Medical Diagnosis Using Hybrid of Machine Learning and Deep Learning Techniques

Raed Alazaidah[1], Moath Alomari[2], Hamza Mashagba[3], Musab Iqtait[4], Azlan B. Abd Aziz[5]*,
Hayel Khafajeh[6], Omar Khair Alla Alidmat[7], Ghassan Samara[8], Haneen Alzoubi[9], Samir Salem Al-Bawri[10]*
Department of Data Science and Artificial Intelligence-Faculty of Information Technology, Zarqa University, Zarqa, Jordan[1,4,6]
Department of Computer Science-Faculty of Information Technology,
Zarqa University, Zarqa, Jordan[2,7,8]
Faculty of Engineering and Technology-Centre for Wireless Technology (CWT),
Multimedia University, Melaka, Malaysia[3,5]
Department of Applied Sciences-Faculty of Sciences, Applied Science Private University, Amman, Jordan[9]
Space Science Centre-Institute of Climate Change, Universiti Kebangsaan Malaysia (UKM), Malaysia[10]

***Abstract*—The rapid development of medical practices and imaging technology tools creates substantial growth in the amount of medical image data each year in our present era. This research aims to develop a hybrid approach that integrates Machine Learning (ML) and Deep Learning (DL) techniques to enhance the accuracy and reliability of medical image classification for diagnostic purposes. Medical imaging data complexity and growing volume serve as the research motivation, which leads to an investigation of standalone ML or DL limitations and their combination into a single framework. The medical image processing starts with normalization, then noise reduction, and continues to grayscale conversion before performing histogram equalization. This research uses VGG16 and ResNet50 alongside MobileNet and InceptionV3 for feature extraction, then applies ten different ML algorithms, including SVM and MLP, and Random Forest, for classification. Five public medical image datasets from Kaggle are used: COVID-19 chest X-rays, melanoma skin lesions, pneumonia chest X-rays, acute stroke facial images, and various eye diseases. Hybrid models display superior performance compared to stand-alone ML or DL models based on accuracy, precision, recall, and F1-score evaluation measures. Multiple datasets demonstrate that the MobileNet+MLP combination delivers the most accurate results, which demonstrates its reliable and efficient performance. The developed AI diagnostic tool presents a scalable system alongside accuracy and interpretability to enhance clinical decision outcomes.**

Keywords*—Classification; deep learning; feature selection; hybrid models; machine learning; medical diagnosis; medical image classification

I. Introduction

Medical imaging is an essential aspect of our time, employing various technologies to obtain images of the human body or specific parts of it [1]. Medical imaging considers a crucial component of medical diagnosis, since it is widely applied in hospitals, medical laboratories, and other facilities due to its importance in diagnosing various diseases, as well as aiding in monitoring patient treatment [2]. The most important medical imaging modern techniques are X-ray radiography, endoscopy, Magnetic Resonance Imaging (MRI), magnetic resonance spectroscopy, thermography, electrical source imaging and several other techniques [3].

*Corresponding authors.

Several limitations in human analysis of medical images compared to modern techniques, such as the fact that the image analysis process takes time due to the complexity of the data, in addition to the fact that the analysis depends mainly on the experience of the image analyst, so there is a possibility of human errors. Moreover, there are some effects that occur to medical images, whether modern or traditional, including noise [4].

According to [5], noise is defined as the random change in the original pixel value that reduces image quality. Therefore, it is important to use modern medical imaging techniques to remove noise, especially compared to conventional images, to enable accurate disease diagnosis without any issues that may affect the analysis of medical images. Consequently, several artificial intelligence techniques have been developed to address these issues in medical imaging and assist in medical diagnosis, including Machine Learning (ML) and Deep Learning (DL) technologies [6], [7], which are important for processing complex data, identifying problems that the human eye might miss, and reducing noise in medical images to enhance their quality [8], [9].

A. Machine Learning

Machine Learning (ML) aims to study statistical algorithms that can learn from the input data to perform tasks that are automatically improved through experience [10]–[12]. All of this leads to increasing the efficiency of the algorithm that was built, which reduces human errors in entering or analyzing data [13], [14]. ML is used to make decisions and predictions based on learning from data. In [15], the authors categorize ML as following:

- Supervised Learning: The aim of this type is to predict outcomes using labeled data.
- Unsupervised Learning: This type learns from unlabeled/unstructured data.
- Semi-Supervised Learning: A large amount of unlabeled data and a small amount of labeled data are combined to improve the performance of the model.

In order to successfully implement machine learning techniques, several important aspects must be carefully considered [16]:

1) High-quality input data: It's important to have the proper high-quality input data, which has minimal artifacts and noise, to achieve the best possible machine learning performance and to maintain low noise levels, but ground truth labels are even more important as erroneous labels severely damage model performance and tend to be hard to catch in training.

2) Accuracy of ground truth labels: Ground truth labels must be accurate, because it involves confirming and checking the correctness of the diagnostic label and of the diagnosis itself. Other types of noise have a more dramatic effect on model accuracy than errors in ground truth labels, emphasizing the need for strict ground truth labeling.

3) Completeness of training sets: The best performance results from machine learning models exist when training datasets contain no missing features while also being complete. While there are data augmentation techniques like random imputation or sophisticated algorithms, which can help fill up these gaps, but have less effectiveness as we train on full datasets.

4) Dataset size: Machine learning models tend to prefer larger datasets so that they can capture the true variability of the data, which helps reduce the risk of performance issues caused by a small number of outliers. Nevertheless, obtaining large enough datasets remains difficult, in particular for precision medicine and disease studies with low frequency.

5) Precision and accuracy in classification algorithms: To obtain strong model performance, the optimization process has to balance both precision and accuracy, as both are indispensable for reliable and effective classification algorithms.

6) Model validation: Before a machine learning model is deployed to make clinical decisions or computer-aided decisions, it has to go through very thorough validation. For instance, a physician diagnoses cases and the model is trained using those cases and is evaluated by comparing its outputs with assessments of new cases by other physicians. This is done for reliability and accuracy [17].

B. Deep Learning

Deep Learning (DL) is mainly used in large datasets to model complex patterns using multi-layered artificial neural networks, which aim to mimic the human brain [18], [19]. As is known, DL is a powerful and transformative branch of machine learning, and it learns from unlabeled data such as images, audio, and text [20], [21]. Deep learning also has several key benefits that enable it to be very effective for a wide variety of applications. It can achieve high accuracy, especially when it has been trained on large data. One of its most significant features is that it can automatically learn data representation to avoid manual feature extraction [22], [23]. Also, deep learning models are highly scalable and easy to adapt to large data sets and increasing computational power, which makes them suitable for large-scale and complex problems too. Deep learning (DL) has numerous uses in various fields. In image and video processing, it is used for the recognition of specific elements, such as facial detection or facial emotion, and in medicine, it can forecast the kind of diseases. DL also powers voice-to-text systems such as Alexa and Google Voice and enables real-time machine translation from one language to another [24]. The building blocks of deep learning are multi-layer artificial neural networks, and they consist of a few basic components. They are neurons and layers, and each artificial neuron processes inputs by calculating weights, biases, and activation functions to supply the result to the next layer. Activation functions such as Rectified Linear Unit (ReLU), sigmoid, and tanh introduce essential non-linearity to neural networks. Loss functions like Mean Squared Error and Cross Entropy are used to quantify predictive accuracy. Optimization algorithms like Stochastic Gradient Descent (SGD) are used for improving model performance by reducing the loss function by adjusting the network's weights during training. Three main objectives of this study are: 1) identify the most appropriate deep learning model to analyze medical image data, 2) propose an integrated hybrid framework which combines deep learning and machine learning classifiers to boost the diagnostic performance, and 3) review previous research as the comparison studies of ML and DL in medical image classification. This research is organized as follows: In Section II, we briefly review related work; Section III describes the research methodology; Section IV provides the experimental results, and finally, conclusions are presented in Section V along with future scope.

II. Related Work

In this section, we will discuss the previous research about medical diagnosis using hybrid of machine learning and deep learning techniques.

A. Machine Learning Model

This section explains ten machine learning techniques used under the supervised learning category, categorized to clarify how each model approaches learning and prediction as the following:

1) Bayes-based models: Bayesian models are designed with reference to Bayes' Theorem and are widely utilized for probabilistic classification tasks. Naive Bayes (NB) is a well-known example, particularly renowned for its simple architecture, high computational efficiency, and fast processing, particularly in dealing with big data. It needs only estimation of significant parameters—mean and variance of variables—to be performed using a small training set [25]. The second category involves models such as Linear Discriminant Analysis (LDA), which is a statistical technique used in discriminating between two or more event or object classes. LDA computes the linear feature combination that maximally discriminates between the classes and is generally used for feature reduction prior to classification [26].

