Classification of human monkeypox with the Fuzzy C-Means Algorithm using image processing methods and Haralick texture parameters

Görüntü işleme yöntemleri ve Haralick doku parametreleri kullanılarak insandaki maymun çiçeği hastalığının Bulanık C-Ortalamalar Algoritması ile sınıflandırılması

Senem Gonenc[1], Ozge Pasin[2]

[1] Department of Statistics, Faculty of Science, Atatürk University

[2] Department of Biostatistics, Hamidiye Medicine Faculty, Health Science University

Abstract

Aim: Human monkeypox can cause skin lesions in the form of blisters of different shapes on various parts of the body. Due to the fact that the skin lesions caused by human monkeypox have a very similar appearance to lesions caused by chickenpox and measles, the study includes images of chickenpox and measles as well as images of human monkeypox. The aim of this study is to distinguish human monkeypox virus skin lesion images from other viral diseases with similar images.

Methods: For this study, the Monkeypox Skin Lesion Dataset, which consists of binary classification data for monkeypox and non-monkeypox (chickenpox, measles) skin lesions, is accessed from the Kaggle.com website. In total, 228 images are processed, with 101 images in the monkeypox group and 127 images in the non-monkeypox group. The images in the Monkeypox Skin Lesion Dataset are processed using image analysis methods and Haralick texture parameters are calculated to create 13 different features for each image. For the classification process in the statistical analysis part of the study, Fuzzy C-Means algorithm is used.

Results: The images used in the study belong to individuals with varying skin tones and from different parts of the body, and the algorithm provides encouraging results in determining the type of skin lesions in the images. The overall classification accuracy rate is 61.8%, and the highest accuracy (76.2%) is achieved in the monkeypox class.

Conclusion: This study demonstrates that images of viral diseases with similar skin lesions can be classified using various image-processing techniques and different statistical methods.

Keywords: Classification; clustering; monkeypox

Öz

Amaç: İnsanlarda görülen maymun çiçeği, vücudun çeşitli yerlerinde farklı şekillerde kabarcıklar şeklinde deri lezyonlarına neden olabilir. İnsan maymun çiçeğinin neden olduğu cilt lezyonları, suçiçeği ve kızamık kaynaklı lezyonlara çok benzer bir görünüme sahiptir. Bu sebeple çalışmada, insan maymun çiçeği görüntülerinin yanı sıra suçiçeği ve kızamık görüntüleri de yer almaktadır. Bu çalışmanın amacı, insan maymun çiçeği virüsü cilt lezyonu görüntülerini, benzer görüntülere sahip diğer viral hastalıklardan ayırmaktır.

Yöntemler: Bu çalışma için maymun çiçeği ve maymun çiçeği olmayan (suçiçeği, kızamık) cilt lezyonlarına yönelik ikili sınıflandırma verilerinden oluşan Maymun Çiçeği Cilt Lezyonu Veri Setine Kaggle.com web sitesinden erişilmektedir. Maymun çiçeği grubunda 101 görüntü ve maymun çiçeği olmayan grupta 127 görüntü olmak üzere toplamda 228 görüntü işlenir. Maymun Çiçeği Cilt Lezyonu Veri Setinde yer alan görüntüler, görüntü analiz yöntemleri kullanılarak işlenmekte ve her görüntü için 13 farklı özellik oluşturacak şekilde Haralick doku parametreleri hesaplanmaktadır. Çalışmanın istatistiksel analiz kısmında sınıflandırma işlemi için Bulanık C-Ortalamalar algoritması kullanılır.

Bulgular: Çalışmada kullanılan görüntüler, farklı cilt tonlarına sahip bireylerden ve vücudun farklı bölgelerinden alınmış olup algoritma, görüntülerdeki cilt lezyonlarının tipinin belirlenmesinde cesaret verici sonuçlar ortaya koymaktadır. Genel sınıflandırma doğruluk oranı %61.8 olarak bulunmakta ve en yüksek doğruluk da (%76.2) maymun çiçeği sınıfında elde edilmektedir.

Sonuç: Bu çalışma, benzer cilt lezyonlarına sahip viral hastalık görüntülerinin, çeşitli görüntü işleme teknikleri ve farklı istatistiksel yöntemler kullanılarak sınıflandırılabileceğini göstermektedir.

Anahtar Sözcükler: Gruplama; maymun poks; sınıflandırma

Received/*Geliş* : 09.05.2024
Accepted/*Kabul*: 16.09.2024

DOI: 10.21673/anadoluklin.1477313

Corresponding author/*Yazışma yazarı*

Senem Gönenç

Atatürk Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Erzurum, Türkiye.
E-mail: senemgonenc@atauni.edu.tr

ORCID

Senem Gönenç: 0000-0002-6990-1507
Özge Pasin: 0000-0001-6530-0942

INTRODUCTION

As the world continues to grapple with the effects of the ongoing COVID-19 pandemic, the emergence of monkeypox as a rapidly spreading virus has become a new cause for concern. Monkeypox, also known as monkeypox virus disease, has been reported in 75 countries as of July 23, 2022, prompting the World Health Organization (WHO) to declare a global emergency and warn of the potential for the virus to spread to even more countries. So, monkeypox is a viral disease that is similar to smallpox and is primarily found in Central and West Africa. The disease can be transmitted to humans through contact with infected animals, particularly rodents and primates. Symptoms of monkeypox include fever and lesions on the skin. While most cases of monkeypox are mild and self-limiting, severe cases can occur, particularly in individuals with weakened immune systems.

