DİCLE ÜNİVERSİTESİ
İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ
DERGİSİ
Dicle University
Journal of Economics and Administrative Sciences
E-ISSN: 2587-0106

Cilt / Volume: 14, Sayı / Issue: 28, Sayfalar / Pages: 842-855

Araştırma Makalesi / Original Article

Received / Alınma: 02.06.2024 Accepted / Kabul: 22.07.2024

PREDICTING SURVIVAL LIMITATION BY MACHINE LEARNING IN PATIENT WITH CANCER

Cuma ÇAKMAK[1]
Fadime ÇINAR[2]
Mehmet Aziz ÇAKMAK[3]

Abstract

Cancer is an important public health problem, ranking second in terms of burden of disease in the United States and ranking first in the global burden of disease in the world. Cancer, which causes significant mortality and morbidity, is affected by many factors. Researchers are increasingly interested in this field, both in examining the factors that cause the disease and in managing the disease and are conducting research on this disease with new treatment methods, new techniques and technologies. In this study, its aimed to determine survival rates by analysing open access cancer data representing 8.3% of the US population. With the data obtained, it was tried to classify the survival of cancer patients. Within the scope of the research, various confidence levels were obtained with decision trees, random forrest, SVM algorithms, which are data mining tools. The highest confidence level was obtained with the random forrest algorithm with 75.3%. As a result, it was seen that the model was meaningful and usable, and that survival classification could be made with the data obtained. Survival classification can be an important element for health service providers in resource allocation and effective care.

Keywords: KNIME, Data Mining, Cancer, Countries Database, U.S.

Jel Codes: I10, I19, Z19.

[1]Dr. Öğretim Üyesi, Dicle Üniversitesi Sağlık Yönetimi Bölümü, E-posta: cuma.cakmak@dicle.edu.tr, ORCID: 0000-0002-4409-9669.

[2]Doç. Dr., Nişantaşı Üniversitesi Hemşirelik Bölümü, E-posta: fadime.cinar@nisantasi.edu.tr, ORCID: 0000-0002-9017-4105

[3]Arş. Gör., Nişantaşı Üniversitesi Hemşirelik Bölümü, E-posta: mehmetaziz.cakmak@nisantasi.edu.tr, ORCID: 0000–0002-5040–5642

Atıf/Citation
Çakmak, C., Çınar, F., & Çakmak, M. A. (2024). Predicting survival limitation by machine learning in patient with cancer. *Dicle Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14*(28), 842-855.

KANSERLİ HASTALARDA MAKİNE ÖĞRENİMİ İLE SAĞKALIM ORANININ TAHMİN EDİLMESİ

Öz

Kanser önemli bir halk sağlığı sorunu olmakla birlikte ABD'de hastalık yükü açısından ikinci sırada yer almakta dünyada ise küresel hastalık yükü sıralamasında ilk sıralarda yer alabilmektedir. Önemli oranda mortalite ve morbiditeye neden olan kanser hastalığı birçok faktörden etkilenmektedir. Gerek hastalığa neden olan faktörlerin incelenmesi gerek hastalığın yönetilmesi konusunda araştırmacılar giderek artan oranda bu alanla ilgilenmekte yeni tedavi yöntemleri, yeni teknikler ve teknolojiler ile bu hastalık üzerinde araştırmalar yapmaktadırlar. Bu çalışmada ABD toplumun %8,3'ünü temsil eden ve açık erişimli olarak ulaşılabilen kanser verileri analiz edilerek sağkalım oranlarını tespit etmek amaçlanmıştır. Araştırmada veri madenciliği yöntemlerinden biri olan Konstanz Information Miner (KNIME) programı kullanılmıştır. Elde edilen veriler ile kanser hastalarının sağkalımları sınıflandırılmaya çalışılmıştır. Araştırma kapsamında veri madenciliği araçları olan karar ağaçları, random forrest, Destek Vektör Makineleri (Support Vector Machine-SVM) algoritmaları ile çeşitli güven düzeyleri elde edilmiştir. En yüksek güven düzeyi %75,3 ile random forrest algoritması ile elde edilmiştir. Sonuç olarak modelin anlamlı ve kullanılabilir olduğu ve elde edilen veriler ile sağkalım sınıflandırılmasının yapılabildiği görülmüştür. Sağkalım sınıflandırması kaynak tahsisinde ve etkili bakım konusunda sağlık hizmet sunucuları için önemli bir unsur olabilir.

