International Journal of Artificial Intelligence in Medical Issues

Volume 1 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v1i1.84

Research Article

Effectiveness Evaluation of the RandomForest Algorithm in Classifying CancerLips Data

Siti Khomsah [1], Edi Faizal [2]
[1] Institut Teknologi Telkom Purwokerto, Purwokerto, Indonesia, siti@ittelkom-pwt.ac.id
[2] Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia, edifaizal@utdi.ac.id

Correspondence should be addressed to Siti Khomsah; siti@ittelkom-pwt.ac.id

Received 21 January 2023; Revised 1 March 2023; Accepted 7 April 2023; Published 31 May 2023

Abstract:
Lip cancer, though less commonly discussed, remains a significant concern in the realm of oncology. Early detection and diagnosis are paramount for improved patient outcomes. This research evaluated the effectiveness of the RandomForest algorithm in classifying the CancerLips dataset, a collection of lip images processed using the Canny segmentation method and described using Hu moments. Using a 5-fold cross-validation approach, the algorithm achieved an average accuracy of approximately 70.96%. The results highlight the potential of machine learning techniques, specifically RandomForest, in aiding lip cancer detection. However, the choice of preprocessing methods and feature extraction plays a crucial role in determining the outcome. The study underscores the need for further research, focusing on algorithm optimization and comparisons with other datasets or feature extraction methods, to enhance diagnostic precision in medical imaging.

Keywords: Lip Cancer, RandomForest Algorithm, Canny Segmentation, Hu Moments, Medical Image Classification.
Dataset link: https://www.kaggle.com/datasets/shivam17299/oral-cancer-lips-and-tongue-images

1. Introduction

Cancer is among the leading causes of death worldwide, with various types affecting different parts of the body. One of the less commonly discussed, yet significant, types is lip cancer. Early detection and diagnosis of this cancer type can lead to more effective treatments and improved patient outcomes. Recent advancements in medical imaging and computer vision offer promising avenues for early detection using non-invasive methods. Specifically, image segmentation and feature extraction techniques have been recognized as crucial steps in the computer-aided diagnosis of various diseases, including cancer.

While several algorithms and techniques exist for image segmentation and feature extraction, choosing the right combination that delivers optimal performance for specific datasets remains challenging. For instance, while the Canny edge detection method is renowned for its capability to detect a wide range of edges in images, its effectiveness in segmenting lip images for cancer detection has not been extensively studied. Similarly, while Hu moments are powerful shape descriptors, their efficiency in capturing the nuances of lip cancer images needs thorough evaluation. The primary aim of this research is to evaluate the effectiveness of the RandomForest algorithm in classifying the CancerLips dataset. This entails segmenting the image using the Canny method, extracting features using Hu

moments, and subsequently applying the RandomForest classification algorithm[1][2]. By doing so, we aim to ascertain the suitability of this approach for lip cancer diagnosis.

The central research question guiding this study is: How effective is the RandomForest algorithm, following image segmentation with the Canny method and feature extraction using Hu moments, in classifying the CancerLips dataset? We hypothesize that the combination of these methods will offer a robust and efficient classification mechanism, potentially beneficial for early lip cancer detection.

This study focuses solely on the CancerLips dataset, and the findings might not be generalizable to other datasets or cancer types. Additionally, while the research evaluates the combined efficacy of the Canny method, Hu moments, and the RandomForest algorithm, it does not delve into comparisons with other algorithms or feature extraction techniques. The choice of Hu moments was based on its widespread use and recognition, but other descriptors might offer different results.

Through this research, we provide insights into the effectiveness of a specific combination of methods for lip cancer image classification. The findings can serve as a foundation for further studies in this domain, potentially paving the way for the development of more advanced and efficient diagnostic tools. Moreover, by highlighting the strengths and limitations of the approach, we aim to inform future research directions and inspire innovations in early cancer detection techniques.

2. Method

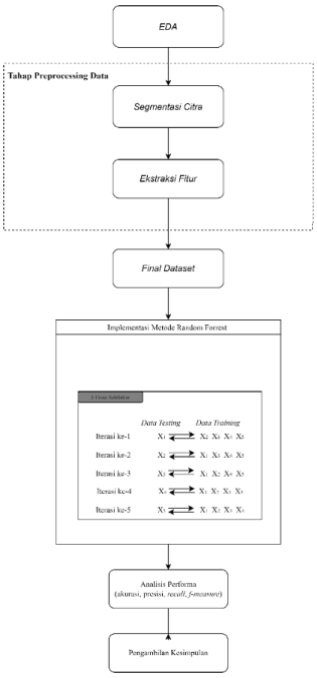

Figure 1: Research Design for Lip Cancer Classification

As illustrated in Figure 1, which depicts the research design, this study adopted a quantitative approach, utilizing a structured methodology to evaluate the effectiveness of image processing techniques combined with machine

learning for lip cancer classification. The investigation was structured in consecutive stages: image segmentation, feature extraction, and classification using the RandomForest algorithm [3][4]. The dataset, referred to as the CancerLips dataset, comprised 131 samples, each described by seven features and a target label. These features represent the Hu moments extracted from lip images, while the target label indicates the presence or absence of lip cancer.