2) Function-based models: Mathematical formula-based classification models are models based on mathematical expressions in order to discriminate between data classes. The Support Vector Machine (SVM) is a strong algorithm used to identify patterns as well as conduct regression and classification functions [27]. It is particularly effective in domains that need high-dimensional space analysis. Logistic Regression is another function-based model used for problems of binary

classification [28]. It works through the use of a database of independent variables and is often used in areas such as diagnosis of disease as well as prediction of risks. In addition, Multi-layer Perceptron (MLP) is an artificial neural network created by a few full connected layers and nonlinear activation functions. MLPs are ideal for classifying complex data sets like intrusion detection and problems where the data is not linearly separable [29].

3) Lazy learning models: Lazy algorithms differ from typical models in that they do not necessarily train a predictive model directly. Instead, they store the training data and make a decision at the moment of prediction. An old prime example is the K-Nearest Neighbors (KNN) algorithm. KNN classifies a query point with the class of the majority of its nearest neighbors in the training data, and it is an easy but effective method for classification tasks [30].

4) Meta-learning models: Meta-learning or ensemble strategies improve prediction performance by combining multiple models [31]. Adaptive Boosting (AdaBoost) is a commonly used ensemble strategy which builds an accurate classifier by iteratively adding up weak base learners. The current iteration allots weights to the classifiers based on their precision, with more accurate learners having greater weightage in the final decision. The final classification emerges through the process of weighted voting, achieving greater overall prediction performance [32]. Tree-Based Models: Tree-based algorithms use a decision tree structure for splitting data according to feature values. Random Forest (RF), introduced by Breiman in 2001, constructs numerous decision trees during training. It votes for the most common class for classification or estimates the average of the individual estimates for regression [33]. Gradient Boosting is another robust method that builds models step by step, each of which corrects the errors in previous models. It excels at boosting accuracy through making better predictions. Last but not least, the Decision Tree (DT) model itself creates a tree where the branches are feature tests and leaves are class labels. DTs are generally used in statistics, data mining, and elsewhere due to their simplicity and interpretability [34].

B. Deep Learning Model

This research explains four deep learning techniques specifically those used in deep convolutional neural networks (CNNs) applied to image classification, object detection, and other related tasks, as follows:

- Visual Geometry Group (VGG16) [35], the number 16 refers to the number of layers that make up this model, comprising 13 are convolutional layers and 3 are connected layers. It applied in medical imaging analysis, face recognition, and the detection of different objects.

- Residual Network (ResNet50) [36], the number 50 refers to the number of layers, comprising 48 convolutional layers and 2 are connected layers. The purpose of using it is residual connections to solve the problem of vanishing gradients in deep networks.

- MobileNet [37], designed specifically for mobile and embedded devices, mobileNet was designed to reduce computational cost and number of parameters by using depthwise separable convolutions.

- InceptionV3 [38], the number three refers to the third version of the model, which released in 2016. It is one of the powerful models in deep learning because of its ability to recognize and classify different images and used in all modern CNN networks.

In [39], the authors explore the application of Machine Learning (ML) models to improve the accuracy of breast cancer detection. They used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, then applied data normalization techniques and handled missing values to ensure data quality. The hybrid feature selection method, combining Correlation based Feature Selection (CFS) and Recursive Feature Elimination (RFE), was chosen to reduce the dataset to 11 key attributes. Different machine learning models, including Random Forest, Gradient Boosting, SVM, ANN, and MLP, were evaluated using 5-fold cross-validation. The Multilayer Perceptron (MLP) achieved the highest accuracy of 99.12. The study focuses of the importance of advanced machine learning algorithms for accurate breast cancer diagnosis in real-world applications. In [40], the authors focused on diagnosing diabetes using machine learning and its various techniques to select the best classifier for predicting the disease. The authors evaluated four feature selection techniques (e.g., Correlation Attribute Evaluator) and 12 classifiers across two datasets using metrics like accuracy, precision, recall, and F1-measure. They concluded that the Correlation Attribute Evaluator was optimal for feature selection, and MultiClassClassifier performed best in diagnosis tasks. This work demonstrates the effectiveness of machine learning in medical diagnostics [41]. Machine learning techniques are used to predict the outcomes of COVID-19 diagnosis and treatment. Different techniques such as decision trees, random forests, and support vector machines used to analyze the symptoms and medical history of patients. They trained the models to give severity predictions of the disease (mild, moderate, severe) and to recommend appropriate treatment for that severity prediction. The results showed that machine-learning models can be used to predict the amount of disease and therefore decreased the workload of healthcare providers and efficiency in the treatment.

In [42], the authors developed a deep learning-based model to diagnose five types of lung and colon tissues (two benign and three malignant), by using deep learning techniques and image processing such as (convolutional neural networks (CNNs)) on histopathological images. Their results showed on their proposed model high classification accuracy of (96.33%) in the early detection of cancer. In [43], the authors address the limitations of traditional machine learning models in medical diagnostics, which often rely on associative patterns between symptoms and diseases. The purpose of this study is to improve diagnostic accuracy by integrating causal reasoning into machine learning models, leading to greater alignment with clinical diagnostic processes. The authors applied the proposed model to clinical vignettes and compared its performance with standard associative algorithms and human doctors. The results showed that their causal model outperformed associative models with an accuracy comparable to the top (25%) of doctors. In [44], the authors proposed a deep learning model for brain tumor classification using both a standard Convolutional Neural Network (CNN) and the ResNet50 model. In

TABLE I. SUMMARIZING THE RECENT REFERENCES: REFERENCE, YEAR, TITLE, METHODS, ADVANTAGES, AND LIMITATIONS

Reference	Year	Title	Methods	Advantages	Limitations
Salma et al. [39]	2025	Leveraging Machine Learning for Effective Breast Cancer Diagnosis	Machine Learning (ML) models including classification algorithms	High diagnostic accuracy for breast cancer; improved efficiency in early detection	Limited to breast cancer data; generalizability to other datasets not evaluated
Alzyoud et al. [40]	2024	Diagnosing Diabetes Mellitus Using Machine Learning Techniques	ML classification techniques (e.g., SVM, Decision Tree)	Efficient identification of diabetic patients; applicable to real-world datasets	Accuracy depends on data quality; lacks deep learning integration
Rudra Kumar et al. [41]	2022	Diagnosis and Medicine Prediction for COVID-19 Using Machine Learning Approach	ML algorithms for diagnosis and treatment prediction	Supports both diagnosis and treatment planning; fast processing	Limited to COVID-19; based on initial pandemic data with possible bias
Masud et al. [42]	2021	A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework	Deep learning (CNN-based classification)	High performance in cancer type classification; robust feature extraction	Computationally intensive; requires large labeled datasets
Richens et al. [43]	2020	Improving the Accuracy of Medical Diagnosis with Causal Machine Learning	Causal ML techniques to improve prediction accuracy	Enhances interpretability and accuracy; reduces bias in predictions	Complex model structure; harder to implement in standard ML workflows
Ali et al. [44]	2025	Learning Architecture for Brain Tumor Classification Based on Deep Convolutional Neural Network: Classic and ResNet50	CNN and ResNet50 architectures	High classification accuracy; effective for medical imaging	Deep models require high computation and large data for training

their work, deep CNNs were shown to accurately distinguish among the different tumor types through the extraction of high-level features from medical imaging data. Comparison of the baseline CNN and ResNet50 revealed that state-of-the-art deep architectures like ResNet50 achieved superior accuracy, underlining the strength of higher-order deep learning models in medical diagnosis and image-based classification.

The comparison table (Table I) highlights a line of recent research studies based on machine learning (ML) and deep learning techniques in medical diagnosis. Various methods were employed throughout the research studies, from popular ML algorithms Support Vector Machines and Decision Trees to more advanced deep learning models such as Convolutional Neural Networks (CNNs) and ResNet50.

These approaches have demonstrated superb performance in the diagnosis of the disease: breast cancer, diabetes, COVID-19, lung and colon cancer, and brain tumors. Notable strengths that they reported are improved accuracy in diagnosis, operational efficiency for early detection, and the capability to process large and multidimensional medical datasets. However, each of the studies came with certain limitations. For instance, models that involve deep learning require a lot of computational power and vast amounts of labeled data, while typical ML approaches may struggle with generalizability or lack the ability to process complex patterns in data. Notably, adding causal machine learning to a single experiment gave better interpretability, but at the cost of more complex models. Overall, the table underscores the growing effectiveness of ML in medicine as well as the necessity for balancing accuracy, interpretability, and computational efficiency.

