rule combination that led to invulnerability to risk cancer.

https://doi.org/10.20428/JST.23.2.1

Table 10: Other interesting rules

Index	Rule	Instance weight %.	Mislass .%	Class
R3	PROVINCE = Cap AND AGE_TESTED = '(69.5-72.5]' AND QAT = True: No (428.0/7.0)	2.01%	1.64%	No
R28	AGE_TESTED = '(31.5-32.5]' AND QAT = True: No (417.0/22.0)	1.96%	5.28%	No
R23	AGE_TESTED = '(36.5-37.5]' AND PROVINCE = Hud: No (399.0/6.0)	1.88%	1.50%	No
R27	QAT = True AND SHAMMA = False AND AGE_TESTED = '(49.5-51.5]' AND PROVINCE = Ibb: No (398.0/11.0)	1.87%	2.76%	No
R77	QAT = True AND PROVINCE = San AND GENDER = M: No (337.0/24.0)	1.77%	6.37%	No
R4	PROVINCE = Cap AND AGE_TESTED = '(60.5-61.5]': No (359.0/11.0)	1.69%	3.06%	No
R13	PROVINCE = Cap AND AGE_TESTED = '(23.5-27.5]' AND GENDER = F AND MARITAL_STATUS = S: No (310.0/9.0)	1.46%	2.90%	No
R6	PROVINCE = Cap AND AGE_TESTED = '(63.5-69.5]': No (304.0/11.0)	1.43%	3.62%	No
R39	MARITAL_STATUS = S AND PROVINCE = Ibb AND AGE_TESTED = '(23.5-27.5]': No (301.0/18.0)	1.42%	5.98%	No
R63	QAT = True AND AGE_TESTED = '(23.5-27.5]': No (291.0/41.0)	1.37%	14.09%	No
R35	MARITAL_STATUS = M AND QAT = True AND AGE_TESTED = '(55.5-56.5]' AND PROVINCE = Dmr: No (263.0/4.0)	1.24%	1.52%	No
R47	AGE_TESTED = '(35.5-36.5]' AND PROVINCE = Hud: No (219.0/10.0)	1.03%	4.57%	No
R18	PROVINCE = Cap AND MARITAL_STATUS = Y AND No_FAMILY_MEMBERS = '(-inf-3.5]' AND AGE_TESTED = '(9.5-18.5]': No (217.0/19.0)	1.02%	8.76%	No

6. Discussion and Conclusion

In this study, Data Mining techniques were used to address the problem of unknown non-human-body risk factors that can contribute to cancer development in Yemen. Decision Tree classification method was selected to classify the NCCF data into two classes: Yes (with caner or malignant) and No (free of cancer of benign). SMOTE technique, which adds synthesized instances

to the dataset, was used to add balance to the NCCF class-imbalanced data. The model produced high true positives (TP) and true negatives (TN), which are measures of high model performance. Low misclassification rate shows that high information gain of attributes selected.

6.1 Summary of Main Findings

The resulting model has shown high performance in terms of classification rate, accuracy, sensitivity, specificity, precision, and area under ROC curve. Accuracy, as the term suggests, shows how the prediction was similar to (or different from) the actuality. Besides, according to [14], sensitivity is a measure of ‹completeness› (i.e., actually positive instances labeled as such after classification), precision can be thought of as a measure of ‹exactness› (i.e., what percentage of instances labeled as positive are actually such). As "area under ROC curve (AUROC) depicts the tradeoff between hit rates (benefit) and false alarm rates (cost)" [13], the benefit value approached 1 (i. e. 100%). Finally, the attribute selected through attribute selection process succeeded in selecting attributes most relevant to DT, which was used here.

6.2 Discussion

The high permanence of the resulting model can be attributed to several factors, but the most important, as we have seen previously, is the right selection of learning algorithm and data preprocessing.

