JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 37 – 44 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3305
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

IMPLEMENTATION OF K-NEAREST NEIGHBOR ALGORITHM FOR CLASSIFICATION OF LUNG CANCER CAUSES

Hanindiya Putri Almeyda[1*], Zidan Fathannul Khoiri[1], M. Sabirin Haris[2], Nabilah Husen Alkaff[1], Sukmadiningtyas[1]
[1]Sistem Informasi, Telkom University Purwokerto
[2]Sistem Informasi, Universitas Hamzanwadi
email: *hanindiyaputri@student.telkomuniversity.ac.id

Abstract: Lung cancer is most deadly cancers in the world. Identification and classification of the causes of understanding lung cancer is essential for developing more effective prevention and treatment strategies. The issue is that a lot of individuals are unaware about the characteristics and causes of lung cancer. The purpose of this study is to apply the K-Nearest Neighbor (K-NN) algorithm in the classification of the causes of lung cancer and provide education to the public must be aware of the traits of lung cancer patients and, to stay away from the causes of lung cancer. The dataset used consists of 309 samples with 16 relevant attributes. The K-NN algorithm was trained and tested to assess its ability to classify the factors that cause lung cancer. The results showed an accuracy of 90.32%, with a precision for the "YES" class of 96% and the "NO" class of 67%. The recall value for the "YES" class was 92% and for the "NO" class was 80%. The implementation of this algorithm gives good results in classification and can help in early detection and prevention of lung cancer which can be used in the development of more effective prevention and early diagnosis strategies.

Keywords: lung cancer; k-nearest neighbor; classification; machine learning

Abstrak: Kanker paru-paru tergolong jenis penyakit kanker yang memperoleh angka kematian paling tinggi di dunia. Identifikasi dan klasifikasi penyebab kanker paru-paru sangat penting untuk pengembangan strategi pencegahan dan pengobatan yang lebih efektif. Masalah yang terjadi adalah banyak orang yang belum mengetahui tentang ciri-ciri dan penyebab-penyebab dari kangker paru tersebut. Tujuan penelitian ini adalah mengimplementasikan algoritma *K-Nearest Neighbor (K-NN)* dalam klasifikasi penyebab kanker paru-paru serta memberikan edukasi kepada masyarakat banyak agar mengetahui ciri-ciri orang yang mengidap kangker paru-paru dan tentunya untuk menghindari penyebab-penyebab dari kangker paru-paru tersebut. Dataset yang digunakan terdiri dari 309 sampel dengan 16 atribut yang relevan. Algoritma K-NN kemudian dilatih dan diuji untuk menilai kemampuannya dalam mengklasifikasikan faktor-faktor penyebab kanker paru-paru. Hasil penelitian menunjukkan akurasi sebesar 90.32%, dengan skor precision untuk kelas "YES" sebesar 96% dan kelas "NO" sebesar 67%. Nilai recall untuk kelas "YES" adalah 92% dan untuk kelas "NO" sebesar 80%. Implementasi algoritma ini memberikan hasil yang baik dalam klasifikasi dan dapat membantu dalam deteksi dini serta pencegahan kanker paru-paru yang dapat digunakan dalam pengembangan strategi pencegahan dan diagnosis dini yang lebih efektif.

Kata kunci: kanker paru-paru; k-nearest neighbor; klasifikasi; machine learning

INTRODUCTION

Lung cancer is a type of cancer that has the highest death rate in the world [1]. The disease can occur if the growth of abnormal cells is also uncontrolled in one or both lungs. Many variables can cause this disease, such as lifestyle, environment, and genetics [2]. The problem is that many people do not know about the characteristics and causes of lung cancer. Identification and classification of the causes of lung cancer are essential for the development of more effective prevention and treatment strategies. In recent years, the use of machine learning techniques has become an increasingly popular approach in the medical field, including for disease classification [3].