III. PROPOSED METHODOLOGY

This section presents the research methodology, including the framework, datasets, evaluation metrics, and a summary of the findings. Section III(A) provides a detailed description of the five selected datasets. Section III(B) presents the methodology.

A. Description of the Dataset

In this study, five publicly available medical image datasets from Kaggle were utilized to ensure comprehensive evaluation across a wide range of clinical scenarios. These datasets include chest X-ray images for COVID-19 and pneumonia detection, dermoscopic images for melanoma skin lesion classification, facial images for acute stroke diagnosis, and a multi-class collection of ocular images for eye disease identification. This diversity provides a robust test bed for assessing the generalizability and effectiveness of hybrid deep learning and machine learning approaches in automated medical diagnosis. Each dataset covers multiple conditions or disease classes with varying sample sizes, enabling a thorough assessment of model performance across different diagnostic challenges. The key characteristics and class distributions of these datasets are summarized in Table II.

B. Methodology

The main idea of this research is to build a hybrid model that combines ML and DL algorithms to determine the best combination suitable for medical diagnosis tasks; Fig. 1 illustrates the general methodology of the current research. The following subsections provide a detailed explanation of the main steps considered in the methodology of this research.

1) Data collection: At this stage, different datasets have been searched to suit the purpose of this research, images containing different diseases have been collected.

Five datasets have been downloaded from the Kaggle website: COVID-19 images, melanoma skin cancer images, chest X-ray images, various eye diseases, and facial images of acute stroke and non-acute stroke. More information regarding the collected datasets is provided later in this section.

2) Image pre-processing: At this stage, the collected images are pre-processed to ensure they are suitable for machine learning or deep learning models. Initial pre-processing steps in image analysis begin with rescaling the image pixel values from 0 to 255 to be OpenCV compatible, as it is a widely used computer vision library. Normalizing the input data using this standard allows for uniform and reliable utilization of OpenCV

functions throughout the processing chain. The second significant step is applying denoising using GaussianBlur, an extremely common technique for reducing noise by smoothing pixel intensity values through averaging neighboring pixels in the context of a Gaussian function. It suppresses random noise, exaggerates prominent image features, and ultimately improves the accuracy and resilience of image analysis models.

Convert the image to grayscale Converting an image to grayscale simplifies the data which removes the unnecessary color information and only leaves the image with the intensity values. It merges the three-color channels (Red, Green, and Blue) into one where a pixel is a value of light intensity. Grayscale format reduces computational complexity and puts the focus on the structural information. Histogram application: Histogram application enhances image contrast, making features more distinguishable and improving visual clarity. Histogram equalization merely reshapes pixel intensity values from 0–255 to distribute them across the full range of 0–255 in order to highlight small details that are a point or two overshadowed by greater attention to detail. This also allows edges and textures which are necessary for tasks such as the edge detection and image segmentation. Fig. 2 show the image pre-processing pipeline starts by rescaling pixel values to the range of 0 to 255. GaussianBlur is then used to smooth pixel intensities and reduce noise, enhancing key features. After that, the pictures are converted to grayscale, which reduces computational complexity by combining color channels into a single intensity channel. Last but not least, histogram equalization is used to enhance image contrast by redistributing pixel intensities, emphasizing textures, edges, and minute details that are essential for precise image analysis. The images are ready for efficient machine learning or deep learning model training thanks to this series of actions. Many studies have highlighted the importance of image pre-processing for boosting the performance of machine learning-based medical diagnosis. As an example, [45] shows that application of Contrast Limited Adaptive Histogram Equalization (CLAHE) improved classification performance of histopathology images in terms of F1-score, which was 98% with pre-processing and 93% without it. Similarly, [46] proposed to use convolutional denoising autoencoders for effective denoising of medical images, thereby, maintaining essential clinical information even in noisy case. Furthermore, the combination of Gaussian smoothing and clustering techniques applied before CNN training, according to [47], enhanced the detection accuracy of tumors. These findings validate the image pre-processing techniques in this study, which include grayscale conversion along with histogram equalization and Gaussian denoising for improving model robustness and feature detection through noise reduction.

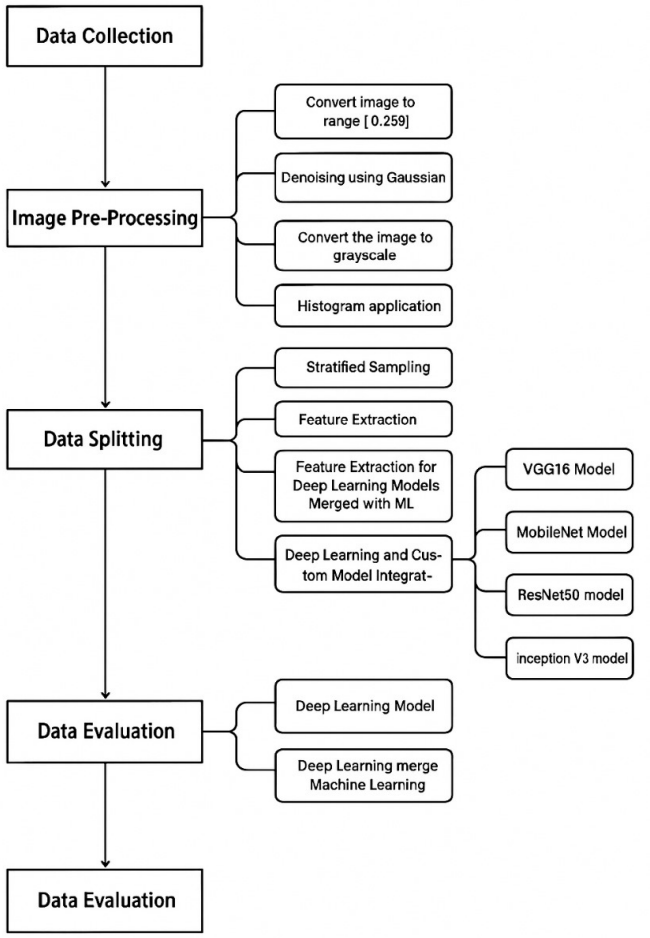

Fig. 1. Research methodology.

TABLE II. CHARACTERISTICS OF THE DATASETS

Dataset	Condition (Class)	Frequency
COVID CXR Image	COVID-19	536
COVID CXR Image	Viral pneumonia	619
COVID CXR Image	Normal cases	668
COVID CXR Image	**Total**	**1823**
Melanoma Skin Cancer	Benign	5000
Melanoma Skin Cancer	Malignant	4605
Melanoma Skin Cancer	**Total**	**9605**
Chest X-Ray	NORMAL	1341
Chest X-Ray	PNEUMONIA	3875
Chest X-Ray	**Total**	**5216**
Face Images of Acute Stroke and Non-Acute Stroke	noStroke_data	2511
Face Images of Acute Stroke and Non-Acute Stroke	stroke_data	1259
Face Images of Acute Stroke and Non-Acute Stroke	**Total**	**3770**
Different Eye Disease	ACRIMA	560
Different Eye Disease	Glaucoma	212
Different Eye Disease	ODIR-5K	7000
Different Eye Disease	ORIGA	517
Different Eye Disease	Cataract	70
Different Eye Disease	retina_disease	70
Different Eye Disease	**Total**	**8429**

3) Data splitting: Once the images are pre-processed, the data is prepared for model training and evaluation. The primary activities of this method start with stratified sampling, ensuring that there is a well-distributed balance of classes between the training, validation, and test datasets, hence enabling representative data splits. Shuffling is then used to avoid bias by the data sample order, enabling randomness in the dataset. Feature extraction plays a vital role in that it utilizes deep

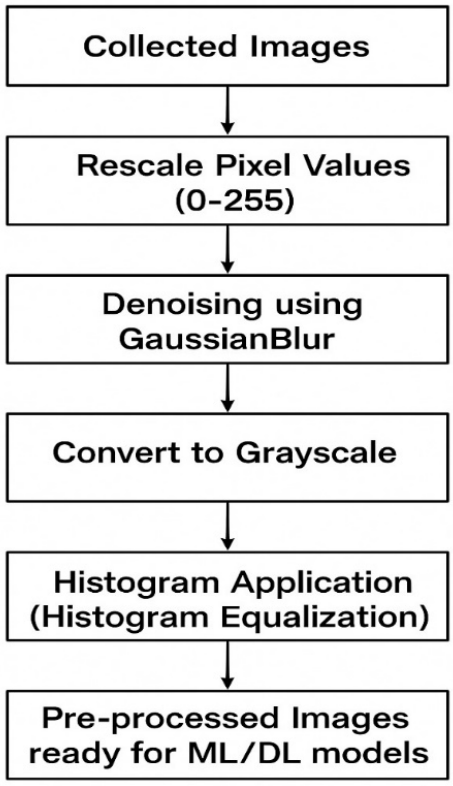

Fig. 2. Images preprocessing flowchart.

