OPEN ACCESS Freely available online

Journal of Cancer Science and Research

Research Article

Analyzing the Impact of Epidermoid and Adeno Tissue on Cancer Incidence with a Data Mining Approach

Mohsen Ghorbian*

Department of Computer Engineering, Islamic Azad University, Qom, Iran

ABSTRACT

Early detection is the only way to effectively control diseases whose treatment can be challenging, expensive, and time-consuming. Identifying the influencing factors in the occurrence of disease can, therefore, reduce the time associated with diagnosis and provide a solid foundation for improving the prognosis and preventing patients' deterioration. Applying data mining techniques as a novel approach to the early detection of disease-causing agents can significantly assist the early detection. In this study, an attempt was made to investigate the effect of epidermoid and adeno tissues on the incidence of cancerous diseases such as bone, bone marrow, lung, and neck cancer by conducting a data mining process on cancer patient data sets. Hence, implementing two data mining techniques, K-Nearest Neighbour (KNN) and decision tree, on the data of patients with these four types of cancer, an attempt was made to evaluate their performance using the three criteria of accuracy, error ratio, and negative prediction value. The implementation of data mining techniques and evaluations of their performance indicates that the decision tree technique performed better with an accuracy of 89.10%, an error ratio of 14.04%, and negative prediction value of 77.71%. Also, based on the findings, contamination of epidermoid and adeno tissues does not affect the early detection of any of the four categories of bone, bone marrow, lung, or neck cancer. In other words, the infection of the two epidermoid and adeno tissues cannot be the cause of the four types of bone, bone marrow, lung, and neck cancer.

Keywords: Cancer diagnosis; Brain cancer; Lung cancer; K Nearest Neighbor (KNN); Decision tree; Data mining

INTRODUCTION

Through the use of computer-related technologies in various fields and the possibility of employing them in a wide variety of applications, it has become possible to produce and utilize more digital data by using these technologies. Additionally, these technologies provide the foundation for generating additional digital data. Hence, for digital information generated through computer technology to be reused, it must be maintained and stored [1]. Parallel to the production of digital data, experts began storing data, so they needed to handle the volume of data, which was increasing daily, and the speed of production as well. The goal of data management, storage, and reuse is to gain access to information hidden in data. By using new technologies in this field, data scientists can identify patterns within stored data to make decisions based on the knowledge contained therein. They will also develop plans based on the knowledge contained within the data [2]. They attempt to decide by gleaning secrets from outdated data and then developing plans based on that knowledge. As a result, data scientists have developed tools and technologies for data analysis. Data mining is an example of such a technology developed by combining different disciplines [3]. By analyzing data, data scientists can identify hidden patterns within the data to make informed decisions. It is achievable to extend data mining to discover knowledge and hidden patterns in data and take critical steps to improve future planning [4]. The use of data mining technology is beneficial in many fields, including medicine, because it can be done on data stored in databases. Thus, using data mining technology can analyze healthcare data [5]. Hence, in fields such as healthcare, data mining is significant for providing physicians with specialized knowledge by revealing hidden patterns and relationships in the available information [6]. The significance of data mining in the medical field can be demonstrated by the fact that doctors or medical specialists in any field have been able to use the data analysis information to accelerate the diagnosis of a difficult-to-diagnose illness, reduce patient costs, or for a wide variety of other reasons [7]. Data mining aims to analyze data in the most efficient and effective manner possible, which is why a wide range of techniques and algorithms have been developed for this purpose. When assessing techniques, it is essential to consider the potential application and structure of the existing data. Hence, it is essential to demonstrate high confidence and accuracy in the analyses presented when choosing data mining techniques [8]. The selection of data mining techniques has a special significance

Correspondence to: Mohsen Ghorbian, Department of Computer Engineering, Islamic Azad University, Qom, Iran, E-mail: ghorbian68@gmail.com

Received: 21-Apr-2023, Manuscript No. JCSR-23-23681; **Editor assigned:** 24-Apr-2023, PreQC No. JCSR-23-23681 (PQ); **Reviewed:** 08-Mar-2023, QC No. JCSR-23-23681; **Revised:** 15-Mar-2023, Manuscript No. JCSR-23-23681 (R); **Published:** 22-May-2023, DOI: 10.35248/2576-1447.23.8.538.