The K-Nearest Neighbor (K-NN) algorithm is one of the most widely used algorithms for classification purposes [4]. The main advantages of K-NN are its simplicity in implementation and its ability to handle irregular data [5].

Several studies have shown that K-NN can be used effectively to classify various types of cancer, including lung cancer. In a study by [6], the K-NN algorithm is applied for early detection of early stage lung cancer through medical data analysis with accuracy results reaching 99.50%, which means that the K-NN algorithm is able to effectively develop a lung cancer prediction model.

Other research by [7] stated that the K-NN algorithm implemented for breast cancer detection achieved high accuracy, namely above 80%, which was proven to be effective for use in breast cancer detection. In line with previous studies, a study by [8] stated that data mining techniques, especially using the K-NN algorithm, have much higher accuracy compared to the Naïve Bayes algorithm for prostate cancer classification. The K-NN algorithm gets an accuracy score of 90% while the Naïve Bayes algorithm only gets an accuracy score of 80%.

This study aims to implement the K-NN algorithm in the classification of lung cancer causes and provide education to the public so that they know the characteristics of people with lung cancer and of course to avoid the causes of lung cancer. Using the dataset that has been collected, the K-NN algorithm will be trained and tested to assess its ability to classify the factors that cause lung cancer. The results of this study are intended to play a role in further awareness of the causes of lung cancer and offer an approach that can be used in early diagnosis and the development of better prevention strategies.

Data mining is the process of finding new information by finding certain patterns in very large amounts of data [9]. This technique involves using machine learning algorithms to identify trends, patterns, and relationships in large datasets. In the context of lung cancer cause classification, data mining can be used to transform complex, unstructured data into information for medical decision making.

Text mining refers to the branch of data mining. Text mining refers to the stage of extracting information from unsystematic texts [10]. This technique involves several stages such as feature extraction, text modeling, and classification. In a medical context, text mining can help in identifying early symptoms of lung cancer from patient medical records, or in analyzing scientific literature to find frequently mentioned risk factors associated with lung cancer. Several studies have shown the success of using K-NN in other medical data classification, such as heart disease and diabetes prediction. However, its application to lung cancer is still underexplored. Thus, this study focuses on filling this gap by implementing the K-NN algorithm and evaluating its performance in lung cancer cause classification.

METHOD

Image 1 shows the research stages.

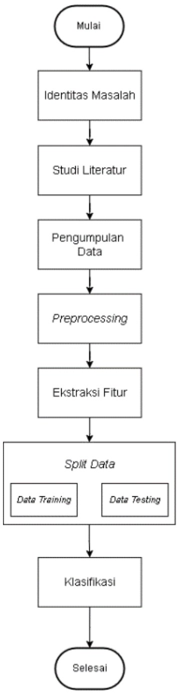

Image 1. Research Flow Diagram

Identification of Problems

This study identifies the problem that it is important to identify and classify the causes of lung cancer in order to develop more effective prevention and treatment strategies.

Literature Study

The stages of literature study in this research are to conduct a review of several previous studies related to the implementation of the K-NN algorithm for classifying the causes of lung cancer.

Data Collection

The dataset used comes from Kaggle which contains data on the types of causes of lung cancer. The dataset contains numeric data.

Preprocessing

Preprocessing is the initial process in data processing which aims to produce new data that is cleaner and ready to be used in the analysis process [11].

Feature Extraction

In this study, feature extraction uses Term Frequency Inverse Document Frequency (TF-IDF). The TF-IDF weighting method distributes values to each attribute in the document to determine the size of the attribute (term) [12]. The following is the TF-IDF equation stated in the formula (1).

$$w_{i,j} = tf_{i,j}\, log\, log \left(\frac{N}{df_i}\right) \qquad (1)$$

Description:
$w_{i,j}$: attribute value i in document j
$tf_{i,j}$: the number of occurrences (frequency) of attribute i in document j
N : total number of existing documents
df_i : number of documents that have attribute i

Split Data

The next stage is split data where the data will be divided into two parts: training data and testing data. Training data is used to create a model in the K-NN method [13]. Test data is used to measure model accuracy[13].

