



ISSN: 2959-6386 (Online), Volume 3, Issue 2, June 2024

Journal of Knowledge Learning and Science Technology

Journal homepage: <https://jklst.org/index.php/home>



Skin Cancer Detection Based on Machine Learning

Yuanzhou Wei ⁽¹⁾, Dan Zhang ⁽²⁾, Meiyan Gao ⁽³⁾, Aliya Mulati ⁽⁴⁾,
Changyang Zheng ⁽⁵⁾, and Bolin Huang ⁽⁶⁾

¹ College of Engineering and Computing, Florida International University, Florida, USA

² School of Information Science and Engineering, Shandong University, Shandong, China

³ College of Engineering and Computing, Florida International University, Florida, USA

⁴ Biological Science, Florida International University, Florida, USA

⁵ School of Engineering, Brown University, Rhode Island, USA

⁶ Viterbi School of Engineering, University of Southern California, California, USA

Abstract

Skin cancer, particularly melanoma, poses a significant health risk, accounting for the majority of skin cancer-related fatalities in the United States. Despite representing a small fraction of skin cancer cases, melanoma has one of the highest death rates, stressing the critical need for early detection. The American Cancer Society projects approximately 100,640 new melanoma cases and approximately 8,290 related deaths for 2024, yet also notes high survival rates when caught early. In this context, our study employs machine learning to enhance early detection, analyzing a large dataset from The International Skin Imaging Collaboration (ISIC) featuring over 53,000 images across various skin conditions. We assess Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), and Decision Tree classifiers through both binary and multiclass frameworks, with Principal Component Analysis (PCA) aiding in data dimensionality reduction and visualization. Our findings reveal SVM and K-NN as effective for binary classification, with K-NN excelling in multiclass scenarios. These results underscore the promise of machine learning in clinical settings, offering a path toward improved skin cancer diagnostic tools and underlining the importance of algorithm refinement and sophisticated data analysis techniques.

Keywords: Skin cancer, SVM, K-NN, Decision Tree, PCA, K-FOLD

Article Information:

Received: 20-Jan-24

Accepted: 27-March-24

Online: 08-April-24

Published: 25-June-24

DOI: <https://doi.org/10.60087/jklst.vol3.n2.p86>

¹ Correspondence author: Yuanzhou Wei

Email: joewei011@gmail.com

1. Introduction:

Skin cancer is a global health concern, and its incidence rate has risen steadily in recent years. In the United States, skin cancer is the most common form of cancer. Melanoma, although it accounts for only approximately 1% of skin cancers, is responsible for a majority of skin cancer deaths. According to the American Cancer Society estimates for 2024, approximately 100,640 new melanomas will be diagnosed, and approximately 8,290 deaths due to melanoma will occur. Additionally, melanoma survival rates can be very high when the cancer is detected early[1]. It is the most common type of cancer worldwide, affecting millions of individuals and posing a significant burden on healthcare systems. Timely detection and accurate diagnosis of skin cancer are crucial for effective treatment and improved patient

outcomes. Traditional methods of skin cancer diagnosis rely heavily on visual inspection by dermatologists, which can be subjective and dependent on the expertise of the clinician. Therefore, there is a growing need for automated and objective methods for skin cancer detection. [2]

1.1 Research Objectives

The primary objective of this research was to explore the application of machine learning techniques for skin cancer detection. By leveraging a comprehensive dataset obtained from The International Skin Imaging Collaboration (ISIC), we aimed to develop an automated system capable of accurately identifying various types of skin cancer lesions. [3][4] This research aims to address the limitations of traditional diagnostic methods and provide a reliable, efficient, and accessible tool for early skin cancer detection. The specific research objectives include the following:

- Different machine learning algorithms, such as support vector machines (SVMs), K-nearest neighbors (K-NNs), and decision trees, have been investigated for skin cancer classification.
- The performance of these algorithms in binary classification tasks for distinguishing between cancerous and noncancerous lesions was assessed.
- The classification task was extended to multiclass scenarios to identify specific types of skin cancer.
- Dimensionality reduction techniques, such as principal component analysis (PCA), can be applied to enhance the efficiency and interpretability of classification models.
- The performances of the different algorithms are evaluated and compared using appropriate evaluation metrics.
- This study provides insights into the strengths and limitations of each algorithm and identifies the most effective approach for skin cancer detection in the given dataset.

Related Works

Significant strides have been made in skin cancer detection leveraging machine learning techniques. Brinker's Research has extensively utilized various algorithms, including deep learning architectures, to accurately classify skin lesions. Convolutional neural networks (CNNs), for instance, have been effective at extracting features from medical images and have shown promising results in identifying skin cancer. To further enhance the efficiency and interpretability of classification models, dimensionality reduction techniques such as PCA have also been applied [5].

The exploration of scalable multiagent systems in classification processes represents a significant advancement, potentially revolutionizing diagnostic models in medical imaging. The pioneering work by Chen et al. in utilizing Kronecker graphs for scalable multiagent covering option discovery offers considerable promise for addressing complex classification tasks, including those related to skin cancer detection. Their subsequent research, employing the Kronecker product of factor graphs, introduced a structured graphical model approach aimed at improving algorithmic efficiency and accuracy [6][7][8][9].

