The application of data mining techniques and feature selection methods in the risk classification of Egyptian liver cancer patients using clinical and genetic data.

Esraa H. Abdelaziz
Faculty of Computer and Information Sciences
Ainshams University, Cairo, Egypt
00201060196461
esraahamdiabbas@cis.asu.edu.eg

Khaled El-Bhanasy
Obour Institutes, Cairo, Egypt
khaled.bahnasy@oi.edu.eg

Sanaa M. Kamal
Faculty of Medicine
Ainshams University, Cairo, Egypt
sanaakamal@ainshamsmedicine.net

Rasha Ismail
Faculty of Computer and Information Sciences
Ainshams University, Cairo, Egypt
rashaismail@cis.asu.edu.eg

ABSTRACT
Data mining techniques has shown great potential in biomedical and health care fields. The objective of this paper is to apply feature selection methods and data mining techniques to Egyptian liver cancer patients' data to predict their prognosis and extract important features that affect the patient's survivability. Genetic and Clinical data from 1541 patients were analyzed. Three feature selection methods and seven data mining techniques were studied and compared. Wrapper Subset method and Random Forest proved to be the best performing feature selection method and data mining technique respectively. Moreover, important genetic features such as p53 gene exon 6 and 9 mutations proved to have a significant impact on patient's overall prognosis.

CCS Concepts
• **Information Systems→Data mining.** • **Applied Computing →Life and medical sciences→Health Informatics.**

Keywords
Medical informatics, Data mining; machine learning; feature selection; liver cancer; cancer prognosis

1. INTRODUCTION

Hepatocellular carcinoma (HCC) is the sixth most common cancer worldwide; it's also one of the main causes of cancer-related deaths in Egypt [1]. The prognosis of the disease is very poor and its treatment is not yet standardized due to its extreme heterogeneity. The incidence of HCC is highest in Asia and Africa, where the endemic high prevalence of hepatitis B and hepatitis C strongly predisposes to the development of chronic liver disease and subsequent development of HCC [2]. The threat of HCC is expected to continue to grow in the coming years [3]. Early diagnosis of HCC and optimization of the timing of various treatments represent the cornerstone for achieving better therapeutic outcomes for HCC [4]. Serum markers for HCC such as alfa-feto protein and des-γ-carboxyprothrombin (DCP) do not have high levels of sensitivity or specificity. A continuing trend in biomarker discovery is the increasing adoption of high-throughput proteomics-based or genomic-based approaches. Therefore, recent more robust, sensitive, specific, cost effective genetic markers are crucially need for prediction of HCC and follow-up response to therapy.

Disease prognosis has multiple aspects, including course of the disease, quality of life, potential for complications and associated health issues, life expectancy and survival. Attempting to predict the prognosis of cancer patients is a very challenging task. Machine learning techniques are usually employed to accurately identify the most significant clinical and genetic features from a patient's complex data that could potentially affect their overall survival duration [5].

In this paper, 1541 Egyptian liver cancer patients' clinical and genetic data were studied and classified into two main classes according to the survival median which was 24 months; patients who lived less than or equal to two years were grouped into a high-risk group and patients who lived more than two years were grouped into a low-risk group, after undergoing the appropriate mode of treatment according to physicians. The study consists of four phases; the first phase consists of preprocessing the data by discretizing the continuous attributes. The second phase is selecting the most significant features to use in classifying the patients; three feature selection methods were tested, which are Correlation attribute evaluation, Information gain attribute evaluation and Wrapper subset evaluation. The third phase is the classification phase in which seven different data mining techniques were applied and compared. The methods that were tested are Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest, REPTree, Decision Tree (J48), JRip and PART. The final phase is studying the outcome of the techniques in the third phase that detected any key features that proved to be significant to the patient's prognosis or produced a set of rules for classifying the patients into the two mentioned groups.

ICSIE 2019, April 9-12, 2019, Cairo, Egypt

ACM ISBN 978-1-4503-6105-7/19/04…$15.00
https://doi.org/10.1145/3328833.3328849

The paper is divided as follow: the related work is discussed in section 2; all of the materials and methods used and how they were evaluated is explained in section 3; the experimental results and the rules produced by the models are presented in section 4 and section 5 respectively; a discussion about the models and the key features discovered is conducted in section 6; conclusions and some future work recommendations are described in section 7.

