Scientific Journal of Computer Science

futuristech
PT. Teknologi Futuristik Indonesia

Received: January 20, 2026; Revised: June 7, 2026; Accepted: July 21, 2026; Published: July 25, 2026; Current version: July 25, 2026. e-ISSN: 3110-3170
Digital Object Identifier 10.64539/sjcs.v2i2.2026.416

Article

Automated Dermatologist-Level Classification of Malignant Melanoma Using Voting Ensemble Learning System

Racheal Shade Akinbo[1,*], Tosin Precious Adeyemi[2], Bamidele Moses Kuboye[2]

[1] Department of Data Science, Federal University of Technology, Akure 340252, Nigeria; e-mail: racheal.akinbo@gmail.com (A. R. Shade).
[2] Department of Information Technology, Federal University of Technology, Akure 340252, Nigeria; e-mail: preciousttheou@gmail.com (A. T. Precious).
* Correspondence

The authors received no financial support for the research, authorship, and/or publication of this article.

Abstract: Skin cancer is a life-threatening dermatological disease, with malignant melanoma representing its most aggressive form. Early detection and accurate classification of skin lesions remain challenging despite advances in computer-aided diagnostic techniques. However, further comparative evaluation is still needed to determine whether a Voting Ensemble framework offers meaningful performance improvements over individual machine learning classifiers for malignant melanoma classification. This study aimed to design and evaluate an automated classification system using a Voting Ensemble framework and compare its performance with individual machine learning classifiers. The International Skin Imaging Collaboration (ISIC) dataset containing 10,000 dermoscopic images was preprocessed, and Principal Component Analysis (PCA) was applied for dimensionality reduction. Four machine learning models, namely Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Logistic Regression (LR), were evaluated alongside the proposed Voting Ensemble model. Experimental results showed that SVM achieved the highest classification accuracy (90.70%), outperforming the Voting Ensemble (89.50%), XGBoost (89.70%), RF (86.40%), and LR (81.60%). Although the proposed Voting Ensemble integrated the predictive strengths of multiple classifiers, it did not surpass the standalone SVM under the current experimental setting, indicating that SVM achieved the best classification performance under the selected dataset and experimental setting. These findings provide useful evidence for selecting appropriate machine learning models for automated melanoma screening and highlight the importance of rigorous comparative evaluation before adopting ensemble approaches in clinical decision-support systems. The results further suggest that ensemble learning does not necessarily outperform a carefully optimized standalone classifier under all experimental conditions.

Keywords: Malignant Melanoma; SVM; LR; RF; XGBoost; Melanin; Ensemble; Dermatologist; Dermoscopic Images; Image Processing; Computer Vision.

1. Introduction

Skin cancer can be identified as the most common skin disease with the highest tendency of spreading to other tissues on the skin. This is attributed to the rising occurrence of prolonged exposure to ultraviolet rays, genetic predisposition, and environmental influences [1]. Early and accurate detection of skin cancer either benign or malignant lesions helps in improving patient survival rates. Often times, melanoma and non-melanoma skin cancers as the most aggressive type [2] and minimal cancer type respectively have been on a high trend. Across the years, there have been up to 1.5 million new cases of skin cancer on annual record with melanoma topping the list with 325,000 cases and 57,000 deaths per year [3].

Some methods used in traditional diagnosis include the use of dermoscope, biopsy and histopathological analysis but all require expertise and are resource-intensive [4]. There has been an occurrence of situations where benign lesions were incorrectly classified as malignant and vice versa resulting in delayed treatment or unnecessary bi-

opsy hence mortality rates increased significantly [5]. Lately, advancements in technology have improved classification of skin cancer for greater results while detecting skin lesion [6].

Further research was performed using International Skin Imaging Collaboration (ISIC) archive. According to Natha *et al*., (2025), these datasets offer a diverse array of dermatological images which enable the development of systems for lesion classification. Many researchers have applied Machine Learning (ML) models which include Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), neural networks and ensemble classifiers especially for differentiating between cancerous and healthy skin images.

They aid the discovery of non-cancerous growth that grow slowly without invading nearby tissues then other parts of the body (benign) and malignant are cancerous tumors that spread to other surfaces via the bloodstream, making them more serious and requires intensive treatment. SVM is a supervised ML algorithm well-suited for separation and classification tasks. With the capacity to handle high dimensional data, it is well-suited for image-based classification [7]. Each tree in the Random Forest model is trained on a random subset of training data, with feature extraction performed using bagging techniques — a bootstrap aggregating approach that promotes diversity among trees and enhances the model's generalization.

It is best fit in handling high dimensional datasets like dermoscopic images where features like texture, color and shape are extracted to differentiate between malignant and benign lesions. It also has the ability to address class imbalance, reduce variance and improve robustness to achieve higher accuracy [8]. LR is a supervised learning algorithm mainly used for binary classification tasks making it well-suited in establishing the difference between cancerous and non-cancerous skin lesions. It utilizes the principle of sigmoid function mapping input feature to either 0 or 1. One of LR's strengths is its simplicity and interpretability as it assigns coefficients to each feature, indicating their contribution to the classification outcome [9].

SVM, RF, XGBoost, and LR serve as critical components in the ensemble learning framework for skin cancer classification. RF's strength in handling non-linear data complements, XGBoost's gradient boosting, LR's probalistic approach and SVM's strength in handling high dimensional data allowing for a hybrid model that balances accuracy and interpretability. Ensemble techniques, such as stacking or boosting can combine these models and leverage on their respective strengths making it well-suited for achieving high accuracy.

Over the years, traditional diagnostic methods have depended on dermatologist expertise, dermoscopic analysis and histopathological examinations. These approaches have been time-consuming, subjective to dermatologist's expertise and prone to inter-observer variability [10]. Recently, Machine Learning (ML) and other intelligent systems have emerged as strong candidates for malignant melanoma classification. Models like SVM, RF, XGBoost and LR have shown promising results in automated lesion detection. However, a single ML model may suffer from generalization issues, particularly due to imbalanced datasets. Ensemble learning techniques which integrate multiple classifiers to enhance classification performance, have gained significant attention as an effective strategy for improving classification accuracy and robustness [11].

This paper presents the design and implementation of a malignant melanoma classification system using a Voting Ensemble framework. The main contributions of this study are threefold: (1) the development and evaluation of a Voting Ensemble model for malignant melanoma classification; (2) a comprehensive comparison of the Voting Ensemble with four widely used machine learning classifiers, namely SVM, RF, XGBoost, and LR, using the ISIC dataset; and (3) experimental evidence showing that the proposed Voting Ensemble does not outperform a carefully optimized standalone SVM under the current experimental setting, highlighting the importance of rigorous comparative evaluation before adopting ensemble methods. The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 describes the proposed methodology, Section 4 presents the implementation and experimental results, and Section 5 concludes the paper.

2. Literature Review

In this section, we study the nature of skin, skin cancer disease, and review related works on skin cancer classification. The skin is the largest organ in the body which covers the entire external surface. Containing three layers (epidermis, dermis, and hypodermis), it has different structures and functions. Furthermore, it can block pathogens based on its structure that forms the first layer. The turgidity of the skin is based on the thickness of the skin layers. The palmar surface of the hand represents the stronger part of the skin. Disruption of the skin barrier function makes it susceptible to inflammatory and infectious conditions [12]. The skin is made up of layers beginning with the epidermis made up of stratum basale, stratum spinosum, stratum lucidum and stratum corneum. Cells on the membrane produce stem cells that regularly generate keratinocytes. Stratum spinosum constitutes 8 to 10 cell layer contains irregular polyhedral cells including cytoplasmic spine that shoot outward with the nearest cells by desmosomes [13].

The function of the skin is vast, it includes maintaining overall health and well-being, protecting the body systems, and being responsible for sensation, thermoregulation, body water resistance, immune defense, synthesis

and storage [14], [15]. Skin diseases encompasses a range of conditions that affects the skin, hair nails and other related glands. It arises from infections, dysregulated immune systems, genetic anomalies, or environmental exposures. Some other types of skin disease includes; cancer of the skin, skin infections, acne, eczema, and psoriasis, among others [16], [17]. In as much as skin cancer begins with growth of skin cells forming malignant tumors, many factors contribute to the development of skin cancer but UV rays is important. Skin cancer is caused by the exposure to ultraviolet radiations, susceptible skin type, age, exposure to certain chemicals, weakened immune system or exposure to radiation [18], [19].

Cancer is the name assigned to various forms of infections such as growth of cells which can affect any part of the body. Malignant tumors and neoplasms are other terms used interchangeably. It is further classified based on tissue type source, histological classification and its primary development site. Some of the classifications includes; carcinomas, sarcomas, leukemias, lymphomas, myelomas, melanomas [20], [21]. Early detection of skin cancer significantly improves treatment outcomes, fresh growths like sores, rapid spot changes, sores that bleed, develop a crust or do not heal are some of the symptoms. Diagnosis of skin cancer involves careful observation of the skin and identification of symptoms by a healthcare professional [22].

Traditional diagnostic methods based on visual inspection and dermoscopy are limited by inter-observer variability and sensitivity constraints. Advances in machine learning have spurred numerous research efforts aimed at developing automated, reliable and accessible systems for skin cancer classification [23].

Natha *et al*., (2025) proposed the boost: diagnostic accuracy of skin cancer using ensemble approach with the aim to improve the classification accuracy of skin cancer lesions by combining multiple machine learning models through an ensemble method and optimize feature selection from dermoscopic images using a Genetic Algorithm (GA) to enhance model performance [24]. The project utilized two publicly available datasets: HAM10000 and ISIC 2018 which contain labeled dermoscopic images of various skin lesions. Relevant features were extracted from the dermoscopic images and GA-based feature vectors to improve the models' predictive power. The ensemble was trained with distinct learning models like Random Forest (RF), Multi-layer Perceptron Neural Network (MLPN) and Support Vector Machine (SVM) using max-voting which combined the predictions of the three models where each of the voted class appearing as the major votes is chosen as the final prediction. The ensemble approach using the Max Voting method achieved an accuracy of 94.7% outperforming individual models which demonstrated better performance as regards F1-measure, recall and precision which presents a robust diversity in cancerous organs where Gradient Boosting has the tendency of improving the model's capabilities.

Kanchana *et al*., (2024) focused on enhancing skin cancer classification using EfficientNet B0-B7 through convolutional neural networks [25]. The study developed a focused image preprocessing pipeline which includes resizing the images to the exact image dimension recognized by EfficientNet models mainly to ensure consistency across the dataset. The dataset were augmented by using flipping, gradual expansion and rotation then unwanted artifacts including hair and markers were removed from the images to prevent them from affecting model training. Also, the model was developed using transfer learning approach and convolutional neural networks (CNN) evaluating eight EfficientNet models (B0-B7) which has revealed that EffcientNet-B7 achieved top-1 accuracy of 84.4% and top-5 accuracy of 97.1%. However, some challenges were identified while classifying eczema, wart molluscum and psoriasis which reflects the need for basic improvements in some areas. Kanchana *et al*., made emphasizes on the importance of integrating the model into a seamless user interface for real-time global accessibility. Highlighting limitations such as generalization and specificity, it was suggested that future research should test the model on diverse datasets beyond local dataset.