The dataset was preprocessed, meaning the raw lip images underwent segmentation using the Canny edge detection method [5][6]. The Canny edge detector works by detecting areas of the image with rapid intensity changes, which correspond to edges. The formula for the Canny method is:

$$G = \sqrt{G_x^2 + G_y^2} \tag{1}$$

Where G_x and G_y are the gradients in the x and y directions, respectively. The results of the Canny segmentation are shown in Figure 2 for the normal class and in Figure 3 for the cancer class.

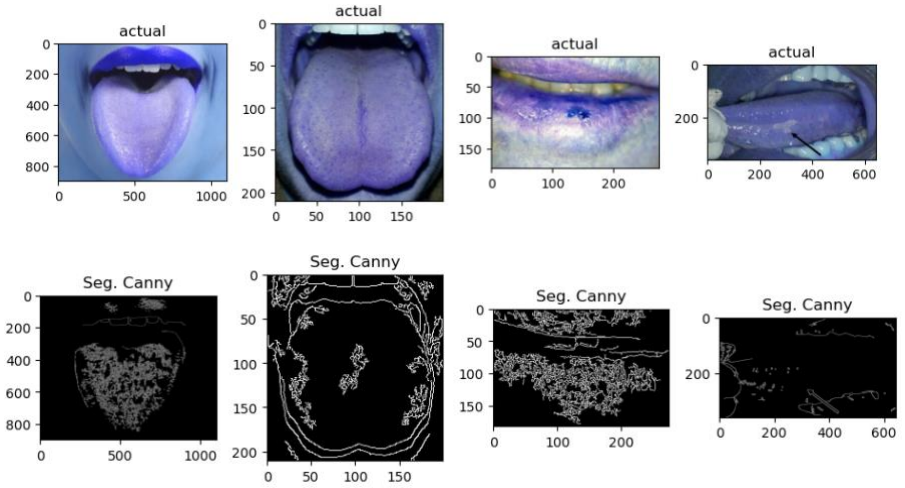

Figure 2: Canny Segmentation Result for Normal Class.

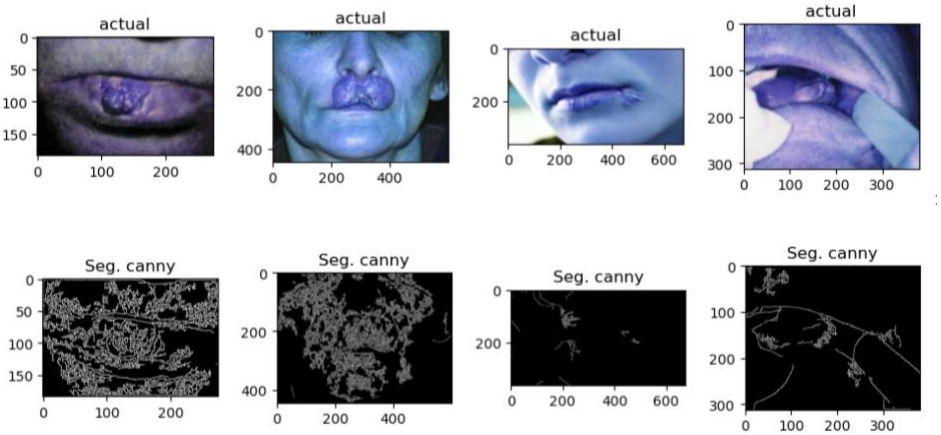

Figure 3: Canny Segmentation Result for Cancer Class.

