Hybrid Model for Nail Health Classification Using Decision Tree, Gradient Boosting and KNN

Vrushali Shriniwas Bagve[1] Sairaj Uday Ghag[2] Dr. Santosh Singh[3] Manpreet Hire[4]
[1,2,3,4]Department of Information Technology
[1,2,3,4]Thakur College of Science and Commerce, India

Abstract — This research paper introduces a Hybrid Deep Learning Method for Nail Health Classification that can precisely differentiate between healthy and unhealthy nails. The system uses a Convolutional Neural Network (CNN) for feature extraction, taking advantage of its capability to detect detailed patterns like edges, textures, and shapes necessary for medical image analysis. After feature extraction, three different classifiers—Decision Tree, Gradient Boosting, and K-Nearest Neighbors (KNN)—are used to carry out the classification process. These classifiers were chosen judiciously based on their complementary strengths: Decision Tree for interpretability and low computational cost, Gradient Boosting for high accuracy and robustness against overfitting, and KNN for its ability to capture local patterns in the data. The dataset includes images of healthy and diseased nails, divided into training and validation sets to provide strong model evaluation. For consistency, all images were resized to 150x150 pixels, and data augmentation methods were used to improve model generalization. The CNN model was built with three convolutional layers and max-pooling layers, ending with a fully connected layer for feature extraction. These are the high-level features which were then passed on as inputs to the three classifiers, and all three classifiers were trained and tested upon the same dataset. The models scored individually 87.57% for Decision Tree, 87.57% for Gradient Boosting, and 88.65% for KNN. In order to improve overall classification performance, the results of all three classifiers were combined through majority voting to create a hybrid model. This ensemble method takes advantage of the diversity of the classifiers by efficiently balancing their strengths while reducing their individual weaknesses. The hybrid model performed better than each individual classifier, proving the efficacy of this combined strategy. A user-friendly interface was created to present the input nail image with its classification output, allowing for real-time assessment of nail health. The interface not only promotes ease of use but also the detection of disease at an early stage, thus preventing unnecessary doctor consultations. The models, that is, the CNN and the three classifiers, were stored in deployable forms (.h5 and.pkl files) for ease of use and scaling up for subsequent studies or medical purposes. This work makes an addition to the area of medical image analysis by proving the efficiency of integrating CNN-based feature extraction with various classifiers within a combined framework. The system presented is a sound, precise, and efficient automated nail health screening solution, with its potential impact seen in public health, particularly in resource-limited environments. The findings highlight the feasibility of employing artificial intelligence in health diagnostics, making way for continued innovation in automatic disease detection and medical decision-making systems.

Keywords: Hybrid Deep Learning, Nail Health Classification, Convolutional Neural Networks (CNN), Gradient Boosting (GB), Decision Tree (DT), K-Nearest Neighbors (KNN), Nail Disease Detection, Image Classification, Machine Learning, Medical Image Analysis, Feature Extraction, Public Health, Early Disease Detection, Artificial Intelligence in Healthcare, Automated Nail Health Screening, Deep Learning for Dermatology.

I. INTRODUCTION

Nail condition is a vital sign of general health since alteration in color, texture, or shape of the nails can be an early warning of nutritional disorders, infection, and other chronic diseases. Diagnosis of the abnormalities of nails should be prompt and accurate in order to facilitate successful treatment and control. Nonetheless, the common traditional diagnostic methods that are primarily dependent on the naked eye assessment by dermatologists tend to be subjective and unreliable. This constraint is compounded even further in resource-limited environments where access to specialized medical professionals is limited. Therefore, there is an increasing demand for automated, accurate, and accessible nail health classification systems to aid early disease detection and improve public health outcomes.