2. RELATED WORK

Data mining techniques have been widely used to identify important features in complex medical datasets and build reliable models that can predict the prognosis of patients with different diseases.

The study in [6] used two data mining methods, namely decision trees and Artificial Neural Networks to determine the prognosis of oral cancer patients. Those methods were applied on 673 patients' clinical and demographic prognostic features; their performances were compared against logistic regression. They both performed better with accuracies 81.7236% and 80.5349% respectively, which is significantly higher than the statistical method's accuracy, which was 63.2987 %. In addition, they discovered important features that were associated with patients that had lower survival rate.

In [7], multiple data sets of lung cancer patients that consisted of gene expression and clinical data of 440 patients were analyzed using ANN. Similar to our study, the patients were divided according to the survival median (36 months); patients who lived longer than that were low-risk groups and the ones who didn't, were considered a high-risk group. After trying several types of ANNs, the method used yielded a strong classification model for patients' survival with overall accuracy 83%.

Also in [8], three data mining models were compared, SVM, ANN, and semi-supervised learning models to find the best model for predicting breast cancer prognosis. Dataset from SEER was used that contain 162,500 records, which include cancer statistics features. The patients were divided based on their survivability, which refers to patients who didn't survive and those who did. Semi-supervised learning model showed superiority over the other techniques with 71% overall accuracy and was recommended by the authors. Five-fold cross validation was conducted as the evaluation method. Some informative features that relates to patient's survivability were also discovered.

In Addition, logistic regression, ANN and decision trees were compared in [9] to predict the prognosis of advanced schistosomiasis in the Hubei province. Demographical data, clinical data, surgical procedures and outcome of 4136 patients were analyzed using the three models. Patients were divided into favorable prognosis; which refers to recovery or improvement and poor prognosis classes, which refers to death or deterioration. Although the accuracies of the models were similar, ANN outperformed the other methods using AUC measure with mean of 0.9267.

Previous studies show a great tendency to rely on older data mining methods like ANN, instead of the methods that can build easily interpreted models [10]. Moreover, publications that analyze Egyptian liver cancer patients' prognosis are very limited. In this study, several data mining techniques that can build complex models such as ANN and easily interpreted ones such as J48 were compared to classify liver cancer patients according to their prognosis and discover informative features that affect their survivability.

3. MATERIALS AND METHODS

3.1 Clinical and Genetic Data

Data of 1541 Egyptian patients with HCC was collected at liver cancer unit, Ainshams, Egypt. Each patient's record consists of 41 attributes; demographic, biochemical, histologic and genetic markers in addition to the class attribute that indicates the patient's survival months.

3.2 Data Preprocessing

Preprocessing was conducted in two forms; data discretization and feature selection. The data mining methods were studied with and without each of these preprocessing steps and the results are detailed in tables 4 and 5.

a. Data Discretization is the process of partitioning continuous features into discrete categories, which might aid into building models with higher accuracy. Some data mining methods necessitates the use of discrete data, while some generally perform better using discrete values [11]. In this study, our data was discretized using WEKA 3.8 supervised discretization function. The resulting categories from the continuous features are listed in table 1. Since this is supervised discretization, the features that don't contribute any information in classifying the class is discretized into one value and was not included in the below table.

Table 1. The categories of the continuous features after being discretized.

Feature Name	Feature Categories
Age	<=65.5, >65.5
Body Mass Index	<=18.5 , >18.5
Red blood cells count	<=2801199.5, >2801199.5
White blood cells count	<=3202.5, >3202.5
Platelets	<=62757.5 , >62757.5 and <=104365.5 , > 104365.5
Bil	<=9.5 , >9.5
ALT (Alanine transaminase test)	<= 58.5, 58.5> and <=104.5, >104.5
AST (Aspartate transaminase test)	<= 58.5, 58.5> and <=117.5, >117.5
ALB (Albumin test)	<=1.5 , >1.5
AFP	<=1616.5, >1616.5 and <=2885, >2885
PIVKA	<=1996.5, >1996.5
Telomerase	<=1621.5, >1621.5
HGF	<=2014 , >2014
TGF b	<=495.5 , >495.5 and <=2012.5 , >2012.5
Survivin	<=532.5, >532.5 and <=2025, >2025
PDGF-ab	<=2019.5 , >2019.5
Survival Months	<= 24 , >24

b. Feature Selection In this preprocessing step, three feature selection methods that are implemented in WEKA 3.8 were studied and compared.

