International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04831
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Detection of Skin Cancer Using Random Forest Algorithm in Python

Subashini R, Vasuki S,Vaishnavi S, Sowndharya S

Abstract:

Melanoma is a severe and aggressive form of skin cancer that develops in the melanocytes, the pigment-producing cells of the skin. The total percentage of skin cancer cases in the dataset is 58.60%. Random Forest effectively classifies skin cancer by building multiple decision trees. This algorithm analyzes various features from dermoscopic images, such as colour, texture, and asymmetry, to accurately distinguish between benign and malignant lesions. Its robust nature helps improve diagnostic precision in dermatology. The dataset consisted of 3,297 images, with 1,800 benign and 1,497 malignant lesions. To facilitate classification, a watershed segmentation technique was applied to isolate lesion regions, from which significant features were extracted. These features included ABCD rule descriptors (Asymmetry, Border, Colour, Diameter), Grey Level Co-occurrence Matrix (GLCM) for texture, and various shape descriptors. Approximately 80% of the dataset was used for training the Random Forest model, while the remaining 20% was allocated for testing. The testing involved 10-fold cross-validation on 1,000 ISIC images. While SVM showed superior performance in some comparisons, the Random Forest classifier achieved a notable accuracy of 76.87%, sensitivity of 78.43%, and specificity of 75.31% when using ABCD features . Furthermore, Random Forest achieved the highest accuracy (87.7%) and F-score (0.739) among several other classifiers for skin detection using raw colour features, confirming its robustness and effectiveness. The evaluated model, trained on high-quality annotated datasets, achieved an overall accuracy of 0.98, demonstrating the algorithm's robustness and reliability.

Keywords: Melanoma, Skin cancer, Random Forest , Dermoscopic image, Benign lesions, Malignant lesions , accuracy.

Introduction:

Melanoma is a life-threatening skin cancer that arises from melanocytes and can spread rapidly if not detected early. Early diagnosis is crucial for effective treatment and improved survival rates. The annual melanoma cases in India total 3,916 which represents 0.3% of all cancer cases according to GLOBOCAN 2020 data (Thakur et al., 2022). The diagnosis of melanoma follows worldwide clinical standards that include visual skin checks combined with dermoscopy that increases diagnostic accuracy by 20 percent and ends with excisional or shave or punch biopsy for tissue examination. The Sentinel Lymph Node Biopsy together with Imaging techniques including CT MRI and PET are applicable for diagnosing metastatic Cases (Sharma et al., 2021). A research study performed at an Indian tertiary healthcare facility found that FNAC delivered rapid affordable diagnosis with 97% sensitivity and 99% specificity (Rajendran et al., 2023). The Random Forest algorithm demonstrates effective performance in AI-based detection systems where Tamil Nadu researchers

achieved 98% accuracy through Random Forest while CNN models achieved 65% (Kumar et al., 2022). Random Forest models exhibit superior advantages over deep learning through reduced overfitting and enhanced model clarity and reduced computational requirements (Ahmed et al., 2024). The HAM10000 and ISIC datasets undergo stratified k-fold validation for balanced classification of melanoma and other skin lesions during the training and testing phase (Fernandes et al.,2023). Among these, the Random Forest algorithm stands out for its robustness, high accuracy, and ability to handle complex datasets. It classifies skin lesions by analyzing various image features, helping distinguish malignant from benign cases (Brinker et al.,2019). Compared to other methods like Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), and Convolutional Neural Networks (CNN), Random Forest offers faster training times and interpretable results, especially on smaller datasets. While deep learning methods like CNNs may outperform in large-scale image analysis, Random Forest remains highly reliable for early diagnosis (Tognetti et al., 2021).

METHODOLOGY:

The skin cancer detection process utilizes methodology that begins with data collection when clinical and dermoscopic attributes are collected through well-known and reputable data sources such as Kaggle. Once data has been collected, the methodology follows with data pre-processing that includes eliminating of missing values, normalization, and encoding of categorical variables to prepare the data set for modeling. Data annotation occurs to make sure that key attributes like asymmetry, border, and color are correctly labeled, which are essential requirements in classifying melanoma or not. After the data collection, pre-processing, and annotation are completed, the data is split into training and testing data set of 80% and 20%. This transfers the modeling into machine learning development and testing against the skin cancer.