Following segmentation, the Hu moments, which are moment invariants and serve as shape descriptors, were extracted from the segmented images [7][8]. The Hu Moments can be represented by a series of equations, as shown in Equation 2.

$$
\begin{aligned}
\phi\_1 &= \eta\_\{20\} + \eta\_\{02\} \\
\phi\_2 &= (\eta\_\{20\} - \eta\_\{02\})\text{^}2 + 4\eta\_\{11\}\text{^}2 \\
\phi\_3 &= (\eta\_\{30\} - 3\eta\_\{12\})\text{^}2 + (3\eta\_\{21\} - \eta\_\{03\})\text{^}2 \\
\phi\_4 &= (\eta\_\{30\} + \eta\_\{12\})\text{^}2 + (\eta\_\{21\} + \eta\_\{03\})\text{^}2 \\
\phi\_5 &= (\eta\_\{30\} - 3\eta\_\{12\})(\eta\_\{30\} + \eta\_\{12\})[(\eta\_\{30\} + \eta\_\{12\})\text{^}2 - 3(\eta\_\{21\} \\
&\quad + \eta\_\{03\})\text{^}2] + (3\eta\_\{21\} - \eta\_\{03\})(\eta\_\{21\} + \eta\_\{03\})[3(\eta\_\{30\} \\
&\quad + \eta\_\{12\})\text{^}2 - (\eta\_\{21\} + \eta\_\{03\})\text{^}2] \\
\phi\_6 &= (\eta\_\{20\} - \eta\_\{02\})[(\eta\_\{30\} + \eta\_\{12\})\text{^}2 - (\eta\_\{21\} + \eta\_\{03\})\text{^}2] \\
&\quad + 4\eta\_\{11\}(\eta\_\{30\} + \eta\_\{12\})(\eta\_\{21\} + \eta\_\{03\}) \\
\phi\_7 &= (3\eta\_\{21\} - \eta\_\{03\})(\eta\_\{30\} + \eta\_\{12\})[(\eta\_\{30\} + \eta\_\{12\})\text{^}2 - 3(\eta\_\{21\} \\
&\quad + \eta\_\{03\})\text{^}2] + (\eta\_\{30\} - 3\eta\_\{12\})(\eta\_\{21\} + \eta\_\{03\})[3(\eta\_\{30\} \\
&\quad + \eta\_\{12\})\text{^}2 - (\eta\_\{21\} + \eta\_\{03\})\text{^}2]
\end{aligned}
\tag{2}
$$

Where η represents the normalized central moments. For a visual representation of the extracted features, Figure 4 showcases a scatter plot visualization of the Hu Moments derived from the brain images.

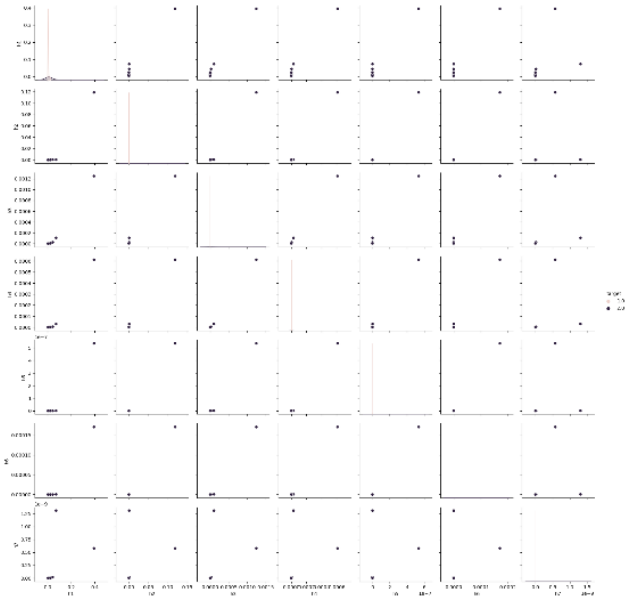

Figure 4: Scatter Plot Visualization of Extracted Hu Moments Features

Upon obtaining the dataset with Hu moments as features, we applied the RandomForest classifier[9]. RandomForest operates by constructing multiple decision trees during training and outputs the mode of the classes for classification [10]. The decision to use RandomForest was based on its capability to handle high-dimensional data and its robustness against overfitting.

For validation purposes, a 5-fold cross-validation was applied, dividing the dataset into five subsets. In each iteration, four subsets were used for training, and the remaining subset was used for validation [11][12]. This ensured that every sample in the dataset was used for both training and validation[13][14].

Performance metrics, including accuracy, precision, recall, and F-measure, were employed to evaluate the model's efficiency [15][16]. The F-measure, for instance, is given by:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ F1 &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{3}$$

3. Result and Discussion

The CancerLips dataset underwent a structured analysis using the RandomForest algorithm combined with a 5-fold cross-validation approach. The primary goal was to evaluate the effectiveness of the chosen classification method on the dataset, and the results were measured across multiple performance metrics: Accuracy, Precision, Recall, and F-Measure [17][18].