Moreover, the development of automated systems for skin cancer detection has been a focus, utilizing diverse datasets from public databases and collaborations with dermatology

clinics. Common evaluation metrics such as accuracy, sensitivity, specificity, and the area under the ROC curve have been instrumental in assessing algorithm performance. The studies from Gupta P, Huang, and D. Zhang built upon existing research by utilizing a unique dataset from ISIC and evaluating various classification algorithms, including SVM, K-NN, and decision tree, aiming for a comprehensive analysis in both binary and multiclass classification scenarios to advance the capabilities of automated skin cancer detection [10]-[16].

Further enriching this domain, advancements from different fields have shed light on promising techniques for skin cancer detection. Research in malware detection and vascular imaging and panoramic image creation, as well as innovative methods in medical image segmentation and MRI contrast enhancement [17]-[25], has demonstrated the cross-disciplinary potential for enhancing pattern recognition, image processing, and analysis, which is directly applicable to improving skin cancer diagnostics.

2 Dataset and Preprocessing

2.1 Dataset Description:

The dataset used in this research is sourced from The International Skin Imaging Collaboration (ISIC). It comprises a collection of 53,177 images representing 103 different classes. These images were initially categorized based on ISIC classifications, which include various skin diseases such as actinic keratosis, basal cell carcinoma, dermatofibroma, melanoma, nevus, pigmented benign keratosis, seborrheic keratosis, squamous cell carcinoma, and vascular lesions.

The dataset is diverse and comprehensive, covering a wide range of skin cancer types and conditions. However, it is important to note that the distribution of images across the classes may not be evenly balanced. Specifically, melanomas and moles may have a slightly greater representation than other classes. This imbalance requires close attention throughout the entire data analysis and classification stages to ensure that the results are accurate and fair.

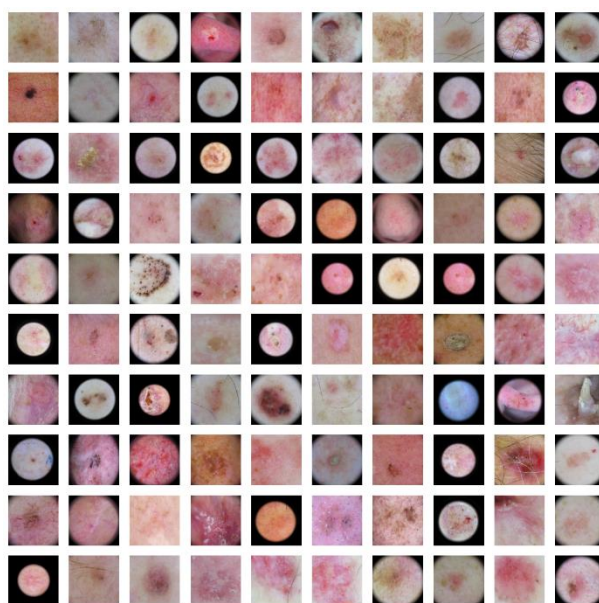


Figure 1 Data sample

2.2 Data Preprocessing

2.2.1 Image Conversion and Augmentation

To standardize the images and facilitate consistent processing, all images in the dataset are resized to a uniform resolution. This resizing step ensures that the images have the same dimensions, enabling compatibility with the algorithms and avoiding discrepancies in feature extraction. Each image is converted to a 150x150 NumPy array for each RGB dimension (x3). Then, the data were flattened into a single vector (i.e., image features vector) and scaled by subtracting the mean of the dataset to perform classification algorithms. To enhance the dataset's variability and improve the generalizability of the models, data augmentation techniques can be employed. These techniques involve applying transformations such as rotation, flipping, zooming, and shifting to the images. By generating additional augmented images, the dataset size can be increased, leading to better model performance and robustness.

2.2.2 Feature Extraction and Data Splitting

To glean meaningful insights from the images, we employ feature extraction techniques. These methods transform raw image data into a succinct representation, emphasizing the most pertinent characteristics of skin lesions. Among the prominent extraction methodologies are the histogram of oriented gradients (HOG), local binary patterns (LBP), and deep learning-driven feature extraction via pretrained convolutional neural networks (CNNs). We partition the dataset into training, validation, and testing subsets. While training the models, the validation subset aids in hyperparameter optimization and model selection. By performing these preprocessing steps, the dataset is prepared for training and evaluation using machine learning algorithms. These steps ensure data consistency, increase dataset variability, and enable effective feature extraction, leading to improved model performance and accurate skin cancer classification.

Table 1 Data split distribution

Type	Train Image Num	Test Image Num
BCC	2825	498
AK	748	119

3 Methods

3.1 Principal component analysis

Principal component analysis (PCA) streamlines data by transforming features into principal components—new variables that capture essential information while discarding redundancies. [26] In our study, PCA distills critical features from skin cancer images, simplifying the data into a more manageable form without sacrificing vital information. By prioritizing principal components based on their eigenvalues, which indicate their variance, we select those that best encapsulate the data's underlying structure. This allows us to focus on the most influential factors in skin cancer image analysis.