Kandhro *et al*., (2024) experimented the E-VGG19 Model: Improvement of real-time skin cancer detection and classification [26]. It focuses on improving accurate detection of malignant melanoma through an enhanced version of VGG19 model, termed E-VGG19 a pre-trained model which combines ML classifiers (SVM). The researchers utilized a complex dataset of diagnostic images, encompassing various types of skin images. Each of which undergoes preprocessing steps which include resizing, normalization and artifact removal to achieve uniform dimensions for a consistent dataset. This was achieved by integrating additional dense layers and applying max pooling techniques. These enhancements help in extraction of features and for improved performance in classifying malignant melanoma. Features were extracted from the enhanced E-VGG19 model and were fed into carious ML classifiers including SVM, KNN, Decision Tress and LR leveraging the strength of both deep learning and traditional machine learning techniques to enhance classification accuracy. The model's performance was validated using standard metrics, achieving an accuracy of 84.22% and AOC-ROC curve 0.94 for classification of malignancy. After undergoing 50 iterations of training, the work faced issues of generalization and optimization and improvements such as deployment of the model for real-time use were suggested, user-friendly systems to enhance accessibility and practical application in clinical settings.

Jaberi *et al*., (2024) conducted research on skin cancer image classification using two-level ensemble deep-learning with the aim of improving the accuracy and consistency of benign versus malignant classification through a novel two-level ensemble approach [27]. The research utilized a subset of ISIC public dataset containing images labeled as positive or negative (benign and malignant respectively). The first ensemble combined the VGG-16 and VGG-19 models, both renowned for their deep architectures and effectiveness in image classification and the other combined ResNet-50 and Res-Net-152 using inception models to leverage their strength in handling deeper networks and capturing intricate image features. At each ensemble level, CatBoost algorithm (gradient boosting method optimized for categorical features) was employed to combine the outputs of the individual models. The research aimed to enhance prediction performance by effectively managing the contributions of each model within the ensemble. The outputs from the VGG and combination of ResNet were further concatenated with the use of CatBoost algorithm. A two-level ensemble strategy was designed to capitalize on the diverse feature extraction capabilities of the different models thereby improving classification accuracy. The model was evaluated using standard metrics, demonstrating higher accuracy and consistency in malignancy classification. The current work aims to leverage ensemble techniques with a stacking algorithm to achieve comparable results across diverse clinical settings.

Hassan (2024) proposed an ensemble system that can efficiently diversify skin disease types using dermoscopic images [28]. The approach involved training CNN and Transfer Learning models and combining their outputs through a voting mechanism to classify lesions into various categories. High-quality dermoscopic images representing various skin cancer types, acquired from clinical settings, underwent preprocessing steps beginning with contrast enhancement and noise removal. The limited dataset was augmented and enhanced for model generalization capabilities using rotation, zooming, flipping and color adjustment to artificially increase the size and diversity of the training dataset. Relevant features were extracted from the images using CNNs, capturing essential information for accurate skin cancer detection. Bagging techniques with SVMs and RFs were applied to the base learners. The outputs of the individual CNN models were combined using majority voting scheme (each model cast a vote for a class and the majority vote was selected as the final prediction. The ensemble model's performance was evaluated using metrics such as accuracy and other standard metrics and achieved overall accuracy of 71% outperforming individual CNN models. The research encountered challenges such as dependency on high-quality annotations in the training data and some skin cancer types were under-represented in the dataset.

Adla *et al*., (2023) developed a full-resolution convolutional network using a dynamic graph cut algorithm for skin cancer classification [29]. The research aimed to optimize hyperparameters for a full-resolution convolutional network (FrCN). It is capable of accurately identifying disease types from diagnostic images while addressing common segmentation challenges, such as over-segmentation and under-segmentation by integrating a dynamic graph cut algorithm. FrCN architecture was employed for semantic segmentation of skin lesions. Unlike traditional convolutional networks that may lose spatial resolution due to subsampling to maintain full image resolution then enable precise pixel segmentation. Dynamic graph cut algorithm was used in the hyper parameters optimization of the FrCN (this method combines individual and global search strategies to balance exploration and exploitation, facilitating effective hyperparameter tuning. The dataset was augmented during training and test phase improving model performance and generalization capabilities. The model was evaluated through multiple experiments including various transfer learning models. Performance metrics can also assess the effectiveness of the FrCN combined with the dynamic graph cut algorithm. Though the mode achieved an accuracy of 97% in classifying skin lesions, the study does not specify the diversity of the datasets used for evaluation. Evaluating the model's performance on larger datasets is essential to ensure its robustness and generalizability. Additionally, clinical validation with contributions from specialist healthcare practitioners is necessary to ensure practicality and authenticity in diagnostic settings.

Kalpana *et al*., (2023) studied OESV-KRF-optimal ensemble Support Vector kernel Random Forest based early detection and classification of skin diseases which aimed to combine SVM and RF classifiers to improve the efficiency and adaptability of skin disease diagnosis in real-time and resource-limited settings [30]. The researchers collected a dataset comprising dermoscopic images of various skin disease, used preprocessing steps including contrast enhancement, noise reduction, and normalization to ensure uniformity and improve image quality. Texture analysis, color histogram analysis and shape descriptors were used as techniques for feature extraction. An SVM classifier with an optimal kernel function was trained on the extracted features to maximize the margin between different classes, together with an ensemble of decision trees to improve classification robustness. The output of the SVM and RF classifiers were combined using an ensemble strategy to leverage the strengths of both models. The OESV-KRF model demonstrated an efficiency rate of 70% alongside other standard metrics. Limitations encountered include the absence of a user-friendly interface, which is crucial for practical implementation and usability by healthcare professionals. Furthermore, the research

lacked a detailed analysis of classification specificity across different skin disease categories, which is essential for evaluating the model's effectiveness.

Teodoro *et al*., (2023) studied dermatological disease classification approach using GAN and RoI-based attention mechanism proposing a novel approach using EfficientAttentionNet, a Convolutional Neural Network (CNN) architecture, for detecting cancerous lesions classified as positive or negative [31]. The project developed an automated system to enhance skin cancer detection by reducing diagnostic errors associated with symptoms similar to those of other diseases. The work also balanced the training dataset by generating synthetic images to address class imbalance and implemented a mask-based attention mechanism focusing on the region of interest (RoI) within lesion images while it improves classification accuracy. The work applied techniques to remove hair and other artifacts from skin images to ensure clearer lesion visualization. It also employed Generative Adversarial Networks (GANs) in the generation process of synthetic images effectively balancing the number of samples across different classes in the training set. They also used a region of interest (RoI) approach to train a U-Net model, generating a mask that highlights the RoI within each skin image, thereby emphasizing the lesion areas. Model used is EfficientAttentionNet, a CNN architecture incorporating as mask-based attention mechanism, it was trained using the preprocessed and augmented dataset and evaluated using standard metrics to assess effectiveness. While the attention mechanism provides insights into features influencing classifications, conducting clinical trials to validate model performance in practical settings would further strengthen confidence in its utility for early skin cancer detection.

Tang *et al*., (2023) made a study on joint-individual fusion structure targeting fusion attention module for multi-modal skin cancer classification which introduces an innovative approach to enhance skin cancer classification by integrating dermoscopic images with patient metadata [32]. This study aimed to develop a fusion method that improves skin cancer classification accuracy and to introduce a fusion attention module to emphasize the most relevant features from both image and metadata inputs, thereby enhancing the decision-making process. The Joint-Individual Fusion structure concurrently extracts shared patterns across multiple modalities while preserving unique characteristics specific to each modality with the purpose of capturing comprehensive information from both dermatological images and patient metadata. The Multi-Modal Fusion Attention (MMFA) module used employs both self-attention, and mutual-attention strategies to highlight the most pertinent features within the image and metadata inputs. The proposed JIF and MMFA components are integrated with various Convolutional Neural Network (CNN) architectures to evaluate their effectiveness across different model frameworks and it improved classification outcomes and outperformed existing state-of-the-art fusion techniques using the three public datasets. After evaluation, public datasets including PAD-UFES-2, Seven-Point Checklist (SPC) and ISIC-2019, further testing on more diverse and larger datasets is necessary to ensure its applicability across different populations and skin types. Beyond quantitative metrics, clinical validation involving dermatologist assessments is necessary to confirm the practical applicability and reliability of the model in real-world diagnostic settings.

Huang *et al*., (2023) explored classification of skin cancer using novel hyperspectral imaging engineering via YOLOv5 [33]. This study applied Hyperspectral Imaging (HSI) together with deep learning techniques to identify malignant skin lesions, with the aim of investigating the efficacy of hyperspectral imaging for improving skin cancer detection and classification, utilize YOLOv5 architecture (You Only Look Once version 5), and to compare the efficiency of models trained on hyperspectral narrowband lesions (HIS-NBI) compared to traditional RGB images. The approach utilized ISIC dataset focusing on images labeled as BCC, SCC, and SK with resolution of 640x640 pixels to maintain consistency and reduce computational load. The hyperspectral image generation analyzed the normalized reflection spectra of the lesion areas, converted RGB images into hyperspectral narrowband images (HIS-NBI) and applied PCA to reduce the dimensionality of the hyperspectral data. YOLOv5 architecture was implemented with two models: one using the original RGB images and other generated HIS-NBI images, combined with data augmentation to improve model generalization and prevent overfitting. The HIS-NBI model demonstrated a recall rate improvement for SCC detection increasing from 0.722 (RGB model) to 0.794 indicating a 7.5% improvement, while the overall accuracy was 78.7% compared to 79.2% for the RGB model. While the HIS-NBI model showed improvements in specific areas, overall accuracy remained similar to the RGB model. Hence, there is a need to better understand the contexts in which hyperspectral imaging provides significant advantages.

This research work aims to apply ensemble learning techniques to accurately classify skin cancer lesions into malignant and benign categories from dermoscopic images.