models to extract useful image features, which are blended with machine learning to increase prediction precision. Several deep learning architectures are blended and adjusted to provide optimal performance: VGG16 is blended with ten different machine algorithms to optimize predictive capacity; MobileNet is customized to provide light-weight and efficient image processing; ResNet50, a residual very deep network, is utilized to provide effective image classification and prediction; and Inception V3 is utilized due to its advanced structure to provide high-performance predictions in complex tasks.Many researchers endorse the efficacy of hybrid and ensemble methods in medical AI applications, combining deep learning capabilities with conventional machine learning classifiers has shown good promise in enhancing the precision of medical image classification. A comparative study showed that hybrid models where pre-trained convolutional neural networks like MobileNet, VGG16, and Inception V3 are used as feature extraction and then machine learning algorithms are used for classification, attained very high accuracy [48]. Furthermore, ensembling techniques that integrate predictions from different classifiers showed the ability to increase the robustness and reliability of models in clinical environments, as shown by an evaluation of a revised medical image classification process [49]. Fig. 3 shows the data splitting diagram.

4) Data training: Data training involves running deep learning models with preprocessed data for improved predictions. This section explains details about the specific models, their configurations and how deep learning is merged with a machine learning algorithm. Fig. 4 shows the data training process.

- Deep Learning Model. We train on the preprocessed images in this phase to extract our feature using different deep learning models for prediction. The models used in this research include: *VGG16*, *MobileNet*, *ResNet50*, and *Inception V3* models.

- Deep Learning merge with Machine Learning. Features from the deep learning models are merged with various machine learning algorithms in order to better increase the predictive power of the deep learning models.

  This hybrid approach involves the following steps:
  - Feature Extraction: Using the deep learning models (such as VGG16, MobileNet, ResNet50 and Inception V3) are used to automatically extract relevant features from the images. Features that typically include textures, edges, and patterns which are essential for disease classification can be included in these features.

  - Using Machine Learning Algorithms: When features are extracted, we input them into various machine learning algorithms for classification and prediction. Ten machine learning algorithms are used including: (Naive Bayes (NB), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Logistic Regression (LG), Multi-layer Perceptron (MLP), K-Nearest Neighbors Algorithm (KNN), Adaptive Boosting (AdaBoost), Random Forest (RF), Gradient Boosting, Decision Tree (DT)). The deep learning models tries to create predictions while these algorithms optimization for the predictions in order to increase accuracy and performance of the overall system.

  - Hybrid Model: This method combines the abilities of deep learning to extract features with the flexibility and efficiency of machine learning for prediction to improve the overall predictive ability of classification, especially in classifying images with high accuracy.

5) Evaluation phase: The hybrid model performance is evaluated in terms of accuracy and efficiency using evaluation metrics in classifying medical images accurately and efficiently. In [42], the authors presented a hybrid model that is a combination of a deep CNN and SVM classifier to improve brain tumor classification. They used common metrics, were Accuracy, Precision, Recall, F1-score, and Dice Similarity Coefficient (DSC) for the assessment of the model. The model achieved an accuracy of (98.3%) and DSC of (97.8%) on the Brain Dataset-1. Their finding showed the efficiency of combining deep learning for feature extraction with machine learning for classification in medical imaging tasks. In this research, four evaluation measures have been used as the following:

C. Evaluation Metrics

Accuracy: Calculates the proportion of correctly classified instances (both positive and negative) out of all instances; it

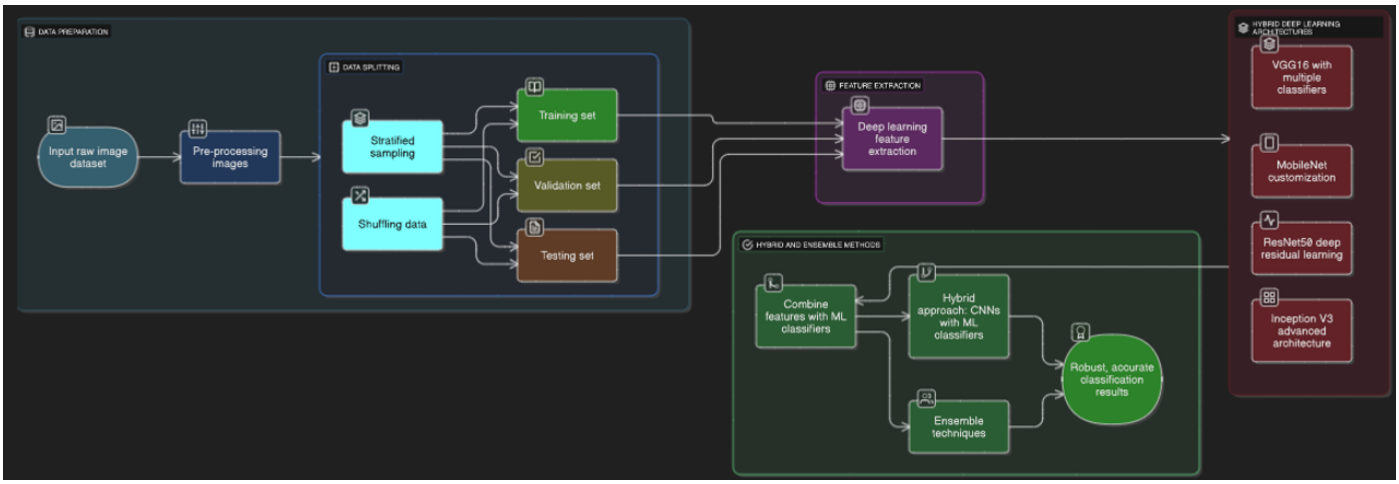

Fig. 3. Data splitting process.

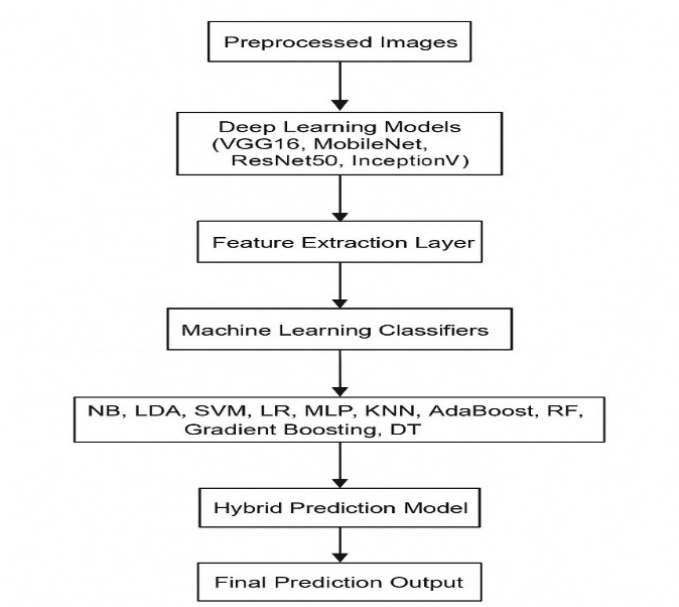

Fig. 4. Data training process diagram.

measures the model's overall correctness.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{1}$$

Precision: Measures how many of the predicted positives are actually positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2}$$

Recall: Measures the model's ability to capture all actual positive cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

F1-score: The harmonic mean of Precision and Recall.

$$\text{F1} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

IV. Experimental Results

This section presents the evaluation results of the experiments on developing an efficient medical image classification model. The model is a hybrid that combines machine learning and deep learning algorithms. This model aims to enhance the predictive performance by taking advantage of the strengths of both techniques.

- Identifying the Most Appropriate DL Model That Suits Medical Data. In this phase, which is considered the most important step of this research, the predictive performance of the four deep learning models (*VGG16*, *ResNet50*, *MobileNet*, and *InceptionV3*) is evaluated across five different datasets related to medical image classification, considering four evaluation metrics (Accuracy, Precision, Recall, and F1-Score), as shown in Table III.

The most significant key insights that could be concluded from Table III are listed next with respect to the five considered datasets.