Anahtar Kelimeler: KNIME, Veri Madenciliği, Kanser, Ülke Veritabanı, ABD.

Jel Kodları: I10, I19, Z19

1. INTRODUCTION

Cancer is a broad group of diseases that can start in almost any organ or tissue of the body when abnormal cells grow uncontrollably, grow beyond their usual boundaries, invade adjacent parts of the body and/or spread to other organs. The latter process is called metastasis and is a major cause of death from cancer. A neoplasm and malignant tumor are other common names for cancer (World Health Organization, 2024). Cancer is one of the deadliest diseases of the 21st century, causing a large number of deaths each year. Disease variations in different parts of the world, the impact of available medical facilities and other socio-economic factors significantly affect the proper management of this disease (Chhikara & Parang, 2022). Cancer is an important public health problem worldwide and is the second leading cause of death, especially in the United States of America. In 2023, 1,958,310 new cancer cases and 609,820 cancer deaths are predicted to occur in the United States. Cancer incidence increased by 3% annually from 2014 to 2019 to 99,000 new cases for prostate cancer after two decades of decline, but otherwise incidence trends are more favorable in men than in women (Siegel et al., 2023).

Cancer is the second leading cause of death worldwide, accounting for an estimated 9.6 million deaths in 2018, or 1 in every 6 deaths. The most common types of cancer in men are lung, prostate, colorectal, stomach and liver cancer, while in women it is breast, colorectal, lung, cervical and thyroid cancer. The burden of cancer continues to rise globally, placing enormous physical, emotional and financial pressure on individuals, families, communities and health systems. Many health systems in low- and middle-income countries are the least prepared to manage this burden, and too many cancer patients worldwide lack access to timely quality diagnosis and treatment. In countries with strong health systems, survival rates for many types of cancer are increasing due to accessible early detection, quality treatment and survivorship care (World Health Organization, 2024). Cancer is classified as a disease that needs to be prioritized globally in terms of reducing morbidity and mortality that may occur due to the fact that cancer is a major leading disease in the world and reduces life expectancy (Sullivan et al., 2011; Bray et al., 2018).

With the rapid development of computer software/hardware and internet technology, the amount of data has increased at an incredible rate. As an abstract concept, "big data" currently affects all areas of life (Herland et al., 2024). As in all areas of life, a large amount of data is produced in the health sector and it is important to analyze the data produced and transform them into meaningful information. It is seen that data analysis techniques have not yet become widespread in medical research (Bellazzi & Zupan, 2008) Therefore, computer scientists have made important contributions with big data applications in analyzing such dense data and have introduced the concept of data mining to solve the difficulties related to such applications. Data mining (also known as knowledge discovery in databases) refers to the process of extracting potentially useful information and knowledge hidden in large amounts of incomplete, noisy, fuzzy and random practical application data (Sahu et al., 2011).

There are various studies on cancer data mining in the literature. Although cancer research has traditionally been clinical and biological in nature, data-driven analytical studies have become widespread in recent years. In a study using three popular data mining techniques (decision trees, artificial neural networks and support vector machines), 120,000 records and 77 variables to develop prostate cancer survival prediction models, k-fold cross-validation methodology was used for model building, evaluation and comparison. As a result, it was revealed that support vector machines were the most accurate predictor for this field (with 92.85% test set accuracy), followed by artificial neural networks and decision trees (Delen, 2009). Another study focusing on breast cancer patients with data mining also focused on

current research using data mining techniques to improve breast cancer diagnosis and prognosis (Karya, 2012). Similar studies and data mining applications for different cancer types can be found in the literature (Li et al., 2004; Delen et al., 2005; Liou & Vhang, 2015).

In this study, it is aimed to predict survival with machine learning algorithms using KNIME program of cancer data provided by the US National Cancer Institute as open access, obtained by machine learning method.

2. MATERIALS AND METHODS

The materials and methods used within the scope of the research are tried to be explained in the form of sub-headings. Such a way was followed due to the complexity of the method used.

Type of this research

The study is descriptive and retrospective. The analytical method used is based on knowledge of the literature and machine learning techniques whose effectiveness is fixed.

