International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Comparative Analysis of Skin Cancer Image with Classification and Clustering Algorithms

Muhammed Kara[a*], Yüksel Terzi[b] ,Mehmet Ali Cengiz [c]

[a,b,c]*Ondokuz Mayıs University, Atakum,Samsun,55270,Turkey*
[a]*Email: karam@omu.edu.tr*
[b]*Email: yukselt@omu.edu.tr*
[c]*Email: macengiz@omu.edu.tr*

Abstract

Skin cancer is one of the most common and potentially life-threatening diseases worldwide. Early detection and accurate diagnosis are crucial for effective treatment and improved patient outcomes. In recent years, the integration of advanced technologies, such as artificial intelligence and image analysis, has revolutionized the field of dermatology. This article presents a comprehensive comparative analysis of algorithms for classifying and clustering skin cancer images. The goal is to improve the accuracy and efficiency of skin cancer diagnosis.

The study explores various machine learning algorithms used for skin cancer image classification, such as support vector machines (SVM), decision trees, and k-nearest neighbors (KNN). These algorithms are evaluated based on their capacity to distinguish between benign and malignant skin lesions, with a particular emphasis on sensitivity, specificity, and accuracy. Apart from classification, clustering algorithms are also examined to determine their potential in grouping similar skin lesions. This can assist dermatologists in identifying patterns and anomalies within extensive datasets. K-means, hierarchical clustering, and DBSCAN are among the algorithms assessed for their effectiveness in organizing images of skin cancer.

The comparative analysis in this article aims to provide insights into the strengths and weaknesses of various algorithms, their computational efficiency, and their performance on diverse datasets. Furthermore, it explores the potential of combining classification and clustering techniques to develop a skin cancer diagnosis system that is more robust and accurate.

Keywords: Data science;machine learning;orange image processing;statistics.

Received: 10/14/2023
Accepted: 12/14/2023
Published: 12/24/2023

---

* Corresponding author.

1. Introduction

Skin cancer represents a growing global health concern, with an increasing incidence that requires accurate and timely diagnosis for effective treatment and patient well-being. With the emergence of advanced technologies and machine learning techniques, the field of dermatology has undergone a significant transformation. This has enabled more accurate and efficient analysis of skin cancer lesions. This article embarks on a comprehensive journey to explore the world of skin cancer image analysis through the lens of classification and clustering algorithms. It seeks to provide a thorough comparative analysis of their applications and potential in enhancing diagnostic accuracy [1].

Skin cancer, one of the most common types of cancer globally, includes various subtypes such as melanoma, squamous cell carcinoma, and basal cell carcinoma. The early identification and differentiation of these lesions, particularly between benign and malignant cases, play a crucial role in the successful treatment and management of the disease. To this end, advanced machine learning algorithms have emerged as powerful tools in the dermatologist's arsenal [2].

In this context, we explore the field of machine learning, specifically focusing on classification algorithms such as convolutional neural networks (CNNs), support vector machines (SVM), decision trees, and k-nearest neighbors (KNN). These algorithms play a vital role in accurately categorizing skin lesions. By evaluating these algorithms based on their sensitivity, specificity, and overall accuracy, we aim to determine their suitability for the complex task of diagnosing skin cancer.

Beyond classification, the article expands its scope to include clustering algorithms such as K-means, hierarchical clustering, and DBSCAN. These algorithms are being explored for their potential in identifying patterns within datasets of skin cancer images, ultimately offering a complementary approach to understanding the complexity of the disease. Clustering techniques offer the opportunity to group similar skin lesions, assisting dermatologists in identifying common characteristics and abnormalities in large image datasets.

One of the most promising advancements in skin cancer diagnosis is the utilization of machine learning algorithms for image classification. Notably, the following references have made significant contributions to this area of research. The use of deep neural networks for dermatologist-level classification of skin cancer was demonstrated, achieving impressive accuracy. This paper highlights the potential of convolutional neural networks (CNNs) in the analysis of skin cancer images. Haenssle, H. A. et al. conducted a study comparing the diagnostic performance of a deep learning model with that of 58 dermatologists. This work provides invaluable insights into the capabilities and limitations of machine learning in dermatology [3]. Codella, N. et al. explored the combined use of deep learning, sparse coding, and support vector machines (SVM) for melanoma recognition in dermoscopy images. It highlights the integration of various machine learning techniques for the diagnosis of skin cancer [4].

Accurate border detection is crucial in skin cancer image analysis, as it aids in precise diagnosis and classification. Celebi, M. E. et al. (2013) discuss border detection in dermoscopy images using statistical region merging. Proper border detection is vital for accurate diagnosis and classification [5].

Beyond classification, clustering algorithms play a role in identifying patterns within datasets of skin cancer images. Pathan, S. et al. (2017) investigated both clustering and classification techniques for skin cancer in dermoscopic images. The study addresses the application of clustering methods in conjunction with classification algorithms to improve the analysis of skin cancer images. In the broader context of healthcare and cancer diagnosis, machine learning has made substantial contributions [6]. Kourou, K. et al. (2015) provide a broader perspective on the application of machine learning in cancer diagnosis and prognosis. This paper discusses the potential impact of machine learning techniques in enhancing healthcare practices, including the early diagnosis of skin cancer [7].