3. Method

3.1. System Architecture

High-quality labeled images of skin lesions were used and underwent a cleaning and quality control process to ensure dataset consistency. The proposed system consists of data acquisition, image preprocessing, feature extraction, data balancing, dimensionality reduction, model

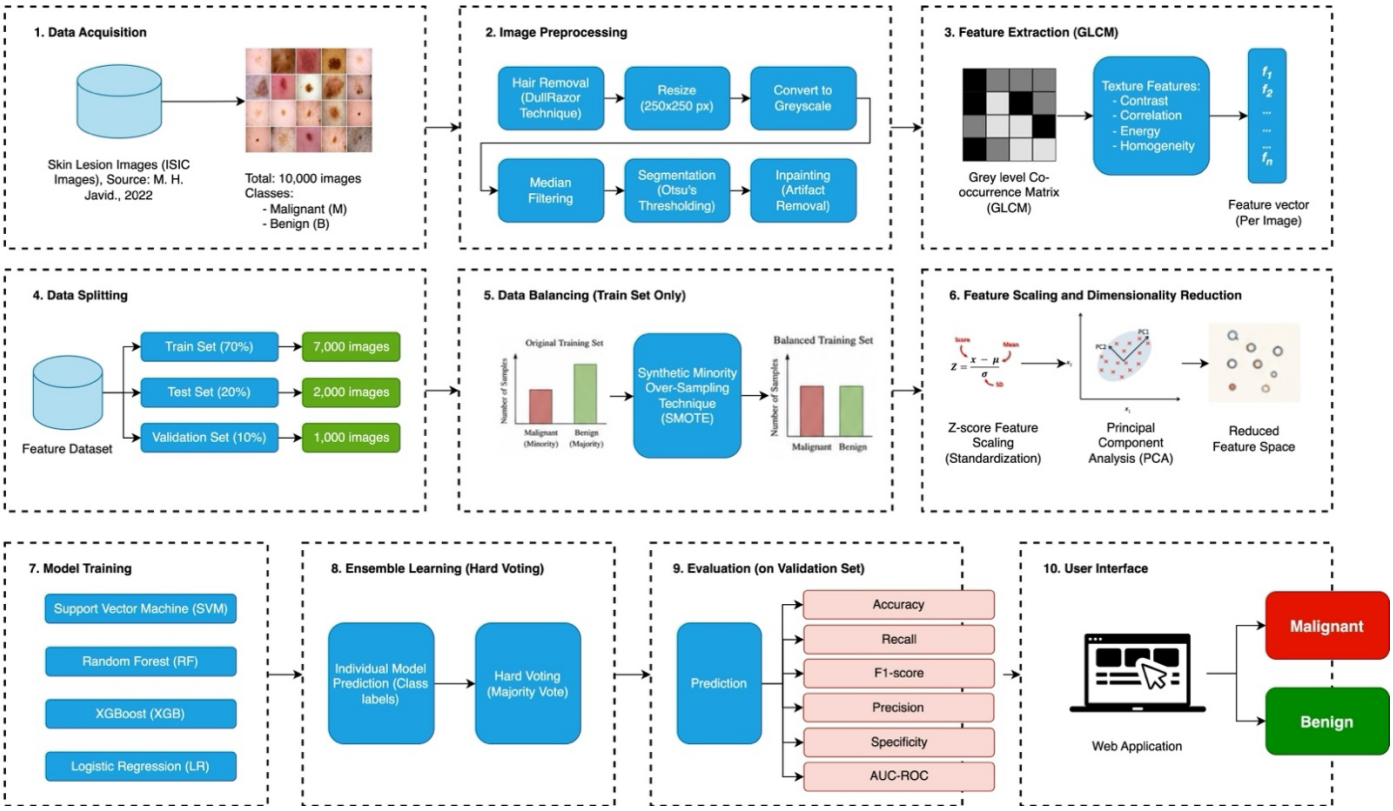


Figure 1. Architecture of the Malignant Melanoma Classification System.

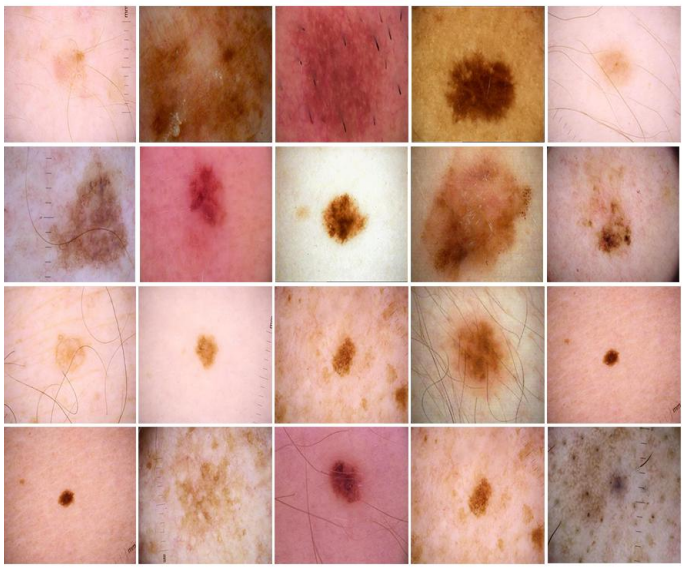


Figure 2. Skin Cancer Image (ISIC Images). Source: M. H. Javid., 2022 [34].

training, ensemble learning, and performance evaluation before generating the final classification outcome, as illustrated in Figure 1.

3.2. Data Collection

The dataset used in this study was obtained from the Melanoma Skin Cancer Dataset by M. H. Javid., 2022 [34] with high resolution dermoscopic images covering a range of skin lesions labeled malignant and benign. This constitutes the most extensive part of the learning approach, as model efficiency is directly dependent on the quantity and quality of the collected dataset. The images were arranged and converted to uniform dimensions to ensure consistency. Sample images are shown in Figure 2.

3.3. System Requirements

The implementation of the classification system was carried out with hardware and software requirements as stated below:

Hardware specification:

a) Processor: a low-power, high-performance processor capable of handling multiple tasks, not lower than Core i5, 8th generation and 1.60GHz.
b) Memory: sufficient RAM and storage to handle data without computational bottlenecks, specifically not lower than 16.0 GB RAM and 500 GB ROM.
c) Power: a reliable power source, such as a battery or direct connection to a power source.

Software specification:

a) Operation system: A lightweight and efficient OS that can handle real-time data processing.
b) Network stack: A stable and secure network stack for internet connectivity and data transmission. A fast internet connection for real time upload is required.
c) Integrated Development Environment (IDE): A platform for model training, validation and analysis.

d) Libraries: libraries imported includes; pip, NumPy as np, matplotlib.pyplot as plt, sklearn.metrics, LabelEncoder, tensorflow.keras.preprocessing.image, sklearn.utils, cv2, tqdm, xgboost, sklearn.decomposition, RandomForestClassifier.

3.4. Data Preprocessing

The image quality was enhanced for clarity, and relevant features were extracted by removing unwanted components from the image. Images were rescaled to 250x250 px to ensure uniform feature extraction, while image transformation introduced controlled deformation so that the model could robustly recognize variations in the shapes of skin lesion images, increasing image size also improves the dataset by simulating varied perspectives and magnifications. Flipping horizontally differentiates the dataset by generating mirrored images. GLCM techniques were used to process the intensity values, the RGB images were transformed to grayscale applying the transformation below:

$$Y = 0.2989\ x\ R + 0.5870\ x\ G + 0.1140\ x\ B \quad (1)$$

where; Y is the grayscale transformation, R as Red, G for Green and B represents Blue.

Reshaping the images ensures dimensional compatibility with the model architecture. Application of median filtering removes noise while preserving edges, enhancing image quality for feature extraction. Segmentation methods like Otsu's thresholding was used to differentiate the lesion area from the surrounding skin focusing the analysis on relevant regions. A mask was generated for inpainting tasks, enabling the algorithm to reconstruct corrupted regions of the images. Inpainting algorithms were applied to seamlessly substitute imperfections, thereby ensuring a robust dataset for cancer detection models. These image processing methods address challenges encountered in real-world scenarios, thereby enhancing the diagnostic capabilities of the system.

3.5. Data Splitting

The dataset was split into training, testing, and validation subsets using a 70:20:10 ratio mathematically expressed as;

Training set D_{train} contains:

$$D_{train} = 0.70D \quad (2)$$

Testing set D_{test} contains:

$$D_{test} = 0.20D \quad (3)$$

Validation D_{val} set:

$$D_{val} = 0.10D \quad (4)$$

Where:

$D_{train} + D_{test} + D_{val} = D$

D is the total dataset

The partitioning ensures sufficient data for model learning, parameter tuning and unbiased performance evaluation.

Let:

x_i as a sample from the inferior class

x_{zi} be one of the KNN nearest neighbor of x_i

λ be a random number such that $\lambda \epsilon$ [0,1]

$$x_{new} = x_i + \lambda \cdot (x_{zi} - x_i) \quad (5)$$

where x_i and x_{zi} are vectors in feature space (e.g GLCM features) and x_{new} is the generated synthetic sample.

3.5.1. Feature Scaling and Extraction

Feature scaling and extraction are of great importance in the classification process, as relevant features must be extracted from the input dataset for use in the machine learning models. Gray Level Co-occurrence Matrix (GLCM) is method that requires deep calculations which places the texture with respect to spatial relationship in image vision. Texture features were extracted from the GLCM computed at four directions (0˚, 45˚, 90˚,135˚) and one-pixel distance. It calculates how often pairs of pixels with specific values occur in a specified spatial relationship within an image, creating a matrix from which texture features are extracted. The statistical features below were captured using GLCM.

a) Contrast: this measures the intensity difference between a pixel and its surrounding neighbors in the image. It is represented as;

$$\sum_{ij=0}^{N-1} P_{ij}(i-j)^2 \quad (6)$$

b) Correlation: measures the linear dependency of grey-level values between neighboring pixels in an image. It is represented below as;

$$\sum_{ij=0}^{N-1} P_{ij} \frac{(i-\mu)(j-\mu)}{\sigma^2} \quad (7)$$

c) Energy: measures the uniformity of the image texture, equating to the sum of squared elements in the GLCM. It is represented below as;

$$\sum_{ij=0}^{N-1} (P_{ij})2 \quad (8)$$

d) Homogenity: this measures the closeness of the distribution of elements in the GLCM to the diagonal. It is represented below;

$$\sum_{ij=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2} \quad (9)$$

The value of *P(i, j)* is given as the GLCM entry at row i and column j, while μ and σ represent the mean and standard deviation of row and column pixel intensities, respectively. These features collectively characterize the texture patterns of skin lesions.

3.5.2. Normalization and Data Balancing

After extracting features using the GLCM, it is essential to normalize them to ensure balanced contributions in the learning process. Some classifiers like SVM are sensitive to feature scales.

a) Min-max normalization:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (10)$$

where scale features ranges [0, 1]

b) Z-score standardization:

$$x_{std} = \frac{x - \mu}{\sigma} \quad (11)$$

This transforms features to have zero mean and unit variance.

Data balancing was employed primarily to prevent bias towards the majority class at the expense of the minority class. SMOTE is an effective technique that addresses this by synthetically generating new, plausible samples for the minority class, thereby reshaping the class distribution to improve model performance. SMOTE fills the gaps between the minority scenario in feature space by interpolation; for each minor sample x_i, get the KNN using *k=5*, which occasionally use a neighbor x_{ij}. Hence synthetic sample can be generated:

$$p_{ij} = x_i + \delta \times (x_{ij} - x_i), \delta \sim Uniform\ (0, 1) \quad (12)$$

This new sample p_{ij} lies somewhere along the line segment connecting x_i and x_{ij}.