Population and sample of the study

The population of the study consists of the annual statistics of patients diagnosed with cancer between 2000 and 2020 in the open-access SEER (The Surveillance, Epidemiology, and End Results) database, which meets the inclusion criteria for the study. When the SEER database is examined, it is seen that it is a database that is updated annually and therefore the data increases cumulatively. As stated in the official guideline of this database, it is reported that 16,683,417 (the total number of cases) have been included in the database so far 14 (SEER Database Guideline, 2024). According to this database, this represents 8.3% of the US population. In this study, no sample selection was made, and the entire population of the entire study constituted the sample of the study. Inclusion criteria;

- Data belonging to the period between 2000 and 2020,
- Availability of data,
- Data should have parameters such as incidence, mortality rate, cancer type, age,
- Applicability for the selected model,

Data collection

The data used in the study consist of data made openly available on the official web page of the US National Cancer Institute (SEER Offical Website, 2024). In terms of the structure of the data recorded on cancer, the data were presented in a distributed manner rather than in related tables. The data in the database were filtered and presented through the website. For example, the opportunity to analyze all cancer types or as a sub-heading was provided in line with the request of the individuals accessing the database. Even in the form of sub-parameters or total parameters, all data are associated with the parameters "SEER Incidence", "U.S. Mortality", "Incidence and Mortality Comparison", "Survival", "Prevalence", "Risk of Diagnosis/Dying". Although this situation seems like an advantage, it makes it difficult to identify meaningful data. This difficulty has been transformed into a problem question in accordance with the nature of data mining and evaluated within the scope of this study.

Used method for data mining

CRISP-DM (Cross Industry Standard Process for Data Mining) is a widely used methodology for data mining projects. This methodology defines the different phases of a project and aims to manage these phases in a sequential manner. CRISP-DM provides a framework for initiating, managing and finalizing data mining projects (Wirth et al., 2000).

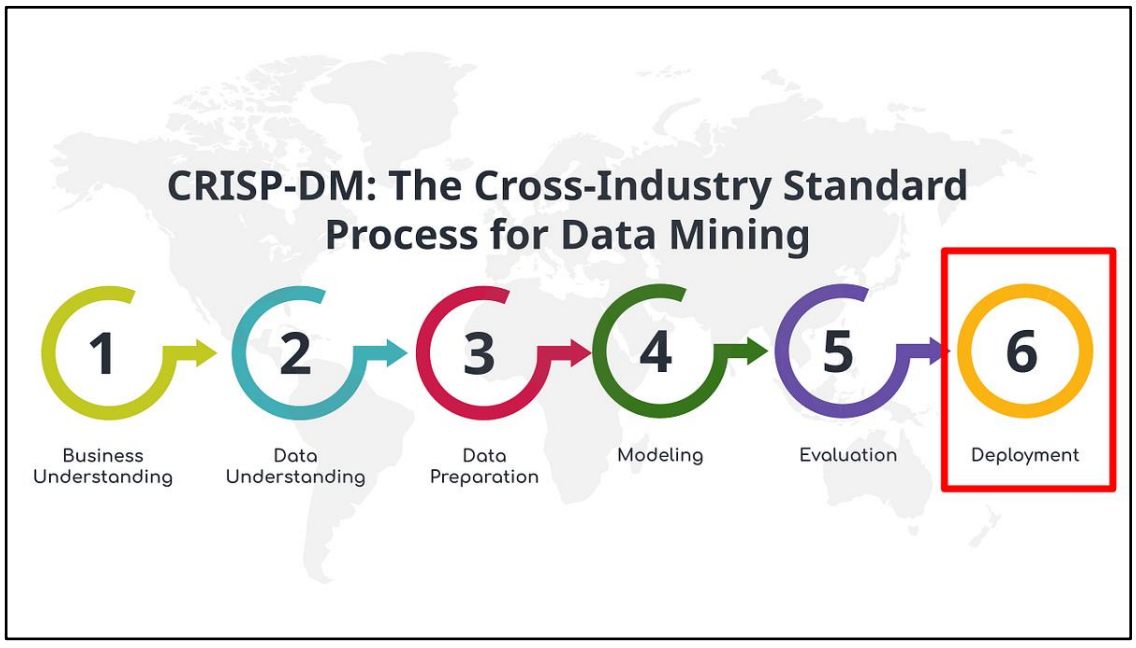

Figure 1. CRISP DM Methodology

Source: Medium

Software

KNIME (Konstanz Information Miner) is a visual programming tool and a data analysis platform. KNIME is an open-source platform for performing various data science tasks such as data analysis, machine learning, artificial intelligence and big data processing. Users can design data analysis workflows using a visual interface and build these workflows from many different stages such as data preprocessing, analysis, modeling, and result visualization (KNIME Software, 2024).