Monkeypox is not a new disease, as it was first identified in 1958 in laboratory monkeys in Copenhagen, Denmark. The disease was named monkeypox virus due to its characteristic skin lesions, but initially, no clinical information about the disease was documented (1). The first recorded case of monkeypox in humans was reported in 1970 when a 9-month-old baby boy in the Democratic Republic of Congo was diagnosed with the disease. Since then, monkeypox cases have been reported in humans in other countries in Central and West Africa, with the majority of cases occurring in the Democratic Republic of the Congo. Monkeypox re-emerged in 2018, with a total of five reported cases in Nigeria, three in the UK, one in Israel, and one in Singapore. Subsequently, in May 2022, numerous cases of monkeypox were identified in regions where the disease is not endemic (2-7).

Monkeypox virus infection produces symptoms that are comparable to smallpox but are usually less severe. The disease is divided into two distinct periods. The first stage is known as the invasion period and is characterized by the onset of fever, severe headache, swollen lymph nodes, backache, muscle aches, and extreme weakness. This stage typically lasts between 0-5 days. Swollen lymph nodes, or lymphadenopathy, are a key clinical feature that distinguishes monkeypox from other diseases that present with similar initial symptoms, such as chickenpox, measles, and smallpox. Typically, a skin lesion appears within 1 to 3 days after the onset of fever, which tends to be most concentrated on the face (in 95% of cases), palms, and soles (in 75% of cases). The number of lesions that appear in monkeypox cases can vary greatly, ranging from only a few to several thousand. In severe cases, the lesions may merge together, causing the affected skin to peel over a large area. Furthermore, severe cases of monkeypox are more frequently observed in children. The Centers for Disease Control and Prevention (CDC) states that there is currently no known cure for monkeypox infection and that treatment involves supportive care. Nevertheless, the CDC website offers detailed information on how to prevent and control the spread of the disease (8, 9).

A review of the literature shows that studies on human monkeypox date back to the 1970s when the first case in humans was reported. In recent years, there has been a significant increase in studies on human monkeypox due to the emergence of the disease in countries where it was previously not observed in 2018, as well as the rapid increase in cases in these countries in 2022. Therefore, there are only a limited number of studies available that focus on the image analysis and statistical classification of the skin lesion symptoms associated with the disease. Ahsan et al. published a preprint article that collected images of monkeypox skin lesions from various sources such as websites, newspapers, and online portals. They obtained a dataset of 43 monkeypox, 47 chickenpox, 17 measles, and 54 normal images, which they augmented to increase the dataset size. The authors aimed to investigate the progression of the monkeypox virus and how its skin lesion findings differ from other similar skin diseases. To achieve this, they performed histogram analysis and analyzed the pixel intensities of the images (10). Haque et al. conducted research on transfer learning-based models using the Monkeypox Skin Lesion Dataset (MSLD) that was also used by Ali et al. in their study. They employed the CBAM attention mechanism and evaluated several models including VGG19, DenseNet121, EfficientNetB3, MobileNetV2, and Xception. The study achieved classification rates of 69.86% for VGG19, 78.27% for DenseNet121, 54.21% for EfficientNetB3, 74.07% for MobileNetV2, and 79.90% for Xception (11). Ali et al. performed a study on the classification of

monkeypox and other similar skin diseases using pre-trained VGG-16, ResNet50 and InceptionV3 models. They used the MSLD developed by them in the study. The results showed that ResNet50 and VGG-16 had correct classification rates of 82.96% and 81.48% respectively. At the end of the study, the authors combined the three models and achieved an accuracy rate of 79.26±1.05 (12). Sahin et al. used the MSLD from Kaggle to classify human monkeypox in their study. The dataset contained binary classification data, with images labeled as either having monkeypox or not (chickenpox or measles). The researchers employed several pre-trained deep learning models on the MSLD dataset and found that the MobileNetV2 and EfficientNetB0 models performed well. Then they converted the entire model into a TensorFlow lite model, allowing it to be adapted for mobile devices. Using this model, they developed a mobile application that could classify images with an accuracy of 91.11% based on test results. The researchers also noted that the application could be used to diagnose other skin diseases, including monkeypox, in a preliminary capacity (13). Sadad et al. developed a new method called FCMRG to identify breast masses in mammograms. Fuzzy C-Means (FCM) and region-growing (RG) algorithms were used in this method. In the feature extraction step of image processing, Local Binary Pattern Gray-Level Co-occurrence Matrix (LBP-GLCM) and Local Phase Quantization (LPQ) were applied to the images. They used machine learning procedures to distinguish between benign and malignant tumors and achieved a high accuracy rate of 98.2% with the proposed method (14). Rohmayani and Rahayu carried out a study to classify lung images into four different categories: normal lung, pneumonia-infected lung, COVID-19-infected lung, and X-ray images of other diseases that involve the lung. They obtained the dataset from Kaggle.com, which consisted of 120 training images and 40 test images. Image processing techniques were applied, and the mean and standard deviation values were used for feature extraction. FCM algorithm was used for classification analysis, and the method achieved an accuracy rate of 65% (15). Zayed and Elnemr tested a dataset containing CT images of 37 patients with lung tumors or pulmonary edema. After performing image processing techniques, they extracted haralick texture features to identify the type of abnormality in the lungs. They found that their proposed method showed potential in detecting lung abnormalities (16).

Our study has three main aims. The first and most important of these is to classify images of virus-induced diseases such as monkeypox, chickenpox and measles. To achieve this, we use the MSLD available on Kaggle website. The second aim is to use the FCM algorithm, a statistical approach that has not been used before in this field, when classifying images of diseases. Finally, using the 13 different statistical features (haralick texture parameters) that we obtained at the end of various pre-processing steps in the image processing section as input variables in FCM analysis. In addition, running the FCM algorithm according to the haralick texture parameters calculated for each image in MSLD is the most innovative aspect of the study.