3.5.3. Principal Component Analysis (PCA)

This is an unsupervised dimensionality reduction method used to transform a high-dimensional dataset into a set of uncorrelated components. They are regarded as the main determinants as it retards much variance as required. Before applying PCA, features must be standardized to zero mean and unit variance to avoid scale dominance:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (13)$$

where μ_j and σ_j are the standard deviation and mean of feature j. Covariance matrix $p \times p$ written as:

$$\sum = \frac{1}{n-1} X^T X \quad (14)$$

where X is given as the standardized data matrix.

Eigenvalue decomposition was performed as:

$$\sum = W \Lambda W^T \quad (15)$$

where W contains eigenvectors (loadings), and Λ is the diagonal matrix of eigenvalues representing variance explained.

The principal component score is then discovered by $T = XW$ where each row in T represents principal component score for a sample.

3.5.4. Dimension Reduction and Feature Selection

As regards Principal Component Analysis (PCA), only the main component containing the highest cumulative variance were reserved for model training. It removed redundant and correlated features while preserving the discriminative information contained in the original dataset. The feature space that was reduced decreased computational complexity, reduced overfitting and improved classifier efficiency without significant loss of information.

3.6. Model Training

The design utilizes SVM, RF, XGBoost, and LR to effectively classify skin lesions into malignant and benign categories.

3.6.1. Support Vector Machine (SVM)

SVM is a supervised learning model trained on labelled data, used for classification and prediction tasks. It identifies a hyperplane in an n-dimensional space that optimally separates data into distinct classes, it uses the closest data points to the decision boundary as support vectors to determine the optimal separating hyperplane. Furthermore, the algorithm derives a hyperplane from the labeled training data to classify new sample points. For a linearly separable dataset, SVM seeks the hyperplane that differentiates classes with the maximum margin. The general architecture of SVM is illustrated in Figure 3.

The training dataset is given as $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in R^d$ and $y_i \in \{-1, 1\}$. The objective is to find $\omega \in R^d$ and $b \in R$ such that the hyperplane $\omega^T x + b = 0$ separates the classes with the classes with the maximum margin. Optimization problem:

$$\min_{w,b} \frac{1}{2} ||w||^2 \quad (16)$$

subject to:

$$y_i (\omega^T x_i + b \geq 1, \forall\, i = 1, \ldots, n \quad (17)$$

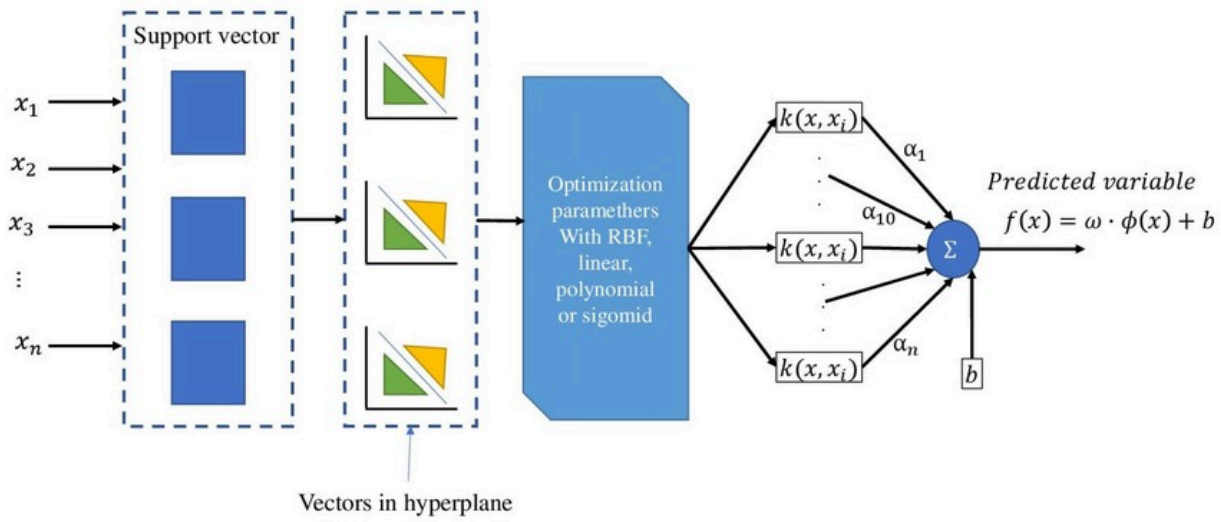


Figure 3. SVM General Architecture. Source: Alvarez A *et al*., 2021 [35].

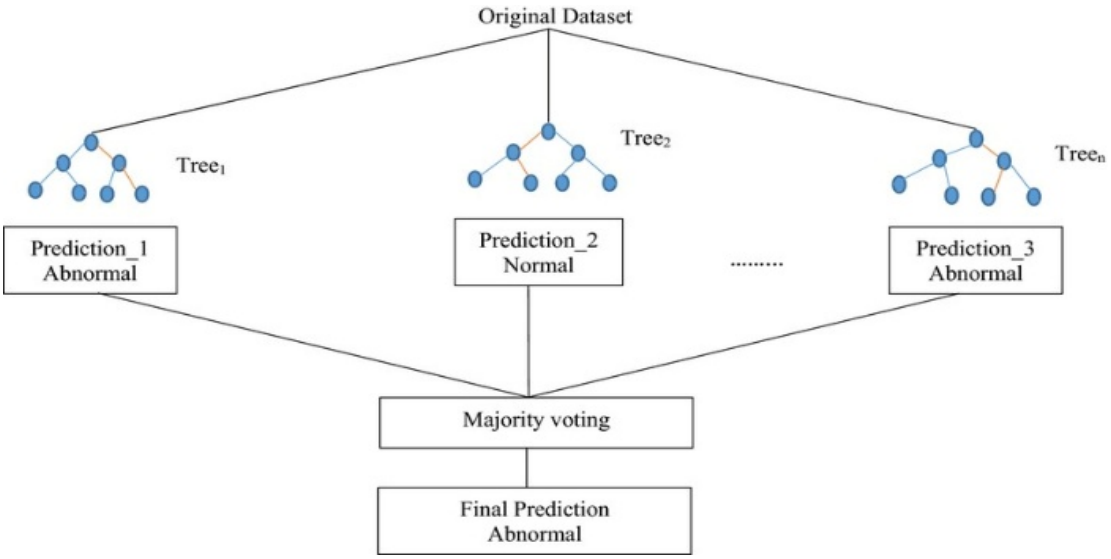


Figure 4. Architecture of Random Forest Classifier. Source: Shafi A. S *et al*,. 2020 [36].

The optimization problem equation shows that all data points are correctly classified and maximizes the margin.

For data which are not perfectly separable, slack variables given as $\varepsilon_i \geq 0$ were introduced to handle the misclassifications. Optimization problem is then modified as:

$$\min_{w,b,\varepsilon} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \varepsilon_i \quad (18)$$

subject to:

$$y_i\,(\omega^T x_i + b) \geq 1 - \varepsilon_i, \forall\, i = 1, \ldots, n \;\; \varepsilon_i \geq 0, \forall\, i = 1, \ldots, n \quad (19)$$

Equation 19 shows $C > 0$ is a regularization parameter that for balancing and maximizing the margin and minimization of the classification error.

If the data is not linearly separable in the original feature space, SVM utilizes functions, kernel functions to be precise which is given as $K(x_i, x_j)$ to align data with a higher dimensional space where linear division is possible.

Dual formulation:

$$\max_{\propto} \sum_{i=1}^{n} \propto_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \propto_i \propto_j y_i y_j K(x_i, x_j) \quad (20)$$

subject to:

$$0 \leq \propto_i \leq c, \forall\, i = 1, \ldots, n \quad (21)$$

$$\sum_{i=1}^{n} \propto_i y_i = 0 \quad (22)$$

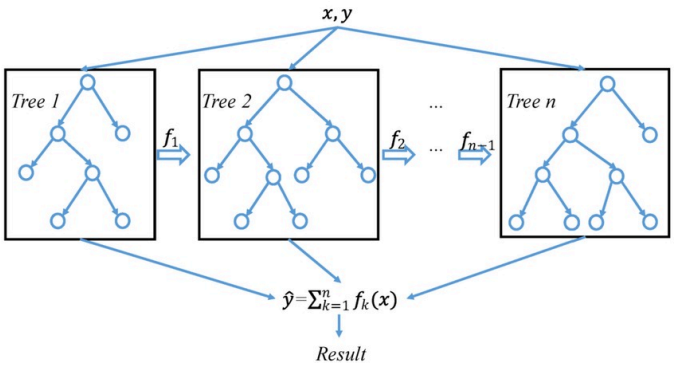


Figure 5. A general architecture of XGBoost. Source: Wang *et al*,. 2019 [37].

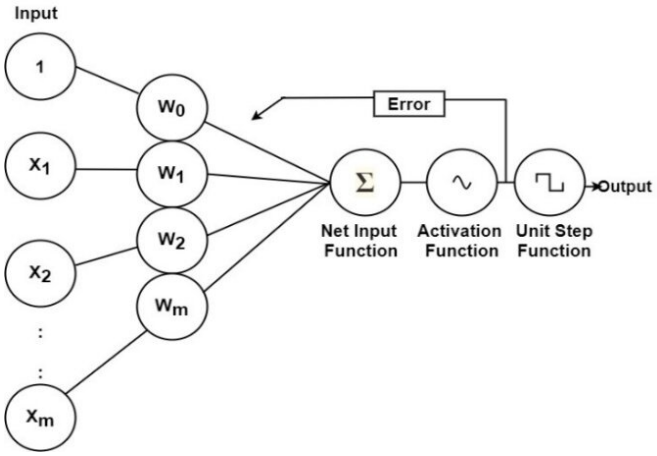


Figure 6. Architecture of Logistic Regression. Source: Biswas *et al*., 2023 [38].

The decision function therefore results to:

$$f(x) = sign(\sum_{i=1}^{n} \propto_i y_i K(x_i, x) + b) \quad (23)$$

Linear: $K(x_i, x_j) = x_i^T x_j$
Polynomial: $K(x_i, x_j) = (x_i^T x_j + r)^d$
Radial Basis Function: $K(x_i, x_j) = exp\ (-\gamma\ ||x_i - x_j||^2)$

The above is the kernel function written as liner, polynomial and Radial Basis Function (RBF) respectively.

3.6.2. Random Forest (RF)

RF is a powerful ensemble learning algorithm that constructs multiple decision trees during training. It uses feature randomness and bagging to create a diverse set of trees, producing superior results compared to a single decision tree. The general architecture of RF is illustrated in Figure 4.

Given a training dataset:

$$D = \{(x_i, y_i)\}_{i=1}^{n} \quad (24)$$

where; $x_i \epsilon R^d$ represents the feature vector y_i as the target variable

Steps:

1) Draw a boostrap sampling D_b of size n from D
2) From each node in the tree, a random subset of m features is selected from the total d features.
3) Then, the best split among the m features is determined based on the chosen criterion (e.g Gini impurity, information gain).
4) Lastly, the tree is grown to its maximum depth without pruning.