Algorithms

Decission Three

Decision Tree is a classification method frequently used in modeling such as data mining and machine learning. It is an important method used in areas such as scientific discovery, banking applications, sales forecasting and marketing (Patel & Rana, 2014). In medical fields, making the right decisions based on correctly observed data is more important than in other disciplines. Decision trees are highly usable in the medical field, provide high classification accuracy according to the effectiveness of the model developed, and have high reliability. In this respect, decision tree has been used in medical decision making (Vens et al., 2008).

Random Forrest

Random Forests, or in other words, random decision trees, is a machine learning method that performs optimal classification by creating more than one decision tree in the modeling phase. In this study, the random decision tree method was applied to obtain a higher reliability rate than that obtained with decision trees. The reliability rate was calculated according to the estimator criterion (estimator) and the optimal criterion was selected. An attempt was made to prevent overlearning or underlearning of the model (Charbuty & Abdulazeez, 2021).

SVM

Support Vector Machines (SVM) is a powerful machine learning method used to solve classification and regression problems. SVM is particularly effective for processing non-linear datasets and generally performs well in high-dimensional feature spaces. The main goal of SVM is to optimally separate data points by creating a decision boundary between the identified classes. This decision boundary is a hyperplane that passes through the closest points (support vectors) of the data points. One of the advantages of SVM is that it can handle

both linear and non-linear separable datasets. This flexibility allows SVM to be used in a variety of application areas, for example, it is often preferred in biomedical imaging, biological data analysis, text classification and image recognition. The training of SVM may require complex mathematical calculations during the solution process, and the computational cost may increase when working with large datasets. However, the high accuracy, reliability and overall performance of SVM make it preferred in many application areas (Zhang & O'Donnell 2020; Charbuty & Abdulazeez, 2021).

Flowchart

The methodological design designed within the scope of the study is designed as a workflow to the extent allowed by the KNIME program. The design is shown in figure 2.

