Prediction of mortality and occurrence of complications for gastric cancer patients

1st Maria Ana de Brito
Algoritmi Research Center
University of Minho
Braga, Portugal
a73580@alunos.uminho.pt

2nd Cristiana Neto
Algoritmi Research Center
University of Minho
Braga, Portugal
a72064@alunos.uminho.pt

3rd José Machado
Algoritmi Research Center
University of Minho
Braga, Portugal
jmac@di.uminho.pt

***Abstract*—Gastric cancer is one of the most prevalent types of cancer in the whole world, affecting millions of people over the last decades. Its symptoms are ambiguous, which leads to late diagnoses, reducing the patients' chances of survival. In most countries, routine screenings are not usual, which also contributes to the detection of this gastric malignancy in later and more dangerous (and often fatal) stages. One of the main focus of improving healthcare services related to gastric cancer relies on increasing the survival rates. This and predicting if a patient will suffer from any complication following the surgery can aid the healthcare professionals in selecting better and more efficient treatment strategies. Thus, this constitutes as the aims of this study which will test and compare a set of classification models in order to improve the prediction accuracy. Data mining techniques will be put into use, since it's been proved they are one of the best ways of producing useful information for many businesses, including healthcare.**

***Index Terms*—healthcare, gastric cancer, data mining, classification, prediction, mortality rates, complication occurrence**

I. Introduction

Data generated by healthcare gives insight into many aspects that were previously unknown to healthcare professionals and can be potentially useful for improving the quality of medical procedures or treatment strategies [1].

Large amounts of data are produced everyday by hospitals and other medical facilities. One of the big characteristics of healthcare data is its heterogeneity, since it includes diverse sources, data types and formats. A careful observation is required in order to assess its quality and identify possible problems that need be to solved.

Since the data is so complex, it's practically impossible to analyze it with traditional tools and methods [2]. This complexity calls for more sophisticated techniques that are able to manage and produce meaningful knowledge. Like this, the healthcare services records can serve as a way of assessing their quality and the patient's satisfaction [3].

Data mining (DM) is a process that refers to the extraction of useful information from vast amounts of data [4]. It's used to find hidden patterns and uncover unknown correlations that are not obvious when observing the data with the naked eye [5]. There are many applications for DM, since it's greatly adaptable to distinct businesses and goals. They can go from retail stores, hospitals and banks to insurance or airline companies. Thus, data mining can greatly benefit the healthcare industry by creating an environment rich in meaningful knowledge [4].

Gastric cancer is one of the most common causes of death worldwide. It's the fourth most frequently occurring cancer in men and the seventh most commonly occurring cancer in women. According to the World Cancer Research Fund (WCRF) [6] there were over 1 million new cases and an estimated 783.000 deaths related to gastric cancer in 2018. The greatest incidence rates are recorded in Eastern Asia (countries like South Korea, Mongolia and Japan occupy the first three spots), whereas in Northern America and Europe the rates are generally low [7].

Despite big advances in technologies and healthcare that provide better and more accurate diagnoses, this cancer, while registering a decreasing trend worldwide, continues to be among the first places of most deadly malignancies. There are a lot of factors that may influence the occurrence of this type of cancer. It's strongly suggested that general bad eating and drinking habits contribute to it. The consumption of alcoholic drinks and a diet rich in salty foods are among the most dangerous causes for gastric cancer. A greater body fatness and smoking also play a part in raising the risk of its occurrence.

The symptoms are often overlooked since they are not specific and can have other reasons besides gastric cancer. Early signs can be indigestion, feeling bloated, slight nausea and loss of appetite. As the tumor grows, other symptoms, often more serious, start to manifest, such as stomach pain, vomiting, weight loss and constipation.

Early diagnosis can save a lot of lives, because there are more treatment opportunities to fight the cancer. However, since the symptoms are considered ambiguous and it's not typical to do routine screenings, this cancer is often detected at later stages. This fact strengthens the high mortality rates all over the world.

The focus of this study is in the prediction of the mortality of patients that suffer from this gastric malignancy and of the occurrence of complications after the patients' hospital stays. The goal is to analyze the data available and the results obtained and make comparisons among different classifiers as to draw conclusions about them.

This paper is divided in five sections. After the current introduction, some works related to gastric cancer are men-

tioned. The third section includes the phases that constituted the knowledge discovery process. In the fourth section, the results obtained are compared and discussed. Finally, in the last section, some conclusions are drawn and the future work that will entail this data mining project is revealed.