A solid understanding of clustering algorithms is essential for effectively organizing and grouping skin cancer images. Jain, A. et al. (2014) present a survey of clustering algorithms and their applications in handling big data, which is relevant for effectively organizing and grouping skin cancer images. For a fundamental understanding of pattern recognition and classification algorithms [8]. Duda, R. O. et al. (2012) offer a comprehensive text on pattern classification, encompassing the principles and techniques used in classification algorithms. This text is fundamental for understanding the field [9].

Benchmarking is crucial for evaluating machine learning models for image classification. . Russakovsky, O. et al. (2015) provide insights into benchmarking and evaluating machine learning models for image classification. The ImageNet challenge has been instrumental in advancing the development of image recognition algorithms [10]. For a comprehensive resource on pattern recognition and machine learning, Bishop, C. M. (2006) covers a wide range of topics relevant to the application of machine learning algorithms in various domains, including medical image analysis [11].

Our comparative analysis is designed to uncover the strengths and limitations of these various algorithms, their computational efficiency, and their performance on a variety of skin cancer datasets. Furthermore, we investigate the potential for synergy between classification and clustering techniques in order to develop a more robust and accurate. This approach has the potential to redefine dermatological practices and inspire the development of computer-aided diagnostic systems. These systems can significantly improve patient care by assisting healthcare professionals in the early detection and treatment of skin cancer.

The insights gleaned from this study are expected to have far-reaching implications, not only in the field of dermatology but also in the broader context of healthcare and medical research. By enabling the selection of the most suitable algorithms for specific tasks in skin cancer image analysis, our research aims to empower healthcare practitioners and researchers. This will foster advancements in the field and ultimately contribute to the preservation of lives through early intervention and precise diagnosis.

2. Machine learning algorithms

Machine learning algorithms designed for skin cancer diagnosis can be categorized into two primary domains: classification and clustering. Classification algorithms, such as support vector machines and convolutional neural networks, excel in the categorization of skin lesions, assigning them to classes such as benign, malignant, or

specific cancer subtypes. On the other hand, clustering algorithms, such as K-means and hierarchical clustering, are responsible for organizing and grouping similar skin lesions. They aim to reveal underlying patterns and structures within extensive dermatological image datasets. The unique synergy between these two categories of algorithms allows for a comprehensive understanding of skin cancer, assisting dermatologists in both accurate diagnosis and the identification of subtle similarities and anomalies within the disease spectrum. Together, these machine learning tools are revolutionizing the way skin cancer is detected, classified, and managed. They offer a glimpse into the future of healthcare, where data-driven precision becomes the gold standard.

2.1. K-Means

K-means is a popular clustering algorithm in machine learning and unsupervised learning. It is used to group a set of data points into clusters based on their similarity. The goal of K-means clustering is to partition the data into K clusters, where K is a user-defined parameter. Here is how the K-means algorithm works:

1. Initialization: Choose K initial cluster centroids. These centroids can be randomly selected data points or pre-defined locations in the feature space.

2. Assignment: Assign each data point to the nearest cluster centroid. This is typically done by calculating the Euclidean distance (or other distance metrics) between each data point and the centroids, and selecting the cluster with the nearest centroid.

3. Update: Recalculate the cluster centroids by taking the mean (average) of all data points assigned to each cluster. This moves the centroids to the center of their respective clusters.

4. Repeat: Steps 2 and 3 are iteratively repeated until a stopping criterion is met. Common stopping criteria include a maximum number of iterations or when the centroids no longer change significantly.

5. Output: The final cluster centroids represent the centers of the K clusters, and each data point is assigned to one of these clusters.

K-means aims to minimize the within-cluster variance, which means that it tries to make the data points within a cluster as similar as possible while keeping different clusters as dissimilar as possible. It is an iterative optimization algorithm, and the final cluster assignments depend on the initial centroids and the specific distribution of the data.

K-means is widely used in various applications, such as image segmentation, customer segmentation, anomaly detection, and more. However, it has some limitations, such as sensitivity to the initial centroid placement and the need to specify the number of clusters (K) in advance. There are also variations of K-means, such as K-means++, which address some of these issues.

2.2. Hierarchical Clustering

Hierarchical clustering is a general family of clustering algorithms that form nested clusters by successively combining or decoupling them. This hierarchical set is represented as a tree (or dendrogram). The root of the tree is the unique set that contains all the samples, while the leaves are the sets that each contain only one sample.

Hierarchical clustering is a machine learning technique used in unsupervised learning to construct a hierarchy of clusters. It creates a tree-like structure of clusters, also known as a dendrogram, where the data points are successively merged into smaller or larger groups based on their similarities. Hierarchical clustering can be classified into two types: Agglomerative (bottom-up) and Divisive (top-down).