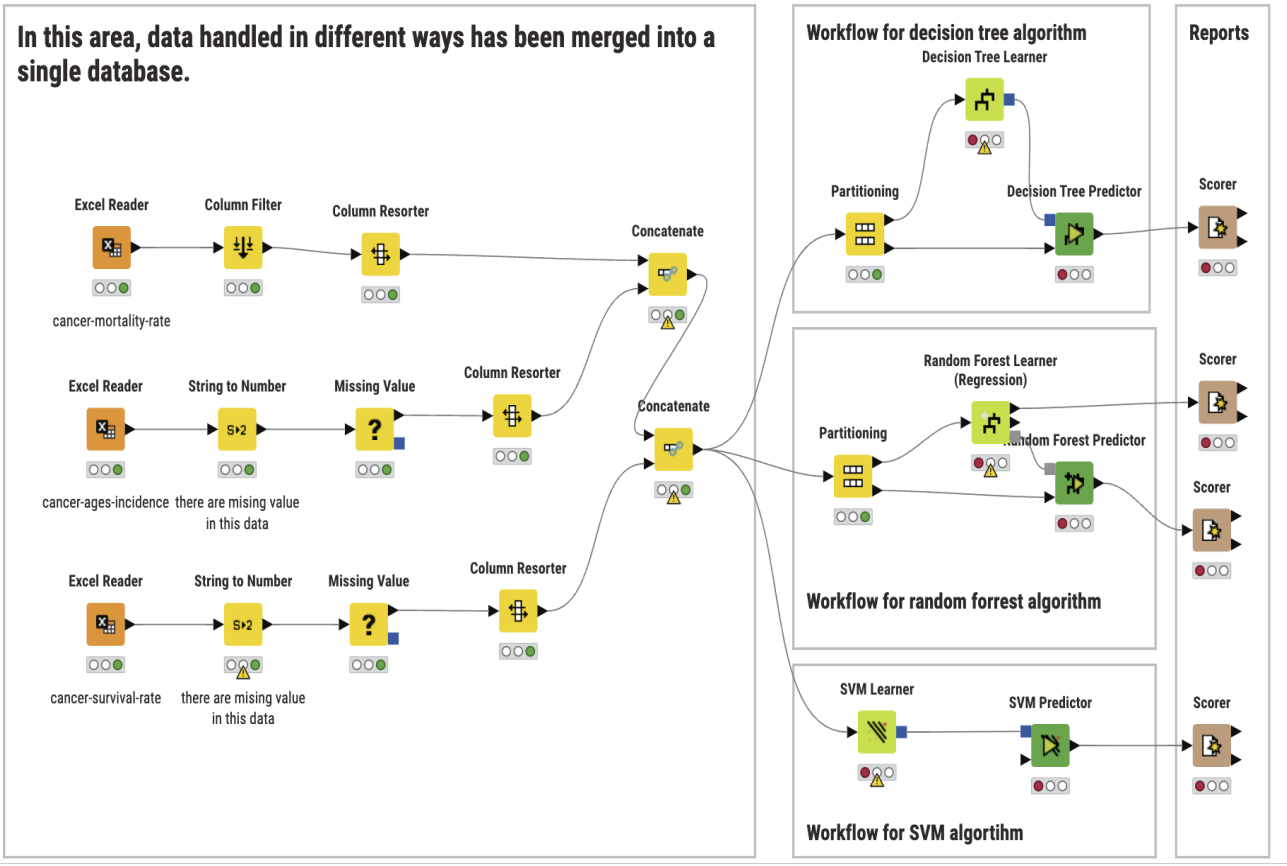

Figure 2. Workflow Created within the Scope of the Study by authors

Metrics for Appropriate Model Selection

Accuracy Calculation

- True Positive (TP): True positive, correctly predicted positive values
- False Positive (FP): False positives, values that cannot be predicted correctly

- True Negative (TN): True negative, correctly estimated negative values.
- False Negative (FN): False negative, incorrectly predicted negative values.

Accuracy/Reliability: The accuracy of a test can be measured by its ability to accurately distinguish between diseased and healthy cases. To estimate reliability, the proportion of true positives and true negatives in all cases evaluated must be calculated. Mathematically, it can be expressed as follows;

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

Sensitivity: The sensitivity of a test is its ability to accurately identify sick cases. To estimate this, the true positive rate in sick cases must be calculated. Mathematically, it can be expressed as follows;

$$Sens. = \frac{(TP)}{(TP + FN)} \quad (2)$$

Specificity: The specificity of a test is its ability to accurately identify healthy cases. To estimate this value, the true negative rate in healthy cases must be calculated. Mathematically, it can be expressed as follows;

$$Spec. = \frac{(TP)}{(TP + TN)} \quad (3)$$

Limitation

Healthcare is a science where errors should be minimal due to its nature. Machine learning, data mining and all innovative methods in the artificial intelligence cluster are used in almost all fields, especially in healthcare. Although useful information can be extracted for healthcare and cumbersome systems can be automated, it is worth remembering some important limitations. When evaluating all other innovative artificial intelligence research similar to this study, the generalizability of the data set, whether the data set consists of sufficient parameters, and whether the artificial intelligence method used is applied correctly are extremely important. In this direction, the limitations of our study can be listed as follows;

- All values in the data set characterize the USA. In this respect, it is not possible to interpret the results globally,

- As a retrospective study, data between 2000 and 2020 were used.

Ethical Approval

Ethical approval was not required since the data used within the scope of the study were shared through an open access database and did not have any experimental concept.

3. RESULTS

When the data evaluated within the scope of the study were collected in the form of different segments from different databases, the data in question were completed with average values since the incorrect and missing values were less than 1% in the pre-processing phase of the data in question.

Table 1. Descriptive table

Year	All Cancer Type	Total Mortality (All Ages).	Total Incidence (All ages)	1-year realative Surival Rate (Total)	3-year realative Surival Rate (Total)	5-year realative Surival Rate (Total).	10-year realative Surival Rate (Total).
2000	353,3	161,6	415,9	81,4	71,4	67,3	62,3
2001	394,2	248,1	506,5	74,1	61,1	56,1	50,1
2002	396,7	169,2	456,6	80,7	70,4	66,2	61
2003	399,3	244,3	513,1	74,6	61,8	56,8	50,9
2004	399,6	173,7	459,6	80,3	69,8	65,6	60,4
2005	402,5	218,9	496,5	77,2	65,4	60,7	55
2006	403	401,5	971,4	156,8	134,2	125,1	114,3
2007	403,8	240,8	508,1	75,1	62,6	57,6	51,7
2008	404,1	181,2	459,7	79,6	68,8	64,4	59,1
2009	404,5	229	504,2	76,2	64	59,2	53,4
2010	405,4	178,4	462,3	80	69,3	65	59,8