In this study, we suggest a classification based on a different functioning for MSLD within the framework of our 3 main aims. It can be seen that this study also stands out as a multi-disciplinary study. The use of image processing methods in the study can be explained as the use of computers in health. Classification of haralick texture parameters with the FCM algorithm also includes the science of statistics in the use of computers in health. Thus, this study will make a significant contribution to the literature.

MATERIALS AND METHODS

Early diagnosis is vital for the human monkeypox outbreak, which has become a global health problem in recent years. The disease is diagnosed using Transcription-Polymerase Chain Reaction (TT-PCR) test, like other viral diseases. However, the sensitivity of this test to detect viruses is low and it takes a long time for the test to be completed. These disadvantages of the test require the development of alternative diagnostic systems. Monkeypox skin lesions are considered an important clinical feature in distinguishing the disease from some other viral diseases such as chickenpox and measles. Therefore, computer-aided preliminary evaluation of images of monkeypox skin lesions may be a useful alternative diagnosis. For this reason, the MSLD is created with images collected from relevant news portals, various websites and some publicly available

case reports. It is worth noting that the disease mostly occurs in African countries where access to healthcare is quite difficult. Therefore, a large data set has not been created so far. Of course, more images of chickenpox and measles skin lesions could be collected. However, facing data imbalance is not a desirable situation and may cause other problems because only monkeypox skin lesion images are limited.

For this research, we obtained the MSLD from the Kaggle.com website which consists of binary classification data for monkeypox and non-monkeypox (chickenpox, measles) skin lesions (17). The dataset contains images from the monkeypox class, as well as similar skin lesions caused by chickenpox and measles. Here, there are 228 images in total, 101 images in the monkeypox group and 127 images in the non-monkeypox group. The images in the MSLD are processed using image analysis methods and Haralick texture parameters are calculated to create 13 different features for each image. Additionally, we use the FCM algorithm for the classification process in the statistical analysis part of the study. In this context, we examine the resulting cluster structure and the distribution of images into clusters. Since the disease type of each image is known in advance, we present a detailed evaluation of the classification success of the method.

Prior to starting the coding stage, the algorithm of the study is developed and the procedures to be followed during the implementation phase are defined. The summarized steps are presented as follows.

1. Download the Monkeypox Skin Lesion Dataset (MSLD) from the Kaggle website
2. Transferring image files to the system and reading them as a matrix
3. Histogram equalization
4. Converting to grayscale image format and invert the image
5. Thresholding the image with the appropriate threshold value-converting to binary format
6. Applying morphological operators to the image
7. Calculating Haralick texture parameters
8. Classification using the Fuzzy C-Means algorithm

For all the transactions and analyses in the coding stage, we use the R package program (version 4.1.3), a free and open-source statistical software development environment that includes various original packages added by users. This program is widely accepted for use in academic studies.

Image Processing

Image processing involves various operations on digital images such as segmentation, merging, rotation, adjustment of brightness, contrast or sharpness, removal of defects or unwanted objects, and enhancing the visibility of important objects according to the researcher's requirements. It is a critical technique for obtaining meaningful data from images and gathering diverse information about objects in the image. The process starts with capturing an image, displaying it on a computer, and then digitizing it for further analysis.

When images are digitized, they are represented as a matrix with rows and columns of pixels. Each pixel in the matrix has a color value that corresponds to its color in the original image. The range of numerical values for the color values of the pixels varies depending on the type of image. There are three types of images used in image processing: color images, grayscale images, and binary images. Color images have multiple color channels, usually red, green, and blue (RGB), while grayscale images have a single color channel representing the intensity of the image. Binary images have only two possible color values, usually black and white, and are used for simple image processing tasks such as edge detection and shape recognition. Digital images are represented as a two-dimensional matrix consisting of pixels in rows and columns, with each pixel being assigned a numerical value that represents its color. The range of numerical values assigned to pixels depends on the type of image. There are three main types of images used in image processing: color images, grayscale images, and binary images. Color images have pixel values in the range of 0 to 255, corresponding to the three colors of Red, Green, and Blue. Grayscale images have pixel values ranging from 0 to 1 and consist of black, white, and shades of gray. Binary images have only two colors, black and white, with pixel values of either 1 or 0. Since most image processing algorithms do not work on color images, it is common to convert color images to binary images.

Image thresholding

Thresholding is a fundamental method in image processing that is used to extract information from an image by separating the objects from the background. This process involves converting a grayscale image to a binary image using a threshold value, which is determined by the researcher. Each pixel in the grayscale image is then compared to this threshold value, and if the comparison result is greater than the threshold, it is considered as an object and evaluated as white, otherwise, it is expressed as black and considered part of the background.

Morphological operations

Mathematical morphology is a technique that uses set theory and is based on the shape structure of an image. It involves two fundamental operations: dilation and erosion, and other operations can be obtained through various combinations of these two operations. However, before these operations can be applied, the image must be converted to binary format (18). Expansion is a type of morphological operation which is used to increase the size or thickness of object outlines in a binary image. Its main objective is to fill small gaps and close holes in the image. Erosion is a morphological operation that has the opposite effect of dilation. It reduces the size of objects or thins the boundaries of objects in a binary-converted image. As a result, it separates objects and increases holes in the image. Applying only expansion or erosion operations to an image can cause significant distortions. Therefore, opening and closing operations are often used to eliminate these distortion problems in the image. Opening is performed by applying erosion first, then expansion sequentially to the image. This operation separates objects that are close to each other in the image while causing minimal change. On the other hand, closing is performed by applying expansion first and then erosion to the image. This operation merges objects that are close to each other in the image while causing minimal change.