1. Correlation Attribute Evaluation: This method measures Pearson's correlation between the attribute and the class. The

search method used is Ranker method, which ranks the attributes according to their evaluation. A cutoff value of 0.2 was used, so the attributes with correlation value less than 0.2 were removed.

2. Information Gain Attribute Evaluation: This method measures that amount of information that is gained from the values of the attribute with respect to the class. The search method used is Ranker method. The cutoff value used is 0.05, so the attributes with information gain less than 0.05 were discarded.

3. Wrapper Subset Evaluation: This method uses a learning scheme to evaluate the features and uses cross validation method to predict the accuracy of that scheme using those features. The search method used is greedy stepwise which performs a greedy search through the attributes forward and backward. J48 algorithm was used as the classifier. Table 2 lists the attributes that were selected by each feature selection method.

3.3 Data Mining Methods

Data mining is the process of analyzing sets of data in order to predict a future unknown outcome [12]. It has been widely used in the field of medical informatics to aid in the prediction of patient's diagnosis and prognosis. Moreover, data mining techniques can extract interesting features from a patient's data that are related to their prognosis and might be independent of their treatment [13].

Table 2: A description of the attributes that were selected by each feature selection method. The Correlation value (CV), and the Information Gain (IG) are included for the first two methods.

Correlation Attribute		**Information Gain Attribute**		**Wrapper Subset**
CV	Attribute	IG	Attribute	Bil
0.496	Mutation 6[1]	0.216	Mutation 6[1]	Mutation 9[1]
0.495	Mutation 9[1]	0.202	Mutation 9[1]	AFP
0.386	HP:0001413	0.116	HP:0001413	Telomerase
0.336	Mutation 2[1]	0.105	Mutation 2[1]	HGF
0.334	Mutation 10[1]	0.104	Mutation 10[1]	TGF b
0.318	Mutation 8[1]	0.089	Mutation 8[1]	Survivin
0.259	Mutation 5[1]	0.070	survivin	bFGF
0.243	Mutation 3[1]	0.069	TGF b	HP:0001413
0.208	Platelets	0.066	Telomerase	
0.206	Telomerase	0.065	Platelets	
		0.052	Mutation 5[1]	
		0.052	PIVKA	

In this study, several data mining methods have been utilized and compared according to their accuracy in classifying the patients into a High-Risk group and a Low-Risk Group. WEKA's 3.8 implementation was used for the following data mining methods:

1. Artificial Neural Network (ANN): ANN is a method that simulates the human brain, which is a densely connected network of neurons [14]. The most popular architecture of ANN is multilayer perceptron (MLP) was used. It consists of at least an input layer, a hidden layer and an output layer. The information moves in one direction only starting from the input layer and ending with the output layer.

2. Support Vector Machine (SVM):"It is a machine learning method based on statistical learning theory that transforms original input space into a higher-dimensional feature space to find an optimal separating hyperplane"[15]. John Platt's sequential minimal optimization algorithm for training was used [16]. Nominal features are converted into binary ones.

3. Decision Tree (J48): It's a data mining algorithm that aims to create a pruned tree that helps in decision making. It consists of a root node, leaf nodes that represents classes and internal nodes that represents the conditions applied on the features[17].

4. RepTree: Reduced Error Pruning Tree ("REPT") is a tree-learning algorithm. Its purpose is minimizing the variance, and

[1] P53 gene mutation

is built based on the information gain [18].The algorithm chooses the best tree out of several trees built in multiple iterations, the best tree is considered to be the representative. The tree is pruned using the mean square error measure on its predictions.