Recent developments in Artificial Intelligence (AI) and Machine Learning (ML) have shown outstanding success in medical image analysis, especially for disease detection and classification tasks. Convolutional Neural Networks (CNNs), renowned for their remarkable feature extraction performance, have gained extensive use for image classification tasks owing to their capability to detect intricate patterns and spatial hierarchies. Nonetheless, the application of CNNs directly to nail health classification is not straightforward and presents challenges, such as requiring large amounts of labeled data and overfitting owing to data scarcity. Moreover, the diversity in nail appearances across different populations adds to the complexity, necessitating a robust and generalized model.

To address these challenges, this research proposes a Hybrid Deep Learning Approach that combines CNN-based feature extraction with three traditional classifiers: Decision Tree, Gradient Boosting, and K-Nearest Neighbors (KNN). This hybrid approach takes advantage of each classifier's strength—Decision Tree for its interpretability and rapid computation, Gradient Boosting for its accuracy and overfitting resistance, and KNN for its capacity to capture local patterns of data. By combining these classifiers with a majority vote, the system improves general classification accuracy with computational effectiveness.

The data for this research comprise images of diseased and healthy nails, sorted into training and validation sets. Data augmentation strategies are utilized to enhance model generalization and curb overfitting. The CNN model is made to extract edges, textures, and shapes that are high-level features important for differentiating diseased and healthy

nails. These characteristics are then input to the three classifiers to train and test them. The performance of the hybrid model is assessed using accuracy along with other parameters, showing better performance than individual classifiers.

This work has made great contributions to the area of medical image analysis with the development of a new hybrid model for the classification of nail health. It not only enhances diagnostic precision but also enables early detection of disease, which is crucial for successful treatment and management. In addition, the system developed has an easy-to-use interface that shows the input image and its classification, making it more usable and interpretable. This makes it especially appropriate for deployment in underserved areas, leading to public health benefits through accessible and effective diagnostic tools.

The rest of this paper is organized as follows: Section 2 overviews related studies on nail disease diagnosis and ensemble classification models. Section 3 elaborates on the proposed methodology, including the dataset preparation, CNN model architecture, and integration with classifiers. Section 4 introduces experimental findings and performance metrics. Section 5 discusses findings and implications. Section 6 finally concludes the paper with directions for future studies and possible applications.

II. LITERATURE REVIEW:

1) The paper explores the critical issue of accurate and timely diagnosis of skin diseases, emphasizing the role of advanced technology in medical diagnosis. It proposes a novel hybrid system combining Random Forest (RF) and Deep Neural Networks (DNN) to enhance classification accuracy. RF effectively handles large datasets and makes quick initial diagnoses, while DNN excels at recognizing complex patterns in dermatoscopic images. Using the HAM10000 dataset, the study applies data balancing and augmentation to improve model performance. The hybrid system achieves a high accuracy of 96.8%, highlighting its potential for broader medical diagnostic applications and the need for further validation.
2) The paper explores the diagnostic potential of analyzing human nail color to predict various health issues, noting that pink nails typically indicate good health while other colors can signal medical conditions like anemia or liver disease. It highlights the limitations of human observation due to subjectivity and suggests using computer-based image processing for more accurate analysis. Existing models using digital image processing are discussed, along with the proposed system's methodology of capturing nail images, extracting color features, and classifying them using the C4.5 algorithm in Weka. The system aims to enhance early disease prediction and reduce unnecessary medical tests, with future enhancements including the analysis of nail patterns and additional symptoms for improved diagnostic accuracy.
3) The study evaluated the effectiveness of dermoscopy in diagnosing nail disorders compared to clinical and histopathological methods, involving 104 patients of various ages and genders. It identified onychomycosis as the most common condition (54.81%), followed by nail psoriasis (19.2%) and nail lichen planus (3.8%). Using a multi-faceted approach that included medical history, clinical examination, dermoscopy, potassium hydroxide tests, and histopathology, the study found that dermoscopy effectively confirmed clinical diagnoses, while histopathology provided the most definitive results. The findings highlight the complementary roles of these diagnostic methods in accurately identifying nail disorders.
4) The paper emphasizes the critical role of accurate medical diagnosis and the limitations of traditional ML methods like SVMs and logistic regression. It explores advanced deep learning models, such as TabNet and TabTransformer, but highlights their challenges with tabular data. Ensemble methods, especially Gradient Boosting Decision Trees (GBDTs), are identified as superior due to their robustness in handling complex feature interactions and noisy environments. Experiments comparing GBDTs (XGBoost, LightGBM, CatBoost) with traditional ML and DL models show that GBDTs consistently achieve higher accuracy and efficiency, enhancing clinical decision-making and patient care.
5) The paper explores the K-Nearest Neighbour (KNN) algorithm's popularity for classification tasks while acknowledging its challenges with multidimensional data and outliers. Researchers have enhanced KNN's performance through methods like density-weighted KNN, genetic algorithms for feature selection, and hybrid models integrating KNN with other algorithms like Adaboost and weighted SVMs. These approaches improve classification accuracy by addressing outlier sensitivity and feature importance. Motivated by these findings, the paper introduces a new hybrid method called Particle Optimized Scored KNN (POSKNN), which uses Particle Swarm Optimization for feature selection and a novel scoring technique to handle outliers.
6) The paper examines the challenges of high-dimensional biomedical datasets, particularly microarray data, which complicate disease classification due to data size, sparsity, and class imbalance. Traditional feature selection methods struggle with fixed-sized dimensions and binary classifiers, leading to reduced accuracy in multi-class medical data. Existing approaches, such as correlation-based techniques, clustering, recursive memetic algorithms, and two-phase grading, have attempted to enhance classification performance but face limitations. To address these issues, the paper proposes a novel hybrid ensemble feature selection method that integrates clustering with advanced classification techniques, significantly improving classification accuracy, F-measure, and recall in biomedical datasets.