Classification

The next stage is to build a classification model using the K-Nearest Neighbor (K-NN) method. Classification refers to a data mining technique that is tasked with classifying new data that has the same characteristics in several classes. Classification will organize and group the number of existing text documents [14].

K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) refers to one of the data mining classification methods, where this algorithm will classify a number of data referring to training data that has been classified or labeled. This algorithm is included in the supervised learning group, namely search results that are categorized based on the majority proximity in K-NN. [15]. The following is the K-NN algorithm formula [14].

$$\boldsymbol{Euclidean} = \sqrt{\sum_{i}^{n} = \mathbf{1}(\boldsymbol{pi} - \boldsymbol{qi})^{\mathbf{2}}} \quad (2)$$

Description:
pi : train data
qi : test data
i : data variable
n : data size

RESULTS AND DISCUSSION

This study implemented a dataset consisting of 309 samples with 16 attributes, namely gen, umur, merokok, jari_kuning, kecemasan, tekanan_teman, penyakit_kronis, kelelahan, alergi, mengi, mengonsumsi_alkohol, batuk, sesak_nafas, kesulitan_menelan, nyeri-dada, dan kanker_paru_paru. There are no missing values in this dataset, but there are 33 duplicates that were removed to ensure good data quality.

The dataset has information for gender Male (M), Female (F), No (1), Yes (2). The existing dataset has been labeled by giving the label YES to patient data indicated as lung cancer and NO to patients not indicated as lung cancer.

In the preprocessing stage, the data is separated into two parts: training data (80%) and test data (20%). The preprocessing stage includes removing duplicate data or cleaning, and normalization. Data features are normalized using StandardScaler to ensure all features have the same scale. The results of the confusion matrix show that the accuracy score of the classification model using the K-NN algorithm is 90.32% where 90.32% of the total model predictions are correct. While the precision value for the YES class is 96%, for the NO class it is 67%. The recall value for the YES class is 92% and for the NO class it is 80%. The f1-score value for the YES class is 94% while the NO class is 73%. The following is a data visualization image of the confusion matrix:

Here is a sample of the data:

Table 1. Lung Cancer Dataset

GEN	UMU	MER	JAR	KEC	TEK	PEN	KEL	ALE	MEI	MEN	BAT	SES	KES	NYE	KAN
M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO

Image 2 shows the visualization result of the confusion matrix in this study.

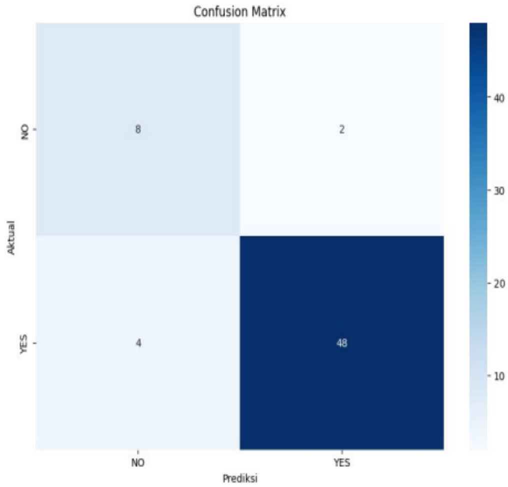

Image 2. Confusion Matrix

Table 2 shows the results of grouping patient data indicated as lung cancer based on gender with the number of details. For the YES class, it is the data on the number of patients identified as having lung cancer, while for the NO class, it is the data on the number of patients not identified as having lung cancer.

Table 2. Patient Identification Data Based on Gender

Gender	Identification	
	YES	NO
Female	125	22
Male	145	17

The classification results show that from the total dataset, 87.38% of the population was identified as lung cancer with around 270 patients indicated as lung cancer.