5. Random Forest: A tree-learning algorithm with the aim of creating accurate predictions while avoiding data overfitting. It is a combination of tree predictors in which each tree is built using the values of a random sample vector from the data and all trees have the same distribution (Breiman 2001, 2002) [19].

6. JRip: It's a rule-learning algorithm that is based on the algorithm RIPPER (Repeated Incremental Pruning to Produce Error Reduction), which was proposed by William Cohen [20]. The algorithm builds a set of rules that are formed by adding rules to an empty rules set. A rule grows by adding conditions in a greedy fashion. It tries all possible values for each feature and chooses the ones with the highest information gain. After the rule set is created, it's optimized to enhance its fit to the data and minimize its size [21].

7. PART: It's an algorithm that uses the separate-and-conquer technique and was proposed by Frank and Witten [22]. It combines between the C4.5 decision tree and the RIPPER rule generator. It creates a decision list by generating partial C4.5 decision tree repeatedly and using the best leaf as a rule in each iteration.

3.4 Evaluation

In this study, 5-fold cross validation was conducted to evaluate the accuracy of each model. On each round of cross validation, 80% of the data is divided into a training set and the remaining 20% is used for testing. This process is repeated five times to avoid overfitting.

4. EXPIREMENTAL RESULTS

The cohort of Egyptian liver cancer patients that were described in 3.1 were studied by applying the six data mining techniques mentioned in 3.3 with and without the preprocessing steps presented in 3.2, which include the discretization process and the different feature selection techniques, to compare between their results and figure out the most accurate model and predictors that could classify the patients according to their survival months. The results of the experiment are detailed in table 3 and table 4.

Table 3: The accuracy of the data mining techniques after discretizing the dataset and applying different feature selection techniques.

	Correlation Attribute Evaluation	**Information Gain Attribute Evaluation**	**Wrapper Subset Evaluation**
ANN	82.35%	85.33%	86.57%
SVM	82.48%	85.20%	85.46%

J48	82.28%	85.20%	86.83%
REPTree	82.35%	85.07%	86.18%
Random Forest	82.28%	85.27%	86.96%
JRip	82.28%	85.27%	86.70%
PART	82.28%	85.20%	86.70%

Table 4: The accuracy of the data mining techniques without each preprocessing step.

	Discretization with no Feature Selection	No Discretization and no Feature selection
ANN	86.24%	79.56%
SVM	85.53%	82.54%
J48	87.35%	79.75%
REPTree	85.79%	85.59%
Random Forest	86.37%	87.54%
JRip	86.76%	83.32%
PART	86.76%	80.08%

5. MODEL AND PREDECTION RULES

The data mining techniques that were used in this study, can either produce complex models, equations that could predict the prognosis of a new patient, or a set of rules that can classify a new patient and identify important features that can be considered markers. In this section, we'll discuss three of the data mining techniques that could produce the latter, which are J48, JRip and PART. In section 4, the accuracy of using each method with and without different preprocessing techniques was reported, however, only the experiment that resulted in the highest accuracy for each method will be discussed. Each rule will be represented by a condition or more, followed by the class and two numbers, the first number represents the number of instances that the rule covers and the second number represents the number of instances misclassified by the rule. Note that the absence of a second number indicates that no instances were misclassified by that rule.

1. Decision Tree (J48): The highest accuracy was produced when the dataset was discretized and no feature selection method was used. The accuracy of the classification was 87.3459%. The resulting tree was converted into a set of 16 rules that were ordered according to the number of cases covered by each rule. Figure 1 describes each rule.

2. JRip: This algorithm also produced the highest accuracy when the dataset was discretized and no feature selection method was used. The accuracy of the classification was 86.7618 %, however a close accuracy was obtained when the feature selection method, Wrapper Subset Evaluation was used with accuracy of 86.697%. Both sets of rules are described in figures 2 and 3.

3. PART: Like the previous two methods, the rule-generating algorithm produced the highest accuracy when the dataset was discretized and no feature selection method was used. Just like the JRip algorithm, the accuracy of the classification was 86.7618 % and a close accuracy was also obtained when the feature selection method, Wrapper Subset Evaluation was used with accuracy of 86.697%. Both sets of rules are described in the figures 4 and 5.