- COVID-19 Dataset. Out of the models it was compared with, *MobileNet* performed the most accurately at 0.97, outperforming *VGG16* and *InceptionV3* with accuracies of 0.96 each. *ResNet50* performed the worst at 0.74. Additionally, Precision, Recall, and F1-score indicate that *MobileNet*, *VGG16*, and *InceptionV3* performed comparably, while *ResNet50* lagged on this task.

- Melanoma Skin Cancer Dataset. *MobileNet* was best with 0.93 accuracy, slightly above *VGG16* and *InceptionV3* (0.92 each). *ResNet50* was worst (0.78). Precision, Recall, and F1-score were high for *MobileNet*, *VGG16*, and *InceptionV3*, highlighting their suitability for melanoma identification.

- Chest X-Ray Images. *MobileNet* achieved the maximum accuracy (0.99), followed by *VGG16* (0.98) and

TABLE III. PERFORMANCE COMPARISON OF DEEP LEARNING MODELS ON VARIOUS MEDICAL DATASETS

Dataset	Deep Learning Model	Acc	Prec	Rec	F1
COVID-19 Dataset	VGG16	0.96	0.97	0.97	0.97
COVID-19 Dataset	ResNet50	0.74	0.73	0.74	0.74
COVID-19 Dataset	MobileNet	0.97	0.97	0.97	0.97
COVID-19 Dataset	InceptionV3	0.96	0.97	0.97	0.97
Melanoma Skin Cancer Dataset	VGG16	0.92	0.92	0.92	0.92
Melanoma Skin Cancer Dataset	ResNet50	0.78	0.78	0.78	0.78
Melanoma Skin Cancer Dataset	MobileNet	0.93	0.93	0.93	0.93
Melanoma Skin Cancer Dataset	InceptionV3	0.92	0.92	0.92	0.91
Chest X-Ray Images	VGG16	0.98	0.97	0.97	0.97
Chest X-Ray Images	ResNet50	0.91	0.88	0.90	0.89
Chest X-Ray Images	MobileNet	0.99	0.98	0.99	0.98
Chest X-Ray Images	InceptionV3	0.97	0.97	0.97	0.97
Face Images of Acute Stroke and Non-Acute Stroke	VGG16	0.99	0.99	0.99	0.99
Face Images of Acute Stroke and Non-Acute Stroke	ResNet50	0.76	0.74	0.70	0.71
Face Images of Acute Stroke and Non-Acute Stroke	MobileNet	0.99	0.99	0.99	0.99
Face Images of Acute Stroke and Non-Acute Stroke	InceptionV3	0.98	0.98	0.98	0.98
Dataset for Different Eye Disease	VGG16	0.96	0.78	0.76	0.74
Dataset for Different Eye Disease	ResNet50	0.90	0.44	0.44	0.44
Dataset for Different Eye Disease	MobileNet	0.94	0.76	0.72	0.73
Dataset for Different Eye Disease	InceptionV3	0.96	0.95	0.96	0.95

InceptionV3 (0.97). *ResNet50* scored 0.91. Across metrics, *MobileNet* is the optimal choice for this dataset.

- Face Images of Acute Stroke and Non-Acute Stroke. Both *VGG16* and *MobileNet* obtained the highest accuracy (0.99), with *InceptionV3* at 0.98. *ResNet50* was much lower (0.76). Precision/Recall/F1 confirm that *VGG16*, *MobileNet*, and *InceptionV3* are very effective for stroke classification.

- Dataset for Different Eye Disease. *InceptionV3* and *VGG16* reached the highest accuracy (0.96), *MobileNet* achieved 0.94, whereas *ResNet50* underperformed (0.90) with poor Precision/Recall/F1 (around 0.44). Overall performance was strong, with *InceptionV3* most reliable across metrics.

Main findings (Objective 1):

- *MobileNet* showed consistently high performance across all datasets, often achieving the highest Accuracy.

- *ResNet50* had the worst performance across Accuracy, Precision, Recall, and F1-score.

- *VGG16* and *InceptionV3* performed similarly well, with *InceptionV3* slightly more consistent.

- *MobileNet* and *InceptionV3* appear to be the most reliable for medical image classification, while *ResNet50* is not recommended.

TABLE IV. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH VGG16 FOR COVID-19 DATASET

Model	Accuracy	Precision	Recall	F1-Score
VGG16+SVM	0.89	0.90	0.89	0.89
VGG16+Random Forest	0.95	0.95	0.95	0.95
VGG16+Gradient Boosting	0.94	0.94	0.94	0.94
VGG16+AdaBoost	0.87	0.88	0.87	0.87
VGG16+KNN	0.94	0.94	0.94	0.94
VGG16+Logistic Regression	0.94	0.94	0.94	0.94
VGG16+Decision Tree	0.83	0.83	0.83	0.83
VGG16+Naïve Bayes	0.82	0.84	0.82	0.82
VGG16+LDA	**0.95**	**0.96**	**0.95**	**0.96**
VGG16+MLP	0.94	0.94	0.94	0.94

TABLE V. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH RESNET50 FOR COVID-19 DATASET

Model	Accuracy	Precision	Recall	F1-Score
ResNet50+SVM	0.44	0.25	0.40	0.30
ResNet50+Random Forest	0.89	0.89	0.89	0.89
ResNet50+Gradient Boosting	**0.93**	**0.93**	**0.93**	**0.93**
ResNet50+AdaBoost	0.85	0.85	0.85	0.85
ResNet50+KNN	0.87	0.88	0.87	0.87
ResNet50+Logistic Regression	0.85	0.85	0.85	0.85
ResNet50+Decision Tree	0.79	0.79	0.80	0.79
ResNet50+Naïve Bayes	0.78	0.78	0.78	0.78
ResNet50+LDA	0.91	0.91	0.91	0.91
ResNet50+MLP	0.78	0.81	0.78	0.76

A. Combine Deep Learning Models with Machine Learning Techniques

This section presents the evaluation results of combining deep learning models and machine learning techniques with respect to the five considered datasets.

1) Evaluation results with respect to the COVID-19 dataset: This part evaluates several ML classifiers combined with the four DL models for the COVID-19 dataset.

a) VGG16 performance: In Table IV, *VGG16* with LDA (Linear Discriminant Analysis) achieved the best results, demonstrating strong performance and making it an optimal choice for COVID-19 classification. However, *VGG16* with Naïve Bayes and with Decision Tree performed worst, implying these classifiers may be less appropriate for this dataset.

ResNet50 Performance. Table V shows that Gradient Boosting achieved the highest performance across all metrics (**Accuracy=0.93**, **Precision=0.93**, **Recall=0.93**, **F1=0.93**), highlighting its strength for COVID-19 classification and its ability to learn the underlying patterns in the dataset. In contrast, the SVM variant recorded the weakest results and exhibited an imbalanced precision–recall trade-off.

MobileNet Performance.

Table VI shows that MobileNet + MLP achieved the best results across all metrics (**Accuracy=0.97**, **Precision=0.97**, **Recall=0.97**, **F1=0.97**), making it the most effective combination. In contrast, MobileNet + Decision Tree recorded

TABLE VI. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH MOBILENET FOR COVID-19 DATASET

Model	Accuracy	Precision	Recall	F1-Score
MobileNet+SVM	0.96	0.96	0.96	0.96
MobileNet+Random Forest	0.95	0.95	0.95	0.95
MobileNet+Gradient Boosting	0.95	0.95	0.95	0.95
MobileNet+AdaBoost	0.92	0.92	0.92	0.92
MobileNet+KNN	0.94	0.94	0.94	0.94
MobileNet+Logistic Regression	0.96	0.96	0.96	0.96
MobileNet+Decision Tree	0.85	0.86	0.86	0.86
MobileNet+Naïve Bayes	0.93	0.93	0.93	0.93
MobileNet+LDA	0.92	0.92	0.92	0.92
MobileNet+MLP	**0.97**	**0.97**	**0.97**	**0.97**

TABLE VII. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH INCEPTIONV3 FOR COVID-19 DATASET

Model	Accuracy	Precision	Recall	F1-Score
InceptionV3+SVM	0.93	0.94	0.94	0.94
InceptionV3+Random Forest	0.91	0.91	0.92	0.91
InceptionV3+Gradient Boosting	0.92	0.93	0.93	0.92
InceptionV3+AdaBoost	0.87	0.88	0.88	0.88
InceptionV3+KNN	0.87	0.89	0.88	0.88
InceptionV3+Logistic Regression	0.92	0.92	0.92	0.92
InceptionV3+Decision Tree	0.79	0.81	0.79	0.79
InceptionV3+Naïve Bayes	0.85	0.85	0.85	0.85
InceptionV3+LDA	0.73	0.74	0.73	0.74
InceptionV3+MLP	0.92	0.92	0.92	0.92

the weakest performance (**Accuracy**=0.85, **Precision**=0.86, **Recall**=0.86, **F1**=0.86).