Citation: Ghorbian M (2023) Analyzing the Impact of Epidermoid and Adeno Tissue on Cancer Incidence with a Data Mining Approach. J Can Sci Res. 8:538.

in every field. Still, it is doubly significant in medicine because decisions must be made based on the outcomes of applying these techniques to human health since human health depends on the ability of the techniques to provide the best and most appropriate analysis that is highly accurate and confident [9]. The evaluation of techniques can assist in answering the question of which data mining method performs best and generates the most accurate and comprehensive analysis. Consequently, the related data must now be subjected to data mining techniques using the predefined criteria, and the results must be analyzed. Data mining methods also have predefined evaluation standards [10]. This study evaluated K-Nearest Neighbor (KNN) and decision tree data mining techniques for their effectiveness in analyzing information about cancer patients based on factors considered.

Related work

Since many scientists from many disciplines and sciences contributed to the development of widely used data mining techniques, this can be said that data mining is a technology that evolved by combining various disciplines. Data mining could be employed wherever there is information. However, the significance of this mechanism increases when applied to health-related data. As a result, the researchers will face several challenges at this moment. These challenges include difficulties obtaining patients' information because it is confidential and private, making its acquisition challenging. It is also necessary to maintain and be cautious about information. In addition, medical professionals do not accept this method as risky and prefer the more conventional, time-consuming, and expensive methods because they understand human health is crucial. This will be expensive, so researchers and data mining experts are working with medical professionals to research this topic. Predicting certain diseases or identifying individuals predisposed to them can increase the amount of maintaining a person's health or decrease the costs associated with the disease. In this regard, researchers rely on the findings' accuracy, reliability, and correctness. Therefore, the analysis techniques employed must also produce the desired results. It is evident that research has been conducted in this area, and the results will undoubtedly greatly assist scientists working on health-related projects. Prather, et al. [11], analyzed the clinical information of 3902 pregnant women using data mining techniques to predict the factors leading to premature birth. Bandyopadhyay, et al. [12], achieved a model by employing the Bayesian network technique on the information related to the individual characteristics of heart patients, through which they could predict the risk of cardiovascular diseases. Linden, et al. [13] used vector machine, logistic regression, and random forest techniques to implement them on health data and compare their performance to introduce a technique that has better performance for researchers can use the results to increase the accuracy of diagnosis and identify high-risk patients. Aljumah, et al. [14], attempted to predict how they could achieve effective treatment methods for diabetes by employing the regression method and its implementation on the data related to diabetic patients. Shah, et al. [15], evaluated three techniques of K-neighbors, Bayesian networks, and proximity decision tree using relative absolute error measure and kappa. Delen, et al. [16] implemented three decision tree techniques, vector machine, and artificial neural networks on a set of patient information that included 122,000 fields and 77 variables and evaluated the performance of these three techniques. The aim of his research, while introducing a technique that has better performance, has been to achieve a model through which a method for predicting prostate cancer disease can be achieved. Sarvestani, et al [17], employed techniques related to neural networks to create a model that would help researchers find a way to slow down the progression of breast cancer.