II. Related Work

The improvement of gastric cancer diagnosis, mortality and complications rates have always been one of the most common work themes when it comes to the application of data mining techniques on healthcare. Thus, some of the existing works have been studied prior to the conception of this paper.

Lee et al. [8] applied data mining techniques in order to create a prediction process for the occurrence of postoperative complications on gastric cancer patients. They've developed artificial neural networks (ANN) and compared their results with those of the traditional logistic regression (LR) approach, where they've achieved an average correct classification rate of 84.16% with ANN in contrast with 82.4% of LR.

Polaka et al. [9] planned various approaches for diagnosing gastric cancer using the original dataset and datasets with subsets of features. The best results were obtained for the dataset using attribute subsets selected with the wrapper approach. Four different models were tested, where C4.5 obtained 74.7% of accuracy, as well as CART. The RIPPER algorithm produced an accuracy of 73.9%, while the Multilayer Perceptron got the best results with 79.6%.

Goshayeshi et al. [10] used an optimized MICE technique to predict the chances of survival in gastric cancer patients. Three different techniques were executed, the first one, which consisted in the application of logistic regression, obtained 63.03% of accuracy, while the second technique that used a not optimized MICE algorithm earned an accuracy value of 66.14%. Finally, the third approach with the optimized MICE algorithm produced results with 72.57% of accuracy.

III. Knowledge Discovering Process

The knowledge discovery process model used during the development of this study is Cross-Industry Standard Process for Data Mining, most commonly known as CRISP-DM. This methodology includes six important steps, such as: Business Understanding, Data Understanding, Data Preparation, Evaluation and Deployment [11]. The Machine Learning software Weka was used for analyzing and understanding the data provided, preparing it for the subsequent Machine Learning algorithms and their application in data mining tasks.

A. Business Understanding

Cancer affects millions of people all over the world and is one of the biggest threats to people's lives and life quality. Gastric cancer is one of the most common causes of cancer-related deaths, behind, for example, lung cancer [12]. The prognostic is usually not favourable to the survival of patients, since there's only a probability of less than 30% survival upon diagnosis in Europe [12]. However, in Japan this rate goes up to 90% thanks to early examinations and tumor resections [13].

This malignancy presents no specific symptoms in early stages, which causes delayed diagnoses that lead to the high mortality of patients. In advanced stages, the patient may feel a variety of more serious symptoms, like abdominal pain, indigestion, severe nausea and inexplicable weight loss [13]. By the time these symptoms appear, the cancer has already developed to more dangerous stages. When the tumor is diagnosed, it's often too late for any curative medical procedure to take place.

There are various objectives with this study, such as:

- Promote early examinations among the general population in order to avoid late gastric cancer diagnoses that often lead to the patient's death
- Predict the probability of mortality after the surgery
- Predict the occurrence of complications after in-hospital stays for gastric cancer patients

Thus, this study aims to improve many aspects related to gastric cancer and the way it affects the patients' lives. The focus falls on their hospital admissions and possible complications that may occur related or not to the tumor. The procedures performed and the patient's health status after the hospital stay are also subjects to this work.

The first item is related to the healthcare business goals. The improvement of the quality of the medical services provided is one of the most crucial aspects in this industry. This translates into an increment on the survival rates of patients, in this case patients that suffer from gastric cancer.

The rest of the goals listed are related to the objectives inherent to the data mining process. Through the application and refinement of data mining techniques these objectives will provide a substantial help to healthcare professionals.

B. Data Understanding

The data used for this study was collected from a Portuguese hospital and is related to patients with gastric cancer. It includes over 60 variables with information about the patients' admission, stay at the hospital, possible complications and the result of the performed procedure related to 154 patients.

C. Data Preparation

The dataset provided has a lot of attributes that have high percentages of missing values. When it comes to the numerical variables, half of the attributes have over 45% of missing or null values. This makes them not useful to study or to subject them to Machine Learning algorithms, since they offer little to no meaningful information. Consequently these attributes were removed from the dataset. Moreover, after a careful analysis, it was stated that there are certain attributes that are extremely similar, even presenting the same values. As such, one of the attributes was also taken off of the dataset, leaving only one of them on the dataset as to avoid any redundancy. Some of the attributes refer to technical aspects related to the extraction of the data, so they were removed from the dataset as well. The categorical attributes were submitted to the same process.

