جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة ديالى
كلية العلوم

تصنيف السرطـــان بالأعتـماد على تقنيـات تنقيـب البيانـات

رسالة مقدمة
الى كلية العلوم في جامعة ديالى وهي جزء من متطلبات نيل
شهادة الماجستير في علوم الحاسبات
تقدمت بها الطالبة

هاجـر كامـل احمـد

بإشراف

أ.د. ظاهر عبدالهادي عبدالله

أ.م. د. جمال مصطفى عباس

1440هـ

2019م

Chapter One

General Introduction

Chapter one
General Introduction

1.1 Introduction

The healthcare industry is among the most information intense productions. Medical information, data and knowledge continue growing on a daily basis. It shows a very important role in the usage of clinical data, such discoveries pattern recognition is essential for the diagnosis of new illnesses and the study of various patterns is available when classification of data takes place [1].

Medical informatics has been estimated that an acute care hospital may generate five terabytes of data a year. The capability to utilize these data for extracting useful information for quality healthcare is crucial.

Computer aid information retrieval may assist to support quality decision making and avoid human error. Though human decision-making is often they are perfect, they become bad when there is a need to classify a large amount of data. Also it will reduction accuracy and efficiency of decisions when humans are put into immense work and stress. Imagine a physician who has to examine five patient records, he/she will go through them with simple and ease. But if increases the number of records from five to fifty with a time constraint, it is most sure that the efficiency with which the physician delivers the results won't be as high as the ones acquired when he had just five records to be analyzed, this leads up to utilize the data mining in medical informatics [2].

Data mining is “the process of drilling data for find out latent patterns which can be interpreted into worthy information”. It uses witness unmatched growth in the past few years and has been comprehend in healthcare field of late. This

understood is in the wake of burst of intricate medical data. Medical data mining can utilize the unobserved kinds existent in large medical data which else is left without detection [3].

Data mining techniques which are executed to medical data contain association rule mining to find frequent types, classification, clustering and prediction. Traditionally, data mining techniques were utilized in different fields. Data mining is presented comparatively late into the healthcare field. Though, as on today many of the research exists in the literature. This has led up to the development of decision support systems and smart systems in healthcare field for exact diagnosis of diseases, predicting the force different diseases, and remote health oversight. Specifically, the data mining techniques are more salutary in predicting all kinds of cancer and heart diseases [3].

1.2 An Overview of Cancer Diseases

Cancer is one of the most prevalent diseases in the world that results in majority of dying. Cancer is caused by abnormal growth of cells in any of the tissues or parts of the body. Cancer may occur in any part of the body and may spread to several other parts. Some kinds of tumors are not propagation in the body since it is not needful that all kinds of tumors are cancerous. Cancer has different symptoms such as abnormal bleeding, tumor, more weight loss, long-term cough …etc. The cancers affecting human body there are near about 100 kinds. In medical domain, research on cancer is one of the challenging, attractive and main points of focused area. There is a need for accurate automatic prediction systems for tumor and cancers. Early detection of cancer at the benign phase could save a person's life and prevent from spreading to other parts in the malignant phase [4].

Among different cancers types, breast cancer is the very prevalent cancer which infects females worldwide, representing 25% of all cancers and in 2017 estimated among 1.67 million cancer status diagnosed. Especially, women from less developed countries have a little more number of statuses compared with developed nations (883000 statuses against 794000 in developed nations) [2]. For example, the rate breast cancer in India is lower as compared with United Kingdom where this rate in India (28.8 per 100000) while in United Kingdom (95 per 100000), but death rate is at equal (12.7 vs 17.1 per 100000) with United Kingdom. There is a big increase in the incident and cancer associated morbidity and mortality in India as Indian and global studies describe. Earlier cervical cancer was most popular cancer in India women but currently the breast cancer has override the previous and this leads to increase cancer death. Detection cancer at an early stage gives an opportunity to heal and avoid death [5].