Here is a visualization of the pie chart:

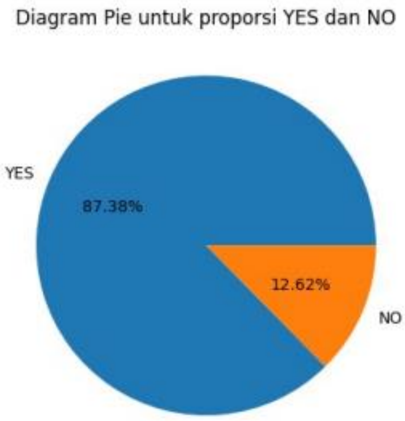

Image 3. Pie Chart for Proportion of YES and NO

The classification stage is carried out by grouping all existing causal attributes. The percentage of characteristics against all attributes is depicted in Image 4.

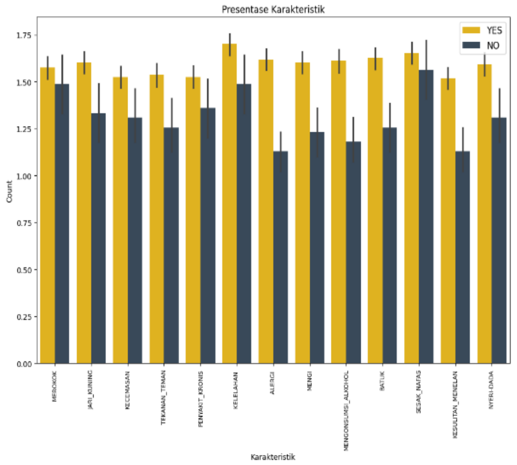

Image 4. Characteristics Percentage

CONCLUSION

The conclusion of the research on the implementation of the K-NN algorithm for the classification of lung cancer causes shows that the K-NN algorithm is effective in identifying lung cancer risk factors with high accuracy. The classification results show an accuracy of 90.32%.

These findings confirm that factors such as smoking, yellow fingers, anxiety, peer pressure, and fatigue have a significant correlation with lung cancer incidence. The implementation of the K-NN algorithm provides good results in this classification, helping in early identification and mitigation efforts of lung cancer. This study plays an important role in understanding the causes of lung cancer and offers an approach that can be used in the development of more effective prevention and early diagnosis strategies. Suggestions for future research to explore the use of other algorithms and optimization of model parameters to further improve the accuracy and effectiveness of prediction.

ACKNOWLEGMENTS

We would like to express our gratitude to God Almighty for His grace and blessings so that this research can be completed properly. We would like to thank Telkom University and Hamzanwadi University for the support of facilities and infrastructure provided during this research.

We would also like to express our deepest appreciation to the supervisors and fellow researchers who have provided guidance, input, and support during the research process. We would also like to express our sincere gratitude to our family and friends who have always provided moral support and prayers throughout the journey of this research. Last but not least, we would like to express our respect and gratitude to fellow researchers who have collaborated and shared their knowledge during this research process. Their productive discussions, inspiring ideas, and collaborative spirit have contributed significantly to the success of this research.

We hope that the results of this research can provide a positive contribution to the development of science, especially in the field of data classification and its application in the world of health.

BIBLIOGRAPHY

[1] R. Rofiani, L. Oktaviani, D. Vernanda, and T. Hendriawan, “Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4.5,” vol. 18, no. 1, 2024.

[2] Mohd Alimin and Ni Putu Rita Jeniyanti, “Pengaruh Penggunaan Fiksasi Masker Paru Terhadap Ketepatan Terget Penyinaran Pada Kanker Paru Teknik Intensity Modulated Radiation Therapy (IMRT) Di Departemen Onkologi Radiasi Rumah Sakit Umum Jakarta,” *Jurnal Ilmu Kesehatan dan Gizi*, vol. 2, no. 1, pp. 216–224, Nov. 2023, doi: 10.55606/jikg.v2i1.2153.