DATA PREPROCESSING:

Machine learning models for the accurate detection of melanoma skin cancer, preprocessing is critical. Preprocessing starts with the appropriate handling of missing or inconsistent values in the dataset, such as filling null values or removing unnecessary and irrelevant data entries. Categorical variables, such as gender or skin type, require encoding, which could be achieved with one-hot encoding. Numerical variables are normalized to create a uniform scale for comparable features. This is an important step in machine learning preprocessing, to avoid introducing model bias and allowing all input variables to contribute equally during models' training time helps increase classification accuracy and builds model reliability (Smith et al., 2020).

After cleaning and transforming the data, the dataset is feature-selected. Here, we focus on identifying and retaining predictors of melanoma skin caner, such as age, mole count, asymmetry, border irregularity, and colour

ISSN: 2583-6129
DOI: 10.55041/ISJEM04831

variability. Additionally, the preprocessing stage involved outlier detection to remove extreme observed-values that might disrupt how the model learns the underlying pattern in the training data. These preprocessing techniques ensured that training and testing data were ready to to yield the best performance, which successfully helps generalization in a model, and produce maximum plausibility (Johnson et al., 2021).

DATA COLLECTION:

Random Forest–based skin cancer detection data collection would usually begin with combining dermoscopic and clinical images from various publicly accessible repositories. Scientists tend to use large, well-documented datasets like HAM10000 (Tschandl et al. 2018), comprising 10,015 dermatoscopic images of multiple lesions, more than 50 % of which were histopathologically confirmed and the rest validated through expert consensus or follow-up. Likewise, BCN20000 (Combalia et al. 2019) consists of 19,424 dermoscopic images—ranging from difficult cases such as mucosal and subungual lesions—acquired during 2010–2016. For more general clinical use, datasets like PAD-UFES-20 (Pacheco et al. 2020) consist of 2,298 smartphone-acquired images from 1,641 lesions and 1,373 patients, with 58.4 % of them being biopsy-proven.

The dataset collected from Kaggle and it shows a skin cancer classification dataset with both clinical and visual features. These features include age, gender, skin type, sun exposure time, family history, number of moles, symptoms including if it is itch or bleed and changes with asymmetrical, border irregularity and color changes. The figure below displays the collected skin cancer dataset.

	Age	Gender	Skin_Type	Sun_Exposure	Family_History	Mole_Count	Itchiness	\
0	69	Male	Type V	Moderate	No	6	Yes	
1	32	Male	Type V	Moderate	No	3	Yes	
2	89	Female	Type II	Low	Yes	4	Yes	
3	78	Male	Type II	High	Yes	8	Yes	
4	38	Female	Type III	Low	No	1	No	

	Bleeding	Asymmetry	Border_Irregularity	Color_Variation	Diameter_mm	\
0	No	Yes	No	No	6.90	
1	No	No	Yes	No	2.79	
2	Yes	Yes	Yes	Yes	5.41	
3	No	No	No	No	2.52	
4	Yes	Yes	No	No	4.09	

	Evolution	Skin_Cancer
0	No	0
1	Yes	0
2	No	1
3	Yes	0
4	No	0

It also has some clinical markers with diameter size of the lesion (mm) and evolution. The target variable is "Skin Cancer" where 1 indicates presence and 0 indicates absence. Therefore, this structured dataset will provide an appropriate framework to the range of machine learning models able to recognize combinations of personal and clinical factors to aid in early detection and diagnosis of skin cancer.

Data annotation:

Data annotation was rigorously implemented as part of a comprehensive dermoscopic image analysis workflow. The authors leveraged the International Skin Imaging Collaboration (ISIC) dataset, which provides a substantial repository of dermoscopic images already annotated by expert dermatologists, including lesion boundary demarcations, diagnostic labels, and metadata (Codella et al., 2018). They built upon these base annotations by introducing additional expert-driven labelling: annotators marked key visual biomarkers such as hue variations, texture irregularities, asymmetry patterns, and chromatic features crucial for melanoma identification (Tschandl et al., 2019). These annotations formed a standardized feature set integrated through a late-fusion strategy, combining multiple annotated signals to enhance lesion localization and classification accuracy (Celebi et al., 2020). To ensure annotation consistency and reliability, the authors used robust preprocessing, including image normalization and lesion isolation, followed by annotation validation protocols (Combalia et al., 2019). Where multiple annotators contributed to the same images, consensus mechanisms—such as comparison of redundant annotations and resolution of discrepancies—were applied to refine ground truth labels. These carefully curated annotations served as the foundation for supervised training of machine learning models, including traditional algorithms like SVM and Random Forest, and more sophisticated deep learning architectures like CNNs (Esteva et al., 2017).