Another type of cancer is the Lung cancer which is the abnormal growth of anomalous cells that begins in one or both lungs, regularly in the cells that line the air passages. The abnormal cells do not broaden into healthy lung tissue, they divide speedily and form tumors. Lung cancer is splitted into two major types are : small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC is further classified into adenocarcinoma, squamous cell carcinoma and large cell carcinoma since treatment differs greatly depending on phase and kind the lung cancer, the diagnostic workup is critical in terms of identifying the specific kind of lung cancer, the phase of the disease, and the ability of the patient to endure treatment. NSCLC represents 80% of all lung cancers, with adenocarcinoma accounting for 40% of all status of lung cancer. Squamous cell carcinoma occurs most often in the central region of the lung while

adenocarcinoma tumors are peripheral in origin, arising from the alveolar surface epithelium or bronchial mucosal [6].

Large cell carcinoma composes only 15% of all lung cancers and seems to be decreasing in happening because of improved diagnostic techniques. The second main kind of lung cancer is SCLC, in which there are also several histologic groupings: pure mixed small cell, small cell and large cell carcinoma as well as combined small cell. SCLC is ordinarily more aggressive than NSCLC [6].

1.3 Related works

This section reviews some of previous studies and explains the different techniques that are used for developing the cancer classification systems.

1- **S. Palaniappan and T. Pushparaj (2013) [7]:** They proposed an automatic diagnosis system for predicting breast cancer based on Association Rules(AR) and Neural Network(NN)algorithm. Where reducing the dimension of breast cancer dataset from nine to four features is performed using AR1 and AR2. Classification is performed using NN. The proposed AR1+AR2+NN system performance is compared with NN model. The accuracy with proposed system is 98.4% while the accuracy with NN is 95.6%.

2- **Shweta Kharya et al. (2014) [8]:** The proposed system in this work is exploiting the ability of Naive Bayes Classifier (NBC) in the classification of cancer data to either benign or malignant tumor. Experimental results uncover that NBC is a proficient methodology for extraction of important patterns from breast cancer dataset. The maximum accuracy of 93% have been achieved .A GUI has been designed to enter the patient's records and

presence of breast cancer for a patient is predicted utilizing the probability of disease being presented in the probability of finding the symptoms.

3- **B. M. Gayathri and C. P. Sumathi (2016) [9]:** They proposed system to classify the breast cancer with few attributes. Reducing the attributes is done by apply Linear Discriminant Analysis (LDA) algorithm. The dataset is passed to LDA repeatedly and the combination of variables which gave the high accuracy is selected. Attributes selected by LDA are classified by Gaussian Naïve Bayes approach .This classification gives a maximum accuracy of 96.6%.

4- **Taysir Hassan et al. (2016) [10]:** They propose a hybrid technique based on Deep Neural Networks (DNN) and multi criteria decision making technique Analytical Hierarchy Process (AHP). DNNs are integrated with AHP to deal with large datasets and improve classification accuracy. They use three different breast cancer datasets for evaluating the performance of the hybrid techniques, involving wisconsin breast cancer datasets. The accuracy obtained of applying the hybrid techniques was 84.33%.

5- **Younus Ahmad Malla et al. (2017) [11]:** In this work, they performed analytical evaluation of certain selected machine learning algorithms it is carried on the breast cancer dataset by using the tool WEKA. Some preprocessing is also done on the input dataset by applying certain WEKA in building filters and its overall effect on the accuracy of the prediction was also noted down. The results showed that the random forest from the decision trees achieved best accuracy with filters the accuracy without filter was found 69%, while after applying the filter its accuracy become 98%. Similarly Logistic Regression got the second rank with 96% but without filters it was 68%. Finally, the accuracy by using Naive Bayes was 91% and without filters 71%.

6- **Vikas Chaurasia et al. (2018) [12]:** They have developed expectation models for breast cancer survivability. By utilizing three common data mining algorithms (RBF Network, J48, Naive Bayes) to develop the expectation models utilizing a large dataset, of breast cancer to predict the survivability of a patient uses ten fold cross validation method to measure the unbiased estimate of the three prediction models for performance comparison purposes. The results indicated that the Naive Bayes is the best predictor where its accuracy was 97.36% with the holdout sample, RBF Network came out to be the second with 96.77% accuracy while J48 came out third with 93.41% accuracy.