2011	405,9	224,8	499,3	76,7	64,7	59,9	54,2
2012	406,3	189,9	466,8	79,1	68,1	63,6	58,2
2013	406,9	202,9	484,5	78,7	67,5	63,1	57,6
2014	407,3	186,6	461,8	79,2	68,2	63,8	58,4
2015	408,1	209,4	499,4	78,1	66,7	62,1	56,6
2016	408,3	216,2	507,1	77,6	66	61,4	55,8
2017	409,5	199,2	481,9	78,8	67,7	63,2	57,8
2018	409,7	194,3	473	79	67,9	63,4	58
2019	415,1	205,4	501,2	78,6	67,4	62,9	57,4
Median	**402,17**	**213,77**	**506,445**	**82,09**	**70,15**	**65,37**	**59,61**

***All indicators are numbers per 100,000**

Table 1 provides details on the incidence of all cancer types, mortality rates by age, incidence, survival rate in one year, survival rate in 3 years, survival rate in 5 years and finally survival rate in 10 years between 2000 and 2019. When the data in the table are analyzed, it is seen that the incidence of the disease has increased over the years. It is seen that death rates have increased at all ages, incidence rates vary by years, but in general there is an increasing trend. Survival rates at 10 years are lower than survival rates at 5, 3 and 1 year.

Table 2. Models Reliability Ratios

ML Algorythm	**Acc**	**Sens.**	**Spec.**	**Detail**
Decission Tree	73.4	73,2	73.2	The dataset is divided into 70% training and 30% test data.
Randonm Forest	75.3	72.4	74.3	The dataset is divided into 70% training and 30% test data.
SVM	71.3	72.2	72.5	The dataset is divided into 70% training and 30% test data.

SVM is a widely used machine learning algorithm for classification and regression problems. In healthcare, it has been used in many areas such as disease diagnosis, medical image processing and genetic data analysis. In particular, SVM is frequently used in cancer diagnosis and screening tests. Artificial neural networks (ANN) are a mathematical modeling of biological neural networks. ANNs are widely used in healthcare, especially in areas such as medical image analysis, EEG signal analysis and patient monitoring. Combined with deep learning techniques, they can be effective in solving more complex health problems. Decision trees and random forests are among the popular algorithms used for classification and regression problems. In healthcare, they are used in many areas such as disease diagnosis, risk factor analysis and determining treatment options. In particular, they can be effective in analyzing clinical data and patient management. KNN algorithm is a simple and effective algorithm used for classification and regression problems. In healthcare, it is especially used in areas such as patient similarity analysis, disease diagnosis and prediction of drug interactions.

4. DISCUSSION AND CONCLUSION

Cancer, which is a global public health problem and ranks first in the world's global disease burden ranking, is one of the leading diseases that cause mortality and morbidity and maintains its importance today. Studies on cancer show that cancer can be caused by many factors. In this case, it is important to address the cancer phenomenon at a multidisciplinary level both in terms of understanding the factors that cause cancer and in managing the disease.

When the literature is examined, it is seen that there are different types and methodologies of research on cancer. Especially with digital opportunities, big data has been interpreted more accurately and has provided valuable findings in terms of cancer management. When the researches are examined, prediction studies blended with machine learning, deep learning and artificial intelligence concepts are frequently encountered in relation to the cancer phenomenon. When these studies are analyzed, it is understood that forecasting studies are conducted with data from different regions and with data of different qualities. For example, in a study conducted by Yue et al. (2018), data of 699 breast cancer patients were used with machine learning algorithms. In another study, Asri et al. (2016) used the Wisconsin Breast Cancer dataset from the UCI Machine Learning Repository and this dataset includes data from 699 breast cancer patients.

Predictions made with data obtained from different regions caused changes in the results. For example, Osman (2017) used data from Irvin, USA. In the study by Tan et al. (2009), data from China were used. When the aforementioned studies are compared, it is understood that local estimation studies were conducted, and therefore, results that can be generalized to the whole society could not be obtained. In this study, data from the US society were used. One of the most important reasons for this use is that the data are both representative of 8.3% of the US population and open access data.