Visualization of the results

The performance metrics for each fold of the cross-validation, as well as their average, offer insights into the robustness and consistency of the Random Forest algorithm applied to the Cancer Lips dataset. These metrics, as tabulated in Table 1 below, serve as a testament to the potential and challenges of employing machine learning techniques in medical image classification. To provide a more intuitive understanding and facilitate easier comparisons, these results are also visually represented in the form of a bar graph in Figure 5. This graphical representation aids in quickly identifying the folds with the highest and lowest performance, offering a clear perspective on the algorithm's strengths and areas for improvement.

Table 1: Performance Metrics for Each Fold of Cross-Validation

K-n	Performa			
	Akurasi	***Presisi***	***Recall***	***F-Measure***
K-1	66.6%	63%	66.6%	62.2%
K-2	73%	71.6%	73%	68.5%
K-3	73%	80.9%	73%	66.8%
K-4	76.9%	82.9%	76.9%	72.8%
K-5	65.3%	60.8%	65.3%	57.3%
$\sum \boldsymbol{Avg}$	70.96%	71.84%	70.96%	65.52%

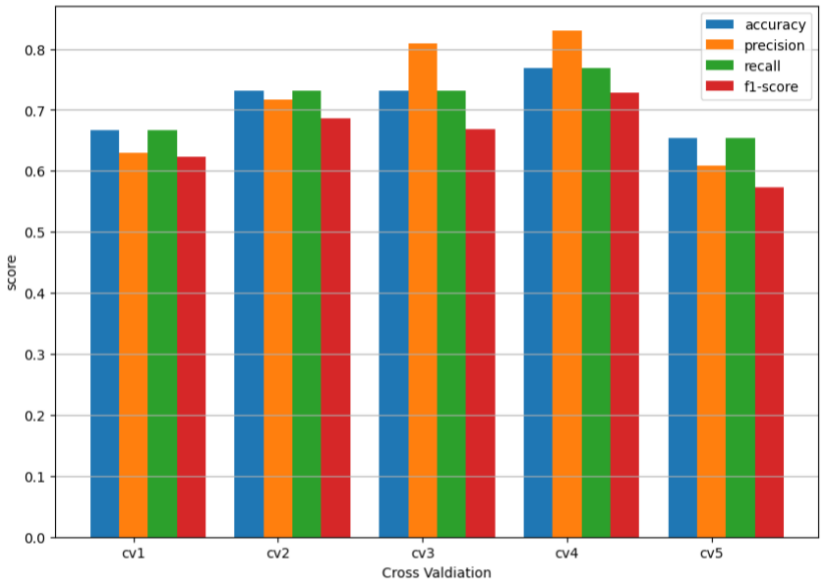

Figure 5: Bar Graph Representation of RandomForest Performance Metrics

From the table, it's evident that the RandomForest algorithm's performance varied across different folds. The highest accuracy achieved was during the fourth fold (K-4) at 76.9%, while the lowest was in the fifth fold (K-5) at 65.3%. The average accuracy across all folds was 70.96%. The RandomForest algorithm showcased a relatively consistent performance across different folds, with minor deviations. The precision metric peaked at 82.9% during the fourth fold, indicating the model's strong capability to correctly classify positive cases [19][20].

Discussion

The variations in performance across different folds suggest that the dataset may have inherent complexities or that certain data partitions were more challenging for the model to classify. While the average accuracy was about 70.96%, there's a noticeable difference in performance metrics across folds, indicating areas for improvement. While the current research focused on the CancerLips dataset, previous studies on medical image classification have also highlighted the challenges posed by the nuances of medical images. The results are in line with the general consensus that while machine learning offers promising results, the choice of preprocessing methods and features play a significant role in the outcome.

The findings underscore the potential of using machine learning algorithms, specifically RandomForest, in assisting with early lip cancer detection. However, it also emphasizes the need for careful data preprocessing and feature selection. The study was confined to the CancerLips dataset, limiting its generalizability. Additionally, while the research evaluated the combined efficacy of the Canny method, Hu moments, and the RandomForest algorithm, it did not compare the results with other algorithms or feature extraction techniques.

Recommendations for further research

Future research could delve deeper into optimizing the RandomForest parameters or explore other machine learning algorithms. Comparing the results with other datasets or considering other feature extraction techniques could also provide a more comprehensive understanding of the model's capabilities.