MATERIALS AND METHODS

Research methodology

Data mining achieves the desired and reliable results by applying the appropriate and reliable techniques per the available information. Therefore, it is essential to use methods that provide accurate information while using a language that is clear and simple. Although applying data mining techniques to a data set and creating a model are essential parts of the process, understanding how to interpret the results is just as important. This study employs K-Nearest Neighbor and decision tree data mining methods. A decision tree is one technique that matters in the classification area. It is one of the most straightforward and practical techniques available in data mining. Understanding and interpreting the results generated using this technique is straightforward, which can explain its popularity [18]. On the other side, the K-Nearest Neighbor (KNN) method divides data into groups that share similar characteristics. As one of the most basic data categorization methods available, this method is used to place this new sample in the group that has the most characteristics in common with the sample under consideration. It uses a new sample and the classification information from the previous data. The K-Nearest Neighbor (KNN) method aims to improve the accuracy and results of the analysis by placing the new samples in a classification close to the previous samples [19]. Different cancers may implicate different tissues or, in other words, may infect different tissues, depending on whether the cancer has tumors. There are two types of tissues involved in tumor infect, known as adeno and epidermoid tissue, and a tumor infect may occur in either of these two types of tissues. However, based on the required tests, a thorough diagnosis of this infect will be crucial, even though the associated costs will vary [20]. In this study, 269 samples were collected from patients with bone, bone marrow, lung, and neck cancers. Oncology samples from Slovenia's Ljubljana Hospital were collected and made available through the UC Irvine Machine Learning Repository. As part of the preparation process for our data, we must apply data mining techniques. This technique can be used by converting data into a format compatible with these techniques to utilize them. Considering that the classification methods employed in this study are classification methods, the input data for these data mining techniques must contain a table-like structure and format to be effective. Given the wide range of data available and the lack of need for some data, selecting the portions of the data on which the research objectives have already been specified will also be necessary. Therefore, some data which have no value will also be deleted at this point. After selecting the required data, defining them in the primary features is necessary. According to the purpose defined in this research, the mentioned features can be seen in Table 1.

Table 1: Features selection.

Properties	Value	Value
Histologic-Type (Label)	Adeno	Epidermoid
Bone	N	Y

Lung	N	Y
Bone-Marrow	N	Y
Neck	N	Y

After selecting the desired characteristics, resolving missing values in the data is necessary. For this purpose, it is possible to utilize the tools provided by data mining. Since all of the data in this study have values, there is no need to quantify non-existent values. Numerous tools can facilitate data mining. Rapid Miner is utilized to perform the data mining procedure in this study. In addition, category-based algorithms will be used during data mining. This study employs a Windows 8.1 64-bit operating system with 4 GB RAM and a 5-core processor.

Evaluation criteria

A data mining model is developed using the chosen techniques and the available data. Nevertheless, based on the objective of this research, which is to evaluate data mining methods, the model is constructed using the available data set and selected methods. An evaluation standard is utilized to compare and evaluate methods. Hence, this study considered Predicting Negative Values (PNV), accuracy, and error ratio as evaluation variables.

Negative Prediction Value (NPV)

A standard measurement criterion is the forecast of negative values. This criterion is calculated by dividing the total number of predicted true negatives by the ratio of predicted true negatives and false negative instances cases. NPV, the calculation method of this criterion is shown in the formula (1).

$$NPV = \frac{TN}{FN + TN} \quad \text{..... (1)}$$

An estimation matrix is created by applying data mining techniques to samples to determine the state of estimations. True Positive (TP) is the number of samples that are correctly detected as positive, True Negative (TN) is the number of samples that are correctly detected as negative, and False Positive (FP) is the number of samples that are incorrectly detected as positive. A False Negative (FN) is the number of samples detected as false negatives [21].

Error rate

An error rate evaluation criterion recreates the contrasting role of its counterpart, an accuracy criterion, in the evaluation process. This evaluation criterion is calculated by dividing the number of samples where the labels were incorrectly estimated by the total number of samples where the model incorrectly and correctly estimated the labels. In this manner, you can determine the percentage of samples with incorrect labels. Thus, the error rate can be computed by subtracting the precision requirement from the number one in that model. ER, the calculation method of this criterion is shown in formula (2) [22].

$$ER = 1 - ACC$$

$$ER = \frac{FP + FN}{TP + TN + FP + FN} \quad \text{..... (2)}$$

Accuracy

The precision assessment criterion is an essential and frequently used criterion for assessing models in data mining methods. This criterion aims to determine how accurately a model depicts the collected data. This evaluation criterion is calculated by dividing the number of samples where the labels were correctly estimated by the total number of samples where the model incorrectly and correctly estimated the labels. ACC, the calculation method of this criterion is shown in formula (3) [23].