Haralick Texture Features

Texture features, also known as second-order feature descriptors, are used to capture patterns and provide statistical correlation between pixels in an image. Unlike first-order features, which only consider the distribution of pixel values, texture features incorporate spatial information and analyze the relationships between gray levels in the image. This allows texture features to capture specific patterns that may be missed by first-order features and can help to minimize information loss in image analysis. The concept of using texture features for image classification was first introduced by Haralick and Shanmugam. Second-order statistical features are similar to the features that are important to clinicians, and they provide quantitatively representative data that can yield more information than what is visible to the human eye. Texture features measure the consistency of patterns and colors in an image, and the Haralick texture is a widely used technique for characterizing texture-based images. Texture analysis is a feature extraction method commonly applied to medical images, which involves analyzing the spatial distribution of neighborhoods at the gray color level within the image (19).

Haralick texture is a crucial technique for characterizing texture-based images and is achieved through the computation of the gray level co-occurrence matrix (GLCM). The GLCM is a popular method for extracting texture features due to its simplicity and the ability to compute numerous features. Textural features can be extracted from a grayscale image by using the GLCM matrix for a particular direction and distance of patterns. The fundamental concept behind GLCM is the pixel neighborhood in an image. The GLCM matrix records the frequency of co-occurrences of gray color levels for each neighbor relationship and direction. It shows how frequently different combinations of gray levels occur in an image. The sum of the number of co-occurrences of a pixel with a value of i in a spatial relationship defined by a pixel with a value of j in the input image is the value of each element in the GLCM (20).

In this research, the characteristics of the Gray Level Co-occurrence Matrix (GLCM) are analyzed for gray level images. The formulas used to extract these features are provided below. The equations describe how to compute a set of statistical features for the co-occurrence matrix.

$P_{ij}=P(i,j)$ is the co-occurrence matrix.

$p(i,j)=P(i,j)/R$, normalized co-occurrence matrix.

N_g, number of discrete intensity levels in the image.

R, normalizing constant.

$p_x(i) = \sum_{j=1}^{N_g} p(i,j)$, marginal row probabilities.

$p_y(j) = \sum_{i=1}^{N_g} p(i,j)$, marginal column probabilities.

μ_x and μ_y are the mean gray level intensities of p_x and p_y, respectively. σ_x and σ_y are also the standard deviations of p_x and p_y, respectively.

Haralick presented a set of 13 statistical features that can be extracted from the Gray Level Co-occurrence Matrix (GLCM), which are commonly known as Haralick texture features. The R Bioconductor package EBImage uses these features and assigns them the names h.asm, h.con, h.cor, h.var, h.idm, h.sav, h.sva, h.sen, h.ent, h.dva, h.den, h.f12, and h.f13. The values for these features are calculated using the equations given in Equation (1) through Equation (13) as provided in references (19, 21, 22).

Angular Second Moment (h.asm)

$$f_1 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(\frac{P(i,j)}{R}\right)^2 = \sum_i\sum_j p(i,j)^2 \quad (1)$$

Contrast (h.con)

$$f_2 = \sum_{k=0}^{N_g-1} k^2 \left\{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\delta_{|i-j|,k}p(i,j)\right\} = \sum_{k=0}^{N_g-1} k^2 p_{x-y}(k) \quad (2)$$

where the Kronecker delta function is defined as;

$$\delta_{m,n} = \begin{cases} 1 & when \quad m = n \\ 0 & when \quad m \neq n \end{cases} \quad (2.1)$$

Haralick and Miyamoto did not use the Kronecker delta function. Instead, they have given the specified summation conditions in the continuation of Equation (9), Equation (12), and Equation (13).

1. Correlation (h.cor)

$$f_3 = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(ij)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y} \quad (3)$$

2. Sum of Squares: Variance (h.var)

$$f_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu)^2 p(i,j) \quad (4)$$

3. Inverse Difference Moment (h.idm)

$$f_5 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{1}{1+(i-j)^2}p(i,j) \quad (5)$$

4. Sum Average (h.sav)

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \quad (6)$$

5. Sum Variance (h.sva)

$$f_7 = \sum_{i=2}^{2N_g}(i-f_8)^2 p_{x+y}(i) \quad (7)$$

6. Sum Entropy (h.sen)

$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i)\log\left(p_{x+y}(i)\right) \quad (8)$$

7. Entropy (h.ent)

$$f_9 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log\left(p(i,j)\right) = HXY \quad (9)$$

$$HXY = -\sum_i\sum_j p(i,j)\log(p(i,j)) \quad (9.1)$$

8. Difference Variance (h.dva)

$$f_{10} = variance\ of\ p_{x-y} \quad (10)$$

9. Difference Entropy (h.den)

$$f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i)\log\left(p_{x-y}(i)\right) \quad (11)$$