For efficient classification, a new input x with each tree T_b provides a class prediction $h_b(x)$ The final prediction $H(x)$ is the majority vote:

$$H(x) = mode\{h_b(x)\}_{b=1}^{B} \quad (25)$$

Regression: Each tree provides a numerical prediction $h_b(x)$. The final prediction H(x) is the average:

$$H(x) = \frac{1}{B} \sum_{b=1}^{B} h_b(x) \quad (26)$$

Gini impurity measures the probability that a randomly selected instance would be incorrectly classified. Gini Impurity G(t) is:

$$\mathrm{G(t)} = 1 - \sum_{k=1}^{k} [p(^k/_t)]^2 \quad (27)$$

where k is taken as the record of class, $p(^k/_t)$ is quantity of instances of section k in node t.

3.6.3. Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced development of gradient boosting machines, designed for speed and efficiency. It builds an ensemble of decision trees sequentially, where each new tree corrects the errors of its predecessors (see Figure 5). The objective function of XGBoost combines a loss function that measures model performance with a regularization term that penalizes model complexity:

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{n} l\,(y_i, Y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (28)$$

Where; $\mathcal{L}$ is given as a distinguishable convex loss function that is mean squared error for regression, logistic loss for classification.

Y_i is the prediction for instance i, $\Omega(f) = \gamma T + \frac{1}{2\lambda}||w||^2$ is the regularization criterion, Given T as the number of leaves in the tree, w the vector of scores on leaves Y and λ are regularization parameters.

The model employs an additive approach. At iteration t, a new function f_t is combined to reduce the details as shown below:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, Y_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (29)$$

XGBoost also employs second-order Taylor approximation to optimize the loss function, as shown in Equation 29 above.

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[l\left(y_i, Y_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \quad (30)$$

Where:

$g_i = \partial_{Y_i^{(t-1)}} l(y_i, Y_i^{(t-1)})$ is the first derivative (gradient).

$h_i = \partial^2_{Y_i^{(t-1)}} l(y_i, Y_i^{(t-1)})$ is the second derivative (Hessian).

Each f_t corresponds to a regression tree. For a tree with P leaves, the structure is defined by a function q: $R^d \rightarrow \{1,2,\ldots,P\}$ that maps an instance to a leaf index, and each leaf j has an output score w_j.

Thus,

$$f_t(x) = w_q(x) \quad (31)$$

The regularization term is then given as:

$$\Omega(f_t) = \gamma P + \frac{1}{2\lambda} \sum_{j=1}^{P} w_j^2 \quad (32)$$

The optimal weight w_j for each leaf j is computed as:

$$w_j^* = -\frac{\sum_{i \epsilon l_j} g_i}{\sum_{i \epsilon l_j} h_i + \lambda} \quad (33)$$

The l_j is the occurrence of scenerios in leaf j. The resulting optimal equivalent of the objective function is given below:

$$\mathcal{L}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \epsilon l_j} g_i)^2}{\sum_{i \epsilon l_j} h_i + \lambda} + \gamma P \quad (34)$$

The quality of a tree structure and guide the tree-building process was evaluated using the formula given in Equation 34.

3.6.4. Logistic Regression (LR)

LR is a supervised learning algorithm used for binary classification tasks (0/1), particularly in applications such as fraud detection, churn prediction, and disease diagnosis. Unlike linear regression, which predicts continuous outcomes, LR models the probability of a discrete event by leveraging the logistic (sigmoid) function to constrain the output between 0 and 1. The general architecture of LR is illustrated in Figure 6.

The core of LR is the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (35)$$

This function maps any real-valued input z to a probability in (0, 1). The logit z, representing the linear combination of input features, is computed as shown in below:

$$Z = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \quad (36)$$

where x_i are feature values, β_i are model coefficients, β_o is the bias/intercept.

Then, the probability estimation is written as shown in below:

$$p = P(y = 1/x) = \sigma(z) = \frac{1}{1 + e^{-(\beta_o + \sum_{i=1}^{m} \beta_i x_i)}} \quad (37)$$

This gives the probability that the target belongs to class 1.

Log-odds can also be alternatively expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \sum_{i=1}^{m} \beta_i x_i \quad (38)$$

As shown here, the left is the log-odd, linear in parameters, Exponentiating coefficient β_j gives the odds ratio for feature x_j: OR = e^{β_j} This represents the multiplicative increase in odds for a unit increase in x_j.

In training the model, model parameters (β) are estimated by maximizing the likelihood:

$$\mathrm{L}(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \quad (39)$$

or equivalently, maximizing log-likelihood which are shown in Equations 39 and 40.

$$\mathrm{L}(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (40)$$

3.6.5. Voting Ensemble

Voting ensemble technique combines predictions from multiple classifiers to improve overall model performance. The general architecture of voting ensemble is illustrated in Figure 7.

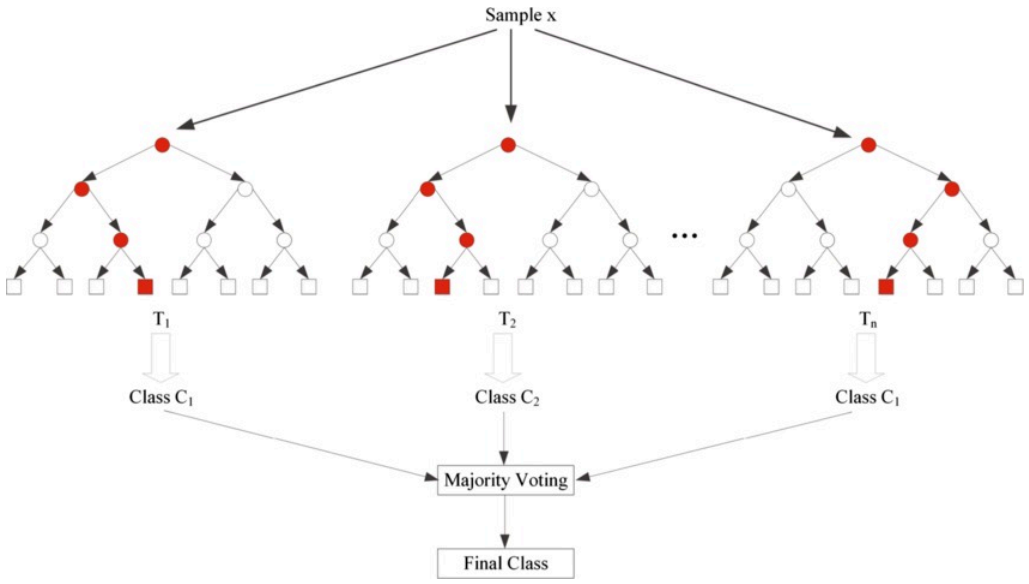


Figure 7. Architecture of voting-ensemble of tree. Source: Zhang *et al*,. 2020 [39].

It is further divided into methods namely below;

a) Hard Voting is a method whereby each classifier casts a vote for a class label, and the class with the highest number of votes is selected as the final prediction.
Let: $c_j\ (x)$ be the assumption of the j^{th} classifier for input x, m be the total number of classifiers and Y be the set of possible class labels.
The final prediction ỹ is:

$$\tilde{y} = \arg max_{y \in Y} \sum_{j=1}^{m} I(c_j(x) = y) \quad (41)$$

where *I* indicates the function, then returns 1 if the condition involved is true, else 0.

b) Soft Voting is a method whereby each classifier provides a probability estimate for each class, and the class with the highest average probability is selected as the final prediction.
Let: $P_j(y/x)$ be the predicted probability of class y by classifier j for input x, w_j be the weight assigned to classifier j.
Hence, the final prediction ỹ is:

$$\tilde{y} = \arg max_{y \in Y} \sum_{j=1}^{m} w_j\ P_j\ (y/x) \quad (42)$$

If all classifiers are equally weighted, then $w_j = \frac{1}{m}$ for all *j*.

3.7. Model Evaluation

The proposed model architecture for classification of malignant melanoma was evaluated using standard classification metrics based on the confusion matrix. True Positive (TP): malignant lesions correctly classified as malignant. True Negative (TN): benign lesion correctly classified as benign. False Positive (FP): benign lesions incorrectly classified as malignant. False Negative (FN): malignant lesions incorrectly classified as benign.

a) Accuracy: the percentage of instances correctly classified. Equation 43 represents the mathematical calculations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (43)$$

b) Precision: this is the proportion of predicted positive instances that were correctly classified, it is expressed below:

$$Precision = \frac{TP}{TP + FP} \quad (44)$$

c) Recall is shown below in Equation 45.

$$Recall = \frac{TP}{TP + FN} \quad (45)$$

d) F1-score: This calculates the harmonic mean of precision and recall.

$$F1 - score\ = 2 \times \frac{(Precision\ \times recall)}{(Precision\ +\ recall)} \quad (46)$$

e) Sensitivity: This is the True Positive Rate (TPR), which measures the proportion of actual positive cases correctly identified by the model.

$$Sensitivity\ = \frac{TP}{TP + FN} \quad (47)$$

f) Area under the curve (AUC): This measures the model's ability to discriminate between positive and negative classes across all classification thresholds. It is shown in Equation 48.

$$\text{AUC} = \int_0^1 TPR\,(FPR)d(FPR) \tag{48}$$

where TPR (True Positive Rate) is derived from Sensitivity as shown in Equation 49.

$$FPR\,(False\ Positive\ Rate): \frac{FP}{FP+TN} \tag{49}$$

4. Implementation and Result

This chapter focuses on the classification of malignant melanoma using ensemble learning techniques, presenting how different libraries and code components are integrated with the required models, which were trained and validated to achieve high accuracy. Models used for this classification includes SVM, RF, LR and XGBoost combined with an ensemble using voting techniques. These models were trained using a publicly available dataset from Kaggle and validated to ensure accurate classification into malignant and benign categories, subsequently integrated into a backend system designed to accept data input for generalization. For the system to be efficient and effective, it has some basic requirements required to be fulfilled.

4.1. Data Analysis

The models used for the classification system were analyzed sequentially on the development IDE. Standard metrics including recall, F1-score, precision, and accuracy were used to evaluate the implementation.

The data reading process involves careful examination and documentation of image details to achieve the desired classification outcome. Dimension 250x250px was used as the image size for proper examination of image features. For optimal classification accuracy, standard scaling, label encoding, and SMOTE techniques were applied for data balancing. It was observed that using larger image dimensions improves pattern recognition.

4.2. Discussion of Model

The models employed for this project are XGBoost, RF, SVM, and LR, combined with a voting ensemble technique chosen for its effectiveness in handling classification tasks through diverse algorithmic approaches.

During the model training, each model was trained on a cleaned and preprocessed dataset comprising up to 7,000 lesion images, representing 70% of the entire dataset. Hyperparameters for each algorithm were optimized using cross-validation techniques such as GridSearchCV to enhance model performance.

The confusion matrix in Figure 8 shows the performance of the SVM model on the validation dataset. Out of the 500 actual positive cases, the model correctly classified 460 as positive, while 40 were incorrectly classified as negative (false negatives). That is, out of 500 actual negative cases, 447 were correctly identified as negative while 53 were incorrectly classified as positive (false positives).