After the data cleaning, three more features were created derived from existing attributes. These new features refer to

the number of postoperative complications registered, to the occurrence of complications 30 days after the in-hospital stay and to the death of patients.

The final result was a dataset with 33 features (4 numeric and 29 categorical). However, in order to analyze alternative approaches with fewer attributes, three more datasets were created.

The first one was created with attribute selection performed by the OneR algorithm, where 19 attributes were selected (1 numeric and 18 categorical). Whereas, the second dataset included a subset of features that were selected using the Relief algorithm. This one was composed of 20 attributes, from which 1 was numeric and the rest categorical. On the other hand, the features selected for the third dataset were chosen based on the Pearson's correlation method. This subset of features was comprised of 21 attributes, where 2 of them were numeric and 19 were categorical.

The summary of the characteristics of the datasets can be checked on the Table I. The first use case (identified with the number 1) refers to the original dataset after the data cleaning and creation of new attributes. The second use case is related to the dataset created with the attribute selection using the OneR algorithm. The dataset that includes the subset of features selected by the algorithm Relief is the third use case. It was assigned to the last use case, Use Case 4, the dataset that is based on the feature selection that measures the attribute's worth with the Pearson's correlation.

It's important to note that the first column refers to the identification of the use case (and consequently the dataset), while the second column indicates the total number of attributes for each dataset. The third column and fourth column present the number of numeric and categorical attributes, respectively.

TABLE I
SUMMARY OF THE DATASETS CREATED FOR THE PREDICTION OF MORTALITY

Use Case	# Attr	# Numeric Attr	# Categorical Attr
1	33	4	29
2	19	1	18
3	20	1	19
4	21	2	19

D. Modeling

The first proposed goal was to predict the mortality of gastric cancer patients that were admitted to the hospital. Based on the health status available, as well as info about the performed surgery and its outcome, the models will predict if it's more likely that the patient will survive or pass away. In this case, two datasets (the original - after the data preparation - and one more that was subjected to feature selection) were tested with cross-validation with 10 folds. Thus, the classification process included two scenarios that contemplated distinct set of features.

On the other hand, the second goal was to predict the occurrence of complications after the hospital stays. In this case, features related to the patients' morbidity and survival, and complications' rank were removed, along with info about the possible existence of complications. These attributes were eliminated in order to ensure an unbiased and correct prediction. These tests were also performed with 10-fold cross-validation.

The classifiers selected were Random Forest, J48, Simple Logistic, Bayes Net and PART. In addiction, the algorithm AdaBoost and Bagging were also executed in conjunction with the first three models already mentioned.

E. Evaluation

Once the modeling phase was concluded, the chosen classifiers were put to test in order to evaluate and compare their results. The metrics used were Accuracy, Precision, F-Measure and Recall. They are defined as such:

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$FMeasure = 2 * ((PR + RC)/(PR * RC)) \quad (3)$$

$$Recall = TP/(TP + FN) \quad (4)$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives PR = Precision and RC = Recall.

1) Prediction results for Original Dataset: The Table II presents the results obtained during the classification process using the original dataset, after the data preparation.

TABLE II
PREDICTION RESULTS FOR THE CLASS "DIEDYN" FOR USE CASE 1

Classifier	Accuracy	Precision	F-Measure	Recall
Random Forest	68.4564	0.670	0.674	0.685
J48 (Pruned)	66.443	0.640	0.647	0.664
J48 (Pruned)[a]	67.1141	0.645	0.652	0.671
Bayes Net	67.1141	0.676	0.672	0.671
Simple Logistic	68.4564	0.683	0.681	0.685
PART	66.443	0.662	0.658	0.664
AdaBoost + RF	69.7957	0.686	0.688	0.698
AdaBoost + J48	64.4295	0.640	0.642	0.644
AdaBoost + J48[a]	65.1007	0.659	0.655	0.651
AdaBoost + SL	61.0738	0.619	0.614	0.611
Bagging + RF	71.8121	0.702	0.705	0.718
Bagging + J48	63.7584	0.605	0.616	0.638
Bagging + J48[a]	65.7718	0.637	0.643	0.658
Bagging + SL	66.443	0.656	0.660	0.664

[a]Using Laplace correction.

2) Prediction process results for dataset with subset of features: The Table III exposes the results obtained for the prediction of gastric cancer patients' mortality using the feature selection technique that evaluates the worth of a feature using the OneR algorithm.

The Table IV presents the results obtained for the prediction of gastric cancer patients' mortality using the feature selection method that evaluates the worth of an attribute using the Relief algorithm.