6. DISCUSSION

As shown in the experimental results in section 4, the accuracy of the studied methods ranged from approximately 80% to 87%.

Figure 1: The rules generated by the decision tree algorithm.


1. IF Tp53 gene exon 6 mutation = No : **Low-Risk Groups** (385/14)
2. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF<=2014 AND HP:0001413 = No AND PIVKA <= 1996.5 AND Tp53 gene exon 9 mutation = Yes : **High-Risk Groups** (731/117)
3. IF Tp53 gene exon 6 mutation = Yes AND Telomerase >1621.5 : **Low-Risk Groups** (91)
4. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF<=2014 AND HP:0001413 = Yes AND AFP>2885 : **Low-Risk Groups** (84/9)
5. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF<=2014 AND HP:0001413 = Yes AND AFP <=1616.5 AND ALT > 58.5 AND<= 104.5 : **High-Risk Groups** (44/15)
6. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF<=2014 AND HP:0001413 = Yes AND AFP <=1616.5 AND ALT <= 58.5 : **High-Risk Groups** (39/10)
7. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 : **Low-Risk Groups** (38)
8. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF<=2014 AND HP:0001413 = No AND PIVKA <= 1996.5 AND Tp53 gene exon 9 mutation = No AND Platelets <= 62757.5 OR > 104365.5 : **Low-Risk Groups** (31/12)
9. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >2025 : **Low-Risk Groups** (30)
10. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF<=2014 AND HP:0001413 = Yes AND AFP >1616.5 AND <=2885 : **Low-Risk Groups** (15/2)
11. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF<=2014 AND HP:0001413 = Yes AND AFP <=1616.5 AND ALT > 104.5 : **Low-Risk Groups** (14/2)
12. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb <=1.5 : **Low-Risk Groups** (12)
13. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF<=2014 AND HP:0001413 = No AND PIVKA <= 1996.5 AND Tp53 gene exon 9 mutation = No AND Platelets >62757.5 AND <=104365.5 : **High-Risk Groups** (10/4)
14. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF>2014 : **Low-Risk Groups** (8)
15. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin <=532.5 : **Low-Risk Groups** (6)
16. IF Tp53 gene exon 6 mutation = Yes AND Telomerase <= 1621.5 AND PDGF-ab >2019 AND survivin >532.5 AND <=2025 AND Alb >1.5 AND HGF<=2014 AND HP:0001413 = No AND PIVKA > 1996.5 : **Low-Risk Groups** (3)

Figure 2: The rules generated by JRIP algorithm without using a feature selection method.

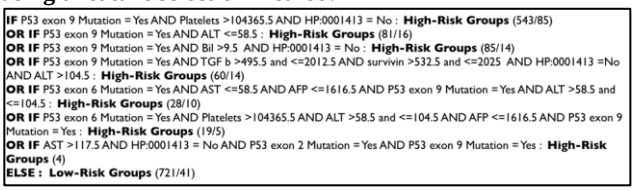

IF P53 exon 9 Mutation = Yes AND Platelets >104365.5 AND HP:0001413 = No : **High-Risk Groups** (543/85)
OR IF P53 exon 9 Mutation = Yes AND ALT <=58.5 : **High-Risk Groups** (81/16)
OR IF P53 exon 9 Mutation = Yes AND Bil >9.5 AND HP:0001413 = No : **High-Risk Groups** (85/14)
OR IF P53 exon 9 Mutation = Yes AND TGF b >495.5 and <=2012.5 AND survivin >532.5 and <=2025 AND HP:0001413 =No AND ALT >104.5 : **High-Risk Groups** (60/14)
OR IF P53 exon 6 Mutation = Yes AND AST <=58.5 AND AFP <=1616.5 AND P53 exon 9 Mutation = Yes AND ALT >58.5 and <=104.5 : **High-Risk Groups** (28/10)
OR IF P53 exon 6 Mutation = Yes AND Platelets >104365.5 AND ALT >58.5 and <=104.5 AND AFP <=1616.5 AND P53 exon 9 Mutation = Yes : **High-Risk Groups** (19/5)
OR IF AST >117.5 AND HP:0001413 = No AND P53 exon 2 Mutation = Yes AND P53 exon 9 Mutation = Yes : **High-Risk Groups** (4)
ELSE : Low-Risk Groups (721/41)

Figure 3: The rules generated by JRIP algorithm after using Wrapper subset evaluation.