When the studies are analyzed, it is understood that the methodologies used at the estimation level also vary. For example, Asri et al. (2016) analyzed the data of 699 breast cancer patients using SVM algorithm and achieved a reliability rate of 97.13%. In another study, Osman (2017) obtained a reliability level of 99.10% with the two-stage SVM method. In another study conducted by Tan et al. (2009) with 122 lung cancer patients using Adaboost machine learning method, a maximum confidence level of 95.7% was achieved. On the other hand, according to the literature cited by Asri et al. Y and Dr. Sivaprakasam stated that these authors achieved 69.23% accuracy in their studies using the decision tree algorithm in breast cancer patients. Based on the highest reliability rate obtained in this study and the method used, it is seen that the results are in parallel with the literature information.

This study emphasizes the need for a multidisciplinary approach to cancer and demonstrates that various research methodologies are used to understand and manage the causes of cancer. It is noted that there are variations in the results of estimates made with data from different regions in the literature and it is pointed out that this study was conducted with data from the US population. When the studies are analyzed, it is stated that the methodologies used at the estimation level vary and different reliability levels are reached. In particular, it was emphasized that the study conducted by Asri et al. reached a reliability rate of 97.13% using SVM algorithm and Osman (2017) reached a reliability level of 99.10% with the two-stage SVM method. In addition, Tan et al. stated that 95.7% confidence level was achieved with the Adaboost machine learning method.

Based on the results of this study, it is stated that the results are in line with the literature and the highest reliability rate is achieved according to the methodology used. These results show that machine learning and artificial intelligence techniques have significant potential in cancer prediction studies. This study emphasizes the importance of using different data sources and methodologies in the fight against cancer and sheds light on future research. Furthermore, the

results of this study suggest that more multidisciplinary studies are needed to develop more effective prevention, diagnosis and treatment strategies for cancer.

REFERANCES

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics, 77(2), 81-97.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6), 394-424.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(1), 20-28.

Chhikara, B. S., & Parang, K. (2023). Global cancer statistics 2022: The trends projection analysis. Chemical Biology Letters, 10(1), 451-451.

Delen, D. (2009). Analysis of cancer data: A data mining approach. Expert Systems, 26(1), 100-112.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 34(2), 113-127.

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. Journal of Big Data, 1, 1-35.

Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. arXiv preprint arXiv:1205.1923.

KNIME Software. (2024). Retrieved February 29, 2024, from https://www.knime.com/

Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., ... & Clark, R. A. (2004). Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine, 32(2), 71-83.

Liou, D. M., & Chang, W. P. (2015). Applying data mining for the analysis of breast cancer data. In C. Perner (Ed.), Data Mining in Clinical Medicine (pp. 175-189). Springer.

Meng, T., Jing, X., Yan, Z., & Pedrycz, W. (2020). A survey on machine learning for data fusion. Information Fusion, 57, 115-129.

Osman, A. H. (2017). An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. International Journal of Advanced Computer Science and Applications, 8(4).

Patel, B. R., & Rana, K. K. (2014). A survey on decision tree algorithm for classification. International Journal of Engineering Development and Research, 2(1), 1-5.

Sahu, H., Shrma, S., & Gondhalakar, S. (2011). A brief overview on data mining survey. International Journal of Computer Technology and Electronics Engineering, 1(3), 114-121.

SEER Database Guideline. (2024). SEER*Stat Databases: SEER November 2022 Submission. Retrieved February 27, 2024, from https://seer.cancer.gov/data-software/documentation/seerstat/nov2022/

SEER Official Website. (2024). Retrieved February 29, 2024, from https://seer.cancer.gov/

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(1), 17-48.

Sullivan, R., Peppercorn, J., Sikora, K., Zalcberg, J., Meropol, N. J., Amir, E., ... & Aapro, M. (2011). Delivering affordable cancer care in high-income countries. The Lancet Oncology, 12(10), 933-980.

Tan, C., Chen, H., & Xia, C. (2009). Early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm. Journal of Pharmaceutical and Biomedical Analysis, 49(3), 746-752.

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. Machine Learning, 73, 185-214.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (Vol. 1, pp. 29-39).

World Health Organization. (2024). Cancer. Retrieved February 14, 2024, from https://www.who.int/health-topics/cancer

Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.

Zhang, F., & O'Donnell, L. J. (2020). Support vector regression. In M. J. Er (Ed.), Machine Learning (pp. 123-140). Academic Press..