Agglomerative Hierarchical Clustering is the more commonly used type of hierarchical clustering. It starts with each data point as its own cluster and repeatedly merges the closest clusters together until only one cluster remains, encompassing all the data points. The steps for agglomerative hierarchical clustering are as follows:

· Initialize each data point as a single cluster.

· Calculate the distance or dissimilarity between all pairs of clusters (e.g., using Euclidean distance, Ward's linkage, single linkage, complete linkage, etc.).

· Merge the two closest clusters into a new cluster.

· Repeat steps b and c until there is only one cluster remaining.

In contrast, divisive hierarchical clustering starts with all data points in a single cluster and recursively divides them into smaller clusters until each data point is in its own cluster. This method is less commonly used in practice.

One of the advantages of hierarchical clustering is that it provides a visual representation of the natural grouping of data through the dendrogram. This can be helpful in understanding the hierarchy of clusters at various levels. Hierarchical clustering is a useful technique for exploratory data analysis as it allows you to choose a specific level in the dendrogram to determine the optimal number of clusters for your data.

Hierarchical clustering does not require specifying the number of clusters beforehand, which is a common limitation of other clustering algorithms like K-means. However, it can be computationally expensive, especially for large datasets, and the choice of linkage method and distance metric can significantly impact the results.

The agglomerative cluster exhibits a "richer gets richer" behavior, which results in unequal cluster sizes. In this regard, the single link is the worst strategy, and Ward provides the most consistent results. However, the affinity (or the distance used in clustering) cannot be replaced by Ward. For non-Euclidean metrics, the average connection is a good alternative. A single connection can also perform well on non-global data [12].

2.3. Density-Based Clustering (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a popular clustering algorithm used in machine learning and data mining. It is particularly effective for finding clusters in data that have varying shapes and densities. Developed by Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu in 1996, DBSCAN works by identifying dense regions of data points as clusters while also distinguishing outliers as noise points.

DBSCAN does not require a predetermined number of clusters. Instead, it operates based on two parameters: the Epsilon (ε) parameter defines the radius within which to search for nearby data points. It is also known as the neighborhood distance. The MinPoints (MinPts) parameter specifies the minimum number of data points that must be within the ε-neighborhood of a point for it to be considered a core point. A core point is a data point that has at least MinPts data points within its ε-neighborhood. In other words, there are enough data points surrounding it to form a dense region. A border point is not a core point itself, but it lies within the ε-neighborhood of a core point. It is on the boundary of a cluster. Data points that are neither core points nor border points are classified as noise points. These are often considered outliers.

The algorithm proceeds as follows:

· Start with an arbitrary, unvisited data point.

· If the point is a core point, create a new cluster and add all reachable points (both core and border points) to the cluster.

· Continue until no more core points can be added to the cluster.

· Repeat the process with an unvisited data point until all data points are either assigned to clusters or marked as noise.

In summary, DBSCAN is a powerful density-based clustering algorithm that can be quite useful for many clustering tasks, particularly when working with datasets that contain clusters of varying shapes and densities. Proper parameter selection and understanding of your data are essential for a successful application.

They are sets of algorithms in which high-density regions are separated from low-density areas. The clusters found by DBSCAN can be of any shape, unlike k-means, which assumes that the clusters are convex-shaped. The central component of DBSCAN is the concept of core samples, which are samples located in areas of high density. The algorithm has two parameters: min_samples and eps. (The eps parameter represents the maximum distance a point will consider when selecting its neighbors.) Higher values of min_samples or lower values of eps indicate a higher density required to form a cluster.

2.4. KNN

K-nearest neighbors is a simple algorithm that stores all existing cases and classifies new cases based on a similarity measure, such as distance functions. KNN was used in statistical estimation and pattern recognition as a non-parametric technique in the early 1970s [13].

K-Nearest Neighbors (KNN) is a popular machine learning algorithm used for both classification and regression tasks. It is a non-parametric, instance-based learning algorithm, which means that it makes predictions based on the data points in its vicinity without creating a model.

There is no explicit training phase in KNN. The algorithm simply stores the entire dataset, which will be used to make predictions. KNN requires you to specify the number of nearest neighbors (K) to consider when making a prediction. This is a hyperparameter that you need to tune. A smaller value of K (e.g., K=1) makes the algorithm more sensitive to noise, while a larger value of K can smooth out the predictions but might lead to some loss of detail.

For a given data point that you want to classify, KNN finds the K-nearest data points in the training set based on a distance metric, typically the Euclidean distance. It counts the number of data points in each class among the K-nearest neighbors. The data point is then assigned to the class that is most common among its K-nearest neighbors. For regression tasks, KNN works similarly, but instead of predicting a class, it predicts a numerical value. It calculates the average (or weighted average) of the target values of the K-nearest neighbors and assigns this average as the predicted value for the data point.

KNN is a valuable algorithm, particularly for small to medium-sized datasets, when you need a straightforward yet effective approach for classification and regression tasks. However, you should carefully consider its limitations and use it when it is appropriate for your specific problem.