TABLE III
PREDICTION RESULTS FOR THE CLASS "DIEDYN" FOR USE CASE 2

Classifier	Accuracy	Precision	F-Measure	Recall
Random Forest	64.4295	0.622	0.631	0.644
J48 (Pruned)	63.7584	0.625	0.629	0.638
J48 (Pruned)[a]	64.4295	0.629	0.634	0.644
Bayes Net	67.7852	0.675	0.676	0.678
Simple Logistic	67.1141	0.662	0.665	0.671
PART	61.742	0.603	0.604	0.617
AdaBoost + RF	64.4295	0.636	0.639	0.644
AdaBoost + J48	64.4295	0.649	0.645	0.644
AdaBoost + J48[a]	63.7584	0.635	0.636	0.638
AdaBoost + SL	63.7584	0.631	0.634	0.638
Bagging + RF	63.7584	0.617	0.624	0.638
Bagging + J48	65.7718	0.629	0.639	0.658
Bagging + J48[a]	65.1007	0.629	0.636	0.651
Bagging + SL	69.1275	0.680	0.684	0.691

[a]Using Laplace correction.

TABLE IV
PREDICTION RESULTS FOR THE CLASS "DIEDYN" FOR USE CASE 3

Classifier	Accuracy	Precision	F-Measure	Recall
Random Forest	68.4564	0.674	0.671	0.685
J48 (Pruned)	63.0872	0.616	0.622	0.631
J48 (Pruned)[a]	63.7584	0.621	0.627	0.638
Bayes Net	68.4564	0.678	0.681	0.685
Simple Logistic	67.7852	0.673	0.675	0.678
PART	68.4564	0.661	0.664	0.685

[a]Using Laplace correction.

In the Table V, the results obtained for the prediction of gastric cancer patients' mortality using the feature selection that evaluates the worth of attributes by using the Pearson's correlation are presented.

TABLE V
PREDICTION RESULTS FOR THE CLASS "DIEDYN" FOR USE CASE 4

Classifier	Accuracy	Precision	F-Measure	Recall
Random Forest	64.4295	0.632	0.636	0.644
J48 (Pruned)	63.0872	0.602	0.614	0.631
J48 (Pruned)[a]	63.7584	0.606	0.619	0.638
Bayes Net	65.7718	0.653	0.655	0.658
Simple Logistic	65.1007	0.638	0.644	0.651
PART	61.0738	0.603	0.606	0.611

[a]Using Laplace correction.

3) Prediction results for the occurrence of complications: The Table VI exposes the results obtained for the prediction of the occurrence of complications for gastric cancer patients after their hospital stay.

TABLE VI
CLASSIFICATION PROCESS RESULTS FOR THE CLASS "COMPLICATION"

Classifier	Accuracy	Precision	F-Measure	Recall
Random Forest	76.4706	0.730	0.730	0.765
J48 (Pruned)	81.6993	0.834	0.777	0.817
J48 (Pruned)[a]	81.6993	0.834	0.777	0.817
Bayes Net	67.3203	0.708	0.687	0.673
Simple Logistic	80.3922	0.807	0.761	0.804
PART	77.7778	0.753	0.754	0.778
AdaBoost + RF	79.7386	0.798	0.750	0.797
AdaBoost + J48	71.2418	0.696	0.703	0.712
AdaBoost + J48[a]	72.549	0.705	0.713	0.725
AdaBoost + SL	71.2418	0.702	0.707	0.712
Bagging + RF	76.4706	0.725	0.698	0.765
Bagging + J48	80.3922	0.793	0.771	0.804
Bagging + J48[a]	81.0458	0.806	0.776	0.810
Bagging + SL	75.817	0.736	0.743	0.758

[a]Using Laplace correction.

IV. DISCUSSION

A. Predict the mortality of gastric cancer patients

a) Use Case 1: After a rigorous analysis of the results obtained, it's possible to verify that the best predictions for the mortality of gastric cancer patients were achieved by the Simple Logistic model with an accuracy of 68.4564%. Another model, Random Forest (RF), obtained the same accuracy value, however the former classifier presented better results for the others metrics in comparison with RF.