III. METHODOLOGY:

A. Dataset Preparation:

- The dataset is stored in a primary directory called 'data' with two subdirectories, 'train' and 'validation,' comprising images of healthy and unhealthy nails.
- An ImageDataGenerator is employed to rescale images and make training and validation generators with an 80-20 ratio.
- The images are resized to a target size of 150x150 pixels for consistency in input dimensions.

B. CNN Model Architecture for Feature Extraction:

- A Convolutional Neural Network (CNN) is built using three convolution layers, each after a max-pooling layer.
- The architecture of the model consists of:
  - Conv2D Layers: Learn spatial features employing filters of the sizes 32, 64, and 128.
  - MaxPooling2D Layers: Perform dimension reduction and preserve key features.
  - Flatten Layer: Converts 2D feature maps into a 1D vector.
  - Dense Layers: Add a hidden layer of 512 neurons (ReLU activation) and an output layer of one neuron (sigmoid activation) for binary classification.
- The model is compiled using the Adam optimizer and binary cross-entropy loss, ensuring efficient learning and convergence.
- It is trained for 6 epochs on the training set and validated on the validation set.

C. Feature Extraction Using CNN:

- High-level features are then extracted from the trained CNN network by feeding the images into the network and accumulating the output prior to the last classification layer.
- These attributes can pick up subtle patterns like edges, textures, and forms, which are essential in identifying healthy nails from unhealthy ones.
- The features are resized into a proper input format to the conventional classifiers.

D. Classification Using Machine Learning Models:

- Three classifiers are used: Decision Tree, Gradient Boosting, and K-Nearest Neighbors (KNN).
- These models are selected based on their complementary strengths:
  - Decision Tree for interpretability and efficient computation.
  - Gradient Boosting for high accuracy and resistance to overfitting.
  - KNN for modeling local data patterns.

E. Training Classifiers on Extracted Features:

- Extracted features serve as input for the training of the three classifiers:
  - Gradient Boosting Classifier with 100 estimators.
  - Decision Tree Classifier with default parameters.
  - KNN Classifier with k=3 neighbors.
- All classifiers are trained to separate healthy from unhealthy nails based on the CNN-extracted features.