The role of the right selection of learning algorithm in high performance of the model is obvious. What has been achieved here support a literature review conclusions, which is that DT is one of the highly reliable learning algorithms. Data preprocessing also had a significant impact on the KDD process. Although C4.5 is designed to deal with missing values, imputation of those values added a great deal of enhancement to C4.5 performance. Other preprocessing tasks such as removing extreme values and outliers and data class-balancing added more enhancement to the model.

The model simplicity and interpretability can be attributed to the useful Weka features such as PART rule algorithm and numerical values discretization that facilitated the conversion of the model tree to a set of rules that are easy to understand to human.

https://doi.org/10.20428/JST.23.2.1

Regarding the tree obtained, the rules extracted from the model tree showed some expectable results such as the high contribution of tobacco smoking in cancer risk, and other unexpected ones, such as neutrality of low connection of chewing qat, the vulnerability of residing in Hajja province and risk of widowhood. However, the prediction power of the model depends, in general, on whether the data obtained from NCCF actually reflects the reality or not. The values of the rules discovered by the model should be considered with some cautiousness. For example, the high risk in Hajja indicates the possibility of existence of environmental problems that pose cancer risk. Similarly, high risk linked to widowhood can imply that there are psychological consequences of widowhood that cause immunity inefficiency and contributes to cancer development. However, there are limitations that need to be considered when interpreting the above rules. For instance, the high cancer risk shown in Hajja province could be because that the NCCF visitors from Hajja are much more than those from the other provinces.

7. Implications

The classifier building method here was limited to Decision Tree (DT). Hence, what would be the outcomes of other methods were used. Of course, DT has proved high reliability in producing high-performance models, however, other methods and algorithms cloud be used for this reason.

Away from the human body and biological processes taking place inside it, there are still external factors that affect its health. Those external factors can be environmental, socials, etc. This is what this study tried to confirm. Thus, cancer treatment should go beyond prescribing medicine or conducting therapy sessions. Cancer spread in Yemen should receive more attention by data miners, as one study is not be enough to decide everything about this kind of risk factors. Further research should address external risk factors that could not be addressed here such as quality of life, job nature, social relations etc. For some restrictions imposed by the type and quality of the data worked on in this study, the model did not specify which risk factor contributes to what type of cancer. Hence, researches are needed to link risk factor directly to particular types of cancer. Similarly, in some provinces, cancer spread is higher than that in others. This reveals environmental/social/economic-based risk factors and the need to data attributes would be needed to work out this ambiguity out.

References

[1] Ada and R. Kaur (2013). "A Study of Detection of Lung Cancer Using Data Mining Classification Techniques." International Journal of Advanced Research in Computer Science and Software Engineering 3(3).

[2] Ada and R. Kaur (2013). "Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient."

[3] Agrawal, A., et al. (2011). A lung cancer outcome calculator using ensemble data mining on SEER data. Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics, ACM.

[4] Ahmed, A., et al. (2013). "Early detection of lung cancer risk using data mining." Asian Pacific Journal of Cancer Prevention 14(1): 595-598.

Al-Bahrani, R., et al. (2013). Colon cancer survival prediction using ensemble data mining on SEER data. Big Data, 2013 IEEE International Conference on, IEEE.

[5] Batista, G. E. and M. C. Monard (2003). "An analysis of four missing data treatment methods for supervised learning." Applied Artificial Intelligence 17(5-6): 519-533.

[6] Bellaachia, A. and E. Guven (2006). "Predicting breast cancer survivability using data mining techniques." Age 58(13): 10-110.

[7] Chaurasia, V. and S. Pal (2014). "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability." IJCSMC) International Journal of Computer Science and Mobile Computing 3(1): 10-22.

[8] Chaurasia, V. and S. Pal (2014). "A Novel Approach for Breast Cancer Detection using Data Mining Techniques." International Journal ofr Innovative in Computer and Communication Engineering & 2(1).

[9] Chawla, N. V., et al. (2002). "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16(1): 321-357.

[10] Dash, R., et al. (2011). "Comparative analysis of supervised and unsupervised discretization techniques." International Journal of Advances in Science and Technology 2(3): 29-37.