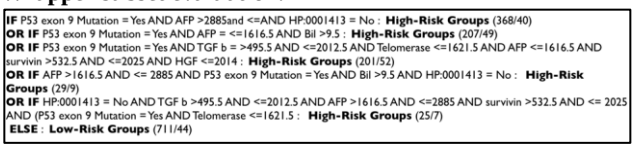
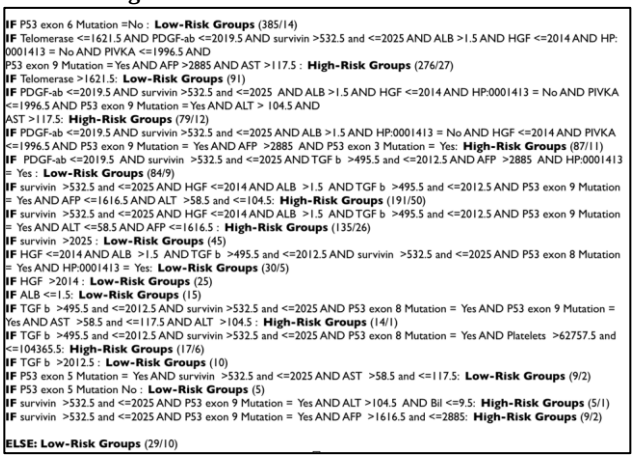

IF P53 exon 9 Mutation = Yes AND AFP >2885and <=AND HP:0001413 = No : **High-Risk Groups** (368/40)
OR IF P53 exon 9 Mutation = Yes AND AFP = <=1616.5 AND Bil >9.5 : **High-Risk Groups** (207/49)
OR IF P53 exon 9 Mutation = Yes AND TGF b = >495.5 AND <=2012.5 AND Telomerase <=1621.5 AND AFP <=1616.5 AND survivin >532.5 AND <=2025 AND HGF <=2014 : **High-Risk Groups** (201/52)
OR IF AFP >1616.5 AND <= 2885 AND P53 exon 9 Mutation = Yes AND Bil >9.5 AND HP:0001413 = No : **High-Risk Groups** (29/9)
OR IF HP:0001413 = No AND TGF b >495.5 AND <=2012.5 AND AFP >1616.5 AND <=2885 AND survivin >532.5 AND <= 2025 AND (P53 exon 9 Mutation = Yes AND Telomerase <=1621.5 : **High-Risk Groups** (25/7)
ELSE : Low-Risk Groups (711/44)

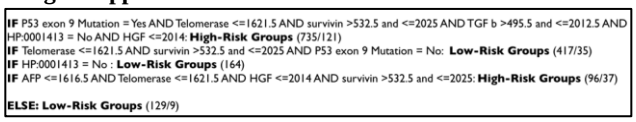
Figure 4: The rules generated by the PART algorithm without using a feature selection method.