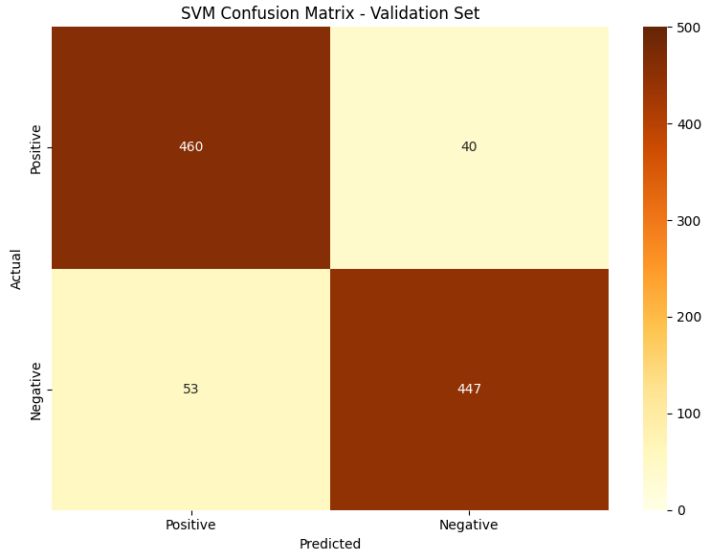
The high number of correctly classified instances (460 true positives and 447 true negatives) compared to the relatively low number of misclassifications indicated that the SVM model achieved strong predictive performance. The model demonstrated a good ability to distinguish between positive and negative classes, although a small number of classification errors were observed. Above all, the confusion matrix suggests that the SVM classifier is reliable and effective for the classification task, with most observations correctly assigned to their respective classes. The result indicates that the SVM model provides a high level of classification accuracy and maintain a balanced performance across both positive and negative classes.

The Figure 9 represents the confusion matrix for Random Forest model on the validation dataset. The model correctly classified 403 out of 500 positive cases as positive (True Positives), while 97 positive cases were incorrectly classified as negative (False Negatives). For the negative class, the model correctly identified 461 out of 500 cases as negative (True Negatives), with only 39 cases incorrectly classified as positive (False Positives).

The results indicate that the Random Forest model exhibits strong classification capability, particularly in identifying negative instances, as reflected by the high number of true negatives and relatively low number of false positives. However, the model produced a higher number of false negatives compared to false positives, suggesting that some positive cases were missed during classification. Despite these misclassifications, the majority of observations were correctly predicted, demonstrating the model's effectiveness and reliability.

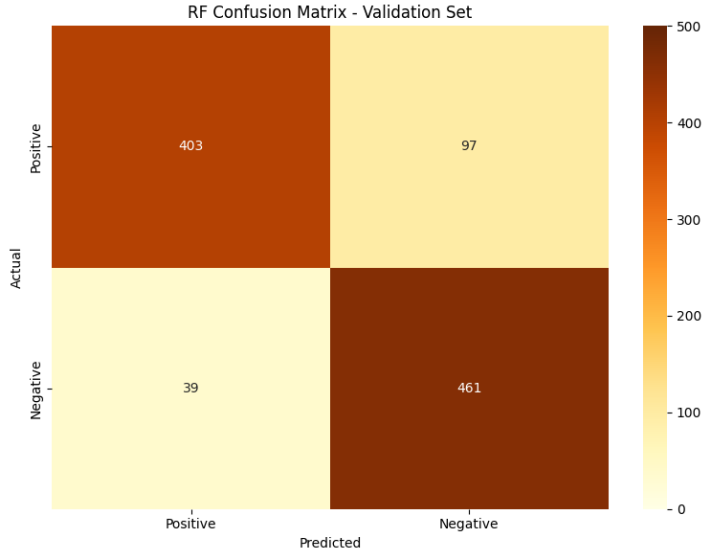
The Random Forest model achieved satisfactory predictive performance, correctly classifying most observations while maintaining a relatively low false positive rate. However, the higher false negative count suggests room for improvement in detecting positive cases, particularly in applications where missing positive instances may have significant consequences.

The confusion matrix in Figure 10 presents the classification performance of the Logistic Regression (LR) model on the validation dataset. The model correctly classified 429 out of 500 positive instances as positive (True Positives), while 71 positive instances were incorrectly classified as negative (False Negatives). For the negative class, 387 out of 500 cases were correctly identified as negative (True Negatives), whereas 113 negative instances were incorrectly classified as positive (False Positives).



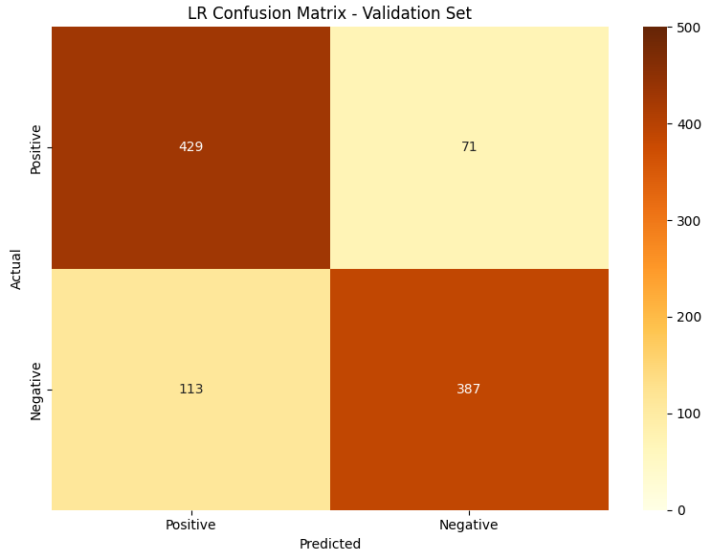
Key Performance Indicators:
- *True Positives (TP): 460*
- *True Negatives (TN): 447*
- *False Positives (FP): 53*
- *False Negatives (FN): 40*
- *Total Correct Predictions: 907 out of 1000 samples*
- *Classification Accuracy: 90.7%*

Figure 8. Confusion Matrix for Support Vector Machine Model Classification.



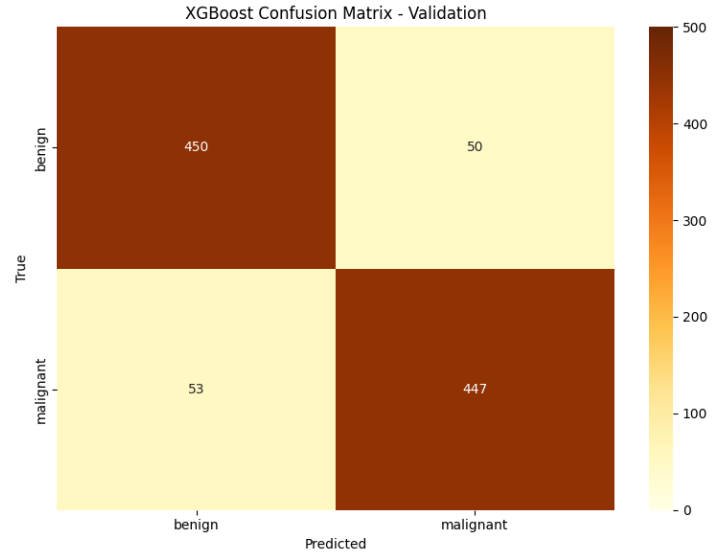
Key Performance Indicators:
- *True Positives (TP): 403*
- *True Negatives (TN): 461*
- *False Positives (FP): 39*
- *False Negatives (FN): 97*
- *Total Correct Predictions: 864 out of 1000 samples*
- *Classification Accuracy: 86.4%*

Figure 9. Confusion Matrix for Random Forest Classification Model.



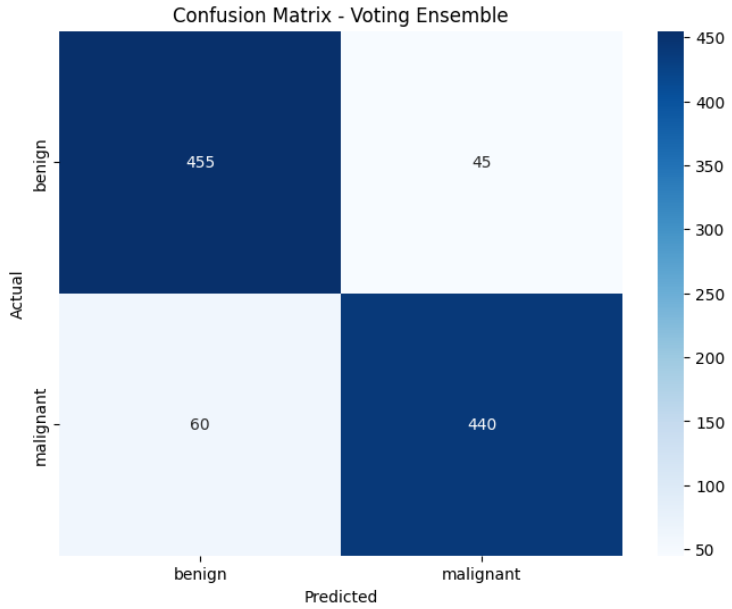
Key Performance Indicators:
- *True Positives (TP): 429*
- *True Negatives (TN): 387*
- *False Positives (FP): 113*
- *False Negatives (FN): 71*
- *Total Correct Predictions: 816 out of 1000 samples*
- *Classification Accuracy: 81.6%*

Figure 10. Confusion Matrix for Logistic Regression Model.



Key Performance Indicators:
- *True Positives (TP): 447*
- *True Negatives (TN): 450*
- *False Positives (FP): 50*
- *False Negatives (FN): 53*
- *Total Correct Predictions: 897 out of 1000 samples*
- *Classification Accuracy: 89.7%*

Figure 11. XGBoost model classification Confusion Matrix.



Key Performance Indicators:
- *True Positives (TP): 440*
- *True Negatives (TN): 455*
- *False Positives (FP): 45*
- *False Negatives (FN): 60*
- *Total Correct Predictions: 895 out of 1,000 samples*
- *Classification Accuracy: 89.5%*

Figure 12. Voting ensemble Confusion Matrix.

Table 1. Classification Performance Metrics Across All Models.

Model	Accuracy	Precision	Recall	Specificity	F1-Score
SVM	0.9070	0.8968	0.9200	0.8940	0.9083
RF	0.8640	0.9117	0.8060	0.9220	0.8556
LR	0.8160	0.7915	0.8580	0.7740	0.8234
XGBoost	0.8970	0.8994	0.8940	0.9000	0.8967
Voting Ensemble	0.8950	0.9072	0.8800	0.9100	0.8934

Table 2. Standard Metrics Result Across All Models.

Model	TP	TN	FP	FN
SVM	460	447	53	40
RF	403	461	39	97
LR	429	387	113	71
XGBoost	447	450	50	53
Voting Ensemble	440	455	45	60

The results indicate that the Logistic Regression model demonstrated a reasonable ability to distinguish between the two classes, with a higher number of correctly classified observations than misclassifications. The model showed better performance in identifying positive cases than negative cases, as evidenced by the higher number of true positives compared to true negatives. However, the relatively higher false positive rate suggests that the model occasionally classified negative cases as positive.