10. Information Measures of Correlation1 (h.f12)

$$f_{12} = \frac{f_9 - HXY1}{\max(HX, HY)} \quad (12)$$

$$HX = -\sum_i p_x(i)\log\left(p_x(i)\right) = entropy\ of\ p_x \quad (12.1)$$

$$HY = -\sum_j p_y(j)\log\left(p_y(j)\right) = entropy\ of\ p_y \quad (12.2)$$

$$HXY1 = -\sum_i\sum_j p(i,j)\log(p_x(i)p_y(j)) \quad (12.3)$$

11. Information Measures of Correlation2 (h.f13)

$$f_{13} = [1 - \exp(-2(HXY2 - f_9))]^{1/2} \quad (13)$$

$$HXY2 = -\sum_i\sum_j p_x(i)p_y(j)\log(p_x(i)p_y(j)) \quad (13.1)$$

Fuzzy C-Means (FCM) algorithm

FCM, which stands for Fuzzy C-Means, is an unsupervised clustering technique that was introduced by Jim Bezdek. Its purpose is to group data points into different clusters in multi-dimensional spaces. The algorithm determines the appropriate cluster for each data point by measuring the distance between the cluster centers

and the data point. In some cases, a data point may belong to more than one cluster. FCM is particularly useful for overlapping data sets, unlike the k-means clustering algorithm. The FCM algorithm assigns each data point a membership value for each cluster center, indicating the degree of membership to the cluster. However, the number of clusters must be specified in advance. Similarity values are calculated based on the cluster center and membership values, which range from 0 to 1, representing the degree of membership between the data and the cluster centers. Similarity and membership values are positively related. FCM clustering is performed using the function given in Equation (14,23-26).

$$min_{U,H} J_{FCM} = \sum_{i=1}^{n}\sum_{g=1}^{c} u_{ig}^{m} d^2(x_i, h_g), \quad (14)$$

$$s.t. \quad u_{ig} \in [0,1], \sum_{g=1}^{c} u_{ig} = 1, \quad (14.1)$$

$d(x_i,h_g)$ is the Euclidean distance. Euclidean distance is a measure of distance between observations and cluster centers. Also, the cluster center for each cluster is called the prototype (27). is the generic element of the matrix U of order (n × c), which takes values between 0 and 1, called the membership degree. are the prototypes (cluster centers) stored in the matrix H of order (c × p) where and p is the number of variables. Finally, the 'm' in the Equation (14) is the weighting exponent for fuzziness.

RESULTS

Initially, the images are transferred to the R system from files downloaded from the Kaggle.com website. The original images used in this study can be viewed in Figure 1. After the transfer process, the necessary packages are installed to perform image processing methods in R. The required packages are library(EBImage), library(raster), and library(png). Thus, R becomes ready for operations to be performed on the images. First, the images are processed using histogram equalization, which is an image enhancement technique. This technique helps to eliminate color distortion caused by the clustering of color values within a specific range and enhances the visibility of unclear details in the images. Histogram equalization is applied separately to each image to create a fair and uniform structure for all the images. The new images obtained after histogram equalization are presented in Figure 2. Following histogram equalization, the new images are converted to grayscale and inverted. The resulting images after this process can be seen in Figure 3.

Figure 1. Original images: (a) Monkeypox (b) Chickenpox (c) Measles

Figure 2. Images after histogram equalization: (a) Monkeypox (b) Chickenpox (c) Measles

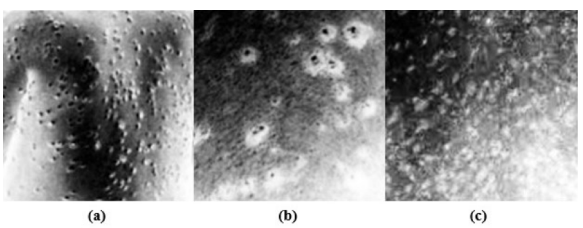

Figure 3. Gray-format inverted images: (a) Monkeypox (b) Chickenpox (c) Measles

The brightness levels of the images are analyzed based on the intensity of light. In order to perform further processing on the images, an appropriate threshold value is determined (0.5) and the images are thresholded accordingly. Thresholding is a critical step in detecting skin lesions in images as it converts the image to binary

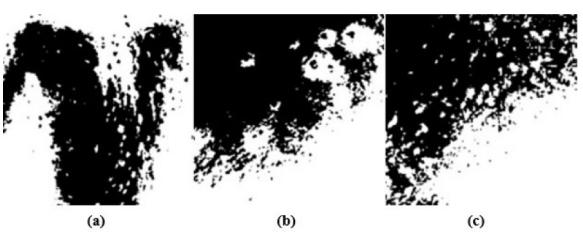

Figure 4. Images after thresholding: (a) Monkeypox (b) Chickenpox (c) Measles

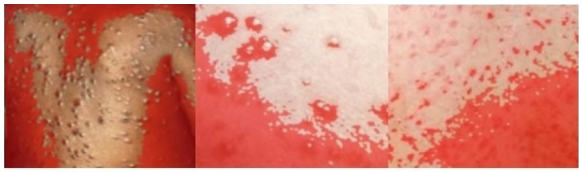

Figure 5. Images after thresholding and after morphological operators: (a) Monkeypox (b) Chickenpox (c) Measles

format. After thresholding, the skin abnormalities and background are completely separated from each other and the images transform into black and white format. The resulting states of the images after thresholding can be viewed in Figure 4. After the mentioned pre-processing steps, we apply some morphological operators to the images to change the pixel sizes of the details for our purpose. We first erode the images and then dilate them which is referred to as the opening process. The application of these morphological operators produces new images, which are shown in Figure 5.

The next step is to calculate the Haralick texture parameters, which are the second-order statistical features of the images and involve 13 of the common

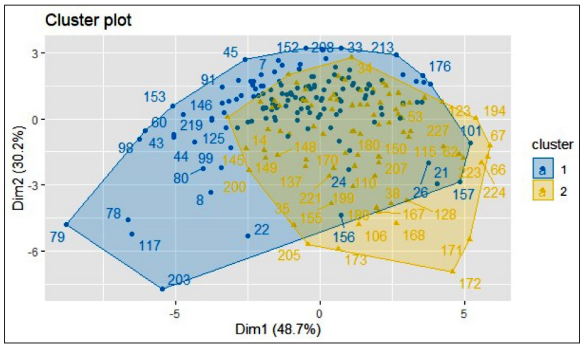

Figure 7. Clusters resulting from analysis with Fuzzy C-Means algorithm

parameters (Details are in the Materials and Methods section). Moreover, a view of the data set created for 13 different Haralick parameters of the images can be seen in Figure 6. These parameters are used as input for the classification process. Based on these parameters, the FCM algorithm analysis results are graphed, which can be presented in Figure 7. It should be noted that the essential packages for fuzzy clustering are used in R. These are: library(cluster), library(fclust), library(dplyr), library(factoextra), library(ppclust).