Yet, when the ensemble techniques Bagging and Boosting were applied, a better accuracy (71.8121%) was obtained by executing the algorithm AdaBoost with Random Forest. Thus, making this the best result achieved for this goal.

b) Use Case 2: Using a dataset with fewer features than the original one, the best results were achieved with the Bayes Net algorithm that produced an accuracy of 67.7852%. However, when the Bagging technique was applied in conjunction with the Simple Logistic model, a better value of 69.1275% was obtained.

c) Use Case 3: With this approach, the results that were obtained from the execution of the selected models showed that Random Forest, as well as Bayes Net (BN) and PART produced the same accuracy of 68.4564%. In spite of the same value, BN presented better results for precision, f-measure and recall - making this the better classifier for this dataset.

d) Use Case 4: Using the feature selection method that measures each attribute's worth with the Pearson's correlation, the results obtained were, in general, inferior to the previous ones. As such, the algorithm Bayes Net produced an accuracy of 65.7718% while the Simple Logistic approach achieved an accuracy value of 65.1007%. These results were very similar and both lower than the best results obtained in the previous tests.

e) Summary: When compared to the results obtained with the datasets that were submitted to feature selection methods, it's possible to conclude that the original dataset produced better overall results for the selected metrics, as can be seen on the Table VII and Table VIII. Anyhow, the results obtained were not very high, due to the multitude of reasons that may lead to a patient's death. These include factors not directly linked to the gastric cancer or deaths that took place because the patient was already palliative.

Only two approaches were selected for the application of the ensemble methods Bagging and Boosting, that aim to reduce bias and variance and then achieve a better performance by the model. The results for the original dataset and the dataset that was submitted to the feature selection method that uses the algorithm OneR were exposed, since these were the ones that

managed to obtain better results than the ones already stated.

TABLE VII
SUMMARY OF THE BEST RESULTS FOR THE PREDICTION OF MORTALITY (ACCURACY)

Use Case	Classifier	Accuracy
1	Simple Logistic	68.4564
1	Bagging + Random Forest	71.8121
2	Bayes Net	67.7851
2	Bagging + Simple Logistic	69.1275
3	Bayes Net	68.4564
4	Bayes Net	65.7728

TABLE VIII
SUMMARY OF THE BEST RESULTS FOR THE PREDICTION OF MORTALITY (PRECISION, F-MEASURE, RECALL)

Use Case	Classifier	Precision	F-Measure	Recall
1	Simple Logistic	0.683	0.681	0.685
1	Bagging + RF	0.702	0.705	0.718
2	Bayes Net	0.675	0.676	0.678
2	Bagging + SL	0.680	0.684	0.691
3	Bayes Net	0.678	0.681	0.685
4	Bayes Net	0.653	0.655	0.658

B. Predict the occurrence of complications after in-hospital stays for gastric cancer patients

When it comes to the prediction of complications after a hospital stay for gastric cancer, the results obtained were more satisfactory. The reason for that is that it's considerably easier to anticipate if a patient will suffer from any complications or disabilities following a surgery by observing the health status available. As such, the best accuracy value was recorded for the J48 algorithm (81.6993%). There were no differences between the pruned and pruned using the Laplace correction models for the metrics used as can be observed in the Table IX. There were others classifiers that came close to this accuracy value, such as Simple Logistic that produced an accuracy of 80.39222% and the ensemble algorithm Bagging with J48 with an accuracy value of 81.0458%.

TABLE IX
SUMMARY OF THE BEST RESULTS FOR THE PREDICTION OF COMPLICATIONS

Classifier	Accuracy	Precision	F-Measure	Recall
J48	81.6993	0.834	0.777	0.817
J48[a]	81.6993	0.834	0.777	0.817

[a]Using Laplace correction.

C. Summary

The best result for the first proposed goal, that is the prediction of mortality in gastric cancer patients, was achieved using the ensemble technique Bagging in conjunction with the algorithm Random Forest. The accuracy of this model was of 71.8121% (Table X), which is in accordance with the reviewed literature. As can be seen on the first row of Table XI, the metrics precision, f-measure and recall all achieved values around 0.7.

The second goal was aiming to predict the possible occurrence of complications among gastric cancer patients after their in-hospital stays. The model that provided the best results was J48 with an accuracy of 81.6993% (Table X). On the other hand, for the precision, f-measure and recall, this classifier obtained values around 0.8 (Table XI).

TABLE X
SUMMARY OF THE BEST RESULTS FOR THE PROPOSED GOALS (ACCURACY)

Goal	Use Case	Classifier	Accuracy
1	1	Bagging + Random Forest	71.8121
2	-	J48[a]	81.6993

[a]Using Laplace correction.