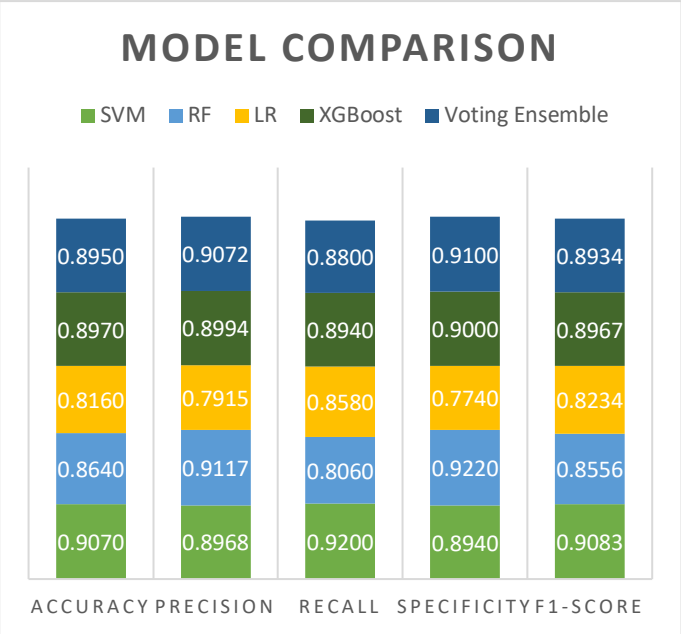


Figure 13. Model comparison.

The Logistic Regression model achieved satisfactory classification performance, correctly predicting the majority of observations. Nevertheless, the higher number of false positives and lower overall accuracy compared to some advanced machine learning models suggest that there is potential for further improvement in predictive performance. This indicates that while Logistic Regression serves as a useful baseline classifier, more sophisticated models may provide better discrimination between the classes.

The confusion matrix in Figure 11 presents the classification performance of the XGBoost (XGB) model on the validation dataset. The model correctly classified 447 out of 500 positive instances as positive (True Positives), while 53 positive instances were incorrectly classified as negative (False Negatives). For the negative class, 450 out of 500 cases were correctly identified as negative (True Negatives), whereas 50 negative instances were incorrectly classified as positive (False Positives).

The results indicate that the XGBoost model demonstrated a strong ability to distinguish between the two classes, with a substantially higher number of correctly classified observations than misclassifications. The model showed comparable performance in identifying both positive and negative cases, as evidenced by the closely matched true positive (447) and true negative (450) counts. The relatively low false positive and false negative rates further suggest that the model maintains a well-balanced classification behaviour across both classes.

Overall, the XGBoost model achieved strong classification performance, correctly predicting the vast majority of observations. The notably higher accuracy of 89.7% compared to the Logistic Regression baseline of 81.6% underscores the advantage of ensemble-based methods in capturing complex, non-linear relationships within the data. The balanced distribution of errors across both classes further highlights XGBoost as a more robust and reliable classifier for this task.

The confusion matrix in Figure 12 presents the classification performance of the Voting Ensemble model on the validation dataset. The model correctly classified 440 out of 500 malignant instances as malignant (True Positives), while 60 malignant instances were incorrectly classified as benign (False Negatives). For the benign class, 455 out of 500 cases were correctly identified as benign (True Negatives), whereas 45 benign instances were incorrectly classified as malignant (False Positives).

The results indicate that the Voting Ensemble model demonstrated a strong ability to distinguish between the two classes, with a considerably higher number of correctly classified observations than misclassifications. The model showed notably superior performance in identifying benign cases (455) compared to malignant cases (440), as evidenced by the higher true negative count. However, the relatively elevated false negative rate of 60 warrants attention in a clinical context, as undetected malignant cases carry significant diagnostic consequences.

Overall, the Voting Ensemble model achieved strong classification performance, correctly predicting the vast majority of observations across both classes. While the accuracy of 89.50% is competitive, the disproportionately higher false negative count (60) compared to false positives (45) suggests the model leans toward predicting benign outcomes, which may limit its reliability in sensitive diagnostic applications where malignant detection is critical.

The classification results presented in Table 1 and 2 reveal notable differences in model performance across all evaluated metrics. The Support Vector Machine (SVM) achieved the highest overall accuracy of 0.9070, with a well-balanced Precision of 0.8968, Recall of 0.9200, and an F1-Score of 0.9083, alongside a Specificity of 0.8940. The strong recall value indicates the SVM's superior ability to correctly identify true positive cases, which is particularly critical in clinical classification tasks swere failing to detect a positive case carries significant consequences.

The XGBoost model closely followed with an accuracy of 0.8970, demonstrating highly balanced performance across all metrics; Precision (0.8994), Recall (0.8940), Specificity (0.9000), and F1-Score (0.8967) reflecting its robustness as an ensemble-based gradient boosting classifier with minimal bias toward either class.

The Voting Ensemble recorded an accuracy of 0.8950 and Precision (0.9072) and Specificity (0.9100) among all models, indicating its exceptional ability to avoid false positive predictions. However, its comparatively lower Recall of 0.8800 and elevated false negative count of 60 suggest a tendency to under-predict malignant cases,

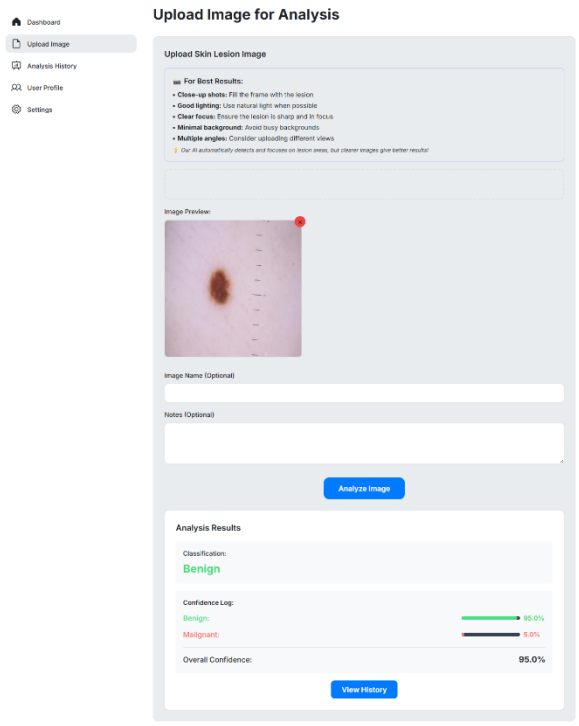


Figure 14. Website user interface showing malignant melanoma.

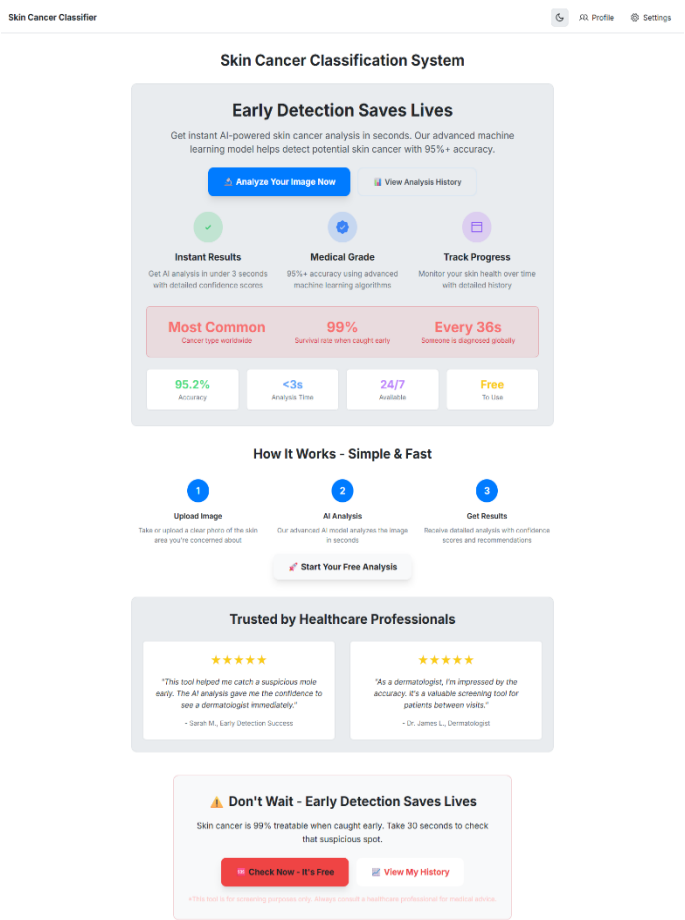


Figure 15. Web app user interface.

which limits its overall sensitivity in positive case detection.

The Random Forest (RF) model achieved an accuracy of 0.8640, with the highest Specificity of 0.9220 among the single classifiers, demonstrating reliable negative class identification. Nevertheless, its Recall of 0.8060 and F1-Score of 0.8556 indicate reduced sensitivity relative to SVM and XGBoost.

The Logistic Regression (LR) model recorded the lowest performance overall, with an accuracy of 0.8160, Precision of 0.7915, and Specificity of 0.7740, consistent with its role as a linear baseline classifier with limited capacity to capture complex, non-linear decision boundaries inherent in the dataset. Figure 13 provides a comparative overview of the performance of all models.

4.3. Discussion

The proposed Voting Ensemble model demonstrated competitive classification performance across all evaluation metrics. It achieved the highest precision, indicating its strong capability to reduce false positive predictions, while also maintaining high specificity. Although the standalone SVM classifier achieved the highest overall accuracy, recall, and F1-score, the performance differences between the two models were relatively small. These findings suggest that the proposed Voting Ensemble provides a balanced and reliable classification framework while remaining competitive with the best-performing individual classifier. The slight performance gap under the current experimental setting, indicating that further optimization of the ensemble strategy could further improve its performance. These findings are consistent with the SVM accuracy of 97% reported by [40] for pneumonia detection, further demonstrating the robustness of SVM across different medical image classification tasks. The interface of the proposed system is shown in Figures 14 and 15.

5. Conclusion and Recommendation

This research has presented the design and implementation of a malignant melanoma classification system using voting ensemble learning techniques. The system aimed to classify skin lesions into malignant and benign categories using multiple machine learning models. During development, dermatological image datasets sourced from Kaggle were preprocessed, and extracted features were used to train and evaluate the listed models. Although the study proposed a Voting Ensemble framework, experimental evaluation demonstrated that the standalone SVM remained the best-performing classifier under the current experimental setting. Therefore, the proposed ensemble approach requires further optimization before it can be considered a superior alternative to SVM.