InceptionV3 Performance.

Table VII shows that *InceptionV3 + SVM* achieved the strongest, well-balanced results across all metrics (**Accuracy**=0.93, **Precision**=0.94, **Recall**=0.94, **F1**=0.94). In contrast, *InceptionV3 + LDA* exhibited the weakest performance (**Accuracy**=0.73, **Precision**=0.74, **Recall**=0.73, **F1**=0.74).

Based on the previous performance metrics, the most suitable machine learning (ML) classifier for the COVID-19 medical data depends on the chosen deep learning (DL) feature extractor. The best combinations by *Accuracy* are:

- If *MobileNet* is used, *MLP* is the best choice (Accuracy: **0.97**).
- If *VGG16* is used, *LDA* is the best choice (Accuracy: **0.95**).
- If *ResNet50* is used, *Gradient Boosting* is the best (Accuracy: **0.93**).
- If *InceptionV3* is used, *SVM* is the best (Accuracy: **0.93**).

For the COVID-19 dataset, **MobileNet + MLP** is the strongest overall. Final model selection should also consider interpretability, computational cost, and dataset-specific requirements, as summarized in Fig. 5.

TABLE VIII. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH VGG16 FOR MELANOMA SKIN CANCER DATASET

Model	Accuracy	Precision	Recall	F1-score
VGG16+SVM	0.84	0.85	0.84	0.84
VGG16+Random Forest	0.89	0.89	0.89	0.89
VGG16+Gradient Boosting	0.90	0.90	0.90	0.90
VGG16+AdaBoost	0.89	0.89	0.89	0.89
VGG16+KNN	0.88	0.89	0.88	0.88
VGG16+Logistic Regression	0.89	0.89	0.89	0.89
VGG16+Decision Tree	0.85	0.85	0.85	0.85
VGG16+Naïve Bayes	0.70	0.76	0.70	0.68
VGG16+LDA	0.90	0.90	0.90	0.90
VGG16+MLP	0.92	0.92	0.92	0.91

TABLE IX. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH RESNET50 FOR MELANOMA SKIN CANCER DATASET

Model	Accuracy	Precision	Recall	F1-Score
ResNet50+SVM	0.62	0.75	0.62	0.56
ResNet50+Random Forest	0.89	0.89	0.89	0.89
ResNet50+Gradient Boosting	0.90	0.90	0.90	0.90
ResNet50+AdaBoost	0.88	0.88	0.88	0.88
ResNet50+KNN	0.87	0.87	0.87	0.87
ResNet50+Logistic Regression	0.89	0.89	0.89	0.89
ResNet50+Decision Tree	0.82	0.82	0.82	0.82
ResNet50+Naïve Bayes	0.63	0.73	0.63	0.60
ResNet50+LDA	0.90	0.90	0.90	0.90
ResNet50+MLP	0.85	0.85	0.85	0.85

2) Evaluation results with respect to the melanoma skin cancer dataset:

a) VGG16 performance: Table VIII shows that the best performing model among all models was *VGG16* with *MLP*, achieving the highest values across all metrics. In contrast, *VGG16* with Naïve Bayes performed the worst, making it the least effective classifier in this comparison.

ResNet50 Performance.

Table IX shows that *ResNet50 + Gradient Boosting* and *ResNet50 + LDA* achieved the highest scores across all metrics (**Accuracy=0.90**, **Precision=0.90**, **Recall=0.90**, **F1=0.90**), indicating strong generalization and a balanced precision–recall trade-off. In contrast, *ResNet50 + SVM* produced the weakest performance among the compared classifiers.

MobileNet Performance. Table X shows that *MobileNet + SVM* achieved the highest scores across all metrics (Accuracy, Precision, Recall, and F1), making it the top-performing configuration for this dataset. In contrast, *MobileNet + Naive Bayes* yielded the weakest results, recording the lowest values on all evaluated metrics.

b) InceptionV3 performance: Table XI illustrates that the *InceptionV3* with *SVM* model achieved the highest Accuracy, Precision, Recall, and F1-score. Meanwhile, the *InceptionV3* combined with the *Decision Tree* model showed the weakest performance across all metrics.

Based on the above performance metrics, the most suitable ML classifier for the *Melanoma Skin Cancer* data depends on the chosen DL feature extractor. The best combinations by *Accuracy* are:

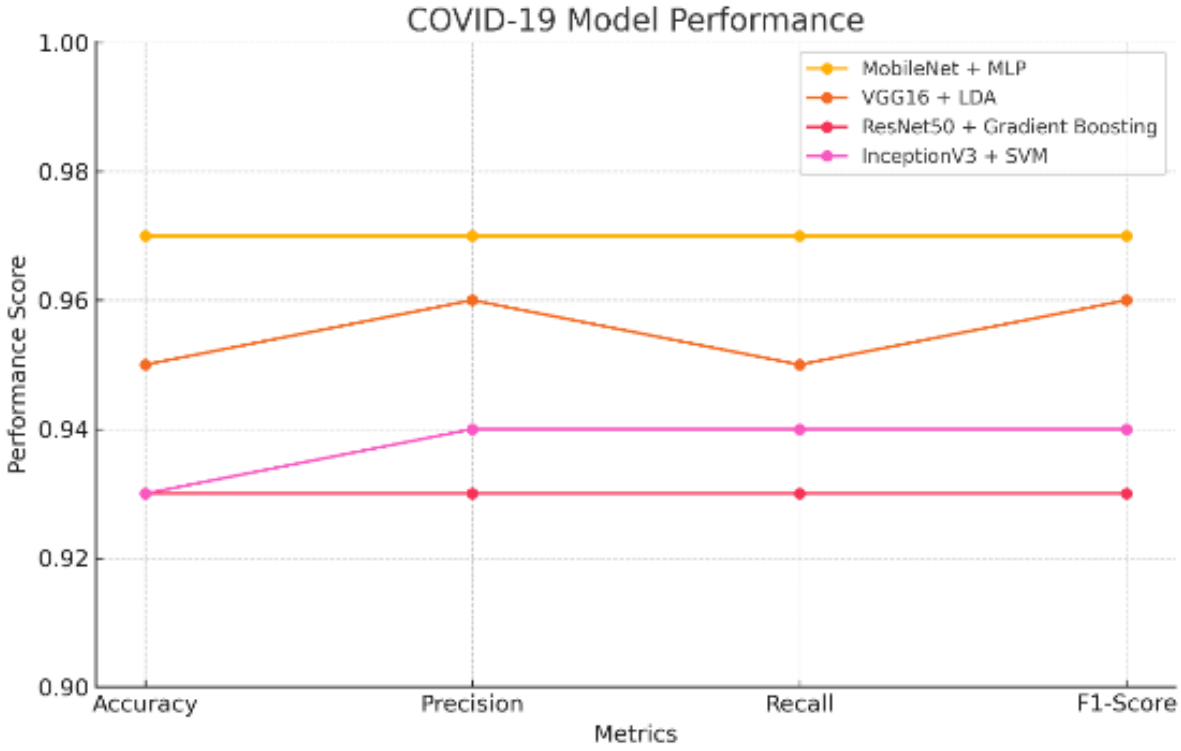

Fig. 5. COVID-19 model performance.