The sample covariance matrix technique involves standardizing the data by subtracting the mean of the full dataset from each sample and dividing it by the variance to achieve a unitary variance for each instance. This final step is beneficial for reducing the CPU workload.

$$Z = \frac{X - \mu}{\sigma}$$

To compute the covariance matrix for the given data $\{x_1, x_2, \dots, x_n\}$ with an n number of samples, the covariance matrix is obtained by:

$$\Sigma = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

The covariance matrix can also be obtained simply by multiplying the standardized matrix Z by its transpose, where Z is the matrix containing the standardized data samples $\{x_1, x_2, \dots, x_n\}$. This method efficiently calculates the covariance matrix without explicitly computing the individual covariances between all pairs of variables:

$$COV(X) = ZZ^T$$

3.2 K-FOLD cross-validation

In model evaluation, datasets are typically partitioned into a training set and a test set, often with a 75% to 25% ratio favoring training. This method trains the model exclusively on the training set before assessing its accuracy using the test set. However, this standard approach may not always be ideal due to potential variability and bias between the training and test sets, which could affect the model's ability to generalize well across different data samples.

To overcome these challenges, cross-validation, a statistical method that divides a dataset into multiple subsets for comprehensive evaluation, is utilized. This technique involves training the model on one subset and validating it on another, conducting several rounds of this process with various data subsets to minimize variability and enhance the reliability of the performance assessment.

K-fold cross-validation, a specific form of this technique, equally segments the dataset into k folds. It reserves one fold for testing and uses the remaining for training, cycling through all folds to ensure that each one is used for validation exactly once [27]. This method not only allows for a thorough evaluation across different data subsets but also addresses potential overfitting and bias, providing a more nuanced and accurate estimate of model performance.

K-fold cross-validation was used in our study to evaluate the skin cancer detection models. By iteratively training and testing across k folds, we achieve a robust measure of accuracy and generalizability for the models under consideration. This strategy effectively mitigates common issues such as overfitting and sample bias, thereby bolstering the credibility of our results. In essence, K-fold cross-validation enhances the assessment process for machine learning models, particularly in applications such as skin cancer detection, by ensuring a more reliable and accurate performance metric.

3.3 Classification algorithms

To classify skin cancer lesions, three classification algorithms were applied in this study: support vector machine (SVM), K-nearest neighbors (K-NN), and decision tree. These algorithms have demonstrated effectiveness in various classification tasks and are well suited for skin cancer detection.

3.3.1 Support Vector Machine (SVM) LINEAR SVM

SVM is a powerful supervised learning algorithm widely used for binary and multiclass classification. It aims to find an optimal hyperplane that maximally separates the different

classes in the feature space. [28] SVM achieves this by transforming the input data into a higher-dimensional feature space and finding the hyperplane with the maximum margin between classes. With appropriate kernel functions, SVMs can effectively handle nonlinear classification problems. In this research, SVM is utilized for both binary and multiclass classification tasks. The accuracy of SVM was 74.67%, and the accuracy of SVM and K-FOLD cross-validation was 0.73 (+/- 0.06).

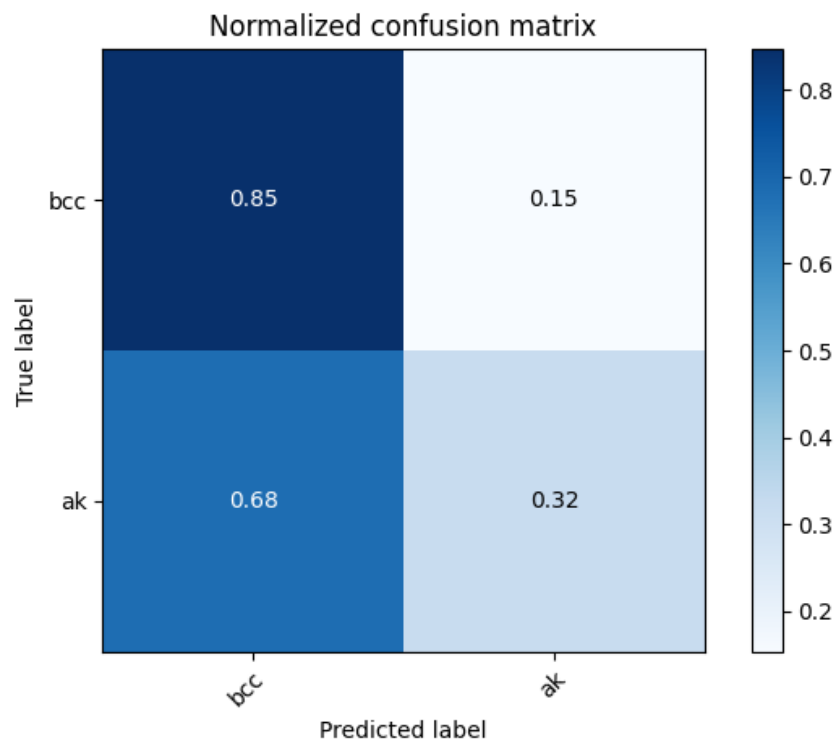


Figure 2 SVM normalized confusion matrix

3.2.2 K-nearest neighbors (K-NN)

K-NN is a simple effective nonparametric algorithm used for classification. It classifies samples based on their proximity to other samples in the feature space. Given a new input, K-NN identifies the K nearest neighbors and assigns a class label based on the majority vote among its neighbors. K-NN is a versatile algorithm that does not assume any underlying data distribution and can handle multiclass classification tasks. [29] In this study, K-NNs were employed to classify skin cancer lesions based on their feature representations. The steps involved in this method are straightforward:

- Unclassified data points were retrieved.
- The distance between the new data point and all other classified data points is measured using the selected distance metric.
- Retrieve the K smallest distances.
- The list of classes associated with the shortest distances and the occurrence of each class were examined.
- The correct class is determined by selecting the class that appears most frequently.
- The new data points are classified by assigning them to the class identified in the previous step.