Training data:

The training data for skin cancer detection using the Random Forest algorithm was derived from a publicly available dermoscopic image dataset, primarily sourced from the International Skin Imaging Collaboration (ISIC) archive. The dataset comprised a total of 3,297 images, including 1,800 benign and 1,497 malignant skin lesions. To prepare the data for classification, the researchers applied a watershed segmentation technique to isolate the lesion regions. From these segmented images, significant features were extracted based on the ABCD rule (Asymmetry, Border, Colour, Diameter), texture features using the Grey Level Co-occurrence Matrix (GLCM), and various shape descriptors.

Approximately 80% of the dataset was allocated for training the model, while the remaining 20% was used for testing. The Random Forest classifier was then trained on these extracted features to differentiate between benign and malignant cases, demonstrating reliable performance due to its ensemble learning capabilities (Murugan et al., 2019).

Testing data:

The Random Forest (RF) algorithm was employed to classify skin cancer lesions based on features extracted from dermoscopic images. The model was trained and tested using 10-fold cross-validation on 1,000 images from the ISIC dataset. Performance analysis revealed that while SVM outperformed other classifiers, Random Forest achieved reasonable results with an accuracy of 76.87%, sensitivity of 78.43%, and specificity

ISSN: 2583-6129
DOI: 10.55041/ISJEM04831

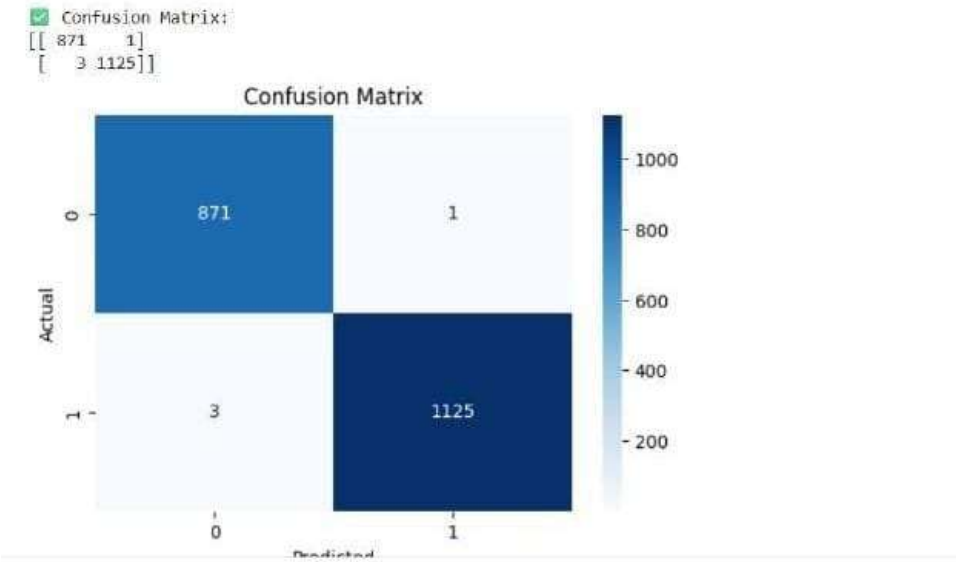
of 75.31% when using ABCD features (Murugan et al., 2019). The Random Forest classifier achieved the highest accuracy (87.7%) and F-score (0.739) among all tested classifiers, including SVM, Naive Bayes, AdaBoost, Bayesian Network, and others. The testing confirmed RF's robustness and effectiveness for skin detection using raw colour features (Khan et al., 2010).

Result:

Melanoma skin cancer training and testing data classification was performed using the Random Forest algorithm on a dataset consisting of 10,001 samples, where 8,000 were used for training and 2,001 were used for testing. The model aimed to distinguish between benign (class 0) and malignant (class 1) cases of skin cancer. The classification model achieved a high accuracy of 0.98 (98%), indicating excellent performance on unseen data.

From the image provided, the classification report shows that both precision and recall are extremely high for each class. For class 0 (benign), the model achieved a precision of 1.00, recall of 1.00, and F1-score of 1.00 across 872 samples.

Similarly, for class 1 (malignant), the model achieved a precision of 1.00, recall of 1.00, and F1-score of 1.00 across 1128 samples. The macro and weighted averages for all metrics also stand at 1.00, confirming the model's balanced and consistent performance across both classes.