7- **S. Karthik et al. (2018) [13]:** They designed a system for breast cancer diagnosis. This system consists of preprocessing for handle missing value and Recursive Feature Elimination (RFE) for feature selection method. Deep Neural Network is used for classification of the data into benign and malignant. The experimental results show that the accuracy obtained from this system was found 97.66%.

1.4 Problem Statement

Cancer is one of the most prevalent diseases in the world that a result in majority of dying.it is abnormal growth of specific cells with the potential to spread to other parts of the body. The actual treatment of this disease has not been found yet. One of methods of getting rid of this disease is the Final and accurate diagnosis at an early stage has been needed for getting rid of this disease but the complexity of the disease made difficult to obtain accurate results on the tumor and in early time. For this reason the physician needs an efficient technique gives diagnosis accurately.

1.5 Aim of Thesis

The aim of this work is to design and implement a cancer classification system able to accurately diagnose the tumors in human body by using Naive Bayes and hybrid algorithm (Class Association Rule and Deep Neural Network), in order to obtain high accuracy to help the physicians to prevent the errors while identifying and classifying the tumors from different datasets.

1.6 Outline of Thesis

The rest chapters in this thesis are organized as follows:

Chapter Two: Theoretical Background

This chapter clarifies the definition and the types of data mining in addition to Knowledge discovery in databases (KDD) steps, it also explains the Data mining techniques used for classification.

Chapter Three: **The Proposed System**

This chapter describes the proposed classification system with its design and implementation.

Chapter Four: Experimental Results and Evaluation

This chapter shows the implementation results of the proposed system steps and evaluates these results.

Chapter Five: Conclusions and Suggestions for Future work

This chapter presents the conclusions of this work. Furthermore, it provides suggestions for future work.

الخلاصـــة

في الوقت الحاضر أصبح السرطان أحد المواضيع المثيرة للباحثين في مجال التكنولوجيا الطبية. السرطان هو واحد من أكثر الأمراض المميتة في العالم و لأن العلاج الفعلي لهذا المرض لم يتم التوصل إليه حتى الآن فهو موضوع مثير للقلق. لا يمكن إنقاذ المرضى الذين يعانون من هذا المرض إلا إذا وجد في مرحلة مبكرة. إذا تم اكتشافه في المرحلة الأخيرة ، فستكون فرصة البقاء على قيد الحياة أقل. ولهذا السبب ، يعد التشخيص المبكر والحقيقي مشكلة مهمة ويلعب دوراً رئيسياً في علاج هذا المرض. يُعد التصنيف أحد القضايا الاساسية في مجالات اكتشاف المعرفة وعلوم القرار حيث يمثل في العثور على مصنف ينتج عن مجموعات البيانات التدريبية ذات الأهداف المحددة مسبقًا واستخدامها لتصنيف مجموعات بيانات أخرى. هناك العديد من أنواع الخوارزميات المستخدمة في بناء مصنفات للكشف عن الأورام.

في هذا العمل بنينا نظام تصنيف سرطان كفوء لزيادة الدقة وتقليل نسبة الخطأ في عملية التشخيص. هذا النظام يتضمن مرحلتين رئيسيتين هما: مرحلة ماقبل المعالجة و مرحلة تصنيف الورم. قد تحتوي مجموعات البيانات على بعض القيم المفقودة في العديد من مهام العالم الحقيقي. هذه القيم المفقودة تؤثر بشكل سلبي على اداء المصنف, ولذلك تتم معالجتها قبل عملية التصنيف. في هذا العمل تم استخدام خوارزمية (Naive Bayes) و تقنية هجينة تتضمن (Class Association Rule and Deep Neural Network).

تم اختبار النظام المقترح باستخدام مجوعتين من بيانات السرطان (سرطان الثدي و سرطان الرئة). تشير النتائج ان النظام المقترح يمتلك دقة عالية مقارنة مع الطرق الموجودة الاخرى حيث ان دقة خوارزمية Naive Bayes باستخدام مجموعة بيانات سرطان الثدي كانت 98% و باستخدام مجموعة بيانات سرطان الرئة كانت99% بينما دقة التقنية الهجينة باستخدام مجموعة بيانات سرطان الثدي كانت98.8% و باستخدام مجموعة بيانات سرطان الرئة كانت95% .