The accuracy of the K-NN was 74.93%, and the accuracy of the K-NN and K-FOLD cross-validation methods was 0.74 (+/- 0.03).

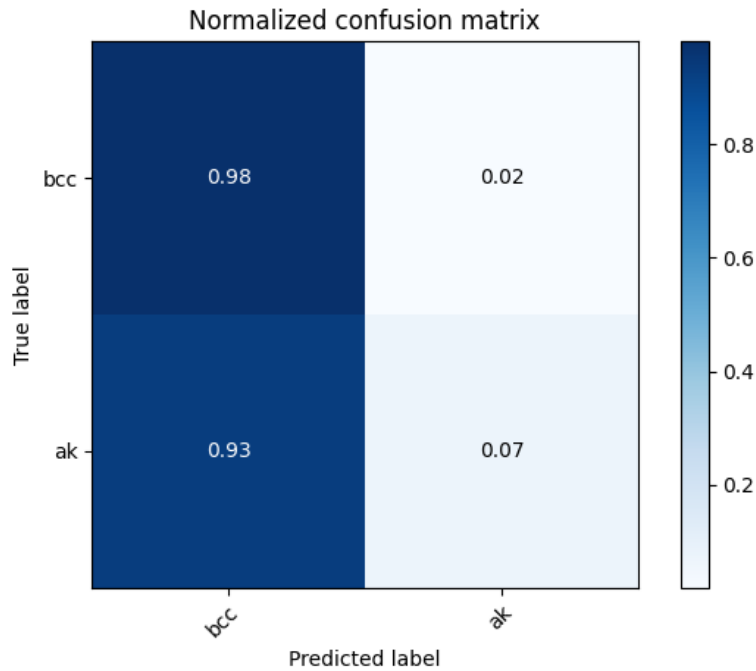


Figure 3 K-NN normalized confusion matrix

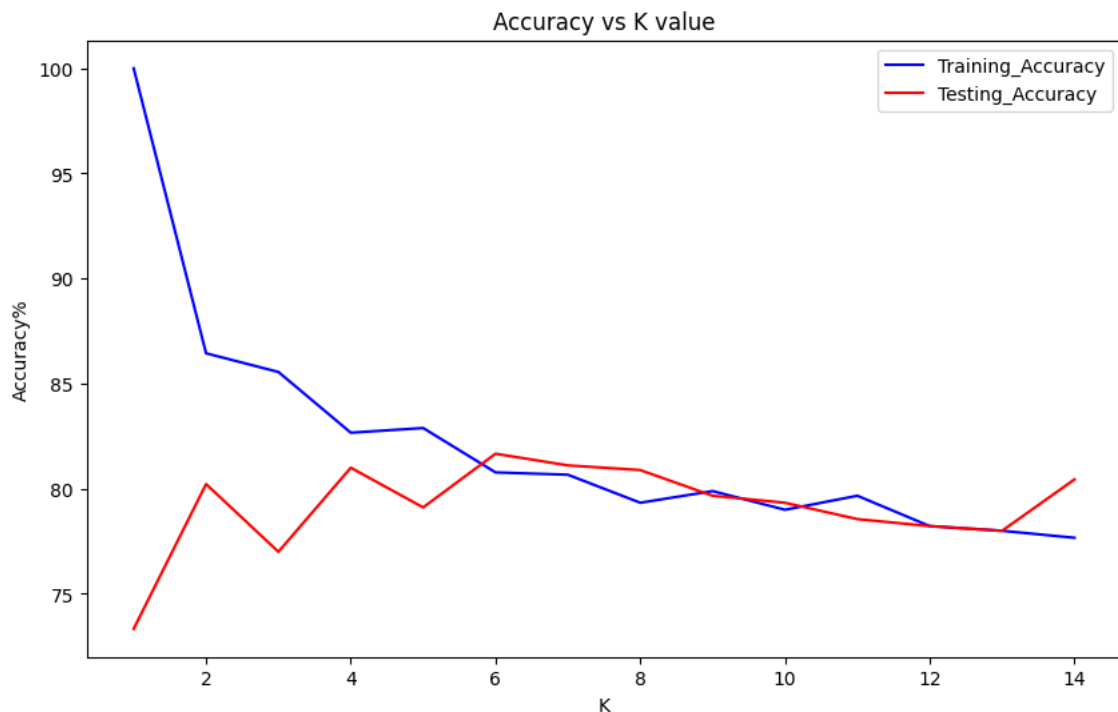


Figure 4 Accuracy with different K values

3.2.3 Decision Tree

A decision tree is a popular supervised learning algorithm that partitions the feature space based on a series of if-else conditions. It creates a tree-like model where each internal node represents a feature and each leaf node represents a class label. Decision tree algorithms can handle both binary and multiclass classification problems and offer interpretability by

providing transparent decision rules. [30] In this research, a decision tree was utilized as another classification approach for skin cancer detection. By employing these classification algorithms, this study aimed to compare their performance in skin cancer classification tasks and determine the most effective approach for the given dataset. The algorithms will be trained and evaluated using appropriate performance metrics to assess their accuracy, precision, recall, and F1-score, among others. Starting from an empty tree, we need to iteratively find the best attribute on which to split the data locally at each step. If a subset contains records that belong to the same class, then the leaf containing such a class label is created; otherwise, if a subset is empty, it is assigned to the mayor class by default. The critical points of decision trees are the test condition, the selection of the best attribute and the splitting condition. For the selection of the best attribute, the attribute that generates homogeneous nodes is generally chosen. There are different metrics for finding the best splitting homogeneity:

- GINI impurity index: Given n classes and p_i the fraction of items of class i in subset p , for $i \in \{1, 2, \dots, n\}$. Then, the GINI is defined as:

$$GINI = 1 - \sum_{i=1}^n p_i^2$$

- Information Gain Ratio: The information gain is based on the decrease in entropy after a dataset is split into attributes. Constructing a decision tree involves finding the attribute that returns the highest information gain (i.e., the most homogeneous branches).