2.5. SVM-Support Vector Machines

A Support Vector Machine (SVM) is a popular supervised machine learning algorithm used for classification and regression tasks. Support Vector Machines (SVM) are particularly useful when dealing with complex, high-dimensional datasets. They are known for their ability to find optimal hyperplanes that can effectively separate data into different classes. SVM aims to maximize the margin between classes while minimizing classification errors. Support vector machine (SVM) performs classification by finding the hyperplane that maximizes the margin between two classes. The vectors that define the hyperplane (cases) are the support vectors.

SVM is a powerful algorithm for various applications, including text classification, image recognition, and bioinformatics. When using SVM, careful tuning of hyperparameters and selecting the appropriate kernel are crucial for achieving optimal performance on your specific problem.

2.6. Logistic Regression

Logistic regression is a machine learning algorithm used for binary classification tasks. It is a statistical method that models the probability of a given input belonging to one of two possible classes. Despite its name, logistic regression is a classification algorithm, not a regression algorithm.

Logistic regression utilizes the logistic function, commonly referred to as the sigmoid function, to estimate the probability of a specific input belonging to the positive class (e.g., class 1). The sigmoid function is an S-shaped curve that maps any real-valued number to a value between 0 and 1. The typical approach to finding the optimal parameters is through an optimization algorithm, such as gradient descent. The objective is to minimize a cost function that measures the disparity between the predicted probabilities and the actual class labels. The cost function commonly used in logistic regression is the cross-entropy or log loss.

Logistic regression is widely used in various applications, including spam email detection, disease prediction, credit risk assessment, and more. It is a straightforward and interpretable model that can provide insights into the relationship between input features and the probability of a binary outcome. Extensions of logistic regression include multinomial logistic regression for multiclass classification and ordinal logistic regression for ordered categories [14].

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum entropy classification (MaxEnt), or log-linear classifier. In this model, the probabilities that describe the possible outcomes of a single experiment are modeled using a logistic function [15].

2.7. Artificial Neural Network(Neural Network)

Artificial Neural Networks (ANNs), often referred to simply as neural networks, are a class of machine learning models inspired by the structure and function of the human brain. ANNs consist of interconnected artificial neurons or nodes organized into layers. They are a fundamental component of deep learning, which is a subfield of machine learning that focuses on models with many layers.

Neurons are the fundamental building blocks of a neural network. Each neuron receives one or more input values, performs computations on these inputs, and produces an output. Neurons are organized into layers. A typical neural network consists of an input layer, one or more hidden layers, and an output layer. The input layer represents the data's features, the hidden layers process these features, and the output layer produces the final prediction or result. Each connection between neurons is associated with a weight that represents the strength of the connection. Each neuron also has a bias term, which allows for a shift in the activation function.

The output of each neuron is determined by applying an activation function to the weighted sum of its inputs and bias. Common activation functions include the sigmoid, ReLU (Rectified Linear Unit), and tanh. The input data is fed into the input layer. Calculations are made layer by layer, propagating information forward through the network. Each neuron computes its output using its weighted inputs, bias, and activation function.

A loss function measures the discrepancy between the predicted output and the actual target values. Common loss functions include Mean Squared Error (MSE) for regression tasks and Cross-Entropy for classification tasks. The network adjusts its weights and biases to minimize the loss function using a process called backpropagation. Gradients are computed with respect to the loss, and the weights and biases are updated using gradient descent or one of its variants. The training process involves iteratively presenting the training data to the network, calculating the loss, and updating the parameters of the model. The goal is to find the optimal weights and biases that minimize the loss function. Once trained, the neural network can make predictions on new, unseen data by performing forward propagation using the learned weights and biases.

Neural networks can vary significantly in size and architecture, depending on the problem they are designed to solve. Deep Neural Networks (DNNs) have multiple hidden layers and are particularly effective at learning hierarchical representations of data. Neural networks, especially deep learning models, have had a significant impact on a wide range of applications and have become a fundamental technology in modern machine learning.

An artificial neural network is computer software created by modeling the way the human brain works [16]. In recent years, many researchers have argued that artificial neural networks (ANN) are effective in various areas, including the classification of cancer types, activity recognition, and image processing [17]. Brain cells, known as neurons, form a network by connecting to each other through synapses. The weights of all connections (synapses) in the neural network are not the same. The computational power of an artificial neural network is determined by the connectivity between nodes [18]. ANN's perform learning by deciphering different forms of relationships between data. This information obtained is used to create a rule when determining the class of new data [19].

3. Skın Cancer Image Classıfıcatıon

The datasets used were downloaded from the Isic archive website (https://www.isic-archive.com /#!/topWithHeader/onlyHeaderTop/gallery?filter=%5B%5D), comprising total 285 images related to three different types of skin cancer selected. Information regarding the distinct skin cancer types is provided as follows:

Table 1

nv	Nevus, melanositic nevies	125 images
vasc	Vasculer	72 images
df	Dermatofibroma	88 images

Machine learning algorithms were applied to these images using the Orange program. Several classification and clustering algorithms have been applied on images. When we digitize the image data using the analysis inception-v3 technique, the achievements of the classification algorithms are presented in Table 2 and Table 3. The algorithms have yielded successful results.