TABLE XI
SUMMARY OF THE BEST RESULTS FOR THE PROPOSED GOALS (PRECISION, F-MEASURE, RECALL)

Goal	Use Case	Classifier	Precision	F-Measure	Recall
1	1	Bagging + RF	0.702	0.705	0.718
2	-	J48[a]	0.834	0.777	0.817

[a]Using Laplace correction.

V. CONCLUSIONS AND FUTURE WORK

Data mining techniques are becoming more than ever more useful for processing and exploiting medical data. These methods can analyze in real-time complex and heterogeneous data and make conclusions about it. This enables the discovery of unknown information that can be used by the healthcare industry in order to improve its quality.

What in the past was an impossible task to manage, now it's feasible to submit millions and millions of medical records to an algorithm and obtain relevant results. There are countless softwares available to the general public that provide tools to process the data. They grant means of reading it, clean it, prepare it for the application of algorithms and even allow to execute and refine the models.

This paper aimed to predict the mortality of gastric cancer patients based on their health status, data about the tumor and surgery info, as well as to make predictions about the possibility of occurrence of complications following a in-hospital stay.

Considering the various reasons that may lead to the patient's death, it becomes challenging to predict if the patient might perish or survive. There are a lot of aspects that influence this outcome that show no direct link to the cancer in question. A lot of patients, due to late diagnosis, face little to no chances of survival since no curative treatment can treat the tumor. These facts contribute to accuracy values around 70%.

On the other hand, it's simpler to predict if a patient will suffer from complications after their hospital stay, since it's

possible to rely more on the data available. Observing the data about the tumor (its localization, stage, size, lymph nodes and metastasis) and analyzing the health status of the patient (given by the ASA score) among other factors, the prediction of the occurrence of complications becomes a more straightforward process. Hence, the accuracy obtained for this goal was around 82%.

The future work will consist in obtaining a larger dataset with more relevant data in order to improve the prediction process for both patients' mortality and occurrence of complications. Others models will also be tested and their results compared with the previous ones already obtained. Techniques such as oversampling will also be put into practice in order to improve the accuracy of the prediction process.

REFERENCES

[1] J. Archenaa and E. A. Anita, "A survey of big data analytics in healthcare and government," *Procedia Computer Science*, vol. 50, pp. 408–413, 2015.

[2] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, p. 3, 2014.

[3] Q. K. Fatt and A. Ramadas, "The Usefulness and Challenges of Big Data in Healthcare," *J Healthc Commun*, vol. 3, no. 2, p. 21, 2018.

[4] C. Neto, H. Peixoto, V. Abelha, A. Abelha, and J. Machado, "Knowledge Discovery from Surgical Waiting lists," in *Procedia Computer Science*, 2017.

[5] Y. Li, "DATA MINING: CONCEPTS, BACKGROUND AND METHODS OF INTEGRATING UNCERTAINTY IN DATA MINING," tech. rep.

[6] "Stomach cancer statistics — World Cancer Research Fund."

[7] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.," *CA: a cancer journal for clinicians*, 2018.

[8] Y.-C. Lee, "Mining the Complication Pattern of Gastric Cancer Patients by Using Artificial Neural Networks and Logistic Regression," tech. rep., 2006.

[9] I. Polaka, E. Gašenko, O. Barash, H. Haick, and M. Leja, "Constructing Interpretable Classifiers to Diagnose Gastric Cancer Based on Breath Tests," in *Procedia Computer Science*, 2016.

[10] R. Hosein Zadeh, L. Goshayeshi, A. Khooie, K. Etminani, Z. Yousefli, S. Nastarani, N. Farhang Nezhad, and A. Golabpoor, "PREDICTIVE MODEL FOR SURVIVAL IN PATIENTS WITH GASTRIC CANCER," *Acta Healthmedica*, 2017.

[11] E. Silva, L. Cardoso, F. Portela, A. Abelha, M. F. Santos, and J. Machado, "Predicting Nosocomial Infection by Using Data Mining Technologies," pp. 189–198, Springer, Cham, 2015.

[12] R. Sitarz, M. Skierucha, J. Mielko, G. J. A. Offerhaus, R. Maciejewski, and W. P. Polkowski, "Gastric cancer: Epidemiology, prevention, classification, and treatment," *Cancer Management and Research*, vol. 10, pp. 239–248, 2018.

[13] P. Correa, "Gastric Cancer. Overview.," *Gastroenterology Clinics of North America*, vol. 42, no. 2, pp. 211–217, 2013.