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{..... (3)}$$

In this study, an attempt has been made to employ a set of significant and frequently employed evaluation criteria, such as accuracy evaluation criteria accuracy, error rate, as well as the negative prediction value of each data mining technique, K-Nearest Neighbor (KNN), and decision tree, in determining the factors that are most likely to affect the incidence of various cancers.

RESULTS AND DISCUSSION

This research considers four types of bone cancer, bone marrow, lung, and neck, as the target variables, and each of these variables is used as a label to identify the type of infected tissue. In light of the conditions considered, the implementation predicted this variable, a type of cancer. The two significant components of acceptance conditions are the confidence interval and the support value. Therefore, the confidence interval for this condition is 80%, and the support value is 50% considered. Now, if the variable that is the type of cancer can be identified with a confidence interval of 80% or higher and a support value of 50% or higher and among the available samples, at least 135 samples from each of the two epidermoid and adeno tissues, which are If the infected tissue has been predicted, then it can be said that in a particular type of cancer, a specific type of tissue is infected. Otherwise, a variable that does not meet the conditions is worthless. The effectiveness of the two data mining techniques, close neighbor and decision tree, has been evaluated in the subsequent sections using criteria for accuracy, error ratio, and negative prediction value. Consequently, the achievement results discuss two types of epidermoid and adeno tissues when diagnosing a specific type of cancer. We attempt to make a technical determination by averaging all the evaluations since a general evaluation of all the tests and evaluations is required. As displayed following, we present the best.

Bone cancer

According to the findings, neither of the two tissues, epidermoid and adeno, contributes significantly to diagnosing bone cancer; in other words, neither can be used to predict the type of infected tissue. Although neither epidermoid nor adeno tissues have been recognized as influential in diagnosing bone cancer.

According to the two data mining technique performances, the decision tree technique performs better than the K-Nearest Neighbor (KNN) technique. In other words, as shown in Figure 1, the decision tree technique performs better than the K-Nearest Neighbor (KNN) technique based on three criteria: accuracy, error rate, and negative prediction value (Figure 1).

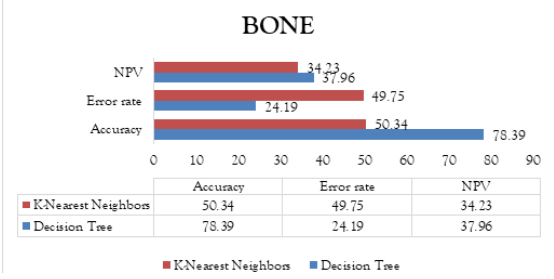

Figure 1: Performance evaluation of K-nearest neighbor techniques and decision tree with bone cancer feature.

Bone marrow cancer

According to the attainment finding, none of the epidermoid or adeno tissues have been recognized as influencing factors in predicting the type of infected tissue in diagnosing bone marrow cancer. Therefore, neither of these two types of tissues can be used as a source of infection in diagnosing bone marrow cancer.

Based on the three criteria of accuracy, error rate, and negative prediction value used to evaluate the two data mining techniques implemented in this case, K-Nearest Neighbor techniques and decision trees, it can be concluded that both methods presented similar results, as shown in Figure 2.

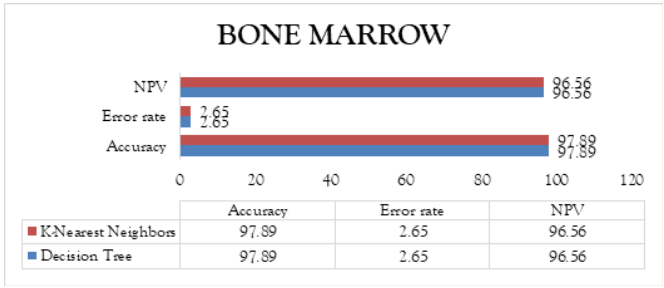

Figure 2: Performance evaluation of K-nearest neighbor techniques and decision tree with bone marrow cancer feature.

Lung cancer

According to the findings, neither epidermoid nor adeno tissue can be considered significant factors in lung cancer diagnosis. Epidermoid and adeno tissues have not been identified as influential factors, nor can they be used as a tissue that can be infected in lung cancer diagnosis.