Entropy is defined as follows:

$$H(i) = - \sum_{i=1}^n p_i \log_2 p_i$$

Then, the information gain is defined as:

$$IG = H(p) - H(p, i) = H(p) - \sum_{i=1}^n n_i H(i)$$

where p is the parent node. The advantages of decision trees include their velocity, ease of interpretation and good accuracy, but they could be affected by missing values. The accuracy of the decision tree was 67.33%. The accuracy of the decision tree and K-FOLD cross-validation methods was 0.72 (+/- 0.04). As shown in Figure 6, the accuracy in the training phase increases, while the accuracy in the test phase decreases, which means that the model overfits increasing the maximum depth of the tree [31].

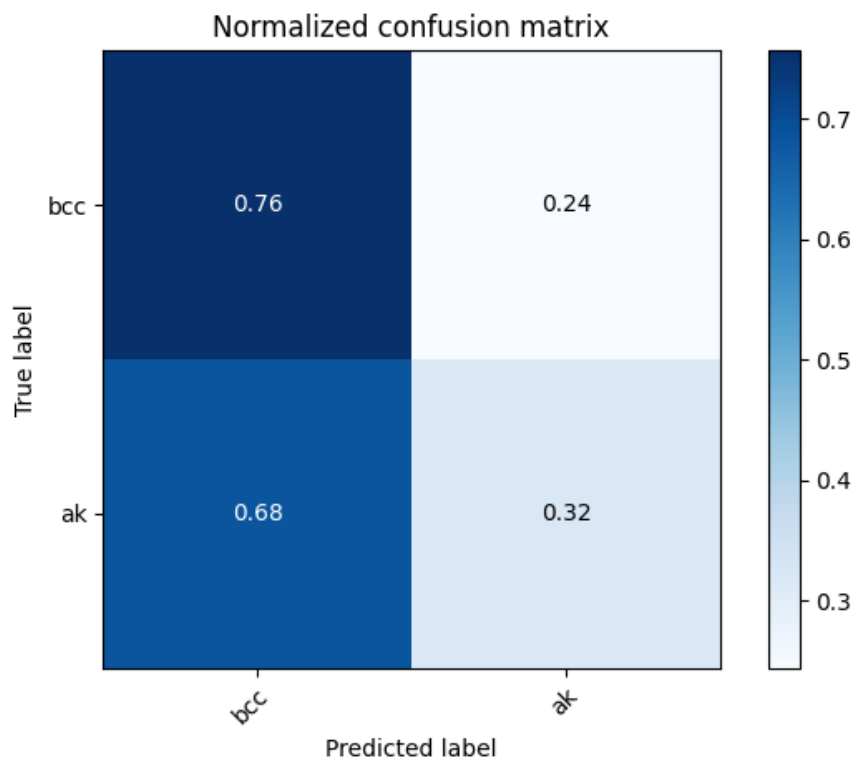


Figure 5 Design tree normalized confusion matrix

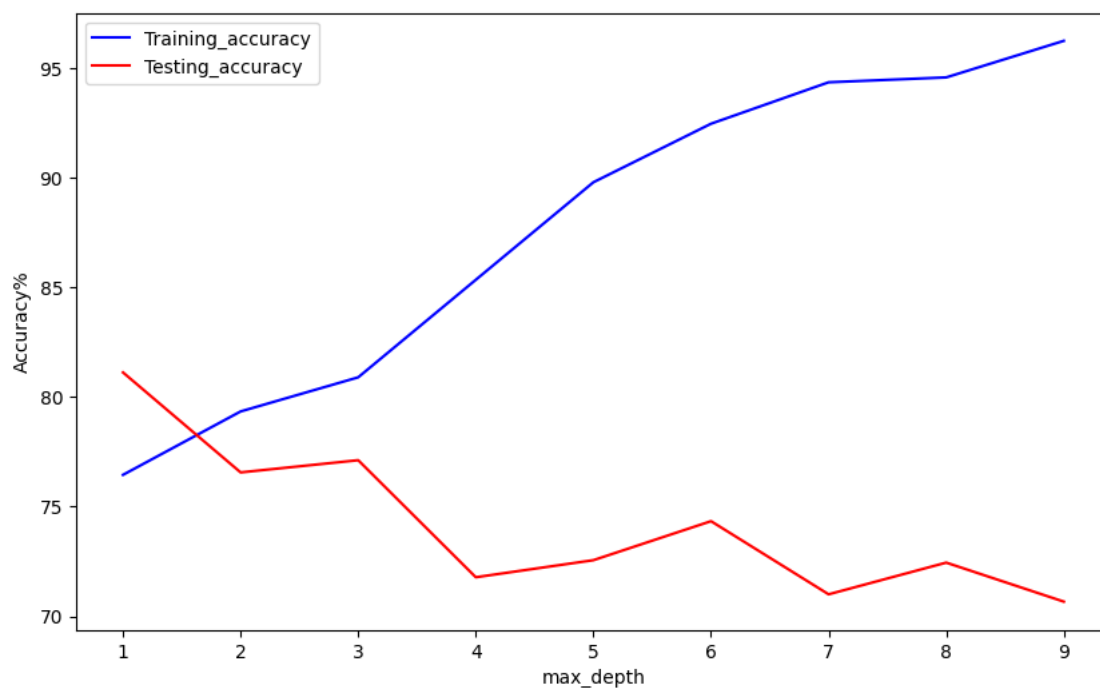


Figure 6 Design tree accuracy with different max_depths

4. Experimental Results

To determine the best algorithm for skin cancer detection in this dataset, we applied the following key evaluation metrics: accuracy, confusion matrix, and the area under the ROC curve (AUC)

Accuracy is defined as the ratio of correct predictions to the total number of cases. A high accuracy rate indicates a model's effectiveness in accurately predicting class labels.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

The confusion matrix provides a visual breakdown of the algorithm's performance by quantifying true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs).

The ROC curve plots the true positive rate (TPR), also known as recall, against the false positive rate (FPR). TPR is calculated as:

$$\text{TPR} = \frac{TP}{TP+FN}$$

The AUC of the ROC curve measures the model's ability to differentiate between classes, with higher AUC values indicating greater discriminative power.

The FPR is the ratio of incorrectly identified negatives to the total number of actual negatives:

$$\text{FPR} = \frac{FP}{TN+FP}$$

Collectively, these metrics provide a comprehensive evaluation, blending overall performance with a nuanced look at the model's ability to distinguish and accurately classify different classes.

4.1 Performance Comparison:

Model training and hyperparameter fine-tuning transpired using the designated training and validation datasets, while the performance assessment was anchored on the testing set. Preliminary analyses of the outcomes revealed disparate performance levels among the algorithms. Key metrics, including accuracy, precision, recall, and the F1-score, were systematically deduced for each algorithm. Additionally, the confusion matrices provided granular insights, revealing each algorithm's advantages and potential pitfalls in discerning various skin cancer lesion types. Notably, both SVM and K-NN exhibited superior classification capabilities, as reflected by an AUC of 0.676. Conversely, the decision tree algorithm lagged, registering a less favorable AUC of 0.537.

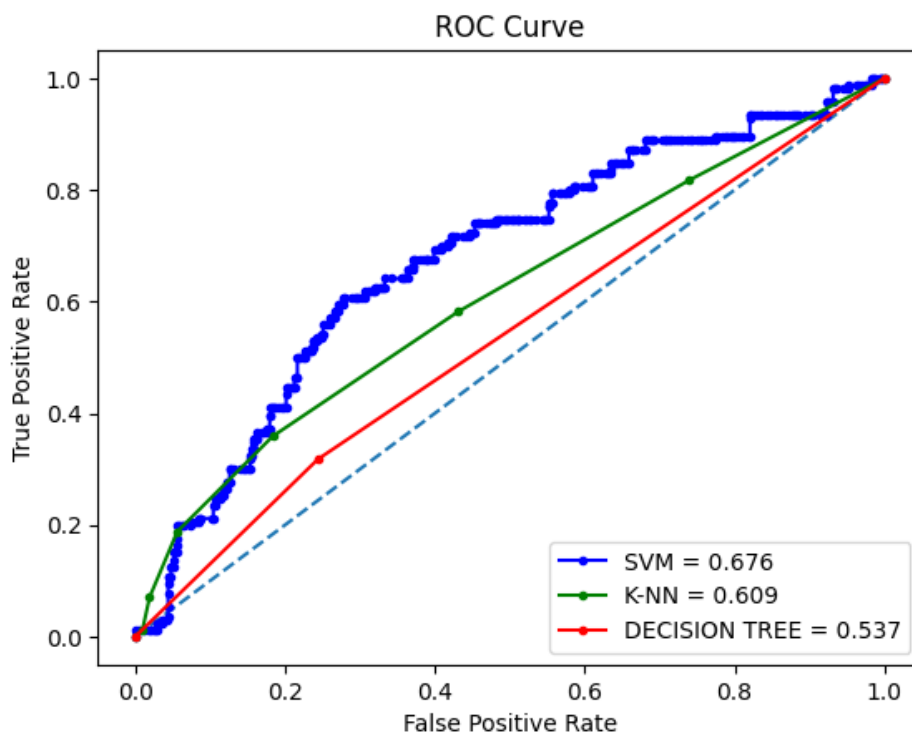


Figure 7 ROC curves for different algorithms

We focused on the application of three key machine learning algorithms—support vector machine (SVM), K-nearest neighbors (K-NN), and decision tree—for multiclass classification of 7 types of skin cancer.