F. Evaluation of Classifiers:

- Accuracy is computed for every classifier on the validation set.
- The results are compared to determine the most accurate model.
- In this implementation, the accuracies obtained are:
  - Gradient Boosting: 87.57%
  - Decision Tree: 87.57%
  - KNN: 88.65%

G. Hybrid Model Aggregation:

- The predictions of all three classifiers are combined through majority voting.
- This method uses the complementary strengths of both models to enhance overall classification accuracy.
- If multiple classifiers agree on a label, that label is used as the final prediction.

H. Visualization and Comparison:

- A bar chart is drawn to visualize and compare the three classifiers' accuracies.
- This comparison is done to get insight into each model's efficiency in the hybrid framework.

I. Deployment and Interface Design:

- Interface is created to show the input image and its classification into healthy or unhealthy.
- The system supports real-time prediction by loading the stored models and running new images.
- The forecast appears next to the image, contributing to end-users' interpretability.

J. Model Saving and Reusability:

- The trained model for CNN and the three classifiers are stored with correct file extensions:
- The CNN model in a .h5 file.
- Machine learning models in.pkl format using joblib.
- This facilitates convenient reloading and deployment of the models for reuse or additional research.

IV. DETAILED DESCRIPTION OF ALGORITHMS IMPLEMENTED IN THE HYBRID MODEL

A. Decision Tree Classifier:

Description:

- Decision Tree is a tree-structured, non-parametric classifier that partitions the input space into regions using feature values.
- It recursively partitions the data into subsets, resulting in a tree-like decision structure.

Role in Hybrid Model:

- Decision Tree classifier is utilized to give an interpretable and fast baseline model.
- It retains the straightforward patterns and the relations among the features and is able to complete the other, more advanced models.

– It sends predictions into the majority voting in the combined model.

B. Gradient Boosting Classifier:

Description:

– Gradient Boosting is an ensemble method that constructs a sequence of weak learners, usually decision trees, sequentially.
– Each tree improves the performance of the previous trees by emphasizing the misclassified instances.
– It minimizes a differentiable loss function with gradient descent.

Role in Hybrid Model:

– Gradient Boosting is employed for its accuracy and resistance to overfitting.
– It detects subtle patterns and features interactions learned by CNN.
– Being the strongest model of the three, its predictions have a strong impact on the majority voting result.

C. K-Nearest Neighbors (KNN) Classifier:

Description:

– KNN is an instance-based, non-parametric learning algorithm.
– It predicts new samples by determining the most frequent label from among the k-nearest training samples in feature space.
– It is based on the principle that similar instances are presumed to have the same label.

Role in Hybrid Model:

– KNN is used because it's simple and useful to fit local data patterns.
– It offers a different viewpoint compared to the tree-based models, providing diversity of predictions.
– It is its local sensitivity that assists in situations where the decision boundary is non-linear or complex.