TABLE X. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH MOBILENET FOR MELANOMA SKIN CANCER DATASET

Model	Accuracy	Precision	Recall	F1-Score
MobileNet+SVM	0.92	0.92	0.92	0.92
MobileNet+Random Forest	0.90	0.90	0.90	0.90
MobileNet+Gradient Boosting	0.91	0.91	0.91	0.90
MobileNet+AdaBoost	0.86	0.86	0.86	0.86
MobileNet+KNN	0.90	0.90	0.90	0.90
MobileNet+Logistic Regression	0.89	0.89	0.89	0.89
MobileNet+Decision Tree	0.82	0.83	0.82	0.82
MobileNet+Naïve Bayes	0.80	0.80	0.80	0.80
MobileNet+LDA	0.90	0.90	0.90	0.90
MobileNet+MLP	0.91	0.91	0.91	0.91

TABLE XI. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH INCEPTIONV3 FOR MELANOMA SKIN CANCER DATASET

Model	Accuracy	Precision	Recall	F1-Score
InceptionV3+SVM	0.91	0.91	0.91	0.91
InceptionV3+Random Forest	0.87	0.87	0.87	0.87
InceptionV3+Gradient Boosting	0.90	0.90	0.90	0.90
InceptionV3+AdaBoost	0.87	0.87	0.87	0.87
InceptionV3+KNN	0.86	0.87	0.86	0.85
InceptionV3+Logistic Regression	0.89	0.89	0.89	0.89
InceptionV3+Decision Tree	0.78	0.78	0.78	0.78
InceptionV3+Naïve Bayes	0.80	0.80	0.80	0.80
InceptionV3+LDA	0.89	0.89	0.89	0.89
InceptionV3+MLP	0.90	0.90	0.90	0.90

- If *MobileNet* is used, *SVM* is best (Accuracy: **0.92**).
- If *VGG16* is used, *MLP* is best (Accuracy: **0.92**).
- If *InceptionV3* is used, *SVM* is best (Accuracy: **0.91**).
- If *ResNet50* is used, *Gradient Boosting/LDA* are best (Accuracy: **0.90**).

Overall, **MobileNet+SVM** is the strongest model for melanoma; **VGG16+MLP** is also a strong alternative, especially if computational efficiency matters (see Fig. 6 and Table XII).

ResNet50 Performance. Table XIII shows that *ResNet50 + MLP* achieved the highest scores across all metrics (**Accuracy**=0.96, **Precision**=0.95, **Recall**=0.95, **F1**=0.95). In contrast, *ResNet50 + Random Forest* exhibited the weakest performance among the compared models.

c) MobileNet performance: Table XIV shows that, across all evaluation metrics, the *MobileNet+MLP* model outperforms the rest (Accuracy 0.98, Precision 0.97, Recall 0.97, and F1-score 0.97). In contrast, *MobileNet+Decision Tree* has the lowest performance (Accuracy 0.90, Precision 0.87, Recall 0.88, F1-score 0.88), indicating notable limitations and making it the least reliable among the tested models.

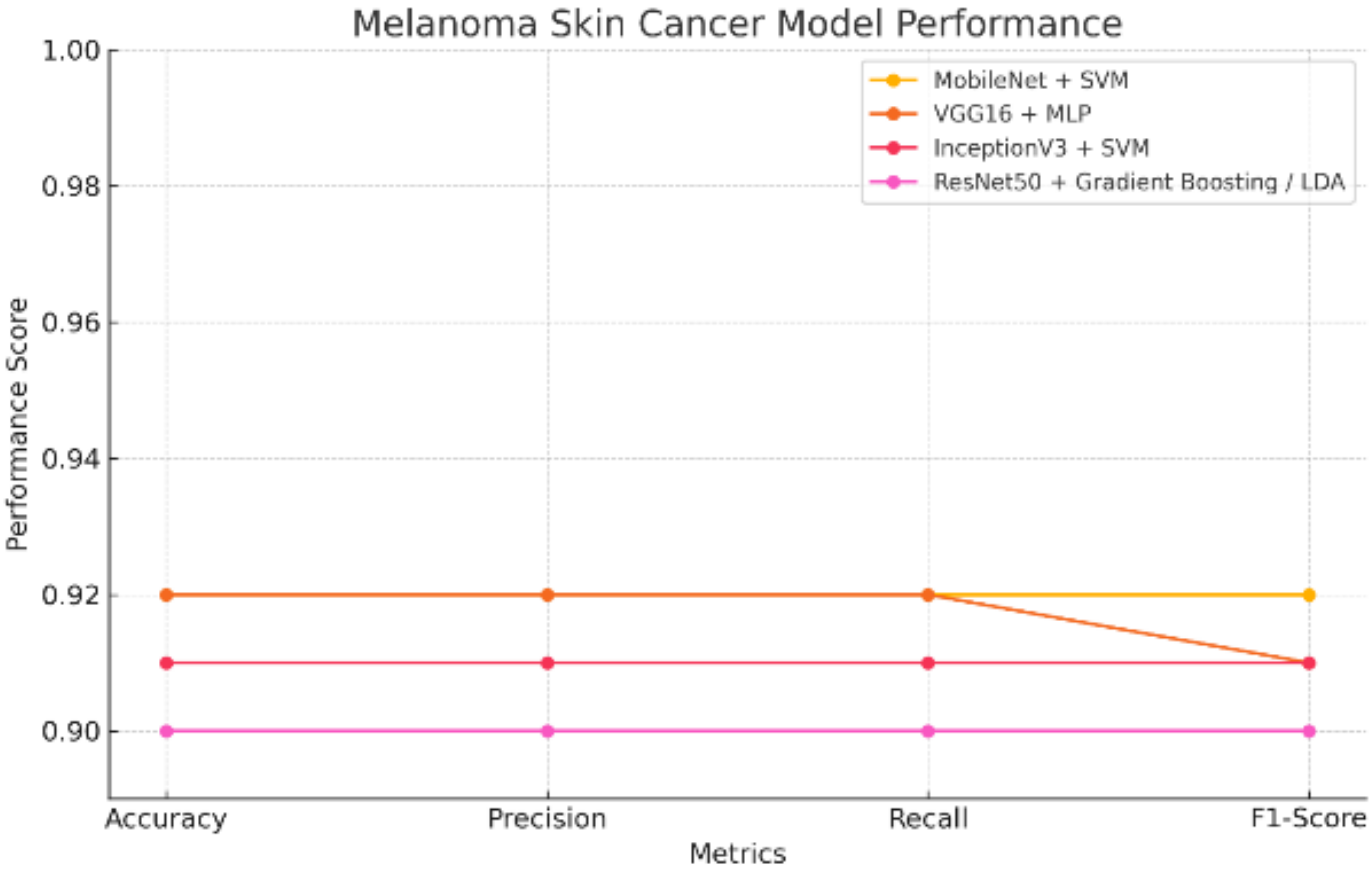

Fig. 6. Melanoma model performance.

TABLE XII. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH VGG16 FOR CHEST X-RAY IMAGES DATASET

Model	Accuracy	Precision	Recall	F1-Score
VGG16+SVM	0.93	0.91	0.91	0.91
VGG16+Random Forest	0.96	0.95	0.94	0.94
VGG16+Gradient Boosting	0.95	0.94	0.94	0.94
VGG16+AdaBoost	0.95	0.94	0.94	0.94
VGG16+KNN	0.95	0.93	0.93	0.94
VGG16+Logistic Regression	0.95	0.94	0.94	0.94
VGG16+Decision Tree	0.90	0.86	0.87	0.87
VGG16+Naïve Bayes	0.93	0.90	0.90	0.87
VGG16+LDA	0.97	0.96	0.95	0.96
VGG16+MLP	0.97	0.96	0.96	0.96

TABLE XIII. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH RESNET50 FOR CHEST X-RAY IMAGES DATASET

Model	Accuracy	Precision	Recall	F1-Score
ResNet50+SVM	0.91	0.88	0.90	0.89
ResNet50+Random Forest	0.74	0.37	0.50	0.43
ResNet50+Gradient Boosting	0.92	0.91	0.88	0.90
ResNet50+AdaBoost	0.94	0.93	0.92	0.92
ResNet50+KNN	0.94	0.96	0.93	0.93
ResNet50+Logistic Regression	0.90	0.86	0.89	0.88
ResNet50+Decision Tree	0.92	0.93	0.92	0.93
ResNet50+Naïve Bayes	0.89	0.86	0.86	0.86
ResNet50+LDA	0.83	0.78	0.90	0.88
ResNet50+MLP	0.96	0.95	0.95	0.95

d) InceptionV3 performance: The performance of models using the *InceptionV3* feature extractor with different ML classifiers is presented in Table XV. The best results were obtained with *InceptionV3+Logistic Regression* and *InceptionV3+MLP*, each achieving 0.96 for Accuracy, Precision, Recall, and F1-score. The weakest model was *InceptionV3+Decision Tree.*

- If *MobileNet* is used, *MLP* is best (Accuracy: **0.98**).
- If *VGG16* is used, *MLP* is best (Accuracy: **0.97**); *LDA* is a strong alternative.
- If *InceptionV3* is used, *MLP/Logistic Regression* are best (Accuracy: **0.96**).
- If *ResNet50* is used, *MLP* is best (Accuracy: **0.96**).