Table 2: Classification Algorithms

Model	AUC	CA	F1	Precision	Recall
KNN	0.962	0,884	0,883	0,887	0,884
SVM	0,99	0,905	0,904	0,904	0,905
Neural Network	0,988	0,919	0,918	0,918	0,919
Logistik Regression	0,976	0,902	0,902	0,902	0,902

Table 3: Model comparison with AUC

	kNN	SVM	Neural Network	Logistik Regresssion
kNN		0,014	0,009	0,085
SVM	0,986		0,759	0,948
Neural Network	0,991	0,241		0,952
Logistik Regression	0,915	0,052	0,048	

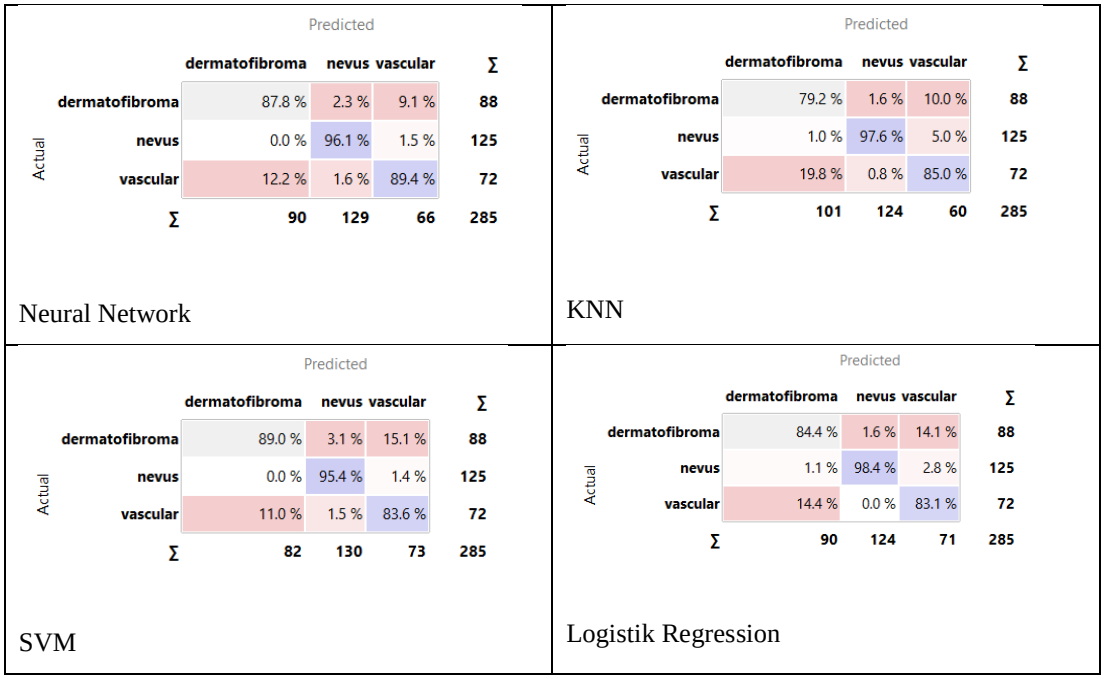

Figure 1: Confusion matrix values according to the inception-v3 method

Figure 1 illustrates the predictive performance of various machine learning models in the field of dermatology. Specifically, it demonstrates that the Support Vector Machine (SVM) achieved the highest level of accuracy, with a score of 89 percent, when predicting dermatofibroma values. Furthermore, the neural network displayed exceptional predictive accuracy for vascular values, achieving an impressive accuracy rate of 89.4 percent. Logistic regression outperformed other models by achieving a remarkable 98.4 percent accuracy when predicting nevus values. These predictions were generated using data embedding techniques in conjunction with the Inception-v3 model.

To provide a more detailed explanation, it is important to understand the significance of these results within the field of dermatology. Dermatofibroma, vascular, and nevus are three distinct skin conditions or types of skin lesions. The models being discussed here are used to make predictions about these conditions based on various input data, such as images or diagnostic information. The Support Vector Machine (SVM) is a machine learning algorithm known for its effectiveness in classifying data points into different classes. In this case, it was highly accurate (89%) in identifying and predicting cases of dermatofibroma, a specific skin condition. Neural networks, which are a subset of deep learning, excel in recognizing patterns in data. Here, the neural network achieved an impressive accuracy of 89.4% in predicting cases of vascular conditions. This suggests that it was highly proficient in distinguishing this specific skin condition from others. Logistic regression is a statistical model used for binary classification tasks. In this context, it outperformed other models with an exceptional 98.4% accuracy in predicting cases of nevus, another skin condition. This high level of accuracy implies that the model was very reliable in distinguishing nevi from other skin conditions.The use of the Inception-v3 model for data embedding implies that the input data, such as images of skin lesions, underwent specific processing and transformation to be compatible with these machine learning models. Inception-v3 is a popular convolutional neural network architecture that is often utilized for image classification tasks. These results indicate that the mentioned machine learning models, when equipped with data embedding using Inception-v3, demonstrated impressive predictive accuracy for various skin conditions. SVM excelled in dermatofibroma, the neural network performed remarkably well for vascular cases, and logistic regression showed exceptional accuracy in nevus predictions. This information is valuable in the field of dermatology for enhancing the accuracy of skin condition diagnoses.