Filter

	h.asm.s1	h.con.s1	h.cor.s1	h.var.s1	h.idm.s1	h.sav.s1	h.sva.s1	h.sen.s1	h.ent.s1	h.dva.s1	h.den.s1	h.f12.s1	h.f13.s1
1	0.06546286	0.4850268	0.9610611	7.228045	0.8071607	46.177385	2045.81657	1.2175209	1.3749193	0.4850268	0.3354051	0.5507689	0.8051878
2	0.04216555	0.7900239	0.9739247	16.148903	0.7277259	35.548558	1227.48995	1.3771609	1.6248833	0.7900239	0.4094041	0.5100886	0.8193262
3	0.07687494	0.3088386	0.9752187	7.231293	0.8615897	34.088541	1106.05236	1.2033459	1.3051514	0.3088386	0.2742401	0.6457757	0.8437948
4	0.09929914	0.2932062	0.9678593	5.561294	0.8708502	31.979186	970.59693	1.1142386	1.2103192	0.2932062	0.2661223	0.6332227	0.8211103
5	0.06452625	0.3006382	0.9872063	12.749437	0.8609336	43.772231	1851.87886	1.2848010	1.3834896	0.3006382	0.2699369	0.6702297	0.8672199
6	0.04264815	0.5220238	0.9873090	21.566741	0.8495264	37.068795	1351.85149	1.4302407	1.5621060	0.5220238	0.3047595	0.6792942	0.8941500
7	0.04021463	0.5453254	0.9914915	33.046101	0.8378357	28.058333	835.49180	1.4528310	1.5895563	0.5453254	0.3128538	0.6765681	0.8961776
8	0.04243274	2.4724536	0.9010292	13.490819	0.6163937	36.591082	1294.35336	1.2801416	1.6996917	2.4724536	0.5753141	0.3003213	0.6719918
9	0.07740995	0.7826117	0.9900293	40.245537	0.8116642	35.744436	1348.45426	1.2153943	1.3836530	0.7826117	0.3580000	0.5602721	0.8120040
10	0.02762434	0.6840411	0.9853499	24.345883	0.7917980	28.258219	807.44122	1.5235652	1.7214387	0.6840411	0.3678367	0.6197701	0.8870752
11	0.05701780	0.6053939	0.9913648	36.053652	0.7977745	36.616869	1380.42755	1.3916178	1.5750726	0.6053939	0.3641731	0.6064279	0.8637957
12	0.07562758	0.6347055	0.9859701	23.619780	0.7903564	40.362041	1617.62248	1.2752388	1.4633718	0.6347055	0.3743910	0.5607354	0.8247881
13	0.07909278	0.4608732	0.9466634	5.320424	0.8007721	37.223757	1319.91804	1.1253189	1.2850875	0.4608732	0.3295623	0.5050669	0.7618142
14	0.05568135	1.5995519	0.9390187	14.115102	0.7261796	44.442684	1913.97657	1.2788602	1.5843427	1.5995519	0.4817017	0.4243133	0.7576280
15	0.05202921	0.5801109	0.9748538	12.534782	0.7977059	39.465191	1500.62304	1.3198899	1.5129789	0.5801109	0.3625926	0.5634820	0.8336397
16	0.03022385	0.7097171	0.9774588	16.742640	0.7744666	25.844691	656.71602	1.4632393	1.6950983	0.7097171	0.3930724	0.5724025	0.8620705
17	0.03406939	0.7927592	0.9737370	16.092720	0.7640970	27.829431	756.13582	1.4370517	1.6830497	0.7927592	0.4070845	0.5487315	0.8484906
18	0.04058255	0.8401308	0.9707768	15.374362	0.7580272	38.780849	1455.01132	1.3862767	1.6330126	0.8401308	0.4136235	0.5279089	0.8306887
19	0.09266633	0.5158531	0.9491354	6.070843	0.8034517	55.272875	2953.13894	1.1122670	1.2803420	0.5158531	0.3450937	0.5013954	0.7585884
20	0.05437139	0.6241637	0.9791238	15.949181	0.7944027	51.926862	2618.01310	1.3418788	1.5283433	0.6241637	0.3623381	0.5742947	0.8414757
21	0.27261905	0.2944350	0.7143251	1.515332	0.8680110	35.484890	1213.87617	0.6695303	0.7671292	0.2944350	0.2688384	0.3371947	0.5170872

Showing 1 to 22 of 228 entries, 13 total columns

Console

Figure 6. Haralick parameters dataset for the images (h.asm.s1: Angular Second Moment, h.con.s1: Contrast, h.cor.s1: Correlation, h.var.s1: Variance, h.idm.s1: Inverse Difference Moment, h.sav.s1: Sum Average, h.sva.s1: Sum Variance, h.sen.s1: Sum Entropy, h.ent.s1: Entropy, h.dva.s1: Difference Variance, h.den.s1: Difference Entropy, h.f12.s1: Information Measures of Correlation1, h.f13.s1: Information Measures of Correlation2)