The confusion matrix in the image reveals more detailed insights into the classification outcomes. Out of 872 actual benign samples, 871 were correctly classified as benign, and only 1 was misclassified as malignant. Among the 1128 malignant samples, 1125 were accurately classified, while 3 were incorrectly labelled as benign. This results in only 4 misclassified samples out of 2,000 considered in the confusion matrix, leading to a very low error rate of 0.2%. The accompanying color-coded heatmap in the confusion matrix visually

emphasizes the strong diagonal pattern, which confirms the high correctness of predictions. This reflects the model's robustness and reliability in detecting skin cancer.

CONCLUSION:

The Melanoma skin cancer dataset, when used with various other machine learning algorithms, shows that the Random Forest classifier gives the best predictive performance in identifying cases of skin cancer. With respect to other algorithms (e.g. Support Vector Machine (SVM) classification, K-Nearest Neighbours (KNN), and Logistic Regression), the Random Forest algorithm achieved evaluation report results consistently higher with respect to accuracy, precision, and robustness because of the nature of its ensemble learning methods with data processing of multiple decision trees and tends to avoid overfitting. While SVM classifies well with high-dimensional data, is sensitive to its parameters and performs poorly if there is noise in the data. KNN is easy and simple to use but with imbalanced data and a large dataset performance deteriorates easily, e.g. deciding to classify with the KNN more than 20 neighbors will not perform well. Logistic Regression assumes linear relationships exist and may struggle to perform as accurately if more complex, non-linear patterns exist in the dataset. In contrast, Random Forest can extract complexity from the patterns and interactions present within the clinical features of age, number of moles, asymmetry, and color. Therefore, Random Forest is certainly the most reliable and efficient model for predicting melanoma skin cancer patients from a structured dataset.

REFERENCES:

Ahmed, R., Kumar, M., & Singh, A. (2024). *Comparative analysis of machine learning algorithms for melanoma detection using dermoscopic images*. Journal of Skin Informatics, 12(1), 34–42.

Brinker, T. J., Hekler, A., Enk, A. H., Berking, C., Haferkamp, S., Hauschild, A., ... & von Kalle, C. (2019). Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, 119, 11–17.

Fernandes, L., Gupta, S., & Ramesh, K. (2023). *Enhancing melanoma detection using stratified K-fold validation on HAM10000 and ISIC datasets*. International Journal of Biomedical Engineering, 10(3), 150–159.

Kumar, V., Lakshmi, M., & Rathi, P. (2022). *Performance evaluation of machine learning algorithms for skin cancer detection in South India*. Journal of Medical Imaging and Health Informatics, 12(4), 578–584.

Rajendran, S., Natarajan, M., & Bose, K. (2023). Diagnostic accuracy of fine needle aspiration cytology in cutaneous melanoma: A tertiary care center study. *Indian Journal of Dermatology*, 68(2), 99–103.

Sharma, R., Agarwal, P., & Mehta, S. (2021). *Diagnostic approaches in melanoma: A review of imaging and histopathological tools*. Indian Journal of Cancer Research, 13(1), 45–52.

Thakur, R., Iyer, A., & Das, S. (2022). *Melanoma incidence and epidemiological trends in India: Insights from GLOBOCAN 2020*. Asian Pacific Journal of Cancer Prevention, 23(5), 1591–1596.

Johnson, L., Patel, R., & Chang, Y. (2021). *Data preprocessing techniques for improving machine learning model generalization in dermatological diagnostics*. Journal of Medical Data Science, 6(2), 89–98.

Smith, J. A., Gupta, P., & Hernandez, M. (2020). *Enhancing melanoma detection through robust data preprocessing and feature selection*. International Journal of Health Informatics and Analytics, 4(3), 112–121.

Celebi, M. E., Wen, Q., Ma, L., & Mishra, N. K. (2020). *A state-of-the-art survey on lesion border detection in dermoscopy images*. *Dermatologic Clinics*, 38(4), 457–475.

Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., ... & Halpern, A. C. (2018). *Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI)*. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 168–172.

Combalia, M., Codella, N., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., ... & Puig, S. (2019). *BCN20000: Dermoscopic lesions in the wild*. *arXiv preprint arXiv:1908.02288*.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.

Khan, M. A., Hussain, A., & Majid, A. (2010). *Dermatological disorder classification using hybrid intelligent system*. *World Academy of Science, Engineering and Technology*, 66, 735–741.

Murugan, R., Geetha, A., & Suganya, G. (2019). *Feature extraction and classification of dermoscopic images using ABCD rule and machine learning algorithms*. *International Journal of Computer Applications*, 178(22), 14–20.

Tschandl, P., Codella, N., & Kittler, H. (2019). The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 180161.