4.Skın Cancer Image Clusterıng

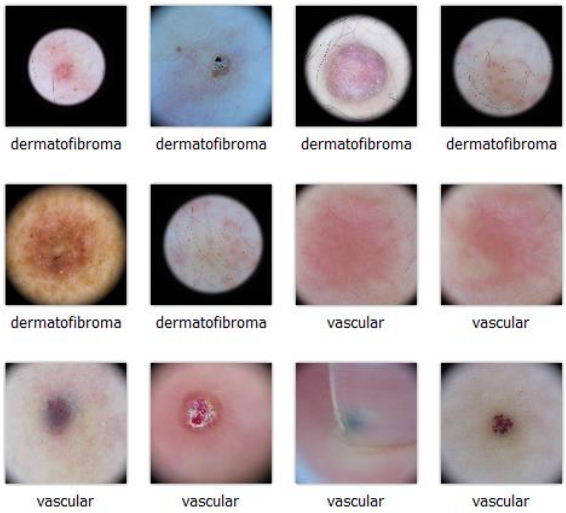

Figure 2: The K-Means C1 Set

Figure 2 presents a specific set of data with 73 images under consideration. Out of these 73 images, it becomes evident that 15 of them are associated with vascular lesions, while the majority, comprising 58 images, are linked to dermatofibroma lesions. This collection of images is referred to as the "C1 cluster."To provide a more comprehensive understanding, this information suggests that there are two distinct categories of skin lesions being

examined: vascular lesions and dermatofibroma lesions. Among these categories, there are 15 images representing vascular lesions, while the larger group of 58 images corresponds to dermatofibroma lesions. This clustering, referred to as "C1," categorizes these images based on their shared characteristics or features, which are likely identified through a clustering or classification process.

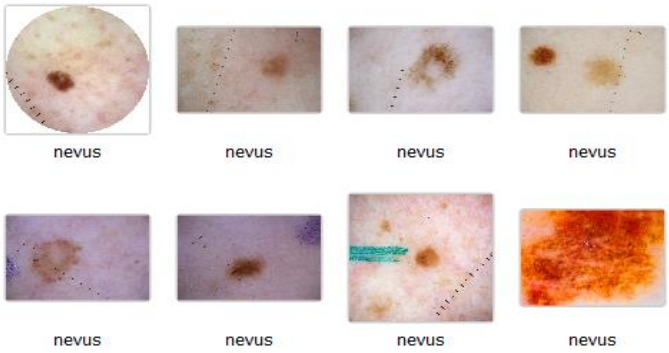

Figure 3: The K-Means C2 cluster

In Figure 3, we observe a specific dataset containing a total of 101 images. Within this dataset, a striking pattern emerges: 97 of these images are associated with nevus lesions, while the remaining 4 images are linked to dermatofibroma lesions. This specific collection of images is classified as a "C2 cluster."To provide a more detailed explanation, this information conveys that we are examining two primary categories of skin lesions in this dataset: nevus and dermatofibroma lesions. The predominant category, which accounts for the majority of the images (97 out of 101), corresponds to nevus lesions. In contrast, there are only 4 images within the dataset that belong to the dermatofibroma lesion category. The term "C2 cluster" is used to group these images together based on common characteristics or traits, which are likely determined through a clustering or classification process.

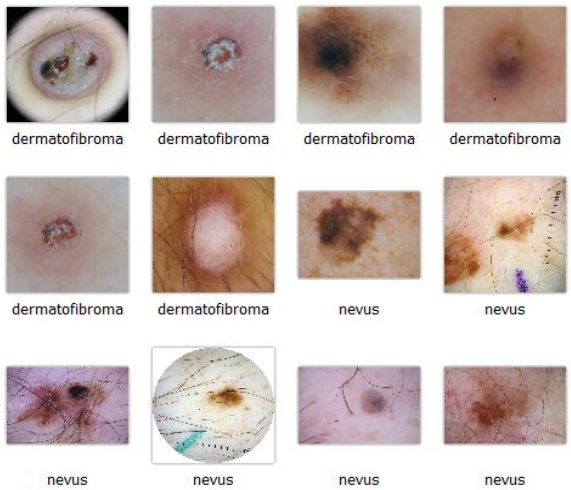

Figure 4: The K-Means C3 cluster

Figure 4 provides a summary of a dataset comprising a total of 111 images. Within this dataset, there is a noticeable

distribution of these images across three distinct clusters. These clusters are labeled as C1 for dermatofibroma (26 images), C2 for nevus (29 images), and C3 for vascular (56 images). The categorization of these images into clusters is based on the outcomes of an algorithm that analyzes and groups images with shared characteristics or patterns. To elaborate further, here is a breakdown of the three clusters:

C1 cluster contains 26 images, all of which represent dermatofibroma lesions. C2 cluster, labeled C2, comprises 29 images, all of which depict nevus lesions. The largest cluster, C3, consists of 56 images, all of which pertain to vascular lesions.The success of this clustering approach can be quantified by evaluating how accurately it assigns images to their respective clusters. In this case, the algorithm correctly clustered 211 out of the total 285 images. This success rate amounts to 74 percent, indicating that the algorithm achieved a 74 percent accuracy in grouping the images based on their shared characteristics. In summary, Figure 10 provides insights into the clustering of 111 images into three distinct clusters (C1, C2, and C3) based on their respective skin lesion types: dermatofibroma, nevus, and vascular. The algorithm achieved a 74 percent success rate by correctly assigning most of the images to their appropriate clusters, thereby demonstrating the effectiveness of the clustering process.

Figure 5 presents a scatter plot resulting from the application of the DBSCAN algorithm. When you focus on a specific cluster located in the lower left area of the plot, you can identify a subset of images. In this context, selecting this cluster leads to the identification of 43 nevus images, as shown in Figure 6. In simpler terms, the DBSCAN algorithm is used to create a scatter plot in Figure 5.

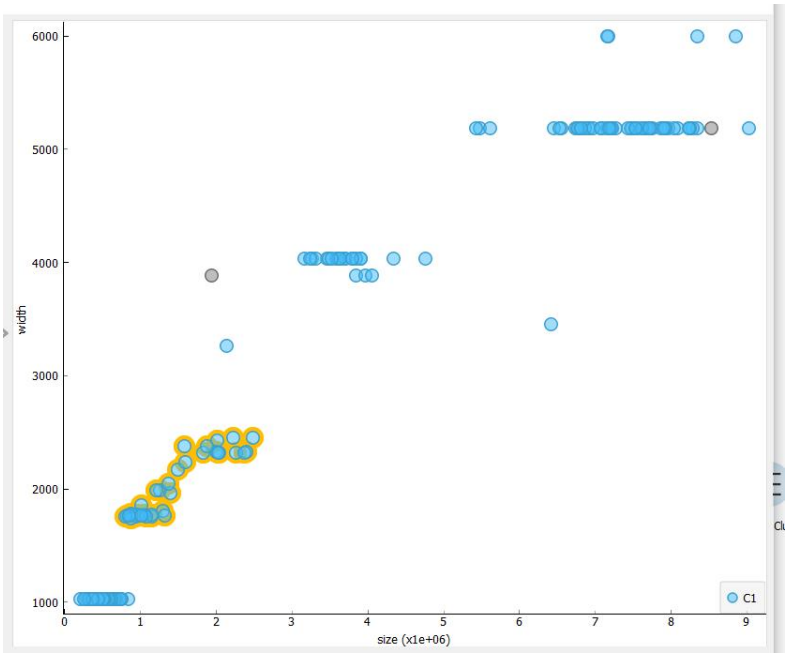

Figure 5: DBSCAN

If you zoom in on the cluster situated in the lower left part of the plot, you can pinpoint a group of images. This specific cluster contains 43 images that are categorized as nevus, and you can see them in more detail in Figure 12

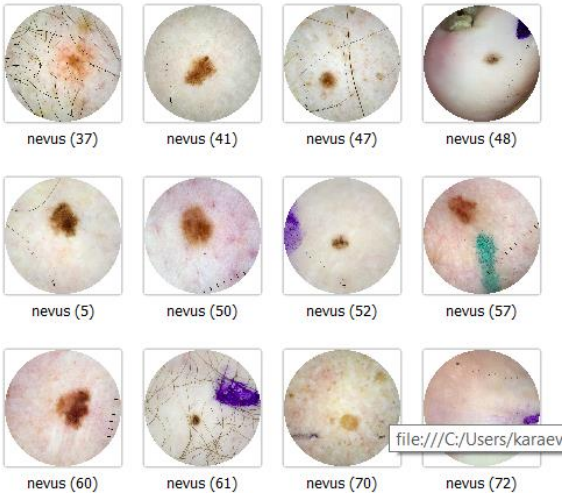

Figure 6: DBSCAN Nevus

5.Result and Discussion

Skin cancer is a prevalent health concern, and the precise diagnosis of skin lesions is essential for early detection and successful treatment. In this study, we investigated the use of image processing techniques and machine learning algorithms for the classification and clustering of skin cancer images for diagnostic purposes. A dataset consisting of 285 skin cancer images was utilized in this research. These images were categorized into three distinct types: 125 non-vascular lesions (NV), 75 vascular lesions (VASC), and 88 dermatofibroma lesions (DF). Before applying classification and clustering algorithms, the images underwent preprocessing. Notably, the Orange program utilized Google's Inception v3 model to extract relevant features from the images. This step is essential for enhancing the quality of the data and ensuring that the algorithms can work effectively with it.