Table 1. Accurate classification rates

Fuzzy C-Means Clustering * Original Group Crosstabulation			Original Group		Total
			Group 1	Group 2	
FCM Clustering	Cluster 1	n(%)	77(76.20%)	64(50.40%)	141(61.80%)
	Cluster 2	n(%)	24(23.80%)	63(49.60%)	87(38.20%)
Total		n	101	127	228

FCM: Fuzzy C-Means

Our study aims to classify skin lesion images caused by monkeypox and other virus-induced diseases such as chickenpox and measles using the FCM algorithm. Table 1 presents the results of the classification of skin lesion images. It is worth noting that Table 1 is a crosstable and shows the relationship between the images for which we have preliminary information about the original groups and the new groups created as a result of the FCM algorithm. According to Table 1, the FCM algorithm achieves an accuracy rate of 61.8% in classifying the skin lesion images of monkeypox and other virus-induced diseases (chickenpox and measles). The accuracy rate for the monkeypox class is the highest at 76.2%, while the accuracy rate for the other class is around 50.4%. The accuracy of classification decreases because the other class includes skin lesions caused by both chickenpox and measles viruses, which have similar characteristics to skin lesions caused by monkeypox virus (Table 1). However, empirical findings reveal promising results in detecting the type of skin lesions in the majority of images, with an overall accuracy of 61.8%.

The scarcity of studies using monkeypox image dataset led us to MSLD. The MSLD used in this study is limited and challenging in terms of discriminating monkeypox, chickenpox and measles lesions from each other. Despite this, we reveal the basic clustering structure of MSLD and look at the images from a statistical perspective. Although the classification rates obtained are not very high, our study shows that the FCM algorithm can be used as a method in this field. Moreover, this situation is directly related to the parameters. Because the classification process is performed with Haralick parameters calculated on the images. Therefore, the unique aspect that differentiates our study from other studies using MSLD is that we can present Haralick parameters.

CONCLUSION AND DISCUSSION

Clustering methods are used to discover unknown behavior of data. Thus, valuable information about the data set can be obtained by revealing complex relationships between observations. When our study is looked at in general terms, we aim to make a classification. We propose to use a very different classification algorithm that operates in the same manner as the general clustering algorithm in an unsupervised machine learning task in statistics. The FCM algorithm, we use for classification is also like this. Firstly, it should be noted that no other study using the fuzzy method for the human monkeypox classification has been found in the literature. More specifically, the unique aspect of this study is that it explores not only the features of a single-pixel but also the link of each pixel with neighboring pixels through Haralick parameters. While studies on Haralick texture parameters are already scarce in the literature, texture analysis of human monkeypox images has not been mentioned at all. Therefore, this study can be an inspiration to many scholars conducting research in different fields with its significant contributions to the existing literature.

Ahsan et al. analyze gray-level histograms of images from monkeypox, chickenpox, and measles datasets. This preprint article is MSLD's first study in this field and has inspired us to plan our workflow. It can be seen that some methods such as deep learning are used to distinguish human monkeypox virus skin lesion images from other viral diseases with similar images. For example; the classification performances of deep learning models are examined in the study of Haque et al. and Ali et al. However, a different approach called the FCM algorithm is used to sort out human monkeypox skin lesion images from other viral diseases with similar appearances in our study. The success of

the fuzzy algorithm, in which cluster memberships are based on certain probabilities, in classifying diseases is presented in detail. Sadad et al. use the FCM algorithm to locate tumors on mammograms, opting for various machine-learning procedures for classifying benign and malignant tumors. However, unlike Sadad et al., our study addresses the FCM algorithm in the process of classifying viral diseases with similar skin lesion appearance. Additionally, there are plenty of studies on classifying images using fuzzy algorithms in X-ray images, as in the study of Rohmayani and Rahayu. In these studies, various image processing techniques are applied to the images, but generally, histogram equalization is not applied at first. It can be noted that, unlike other studies, to be fair to the images in our dataset, histogram equalization is done on the images in RGB format first. Moreover, in Rohmayani and Rahayu's paper, the classification process is based on basic statistical criteria such as the gray level mean value of the images. However, we classify the images according to 13 haralick texture parameters, known as second-order feature descriptors. The MSLD, which we used, is similar to the data set used by Sahin et al. However, we design a workflow that differs from Sahin et al. in terms of both the image processing steps applied and the classification model used. In Zayed and Elnemr's study, the haralick parameters calculated after various image processing steps are tested with ANOVA and statistically significant results are found. Our study contains more Haralick texture features (13 parameters) than in the paper by Zayed and Elnemr. Additionally, we classify these features with the FCM algorithm and mention the performance of the algorithm.

The FCM algorithm is an unsupervised machine learning technique that reveals the basic structure of the data set (numerical, categorical, mixed or image data). Applications of this algorithm abound on numerical and categorical or mixed data sets. However, there is no study in the literature that processes image data with this technique and reveals the basic characteristics of images. The literature mostly constructs the analysis of image data (including monkeypox image data) on various CNN architectures, convolutional artificial neural networks and deep learning. The classification performances of such studies are high, but there is always a hidden issue and a question always comes to mind. The question is, what are the parameters used in these studies and this question remains unanswered. Algorithms do not output which parameters are used for classification, the machine runs the process automatically in the background. This is where the importance of our study comes into play. In our study, the images are digitized and the parameters used for classification are presented as output, which are the Haralick parameters. In addition, until this stage, the images are processed in many image pre-processing steps and we do this manually with R program codes in line with our purpose.