[11] Dougherty, J., et al. (1995). Supervised and unsupervised discretization of continuous features. Machine learning: proceedings of the twelfth international conference.

[12] Fawcett, T. (2006). "An introduction to ROC analysis." Pattern recognition letters 27(8): 861-874.

[13] Han, J., et al. (2012). Data Mining: Concepts and Techniques. 225Wyman Street,Waltham, MA 02451, USA, Morgan kaufmann.

https://doi.org/10.20428/JST.23.2.1

[14] He, Y., et al. (2012). "Using association rules mining to explore pattern of Chinese medicinal formulae (prescription) in treating and preventing breast cancer recurrence and metastasis." Journal of translational medicine 10(supplement 1): S12.

[15] Kharya, S., et al. (2014). "Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer." International Journal of Computer Applications 92(10).

[16] Nookala, G. K. M., et al. (2013). "Performance analysis and evaluation of different data mining algorithms used for cancer classification." International Journal of Advanced Research in Artificial Intelligence (IJARAI) 2(5).

[17] Osei-Bryson, K.-M. (2004). "Evaluation of decision trees: a multi-criteria approach." Computers & Operations Research 31(11): 1933-1945.

[18] Rajan, J. R. and J. J. Prakash (2013). Early Diagnosis of Lung Cancer using a Mining Tool. National Conference on Architecture, Software systems and Green computing-2013 (NCASG2013).

[19] Sharma, C. T. and M. Jain (2013). "WEKA approach for comparative study of classification algorithm." International Journal of Advanced Research in Computer and Communication Engineering 2(4): 1925-1931.

[20] Sharma, N. (2011). Comparing the performance of data mining techniques for oral cancer prediction. Proceedings of the 2011 International Conference on Communication, Computing & Security, ACM.

[21] Sharma, N. and H. Om (2012). "Framework for early detection and prevention of oral cancer using data mining." International Journal of Advances in Engineering & Technology 4(2).

[22] Sharma, N. and H. Om (2013). "Data mining models for predicting oral cancer survivability." Network Modeling Analysis in Health Informatics and Bioinformatics 2(4): 285-295.

[23] Sharma, N. and H. Om (2014). "Extracting Significant Patterns for Oral Cancer Detection Using Apriori Algorithm." Intelligent Information Management 2014.

[24] Sumbaly, R., et al. (2014). "Diagnosis of Breast Cancer using Decision Tree Data Mining Technique." International Journal of Computer Applications 98(10): 16-24.

[25] Upton, G. and I. Cook (1996). Understanding statistics, Oxford University Press.

[26] Viera, A. J. and J. M. Garrett (2005)."Understanding interobserver agreement: the kappa statistic." Fam Med 37(5): 360-363.

[27] Witten, I. H., et al. (2011). Data mining: Practical machine learning tools and techniques, Morgan Kaufman, Boston.

[28] Tseng Chih-Jen, Lu Chi-Jie, Chang Chi-Chang, Chen Gin-Den, Cheewakriangkrai Chalong.Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence.Artificial Intelligence in Medicine http://dx.doi.org/10.1016/j.artmed.2017.06.003.

[29] Iraklis Varlamis , Ioannis Apostolakis , Dimitra Sifaki-Pistolla , Neelanjan Dey , Vassilios Georgoulias , Christos Lionis , Application of data mining techniques and data analysis methods to measure cancer morbidity and mortality data in a regional cancer registry: The case of the island of Crete, Greece, Computer Methods and Programs in Biomedicine (2017), doi:10.1016/j.cmpb.2017.04.011.

[30] Yang, H., & Chen, Y. P. P. Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information. Expert Systems with Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.03.019.

[31] Hamed Majidi Zolbanin, Dursun Delen, Amir Hassan Zadeh, Predicting Overall Survivability in Comorbidity of Cancers: A Data Mining Approach, Decision Support Systems (2015), doi: 10.1016/j.dss.2015.04.003

https://doi.org/10.20428/JST.23.2.1