According to the three criteria of accuracy, error rate, and negative prediction value used to evaluate the two data mining techniques implemented in this section, i.e., K-nearest neighbor and decision tree, as shown in Figure 3, the K-Nearest Neighbor technique performed better.

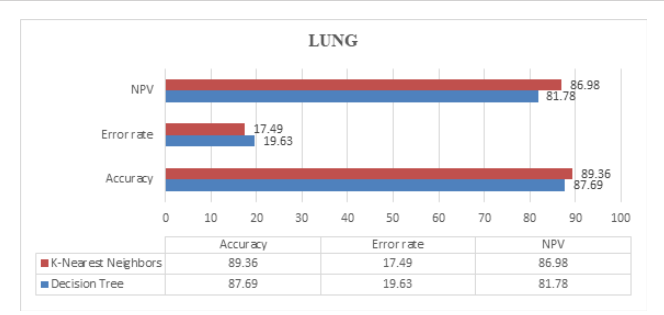

Figure 3: Performance evaluation of K-nearest neighbor techniques and decision tree with lung cancer feature.

Neck cancer

According to the findings, neither epidermoid nor adeno tissue can be considered significant factors in neck cancer diagnosis. Epidermoid and adeno tissues are not valuable indicators in neck cancer diagnosis.

In this case, the decision tree technique performs better than the K-Nearest Neighbor (KNN) data mining technique. In other words, as shown in Figure 4, the decision tree technique performs better than the K-Nearest Neighbor technique based on three criteria: accuracy, error rate, and negative prediction value.

By analyzing and analyzing the results of implementing two data mining techniques, K-Nearest Neighbor and decision tree, on cancer patient data, we determined the decision tree technique was most efficient based on its accuracy, error ratio, and negative prediction value. As a result, it is concluded that the decision tree method performs better than the K Nearest Neighbor (KNN) method. As shown in Table 2, this conclusion is based on averaging the results of the evaluations conducted at various stages (Figure 4).

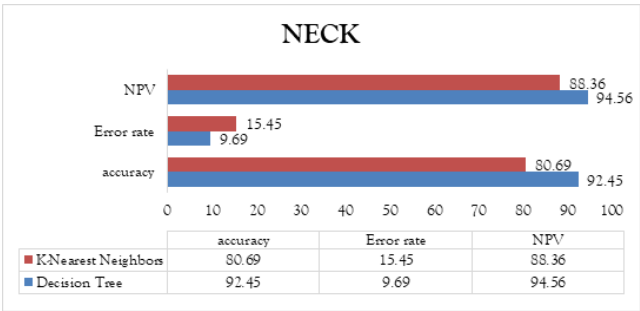

Figure 4: Performance evaluation of K-nearest neighbor techniques and decision tree with neck cancer feature.

Table 2: Final comparison of the performance of K-nearest neighbor and decision tree.

Accuracy	Error rate	NPV	Data mining techniques
89.1	14.04	77.71	Decision tree
79.57	21.33	76.53	K-nearest neighbor

The results of all the evaluations indicate that decision tree perform better than K-Nearest Neighbor, as shown in Figure 5.

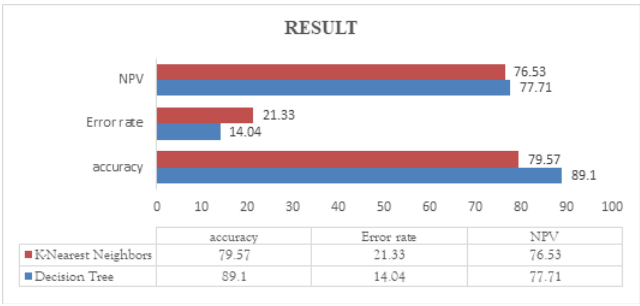

Figure 5: Analyzing the performance of K-nearest neighbor and decision tree.