4. Conclusion

In our examination of the CancerLips dataset using the RandomForest algorithm, we observed an average accuracy of approximately 70.96%, with minor performance variations across different folds of the 5-fold cross-validation. This reinforces our initial hypothesis that the combination of the Canny segmentation method, Hu moments for feature extraction, and the RandomForest classifier would offer a robust mechanism for lip cancer classification. The research contributes to the growing body of literature that explores machine learning techniques in medical image classification, highlighting the importance of preprocessing methods and feature selection in achieving optimal results.

For future endeavors, it is recommended to delve deeper into the parameter optimization of the RandomForest algorithm or consider other machine learning models. Further comparative studies with other datasets or exploring alternative feature extraction techniques could pave the way for more comprehensive diagnostic tools, enhancing early detection and treatment of lip cancer.

References

[1] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart J.*, vol. 6, no. 2, p. 167, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

[2] I. Wardhana, M. Ariawijaya, V. A. Isnaini, and R. P. Wirman, "Gradient Boosting Machine, Random Forest dan Light GBM untuk Klasifikasi Kacang Kering," *J. Resti*, vol. 5, pp. 92–99, 2022.

[3] L. Britanthia, C. Tanujaya, B. Susanto, and A. Saragih, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 68–78, 2020.

[4] F. T. Admojo and S. R. Jabir, "Analisis performa metode Naïve Bayesh Classifier pada Electronic Nose dalam identifikasi formalin pada tahu," *Indones. J. Data Sci.*, vol. 4, no. 1, pp. 1–16, 2023, doi: 10.56705/ijodas.v4i1.67.

[5] E. A. Sekehravani, E. Babulak, and M. Masoodi, "Implementing canny edge detection algorithm for noisy image," *Bull. Electr. Eng. Informatics*, vol. 9, no. 4, pp. 1404–1410, 2020, doi: 10.11591/eei.v9i4.1837.

[6] M. Radhakrishnan, A. Panneerselvam, and N. Nachimuthu, "Canny edge detection model in mri image segmentation using optimized parameter tuning method," *Intell. Autom. Soft Comput.*, vol. 26, no. 6, pp. 1185–1199, 2020, doi: 10.32604/iasc.2020.012069.

[7] G. Xie, B. Guo, Z. Huang, Y. Zheng, and Y. Yan, "Combination of Dominant Color Descriptor and Hu Moments in Consistent Zone for Content Based Image Retrieval," *IEEE Access*, vol. 8, pp. 146284–146299, 2020, doi: 10.1109/ACCESS.2020.3015285.

[8] A. Mustopa, H. M. Nawawi, S. Agustiani, and S. K. Wildah, "Feature Extraction With Forest Classifer To Predicate Covid 19 Based On Thorax X-Ray Results," *Sistemasi*, vol. 11, no. 2, p. 515, 2022, doi: 10.32520/stmsi.v11i2.1966.

[9] Y. W. Sitorus, P. Sukarno, and S. Mandala, "Analisis Deteksi Malware Android menggunakan metode Support Vector Machine & Random Forest," *e-Proceeding Eng.*, vol. 8, no. 6, pp. 12500–12518, 2021.

[10] F. Tangguh and Y. Islami, "Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam

mengklasifikasi tahu berformalin," *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 1–8, 2022, doi: 10.56705/ijodas.v3i1.42.

[11] A. Aisyah and S. Anraeni, "Analisis Penerapan Metode K-Nearest Neighbor (K-NN) pada Dataset Citra Penyakit Malaria," *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 17–29, 2022, doi: 10.56705/ijodas.v3i1.22.

[12] Ericha Apriliyani and Y. Salim, "Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 47–54, 2022, doi: 10.56705/ijodas.v3i2.45.

[13] D. Cahyanti, A. Rahmayani, and S. Ainy, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020.

[14] L. Saiman and R. Satra, "Analisis performa metode Support Vector Machine untuk klasifikasi dataset aroma tahu berformalin," *Indones. J. Data Sci.*, vol. 2, no. 2, pp. 50–61, 2021, doi: 10.56705/ijodas.v2i2.28.

[15] S. Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 79–86, 2020, doi: 10.33096/ijodas.v1i3.20.

[16] A. Maulida, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.

[17] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020, [Online]. Available: file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Azis, Admojo, Susanti - 2020 - Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah.pdf.

[18] A. Nurul, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indones. J. Data Sci.*, vol. 3, no. 3, pp. 115–121, 2022, doi: https://doi.org/10.56705/ijodas.v3i3.54.

[19] H. Azis, F. Fattah, and P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020, [Online]. Available: file:///Users/kbh/Downloads/507-2012-5-PB.pdf.

[20] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019, [Online]. Available: file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Baharuddin, Hasanuddin, Azis - 2019 - Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca.pdf.