This study shows that the proposed approaches can lead to higher accuracy rates in skin lesion classification when applied to more images. Moreover, including images of other similar diseases can improve the study. Choosing different classification algorithms can also affect accuracy levels, and comparing the performance of different algorithms can be a useful approach. Overall, we believe that our research will inspire similar studies using image processing applications and classification algorithms with the help of Haralick texture parameters.

Conflict-of-interest and financial disclosure

The authors declare that they have no conflict of interest to disclose. The authors also declare that they did not receive any financial support for the study.

REFERENCES

1. Magnus P, Andersen EK, Petersen KB, Birch-Andersen A. A pox-like disease in cynomolgus monkeys. Acta Pathol Microbiol Scand. 1959;46:156-76.
2. Beer EM, Rao VB. A systematic review of the epidemiology of human monkeypox outbreaks and implications for outbreak strategy. PLoS Negl Trop Dis. 2019;13(10):0007791.
3. Bunge EM, Hoet B, Chen L, et al. The changing epidemiology of human monkeypox - a potential threat? A systematic review. PLoS Negl Trop Dis. 2022;16(2):0010141.
4. Ladnyj ID, Ziegler P, Kima E. A human infection caused by monkeypox virus in Basankusu Territory, Democratic Republic of the Congo. Bull World Health Organ. 1972;46(5):593-7.
5. Mauldin MR, McCollum AM, Nakazawa YJ, et al. Ex-

portation of monkeypox virus from the African continent. J Infect Dis. 2022;225(8):1367-76.
6. Nguyen PY, Ajisegiri WS, Costantino V, Chughtai AA, MacIntyre CR. Reemergence of human monkeypox and declining population immunity in the context of urbanization, Nigeria, 2017–2020. Emerg Infect Dis. 2021;27:1007-14.
7. Ogoina D, Izibewule JH, Ogunleye A, et al. The 2017 human monkeypox outbreak in Nigeria—Report of outbreak experience and response in the Niger Delta University Teaching Hospital, Bayelsa State, Nigeria. PLoS One. 2019;14(4):0214229.
8. CDC. *Infection prevention and control of monkeypox in healthcare settings*. 2022. Accessed 25 January 2023, https://www.who.int/news-room/fact-sheets/detail/monkeypox
9. WHO. *Monkeypox*. 2023. Accessed 25 January 2023, https://www.who.int/news-room/fact-sheets/detail/monkeypox
10. Ahsan MM, Uddin MR, Luna SA. *Monkeypox image data collection*. 2022. Accessed 25 January 2023, https://doi.org/10.48550/arXiv.2206.01774
11. Haque ME, Ahmed MR, Nila RS, Islam S. *Classification of human monkeypox disease using deep learning models and attention mechanisms*. Electrical Engineering and Systems Science, Image and Video Processing. 2022. Accessed 25 January 2023, https://arxiv.org/abs/2211.15459
12. Ali SN, Ahmed MT, Paul J, Jahan T, et al. *Monkeypox skin lesion detection using deep learning models: A feasibility study*. 2022. Accessed 25 January 2023, https://doi.org/10.48550/arXiv.2207.03342
13. Sahin VH, Oztel I, Yolcu Oztel G. Human monkeypox classification from skin lesion images with deep pre-trained network using mobile application. J Med Syst. 2022;46(11):1-10.
14. Sadad T, Munir A, Saba T, Hussain A. Fuzzy C-Means and region growing based classification of tumor from mammograms using hybrid texture feature. J Comput Sci. 2018;29:34-45.
15. Rohmayani D, Rahayu AH. Classification of X-ray images of normal, pneumonia, and COVID-19 lungs using the Fuzzy C-means (FCM) algorithm. J Appl Intell Syst. 2022;7(1):16-25.
16. Zayed N, Elnemr HA. Statistical analysis of Haralick texture features to discriminate lung abnormalities. Int J Biomed Imaging. 2015; 2015: 1-7.
17. Kaggle. *Monkeypox Skin Lesion Dataset*. 2022. Accessed 27 July 2022, https://www.kaggle.com/datasets/nafin59/monkeypox-skin-lesion-dataset
18. Serra J. Image analysis and mathematical morphology. (1982), New York: Acad. Press.
19. Haralick RM, Shanmugam K, Dinstein IH. Textural Features for Image Classification. IEEE Trans Syst Man Cybern. 1973;3(6):610-21.
20. Zulpe N, Pawar V. GLCM textural features for brain tumor classification. Int J Comput Sci Issues. 2012;9(3):354.
21. Miyamoto E, Jr. Merryman T. Fast Calculation of Haralick Texture Features .2008. Accessed 25 January 2023, https://people.inf.ethz.ch/markusp/teaching/18-799B-CMU-spring05/material/eizan-tad.pdf
22. Pham TA. *Optimization of Texture Feature Extraction Algorithm*. 2010. Accessed 25 January 2023, https://repository.tudelft.nl/islandora/object/uuid%3Aa7924113-c9f8-435d-824f-0232ff6b419c
23. Bezdek JC. Pattern recognition with fuzzy objective function algorithms. (1981) New York: Plenum Press.
24. Bora DJ, Gupta D, Kumar A. A comparative study between fuzzy clustering algorithm and hard clustering algorithm. 2014, Accessed 25 January 2023, https://doi.org/10.48550/arXiv.1404.6059
25. Cinar A, Tuncer T. Segmentation of urban images with Fuzzy C-Means. J Comput Sci Tech. 2021;1(1):01-06.
26. Ferraro MB, Giordani P, Serafini A. fclust: An R package for Fuzzy clustering. The R Journal. 2019;11:2073-4859.
27. Hoppner F, Klawonn F, Rudolf K, Runkler T (1999), Fuzzy cluster analysis: Methods for classification data analysis and image recognition. John Wiley & Sons.