CONCLUSION

Applying data mining algorithms to sensitive data, such as healthcare data, can significantly assist in diagnosing diseases. Early diagnosis of any disease and detection of influential factors can provide the foundation for the speedy recovery of those affected. In this research, an attempt was made to implement data mining techniques such as the K Nearest Neighbor (KNN) and the decision tree on a dataset containing information about individuals with cancer. In assessing each technique's performance, three factors were considered: Accuracy, error ratio, and the ability to negative prediction value. Hence, using the two aforementioned data mining techniques, it would be attempted to determine whether two essential factors, epidermoid and adeno, play a role in the occurrence of bone, bone marrow, lung, and neck malignancies

by analyzing the data of cancer patients. Therefore, based on the results obtained from the evaluation, we found that the data mining method based on decision trees outperformed the K Nearest Neighbor (KNN) technique with 89.10% accuracy, 14.04% error ratio, and 77.71% negative prediction value. In addition, by analyzing the obtained results, epidermoid and adeno factors were not-effective as associated with bone, bone marrow, lung, and neck cancers. Thus, they can be considered non-effective factors in the selection process.

REFERENCES

1. Jun Lee S, Siau K. A review of data mining techniques. Ind Manag Data Syst. 2001;101(1):41-46.
2. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996;17(3):37.
3. Hand DJ. Data Mining. Encyclopedia of Environmetrics. 2006.
4. Chen MS, Han J, Yu PS. Data mining: An overview from a database perspective. IEEE Trans Knowl Data Eng. 1996;8(6):866-883.
5. Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. IEEE Trans Knowl Data Eng. 2013;26(1):97-107.
6. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. In2008 IEEE/ACS AICCSA. 2008(108-115).
7. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. Int J Med Inform. 2008;77(2):81-97.
8. Wu X, Kumar V, editors. The top ten algorithms in data mining. CRC press. 2009.
9. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. Artifi Intell Med. 2005;34(2):113-127.
10. Zurada J, Lonial S. Comparison of the performance of several data mining methods for bad debt recovery in the healthcare industry. Journal of Applied Business Research (JABR). 2005.
11. Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE et al. Medical data mining: Knowledge discovery in a clinical data warehouse. Proc AMIA Annu Fall Symp. 1997 (101).
12. Bandyopadhyay S, Wolfson J, Vock DM, Vazquez-Benitez G, Adomavicius G, Elidrisi M, et al. Data mining for censored time-to-event data: A bayesian network model for predicting cardiovascular risk from electronic health record data. Data Mining and Knowledge Discovery. 2015;29:1033-1069.
13. Linden A, Yarnold PR. Using data mining techniques to characterize participation in observational studies. J Eval Clin Pract. 2016;22(6):839-847.
14. Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old Patients. J King Saud Univ. - Comput. Inf. 2013;25(2):127-1236.
15. Shah C, Jivani AG. Comparison of data mining classification algorithms for breast cancer prediction. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). 2013(1-4). IEEE.
16. Delen D. Analysis of cancer data: A data mining approach. Expert Systems. 2009;26(1):100-112.
17. Sarvestani AS, Safavi AA, Parandeh NM, Salehi M. Predicting breast cancer survivability using data mining techniques. In 2010 2nd International Conference on Software Technology and Engineering .2010(Vol. 2, pp. V2-227). IEEE.
18. Apté C, Weiss S. Data mining with decision trees and decision rules. Future Gener Comput Syst. 1997 Nov 1;13(2-3):197-210.
19. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, 2003. Proceedings 2003 (986-996). Springer Berlin Heidelberg.
20. Auerbach O, Garfinkel L, Parks VR. Histologic type of lung cancer in relation to smoking habits, year of diagnosis and sites of metastases. Chest. 1975;67(4):382-387.
21. Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case–control designs. Biostatistics. 2009;10(1):94-105.
22. Sharma A, Kaur B. A research review on comparative analysis of data mining tools, techniques and parameters. Int J Adv Comput. Res. 2017.
23. Tapak L, Mahjub H, Hamidi O, Poorolajal J. Real-data comparison of data mining methods in prediction of diabetes in Iran. Healthc Inform Res 2013;19(3):177-185.