Despite the promising outcomes, some limitations were encountered. A major challenge was the lack of access to primary labeled clinical data, which limited the generalizability of the models. Furthermore, data com-

plexity and variations in lesion features presented difficulties during preprocessing and model training. For any improvement on this project to have more efficient impact, the following recommendations are proposed:

1) Access to primary clinical data: Further improvements should involve collaboration with medical practitioners or dermatologists to obtain larger, verified, and diverse local clinical image datasets. This would improve the clinical relevance of the model and enhance its performance in real-world applications.
2) Integration of deep learning models such as CNN: Should the project scope extend beyond traditional machine learning, incorporating advanced deep learning techniques, including Convolutional Neural Networks (CNNs), could significantly improve feature extraction, particularly when working with raw dermoscopic images. Multimodal deep learning architectures have shown promise in complex medical classification tasks, including early detection and severity staging of neurodegenerative diseases such as Alzheimer's [41], suggesting that similar approaches could further enhance dermatological classification outcomes.

6. Declarations

6.1. Author Contributions

Racheal Shade Akinbo: Conceptualization, Writing – Review & Editing, Visualization, Supervision; **Tosin Precious Adeyemi:** Writing – Original Draft, Software, Formal analysis and Validation; **Bamidele Moses Kuboye:** Project administration, Formal analysis and Investigation.

6.2. Institutional Review Board Statement

Not applicable.

6.3. Informed Consent Statement

Not applicable.

6.4. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.5. Acknowledgment

Not applicable.

6.6. Conflicts of Interest

The authors declare no conflicts of interest, The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

7. References

[1] R. L. Siegel *et al.*, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, 2023. https://doi.org/10.3322/caac.21763.

[2] American Cancer Society, "Cancer Facts and Figures 2024," Mar. 9, 2024. [Online]. Available: https://www.cancer.org.

[3] World Health Organization, "Cancer Key Facts," Mar. 28, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer.

[4] T. J. Brinker *et al.*, "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *European Journal of Cancer*, vol. 113, pp. 47–54, 2019. https://doi.org/10.1016/j.ejca.2019.04.001.

[5] P. Tschandl *et al*, "Human-computer collaboration for skin cancer recognition," *Nature Medicine*, vol. 26, pp. 1229–1234, 2020. https://doi.org/10.1038/s41591-020-0942-0.

[6] Y. Liu *et al*, "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine*, vol. 26, no. 6, pp. 900-908, 2020. https://doi.org/10.1038/s41591-020-0842-3.

[7] T. G. Debelee, "Skin lesion classification and detection using machine learning techniques: A systematic review," *Diagnostics*, vol. 13, no. 19, p. 3147, 2023. https://doi.org/10.3390/diagnostics13193147.

[8] J. V. Tembhurne, N. Hebbar, H. Y. Patil, and T. Diwan, "Skin cancer detection using ensemble of machine learning and deep learning techniques," *Multimedia Tools and Applications*, vol. 82, pp. 27501–27524, 2023. https://doi.org/10.1007/s11042-023-14697-3.

[9] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, I. Ellinger, "Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification," *Computer Methods and Programs in Biomedicine*, vol. 193, 2020. https://doi.org/10.1016/j.cmpb.2020.105475.

[10] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017. https://doi.org/10.1038/nature21056.

[11] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthcare*, vol. 11, no. 12, art. No. 1808, 2023. https://doi.org/10.3390/healthcare11121808.

[12] H. Yousef, M. Alhaji, A. O. Fakoya, and S. Sharma, "Anatomy, Skin (Integument), Epidermis," *National Library of Medicine (NCBI)*, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK470464.

[13] P. Dodia, M. A. Arida, "Dermatopathology Epidermis Histology," *National Library of Medicine (NCBI)*, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK592423.

[14] W. Lopez-Ojeda, A. Pandey, M. Alhajj, and A. M. Oakley, "Anatomy, Skin (Integument)," *National Library of Medicine (NCBI)*, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK441980.

[15] J. Rajkumar, N. Chandan, P. Lio, and V. Shi, "The skin barrier and moisturization: Function, disruption, and mechanisms of repair," *Skin Pharmacology and Physiology*, vol. 36, no. 4, pp. 174–185, 2023. https://doi.org/10.1159/000534136.

[16] T. C. Karl and T. P. Michael, "Systemic pharmacologic treatments for plaque psoriasis," *American Family Physician*, vol. 109, no. 2, pp. 116–117, 2024. [Online]. Available: https://www.aafp.org/afp/2024/0200/cochrane-plaque-psoriasis.pdf.

[17] National Institute of Arthritis and Musculoskeletal and Skin Diseases, "Skin Disease," 2025. [Online]. Available: https://www.niams.nih.gov/health-topics/skin-diseases.

[18] Cancer Research UK, "Risks and Causes of Skin Cancer," Mar. 24, 2023. [Online]. Available: https://www.cancerresearchuk.org/about-cancer/skin-cancer/risks-causes.

[19] MD Anderson Cancer Center, "Skin Cancer Symptoms," *MD Anderson Cancer Center*, Mar. 27, 2025. [Online]. Available: https://www.mdanderson.org/cancer-types/skin-cancer/skin-cancer-symptoms.html.

[20] A. Mandal, "Cancer Classification," *News Medical Life Sciences*, Jul. 5, 2023. [Online]. Available: https://www.news-medical.net/health/Cancer-Classification.aspx.

[21] M. Ratini, "Understanding Cancer — The Basics," *WebMD*, 2024. [Online]. Available: https://www.webmd.com/cancer/understanding-cancer-basics.

[22] D. S. Sominidi, "Dermatitis — Symptoms and Causes," *Mayo Clinic*, Apr. 1, 2023. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/dermatitis-eczema/symptoms-causes/syc-20352380.

[23] R. B. Tchema, A. C. Polycarpou, and M. Nestoros, "Skin cancer classification using machine learning," *Multimedia Tools and Applications*, vol. 84, pp. 3239–3256, 2025. https://doi.org/10.1007/s11042-025-20595-7.

[24] P. Natha, P. S. Tera, R. Chinthaginjala, S. O. Rab, V. Narasimhulu, and T. H. Kim, "Boosting skin cancer diagnosis accuracy with ensemble approach," *Scientific Reports*, vol. 15, Art. no. 1290, 2025. https://doi.org/10.1038/s41598-024-84864-5.

[25] K. Kanchana, S. Kavitha, K. J. Anoop, and B. Chinthamani, "Enhancing skin cancer classification using EfficientNet B0-B7 through convolutional neural networks," *Asian Pacific Journal of Cancer Prevention*, vol. 25, no. 5, pp. 1795–1804, 2024. https://doi.org/10.31557/APJCP.2024.25.5.1795.

[26] I. A. Kandhro, S. Manickam, K. Fatima, U. Malik, A. Naz, and A. Dandoush, "Performance evaluation of E-VGG19 model: Enhancing real-time skin cancer detection and classification," *Heliyon*, vol. 10, no. 10, p. e31488, 2024. https://doi.org/10.1016/j.heliyon.2024.e31488.

[27] P. Jaberi, S. Nemati, and M. E. Basiri, "Classification of skin cancer images using two-level ensemble deep learning," *Signal and Data Processing*, vol. 21, no. 2, pp. 79-90, 2024. [Online]. Available: http://jsdp.rcisp.ac.ir/article-1-1350-en.html.

[28] B. Hassan, "Ensemble learning for multi-class skin cancer detection," *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/381925068.

[29] D. Adla, G. V. R. Reddy, P. Nayak, and G. Karuna, "A full-resolution convolutional network with a dynamic graph cut algorithm for skin cancer classification and detection," *Healthcare Analytics*, vol. 3, 2023. https://doi.org/10.1016/j.health.2023.100154.

[30] B. Kalpana, A. K. Reshmy, S. S. Pandi, S. Dhanasekaran, "OESV-KRF: Optimal ensemble support vector kernel random forest based early detection and classification of skin diseases," *Biomedical Signal Processing and Control*, vol. 85, pp. 29–30, p. 104779, 2023. https://doi.org/10.1016/j.bspc.2023.104779.

[31] A. A. M. Teodoro *et al.*, "A Skin Cancer Classification Approach using GAN and RoI-Based Attention Mechanism," *Journal of Signal Processing Systems*, vol. 95, no. 2, pp. 211–224, 2023. https://doi.org/10.1007/s11265-022-01757-4.

[32] P. Tang, X. Yan, Y. Nan, X. Hu, X. Hu, B. H. Menze, S. Krammer, and T. Lasser, "Joint-individual fusion structure with fusion attention module for multi-modal skin cancer classification," *Pattern Recognition*, vol. 154, p. 110604, 2024. https://doi.org/10.1016/j.patcog.2024.110604.

[33] H.-Y. Huang, Y.-P. Hsiao, A. Mukundan, Y.-M. Tsao, W.-Y. Chang, and H.-C. Wang, "Classification of skin cancer using novel hyperspectral imaging engineering via YOLOv5," *Journal of Clinical Medicine*, vol. 12, no. 3, p. 1134, 2023. https://doi.org/10.3390/jcm12031134.

[34] M. H. Javid, *Melanoma Skin Cancer Dataset of 10000 Images* [Dataset]. Kaggle, 2022. [Online]. Available: https://doi.org/10.34740/KAGGLE/DSV/3376422.

[35] J. M. Álvarez-Alvarado *et al.*, "Hybrid Techniques to Predict Solar Radiation Using Support Vector Machine and Search Optimization Algorithms: A Review," *Applied Sciences*, vol. 11, no. 3, p. 1044, 2021. https://doi.org/10.3390/app11031044.

[36] A. S. M. Shafi, M. M. I. Molla, J. J. Jui, M. M. Rahman, "Detection of colon cancer based on microarray dataset using machine learning as feature selection and classification techniques," *SN Applied Sciences*, vol. 2, 2020. https://doi.org/10.1007/s42452-020-3051-2.

[37] Y. Wang, Z. Pan, J. Zheng, L. Qian, and M. Li, "A hybrid ensemble method for pulsar candidate classification," *Astrophysics and Space Science*, vol. 364, 2019. https://doi.org/10.1007/s10509-019-3602-4.

[38] S. Biswas, S. Ghosh, S. Roy, R. Bose, and S. Soni, "A study of stock market prediction through sentiment analysis," *Mapana Journal of Sciences*, vol. 22, 2023. https://doi.org/10.12723/mjs.64.6.

[39] F. Zhang, Y. Wang, S. Liu, H. Wang, "Decision-based evasion attacks on tree ensemble classifiers," *World Wide Web*, vol. 23, pp. 2957–2977, 2020. https://doi.org/10.1007/s11280-020-00813-y.

[40] R. S. Akinbo, O. Olabode, O. Daramola, and E. Ibam, "Comparative analysis of machine learning models on the classification of pneumonia disease using chest X-ray images," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 17, no. 2, pp. 23–30, 2025. https://doi.org/10.54554/jtec.2025.17.02.003.

[41] R. S. Akinbo, B. M. Kuboye, A. F. Owoseeni, J. O. Igbogbo, and T. J. Fatoke, "Development of multimodal deep learning models for early detection and severity staging of Alzheimer's disease," *Cureus Journal of Computer Science*, vol. 3, p. es44389-026-00072-4, 2026 https://doi.org/10.7759/s44389-026-00072-4.