[3] R. Dwi Yulian Prakoso, B. Soejono Wiriaatmadja, and F. Wahyu Wibowo, *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS) Sistem Klasifikasi Pada Penyakit Parkinson Dengan Menggunakan Metode K-Nearest Neighbor*. 2020.

[4] M. Yunus and N. K. A. Pratiwi, “Prediksi Status Gizi Balita Dengan Algoritma K-Nearest Neighbor (KNN) di Puskemas

Cakranegara," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 4, no. 4, pp. 221–231, Feb. 2023, doi: 10.35746/jtim.v4i4.328.

[5] Rachmadhany Iman, Basuki Rahmat, and Achmad Junaidi, "Implementasi Algoritma K-Means dan Knearest Neighbors (KNN) Untuk Identifikasi Penyakit Tuberkulosis Pada Paru-Paru," *Repeater : Publikasi Teknik Informatika dan Jaringan*, vol. 2, no. 3, pp. 12–25, Jun. 2024, doi: 10.62951/repeater.v2i3.77.

[6] M. Annan, M. Mustofa, H. N. Wahiid, B. M. Islami, A. Ristyawan, and E. Daniati, "Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi) 485 Penggunaan Algoritma KNN dalam Deteksi Awal Kanker Paru-Paru Menggunakan Data Medis," 2024.

[7] Y. Setiawan, "Data Mining berbasis Nearest Neighbor dan Seleksi Fitur untuk Deteksi Kanker Payudara," vol. 8, no. 2, 2023.

[8] A. Muzakir, A. Desiani, and A. Amran, "Klasifikasi Penyakit Kanker Prostat Menggunakan Algoritma Naïve Bayes dan K-Nearest Neighbor," *Komputika : Jurnal Sistem Komputer*, vol. 12, no. 1, pp. 73–79, May 2023, doi: 10.34010/komputika.v12i1.9629.

[9] C. Zai and T. Komputer, "Implementasi Data Mining Sebagai Pengolahan Data," 2022.

[10] S. Analisis, A. Satusehat, D. Wardhani, R. Astuti, and D. D. Saputra, "Optimasi Feature Selection Text Mining: Stemming dan Stopword," *INNOVATIVE: Journal Of Social Science Research*, vol. 4, pp. 7537–7548, 2024.

[11] J. Homepage, Q. A'yuniyah, and M. Reza, "IJIRSE: Indonesian Journal of Informatic Research and Software Engineering Application Of The K-Nearest Neighbor Algorithm For Student Department Classification At 15 Pekanbaru State High School Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Di Sma Negeri 15 Pekanbaru," 2024.

[12] T. Abdi Mangun, O. Nurdiawan, and A. Irma Purnamasari, "Lung Cancer Analysis Using K-Nearst Neighbor Algorithm," 2023. [Online]. Available: https://ejournal.ubibanyuwangi.ac.id/index.php/jurnal_tinsika

[13] T. K. Ningsih and H. Zakaria, "Implementasi Algoritma K-Nearest Neighbor Pada Sistem Deteksi Penyakit Jantung (Studi Kasus : Klinik Makmur Jaya)." 2024. [Online]. Available: https://journal.mediapublikasi.id/index.php/logic

[14] V. Wulandari, W. J. Sari, Z. Alfian, L. Legito, and T. Arifianto, "Implementasi Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor untuk Klasifikasi Penyakit Ginjal Kronik," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 710–718, Apr. 2024, doi: 10.57152/malcom.v4i2.1229.

[15] A. Naufal Hilmi *et al.*, "Implementasi Algoritma K-Nearest Neighbor (KNN) untuk

Identifikasi Penyakit pada Tanaman Jeruk Berdasarkan Citra Daun," no. 2, pp. 107–117, 2024, doi: 10.62951/router.v2i2.78.