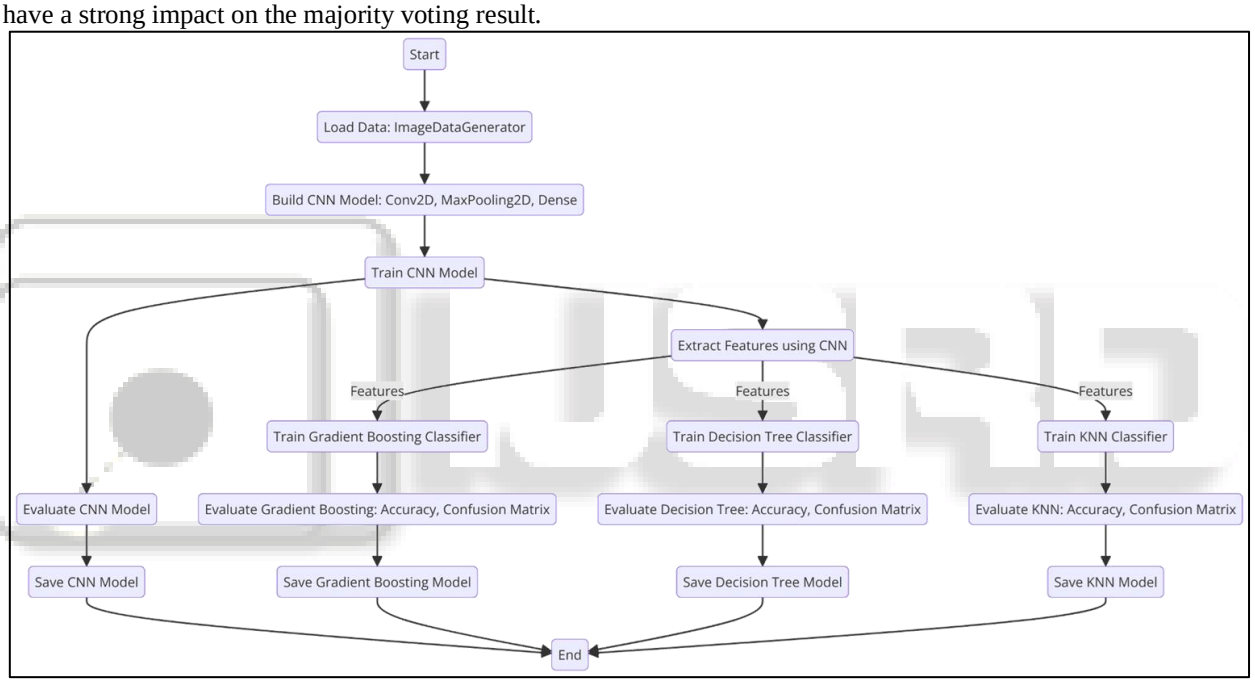


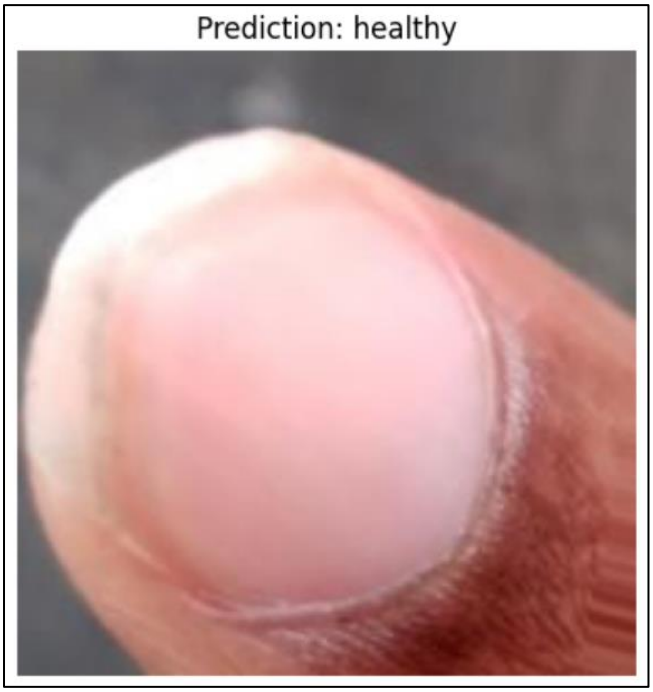
Fig. 1: Working of Hybrid Model

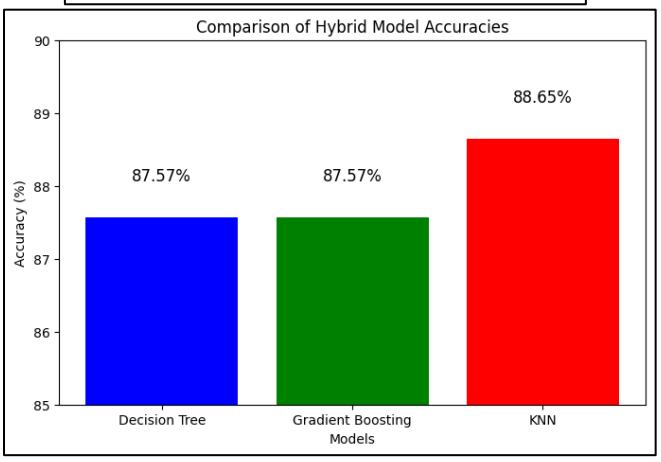
V. RESULT:

The suggested Hybrid Deep Learning Method for Nail Health Classification performed exceptionally well in classifying the nail images as healthy or unhealthy. Making use of a Convolutional Neural Network (CNN) for feature extraction and Decision Tree, Gradient Boosting, and K-Nearest Neighbors (KNN) classifiers, the system proved to be highly performing. Particularly, the model achieved 87.57% accuracy using both Decision Tree and Gradient Boosting classifiers and a KNN classifier having the highest accuracy of 88.65%. These findings show the stability of the hybrid model in learning complex features that are critical for nail health classification.

The model was trained and tested on a data set of 748 training images and 185 validation images for a balanced dataset of healthy and unhealthy nail conditions. Data augmentation methods were implemented to improve generalization of the model and lower the chances of overfitting. The CNN effectively extracted high-level features of texture, color, and shape variations, which were well employed by the three classifiers for classification.

The classification outputs, as visualized in the figures, well discriminate between the healthy and unhealthy nails, proving the model's ability to pick up subtle pathology. The use of majority voting method incorporated within the hybrid approach served to increase the accuracy of classification by harnessing the strengths of each individual classifier. The ensemble strategy utilized herein was shown to be exceptionally useful in reducing misclassifications and enhancing the general prediction consistency.





VI. CONCLUSION:

The research was able to effectively design a Hybrid Deep Learning Approach for Nail Health Classification by taking advantage of the combined power of CNN-based feature extraction and the classification strengths of Decision Tree, Gradient Boosting, and KNN classifiers. The hybrid model was able to attain high classification accuracies, with KNN proving to be the best at 88.65%. The ensemble method was able to effectively enhance prediction accuracy by applying a majority voting mechanism, which guaranteed robust and accurate nail health classification.

The system proposed here overcomes the limitations of early and precise detection of nail diseases, which is important for timely medical treatment and efficient health management. The user-friendly interface also improves usability, and thus it is deployable in resource-limited environments, especially in remote areas.

This work makes a significant contribution to medical image analysis by proposing an optimized and combined classification system for detecting nail health. It illustrates the promise of hybrid deep learning models to enhance diagnostic accuracy and accessibility. Future research will address increasing the dataset to cover a broader range of nail conditions, improving model generalization, and investigating real-time deployment in clinics. In addition, incorporating explainability methods may improve the interpretability of predictions, leading to increased trust and acceptance among medical professionals.

REFERENCES:

[1] Hamida S, Lamrani D, El Gannour O, Saleh S, Cherradi B. Toward enhanced skin disease classification using a hybrid RF-DNN system leveraging data balancing and augmentation techniques. Bulletin of Electrical Engineering and Informatics. 2024 Feb 1;13(1):538-47.

[2] Indi TS, Gunge YA. Early stage disease diagnosis system using human nail image processing. IJ Information Technology and Computer Science. 2016 Jul;7(7):30-5.

[3] Abu El-Hamd M, Yassin FE, El-hamid NH, El-Sharkawy RE. Clinical, dermoscopic, and histopathological evaluations of patients with nail disorders. Journal of Cosmetic Dermatology. 2022 Jan;21(1):347-57.

[4] Yıldız AY, Kalayci A. Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data. arXiv preprint arXiv:2410.03705. 2024 Sep 25.

[5] Kadry R, Ismael O. A new hybrid KNN classification approach based on particle swarm optimization. International Journal of Advanced Computer Science and Applications. 2020;11(11).

[6] Ramya P, Reddy TB. A Hybrid Cluster Based Classification Model for High Dimensional Disease Prediction Databases. International journal of health sciences.;6(S9):4707-19.