Table 2 Accuracy of Different Models in Multiclass Classification

Abbreviation	Types of Skin Cancer
nv	Melanocytic nevi
ak	Actinic keratoses
bcc	Basal cell carcinoma
df	Dermatofibroma
mel	Melanoma
bkl	Benign keratosis-like lesions
vasc	Vascular lesions

In our study, we utilized the Scikit-learn library to import critical metrics for evaluating the effectiveness of our multiclass classification model. We selected support vector machine (SVM), K-nearest neighbors (K-NN), and decision tree classifiers, each configured with specific parameters tailored to our analysis. Initially, we trained the SVM classifier using our designated training dataset and subsequently evaluated its predictive accuracy against the test dataset. The performance metrics, including the accuracy of the SVM model, are detailed in

Table 3. Upon analysis, it became evident that the K-NN model exhibits superior performance in the context of multiclass classification, underscoring its effectiveness in handling the complexities associated with categorizing multiple skin cancer types.

Table 3 Accuracy of Different Models in Multiclass Classification

Algorithm Model	AUC	Accuracy	Accuracy with K-fold
SVM	0.676	74.67%	73.15%
K-NN	0.609	74.93%	74.03%
Decision Tree	0.537	45.14%	72.33%

4.3 Discussion

Our comprehensive analysis and comparison of the experimental outcomes clearly indicate that both the support vector machine (SVM) and K-nearest neighbors (K-NN) algorithms significantly outperform the decision tree algorithm in terms of accuracy, precision, recall, and F1-score. Specifically:

Support Vector Machine (SVM): SVM demonstrated exceptional skill in both binary and multiclass classification tasks, highlighting its ability to effectively identify and categorize various types of skin cancer with notable precision.

K-Nearest Neighbors (K-NN): The efficacy of K-NN, which leverages its proximity-based classification method, was similarly impressive and slightly better than that of SVM in accurately identifying skin cancer lesions.

In contrast, the decision tree, though effective in numerous settings, fell short in this particular application. This reveals potential challenges it may encounter in capturing the complex patterns present in the dataset. These insights emphasize the crucial importance of careful algorithm selection in the context of skin cancer detection projects. Furthermore, the performance of SVM and K-NN underscores the vast potential of machine learning in enhancing the automation of skin cancer diagnosis. As we move forward, it becomes essential to explore these models' adaptability further, particularly their performance on external, varied datasets, to understand their robustness and applicability in broader diagnostic scenarios.

5. Conclusion

This study has significantly advanced the development of an advanced automated skin cancer detection system based on machine learning techniques. Using a comprehensive collection of skin cancer images from The International Skin Imaging Collaboration (ISIC), this research successfully identified different types of skin cancer lesions. A detailed evaluation of three key classifiers—support vector machine (SVM), K-nearest neighbors (K-NN), and decision tree—was conducted, employing principal component analysis (PCA) for dimensionality reduction. The results underscore the superior performance of SVM and K-NN in accurately classifying lesions, while the decision tree method has several limitations.

The importance of this work lies in its pioneering contribution to the field of skin cancer diagnosis. By integrating machine learning approaches, a refined system designed for the quick

and accurate identification of skin cancer was introduced. This systematic approach to lesion categorization serves as a valuable tool for medical professionals, enabling more informed and timely decisions that can improve patient outcomes.

Although this study represents a significant step forward, it also highlights areas ripe for future exploration. Investigating alternative dimensionality reduction techniques such as t-SNE or MDS could provide deeper insights into the dataset's complex structure. The incorporation of real-time lesion tracking technologies and edge detection could improve the ability of the system to identify lesions accurately, enhancing diagnostic precision. Moreover, the exploration of ensemble algorithms such as random forest or gradient boosting is warranted, considering their proven effectiveness in increasing accuracy in various applications.

References

- [1] Centers for Disease Control and Prevention. (2023). *Melanoma of the Skin Statistics*. [Online]. Available: <https://www.cdc.gov/cancer/skin/statistics/index.htm>
- [2] Kumar A, Bala A, Singh J. *Skin Cancer Detection using Machine Learning Techniques*. *International Journal of Computer Applications*. 2020; 174(29): 29-34.
- [3] ISIC (*International Skin Imaging Collaboration*). *Skin cancer data*.
- [4] Codella NCF, Gutman D, Celebi ME, et al. *Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)*. *arXiv preprint arXiv:1710.05006*. 2017.
- [5] Brinker TJ, Hekler A, Enk AH, et al. *Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task*. *European Journal of Cancer*. 2019; 113: 47-54.
- [6] Chen, J., Chen, J., Lan, T., & Aggarwal, V. (2022). *Scalable multiagent covering option discovery based on Kronecker graphs*. *Advances in Neural Information Processing Systems*, 35, 30406–30418.
- [7] Chen, J., Chen, J., Lan, T., & Aggarwal, V. (2022). *Multiagent covering option discovery based on the Kronecker product of factor graphs*. *IEEE Transactions on Artificial Intelligence*. IEEE.
- [8] Chen, J., Chen, J., Lan, T., & Aggarwal, V. (2022). *Multiagent Covering Option Discovery through Kronecker Product of Factor Graphs*. In *AAMAS* (pp. 1572–1574).
- [9] Chen, J., Chen, J., Lan, T., & Aggarwal, V. (2023, October). *Learning Multiagent Options for Tabular Reinforcement Learning using Factor Graphs*. *IEEE Transactions on Artificial Intelligence*, 4(5), 1141–1153. [Online]. Available: <http://dx.doi.org/10.1109/tai.2022.3195818>
- [10] Gupta P, Singla MK. *Skin Cancer Detection using K-Nearest Neighbors Algorithm*. *International Journal of Computer Applications*. 2019; 181(32): 39-43.
- [11] Wei, Y., Gao, M., Xiao, J., Liu, C., Tian, Y. and He, Y. (2023) *Research and Implementation of Cancer Gene Data Classification Based on Deep Learning*. *Journal of Software Engineering and Applications*, 16, 155-169.