IF P53 exon 6 Mutation =No : **Low-Risk Groups** (385/14)
IF Telomerase <=1621.5 AND PDGF-ab <=2019.5 AND survivin >532.5 and <=2025 AND ALB >1.5 AND HGF <=2014 AND HP: 0001413 = No AND PIVKA <=1996.5 AND
P53 exon 9 Mutation = Yes AND AFP >2885 AND AST >117.5 : **High-Risk Groups** (276/27)
IF Telomerase >1621.5: **Low-Risk Groups** (91)
IF PDGF-ab <=2019.5 AND survivin >532.5 and <=2025 AND ALB >1.5 AND HGF <=2014 AND HP:0001413 = No AND PIVKA <=1996.5 AND P53 exon 9 Mutation = Yes AND ALT > 104.5 AND
AST >117.5: **High-Risk Groups** (79/12)
IF PDGF-ab <=2019.5 AND survivin >532.5 and <=2025 AND ALB >1.5 AND HP:0001413 = No AND HGF <=2014 AND PIVKA <=1996.5 AND P53 exon 9 Mutation = Yes AND AFP >2885 AND P53 exon 3 Mutation = Yes: **High-Risk Groups** (87/11)
IF PDGF-ab <=2019.5 AND survivin >532.5 and <=2025 AND TGF b >495.5 and <=2012.5 AND AFP >2885 AND HP:0001413 = Yes : **Low-Risk Groups** (84/9)
IF survivin >532.5 and <=2025 AND HGF <=2014 AND ALB >1.5 AND TGF b >495.5 and <=2012.5 AND P53 exon 9 Mutation = Yes AND AFP <=1616.5 AND ALT >58.5 and <=104.5: **High-Risk Groups** (191/50)
IF survivin >532.5 and <=2025 AND HGF <=2014 AND ALB >1.5 AND TGF b >495.5 and <=2012.5 AND P53 exon 9 Mutation = Yes AND ALT <=58.5 AND AFP <=1616.5 : **High-Risk Groups** (135/26)
IF survivin >2025 : **Low-Risk Groups** (45)
IF HGF <=2014 AND ALB >1.5 AND TGF b >495.5 and <=2012.5 AND survivin >532.5 and <=2025 AND P53 exon 8 Mutation = Yes AND HP:0001413 = Yes: **Low-Risk Groups** (30/5)
IF HGF >2014 : **Low-Risk Groups** (25)
IF ALB <=1.5: **Low-Risk Groups** (15)
IF TGF b >495.5 and <=2012.5 AND survivin >532.5 and <=2025 AND P53 exon 8 Mutation = Yes AND P53 exon 9 Mutation = Yes AND AST >58.5 and <=117.5 AND ALT >104.5 : **High-Risk Groups** (14/1)
IF TGF b >495.5 and <=2012.5 AND survivin >532.5 and <=2025 AND P53 exon 8 Mutation = Yes AND Platelets >62757.5 and <=104365.5: **High-Risk Groups** (17/6)
IF TGF b >2012.5 : **Low-Risk Groups** (10)
IF P53 exon 5 Mutation = Yes AND survivin >532.5 and <=2025 AND AST >58.5 and <=117.5: **Low-Risk Groups** (9/2)
IF P53 exon 5 Mutation No : **Low-Risk Groups** (5)
IF survivin >532.5 and <=2025 AND P53 exon 9 Mutation = Yes AND ALT >104.5 AND Bil <=9.5: **High-Risk Groups** (5/1)
IF survivin >532.5 and <=2025 AND P53 exon 9 Mutation = Yes AND AFP >1616.5 and <=2885: **High-Risk Groups** (9/2)

ELSE: Low-Risk Groups (29/10)

Figure 5: The rules generated by PART algorithm after

using Wrapper subset evaluation.



The best performing method was Random Forest, followed closely by J48 decision tree with accuracies 87.54% and 87.35% respectively. Almost all methods performed better after discretizing the data with the exception of Random Forest which showed a slight increase in the accuracy without that step. In addition, some methods produced higher accuracy when the feature selection step was omitted, however, the best performing feature selection technique was Wrapper Subset Evaluation in all the experiments.

Not all data mining techniques result in a set of readable rules, however, three of the studied methods do. In section 5, the rules with the highest accuracies were reported. The J48 resulted in a tree that was converted into a set of 16 rules. Both JRip and PART produced the same accuracy, however, the number of rules differed. While both of them performed slightly better without the feature selection step, using the Wrapper Subset Evaluation reduced the number of rules in the first algorithm from eight to six rules and the latter from twenty to five rules.

Genetic features were present in the rules with the most coverage, which proves their significance in classifying liver cancer patients according to their prognosis. The algorithms used indicate that the absence of P53 gene exon 6 mutations would likely mean that the patient is a low-risk patient, while the presence of P53 gene exon 9 mutations with other prognostic features presented in figures 2,3 and 5 classifies the patient as a high-risk patient.

7. Conclusion and Future Work

In this study, a cohort of 1541 Egyptian liver cancer patients were studied and divided into a low-risk group and a high-risk group. Multiple feature selection methods were compared and the Wrapper subset evaluation technique proved to be superior on our data in selecting the most significant features. Seven data mining methods were utilized to classify the patients according to their prognosis. Random Forest and J48 resulted in the highest accuracies with the other methods not very far behind. The methods that produced easily interpreted models were reported and informative features such as P53 gene exons 6 and 9 proved to be significant in the classification of patients.