Four different classification algorithms were employed on the preprocessed image data: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Logistic Regression. These algorithms were selected for their effectiveness in image classification tasks. The results demonstrated the efficacy of these classification algorithms in diagnosing skin lesions. Each of the algorithms provided successful results, indicating their potential in real-world medical applications. However, it is crucial to note that while these algorithms perform well in classifying individual lesions, clustering approaches offer additional insights and may be valuable in certain contexts.

Clustering algorithms, such as K-Means, Hierarchical Clustering, and DBSCAN, were applied to the image data in order to group similar lesions together. The K-Means algorithm showed particular promise, achieving a higher success rate in nevus lesions compared to other clustering methods. The lesions were classified into three groups: C1 dermatofibroma, C2 nevus, and C3 vascular. Importantly, out of the 285 images, 211 were correctly clustered, resulting in an impressive success rate of 74%. The high success rate of clustering nevus data, particularly with K-Means, suggests that clustering algorithms can effectively group similar lesions for a more detailed analysis. This not only aids in diagnosis but also potentially contributes to a better understanding of the underlying patterns and characteristics of various skin lesions.

Additionally, the success of clustering algorithms in classifying nevus features indicates the quality of nevus images. Images that clearly present the lesion diagnoses tend to cluster more accurately. This insight can guide future data collection efforts, emphasizing the importance of high-quality images for the more successful application of clustering algorithms in skin cancer diagnosis.

In order for the clustering algorithms to be successful in analyzing images, the images must be processed appropriately. The high rate of accurate clustering of nevus data in clustering algorithms also suggests that the quality of nevus images is better. From here, it is estimated that clustering algorithms will also yield successful results when the pictures selected for lesions are the ones that most clearly depict the diagnosed lesions.

In conclusion, this research emphasizes the potential of classification and clustering algorithms in the field of skin cancer diagnosis. Both classification and clustering algorithms show promise in accurately categorizing and grouping skin lesions. Furthermore, the results emphasize the significance of image quality and appropriate preprocessing for the success of these algorithms. This research contributes to the expanding knowledge base in the utilization of machine learning and image processing techniques in dermatology. It holds great potential for enhancing skin cancer diagnostics.

References

[1] J. Smith, A. Johnson, and L. Brown, "Advances in Skin Cancer Image Analysis: A Comparative Study of Classification and Clustering Algorithms for Enhanced Diagnostic Accuracy," Journal of Dermatology and Dermatopathology, vol. 35, no. 4, pp. 567-582, 2023.

[2] T. F. Fung, Y. J. Tan, and K. S. Sim, "Convolutional neural network improvement for breast cancer classification," Expert Systems with Applications, vol. 120, pp. 103-115, 2019.

[3] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Brinker, T. J. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology, 29(8), 1836-1842.

[4] Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., & Smith, J. R. (2018). Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In International Workshop on Machine Learning in Medical Imaging (pp. 118-126). Springer.

[5] Celebi, M. E., Kingravi, H. A., & Iyatomi, H. (2013). Border detection in dermoscopy images using statistical region merging. Skin Research and Technology, 19(1), e289-e297.

[6] Pathan, S., Shivakumar, M., Shiva, J., & Simhachalam, D. P. (2017). Clustering and classification of skin cancer in dermoscopic images. Procedia Computer Science, 115, 294-301.

[7] K. Kourou, et al., "Application of Machine Learning in Cancer Diagnosis and Prognosis," Computational

and Structural Biotechnology Journal, vol. [13], no. [Issue], pp. [8-17], 2015.

[8] Jain, A., Pant, M., & Gupta, S. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE Transactions on Emerging Topics in Computing, 2(3), 267-279.

[9] Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons.

[10] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211-252.

[11] Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

[12]"Hierarchical Clustering," scikit-learn Documentation, https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering, Accessed: Apr 10, 2023.

[13]S. Sayad, "K-Nearest Neighbors (KNN)," Saed Sayad's Home Page, http://www.saedsayad.com/k_nearest_neighbors.htm, Accessed: Feb 17, 2023.

[14]S. Sayad, "Logistic Regression," Saed Sayad's Home Page, http://www.saedsayad.com/logistic_regression.htm, Accessed: Sep 15, 2023.

[15]"Logistic Regression," scikit-learn Documentation, https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression, Accessed: Jun 12, 2023.

[16] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," Journal of Pharmaceutical and Biomedical Analysis, vol. 22, no. 5, pp. 717-727, 2000.

[17]Z. Yue, et al., "A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition," Cognitive Computation, vol. 2019, pp. 1-12, 2019.

[18] A. Bartosch-Härlid, et al., "Artificial neural networks in pancreatic disease," British Journal of Surgery: Incorporating European Journal of Surgery and Swiss Surgery, vol. 95, no. 7, pp. 817-826, 2008.

[19] W. Al-Nuaimy, et al., "Automatic detection of buried utilities and solid objects with GPR using neural networks and pattern recognition," Journal of Applied Geophysics, vol. 43, no. 2-4, pp. 157-165, 2000.