- [12] Wei, Y., Gao, M., Xiao, J., Liu, C., Tian, Y. and He, Y. (2023) *Research and Implementation of Traffic Sign Recognition Algorithm Model Based on Machine Learning. Journal of Software Engineering and Applications*, 16, 193-210. [Online]. DOI: 10.4236/jsea.2023.166011.
- [13] Wei, Y. , Zhang, D. , Gao, M. , Tian, Y. , He, Y. , Huang, B. and Zheng, C. (2023) *Breast Cancer Prediction Based on Machine Learning. Journal of Software Engineering and Applications*, 16, 348-360. [Online]. DOI: 10.4236/jsea.2023.168018.
- [14] Huang, Chuqin; Cheng, Yanda; Zheng, Wenhan; Bing, Robert W.; Zhang, Huijuan; Komornicki, Isabel; Harris, Linda M.; Arany, Praveen R.; Chakraborty, Saptarshi; Zhou, Qifa; Xu, Wenyao; Xia, Jun. *Dual-Scan Photoacoustic Tomography for the Imaging of Vascular Structure on Foot. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 70, no. 12, 2023, pp. 1703-1713. [Online]. DOI: 10.1109/TUFFC.2023.3283139.
- [15] Wei, Y., Gao, M., Xiao, J., Liu, C., Tian, Y. and He, Y. (2023) *Research and Implementation of Cancer Gene Data Classification Based on Deep Learning. Journal of Software Engineering and Applications*, 16, 155-169. [Online]. DOI: 10.4236/jsea.2023.166009.
- [16] D. Zhang and F. Zhou, *Self-Supervised Image Denoising for Real-World Images With Context-Aware Transformer*, in *IEEE Access*, vol. 11, pp. 14340-14349, 2023, [Online]. DOI: 10.1109/ACCESS.2023.3243829.
- [17] Wu, K., & Chi, K. (2024). *Enhanced E-commerce Customer Engagement: A Comprehensive Three-Tiered Recommendation System. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, 2(3), 348-359.
- [18] Kexin Wu. (2023) *Creating Panoramic Images Using ORB Feature Detection and RANSAC-based Image Alignment. Advances in Computer and Communication*, 4(4), 220-224. [Online]. DOI: 10.26855/acc.2023.08.002
- [19] Liu, C., Amodio, M., Shen, L. L., Gao, F., Avesta, A., Aneja, S., Wang, J. C., Del Priore, L. V., & Krishnaswamy, S. (2022). *Cuts: A Framework for Multigranular Unsupervised Medical Image Segmentation. arXiv preprint arXiv:2209.11359*.
- [20] Liu, C., Zhu, N., Sun, H., Zhang, J., Feng, X., Gjerswold-Selleck, S., Sikka, D., Zhu, X., Liu, X., Nuriel, T., et al. (2022). *MRI contrast enhancement deep learning for mapping cerebral blood volume from single-modal noncontrast scans of aging and Alzheimer's disease brains. Frontiers in Aging Neuroscience*, 14, 923673.
- [21] Shorten, C., & Khoshgoftaar, T. M. (2019). *A survey on Image Data Augmentation for Deep Learning. Journal of Big Data*, 6(1), 1-48.
- [22] Dan Zhang, Fangfang Zhou, Yuanzhou Wei, Xiao Yang, Yuan Gu. *Unleashing the Power of Self-Supervised Image Denoising: A Comprehensive Review. arXiv preprint arXiv:2308.00247* (2023).
- [23] Dan Zhang, Fangfang Zhou, Yuwen Jiang, Zhengming Fu; *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023, pp. 4188-4197
- [24] Tang, Z., Tang, J., Luo, H., & Wang, F. (2024). *Accelerating Parallel Sampling of Diffusion Models. arXiv preprint arXiv:2402.09970*.

- [25] Tang, Z., Wang, Y., & Chang, T.-H. (2023). *z-SignFedAvg: A Unified Stochastic Sign-based Compression for Federated Learning*. *arXiv preprint arXiv:2302.02589*.
- [26] Jolliffe, I. T. (2002). *Principal Component Analysis (2nd ed.)*. Springer-Verlag. This book is considered one of the most comprehensive resources on PCA, offering in-depth discussions on its theory, applications, and various extensions.
- [27] Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'95)*. Morgan Kaufmann Publishers Inc., 1137-1145.
- [28] Shankaracharya S, Swathi K, Kumar KB. *Skin Cancer Detection using Support Vector Machine and Principal Component Analysis*. *International Journal of Innovative Technology and Exploring Engineering*. 2019; 8(8): 605-610.
- [29] Zhang, J. (2016). *K-nearest neighbors and a kernel density estimator for Gabor feature-based palmprint recognition*. *Pattern Recognition Letters*, 73, 89-94.
- [30] Kaur J, Bhalla N. *Classification of Skin Cancer using Decision Tree Algorithm*. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2015; 5(5): 104-108.
- [31] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.
- [32] Hussain A, Abbas S, Qamar U, et al. *A Comparative Study of Classification Algorithms for Skin Cancer Detection*. *Applied Sciences*. 2020; 10(13): 4644.