In the future, acquiring more data would be helpful in building more reliable models and discovering more significant markers. In addition, even though combining data mining methods in techniques such as voting or bagging wouldn't produce easily understandable rules, the combination of the strengths of more than one method could potentially result in an increase in the overall accuracy of the classifying model.

8. REFERENCES

[1] Holah, Nanis S., et al. "Hepatocellular carcinoma in Egypt: epidemiological and histopathological properties." Menoufia Medical Journal 28.3 (2015): 718.

[2] Wallace, Michael C., et al. "The evolving epidemiology of hepatocellular carcinoma: a global perspective." Expert review of gastroenterology & hepatology 9.6 (2015): 765-779.

[3] Grandhi, Miral Sadaria, et al. "Hepatocellular carcinoma: from diagnosis to treatment." Surgical oncology 25.2 (2016): 74-85.

[4] Forner A, Llovet JM, Bruix J. Hepatocellular carcinoma. Lancet. 2012;379(9822):1245–55

[5] Croft, Peter, et al. "The science of clinical practice: disease diagnosis or patient prognosis? Evidence about "what is likely to happen" should shape clinical practice." BMC medicine13.1 (2015): 20.

[6] Tseng, Wan-Ting, et al. "The application of data mining techniques to oral cancer prognosis." Journal of medical systems 39.5 (2015): 59

[7] Chen, Yen-Chen, Wan-Chi Ke, and Hung-Wen Chiu. "Risk classification of cancer survival using ANN with gene expression data from multiple laboratories." Computers in biology and medicine 48 (2014): 1-7.

[8] Park, Kanghee, et al. "Robust predictive model for evaluating breast cancer survivability." Engineering Applications of Artificial Intelligence 26.9 (2013): 2194-2205.

[9] Li, Guo, et al. "Comparison of three data mining models for prediction of advanced schistosomiasis prognosis in the Hubei province." PLoS neglected tropical diseases 12.2 (2018): e0006262.

[10] Cruz, Joseph A., and David S. Wishart. "Applications of machine learning in cancer prediction and prognosis." Cancer informatics 2 (2006): 117693510600200030

[11] Ram ́ez-Gallego, Sergio, et al. "Data discretization: taxonomy and big data challenge." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 6.1 (2016): 5-21.

[12] Crockett, David, and Brian Eliason. "What is data mining in healthcare?." HealthCatalyst,[Online]. Available: https://www. healthcatalyst. com/data-mining-in-healthcare (2014).

[13] Kourou, Konstantina, et al. "Machine learning applications in cancer prognosis and prediction." Computational and structural biotechnology journal 13 (2015): 8-17.

[14] Rakhman, Arief, Goeij Yong Sun, and Rama Catur APP. "Building artificial Neural network Using Weka Software." Information System Department, Sepuluh Nopember Institute of Technology at Surabaya, Indonesia (2009).

[15] Bui, Dieu Tien, et al. "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree." Landslides 13.2 (2016): 361-378.

[16] Keerthi, S. Sathiya, et al. "Improvements to Platt's SMO algorithm for SVM classifier design." Neural computation 13.3 (2001): 637-649.

[17] Bhargava, Neeraj, et al. "Decision tree analysis on j48 algorithm for data mining." Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering 3.6 (2013).

[18] Kalmegh, Sushilkumar. "Analysis of WEKA data mining algorithm REPTree, Simple CART and Random Tree for classification of Indian news." International Journal of Innovative Science, Engineering and Technology 2.2 (2015): 438-46.

[19] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[20] Cohen, William W. "Fast effective rule induction." Machine Learning Proceedings 1995. 1995. 115-123.

[21] Daud, Nor Ridzuan, and David Wolfe Corne. "Human readable rule induction in medical data mining." Proceedings of the European Computing Conference. Springer, Boston, MA, 2009.

[22] Frank, Eibe, and Ian H. Witten. "Generating accurate rule sets without global optimization." (1998).