Overall, **MobileNet+MLP** achieves the highest Accuracy for Chest X-Ray Images, while **VGG16+MLP/LDA** is also a strong choice (see Fig. 7 and Table XVI).

e) ResNet50 performance: Table XVII shows that *ResNet50+KNN* achieved the best performance with the highest Accuracy (0.91), Precision (0.90), Recall (0.89), and F1-

TABLE XIV. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH MOBILENET FOR CHEST X-RAY IMAGES DATASET

Model	Accuracy	Precision	Recall	F1-Score
MobileNet+SVM	0.97	0.96	0.96	0.96
MobileNet+Random Forest	0.95	0.94	0.93	0.93
MobileNet+Gradient Boosting	0.96	0.95	0.94	0.94
MobileNet+AdaBoost	0.92	0.90	0.87	0.90
MobileNet+KNN	0.96	0.94	0.95	0.95
MobileNet+Logistic Regression	0.97	0.96	0.96	0.96
MobileNet+Decision Tree	0.90	0.87	0.88	0.88
MobileNet+Naïve Bayes	0.92	0.89	0.88	0.85
MobileNet+LDA	0.97	0.96	0.96	0.96
MobileNet+MLP	**0.98**	**0.97**	**0.97**	**0.97**

TABLE XV. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH INCEPTIONV3 FOR CHEST X-RAY IMAGES DATASET

Model	Accuracy	Precision	Recall	F1-Score
InceptionV3+SVM	0.95	0.95	0.95	0.95
InceptionV3+Random Forest	0.92	0.92	0.92	0.92
InceptionV3+Gradient Boosting	0.95	0.95	0.95	0.95
InceptionV3+AdaBoost	0.93	0.93	0.93	0.93
InceptionV3+KNN	0.91	0.92	0.91	0.91
InceptionV3+Logistic Regression	0.96	0.96	0.96	0.96
InceptionV3+Decision Tree	0.96	0.86	0.86	0.85
InceptionV3+Naïve Bayes	0.89	0.89	0.87	0.88
InceptionV3+LDA	0.94	0.94	0.94	0.94

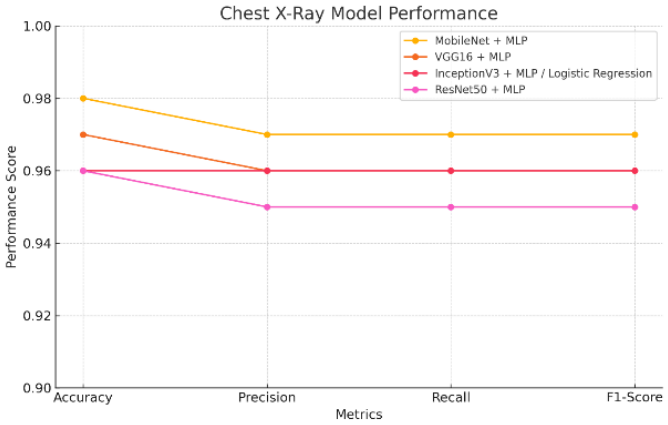

Fig. 7. Chest X-ray model performance.

score (0.89). Conversely, *ResNet50+SVM* recorded the lowest metrics (Accuracy 0.67, Precision 0.33, Recall 0.50, F1-score 0.40).

f) MobileNet performance: Table XVIII shows that the most effective models are *MobileNet+SVM*, *MobileNet+KNN*, *MobileNet+LDA*, and *MobileNet+MLP*, each achieving perfect scores (Accuracy, Precision, Recall, and F1-score of 1.00). In contrast, *MobileNet+Naïve Bayes* recorded the weakest performance (Accuracy 0.81, Precision 0.80, Recall 0.83, F1-score 0.80), making it the least effective classifier in this comparison.

g) InceptionV3 performance: Table XIX presents the results for models using the *InceptionV3* feature extractor.

TABLE XVI. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH VGG16 FOR FACE IMAGES DATASET

Model	Accuracy	Precision	Recall	F1-Score
VGG16+SVM	0.91	0.90	0.89	0.89
VGG16+Random Forest	0.98	0.98	0.97	0.98
VGG16+Gradient Boosting	0.98	0.98	0.98	0.98
VGG16+AdaBoost	0.96	0.96	0.95	0.95
VGG16+KNN	1.00	1.00	1.00	1.00
VGG16+Logistic Regression	0.97	0.97	0.97	0.97
VGG16+Decision Tree	0.90	0.89	0.89	0.89
VGG16+Naïve Bayes	0.84	0.83	0.87	0.83
VGG16+LDA	0.99	0.99	0.99	0.99
VGG16+MLP	1.00	1.00	1.00	1.00

TABLE XVII. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH RESNET50 FOR FACE IMAGES DATASET

Model	Accuracy	Precision	Recall	F1-Score
ResNet50+SVM	0.67	0.33	0.50	0.40
ResNet50+Random Forest	0.88	0.89	0.84	0.86
ResNet50+Gradient Boosting	0.88	0.86	0.85	0.86
ResNet50+AdaBoost	0.84	0.82	0.82	0.82
ResNet50+KNN	0.91	0.90	0.89	0.89
ResNet50+Logistic Regression	0.81	0.79	0.77	0.78
ResNet50+Decision Tree	0.78	0.75	0.75	0.75
ResNet50+Naïve Bayes	0.59	0.70	0.68	0.65
ResNet50+LDA	0.84	0.82	0.83	0.82
ResNet50+MLP	0.83	0.81	0.79	0.80

TABLE XVIII. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH MOBILENET FOR FACE IMAGES DATASET

Model	Accuracy	Precision	Recall	F1-Score
MobileNet+SVM	1.00	1.00	1.00	1.00
MobileNet+Random Forest	0.98	0.97	0.97	0.98
MobileNet+Gradient Boosting	0.99	0.99	0.99	0.99
MobileNet+AdaBoost	0.99	0.99	0.99	0.99
MobileNet+KNN	1.00	1.00	1.00	1.00
MobileNet+Logistic Regression	0.99	0.99	1.00	0.99
MobileNet+Decision Tree	0.88	0.86	0.80	0.87
MobileNet+Naïve Bayes	0.81	0.80	0.83	0.80
MobileNet+LDA	1.00	1.00	1.00	1.00
MobileNet+MLP	1.00	1.00	1.00	1.00

The best-performing classifiers were *SVM*, *KNN*, *Logistic Regression*, and *MLP*. In contrast, *InceptionV3+Naïve Bayes* recorded the weakest performance (Accuracy 0.75, Precision 0.79, Recall 0.75, F1-score 0.75).

Based on the previous performance metrics for the *Face Images of Acute Stroke and Non-Acute Stroke* dataset, the best ML classifier depends on the DL feature extractor:

- If *MobileNet* is used, *SVM*, *KNN*, *LDA*, or *MLP* achieved perfect scores (Accuracy: **1.00**).
- If *VGG16* is used, *KNN* or *MLP* are best (Accuracy: **1.00**).
- If *InceptionV3* is used, *SVM*, *KNN*, *Logistic Regression*, or *MLP* are best (Accuracy: **0.99**).

TABLE XIX. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH INCEPTIONV3 FOR FACE IMAGES DATASET

Model	Accuracy	Precision	Recall	F1-Score
InceptionV3+SVM	0.99	0.99	0.99	0.99
InceptionV3+Random Forest	0.94	0.94	0.94	0.94
InceptionV3+Gradient Boosting	0.95	0.96	0.95	0.95
InceptionV3+AdaBoost	0.93	0.93	0.93	0.93
InceptionV3+KNN	0.99	0.99	0.99	0.99
InceptionV3+Logistic Regression	0.99	0.99	0.99	0.99
InceptionV3+Decision Tree	0.83	0.84	0.84	0.84
InceptionV3+Naïve Bayes	0.75	0.79	0.75	0.75
InceptionV3+LDA	0.95	0.95	0.95	0.95
InceptionV3+MLP	0.99	0.99	0.99	0.99

TABLE XX. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH VGG16 FOR DIFFERENT EYE DISEASE DATASET

Model	Accuracy	Precision	Recall	F1-Score
VGG16+SVM	0.93	0.45	0.49	0.47
VGG16+Random Forest	0.92	0.67	0.62	0.62
VGG16+Gradient Boosting	0.92	0.67	0.61	0.63
VGG16+AdaBoost	0.81	0.30	0.29	0.29
VGG16+KNN	0.95	0.77	0.74	0.73
VGG16+Logistic Regression	0.94	0.80	0.58	0.59
VGG16+Decision Tree	0.88	0.58	0.61	0.59
VGG16+Naïve Bayes	0.62	0.54	0.67	0.54
VGG16+LDA	0.94	0.65	0.72	0.69
VGG16+MLP	0.95	0.75	0.73	0.72

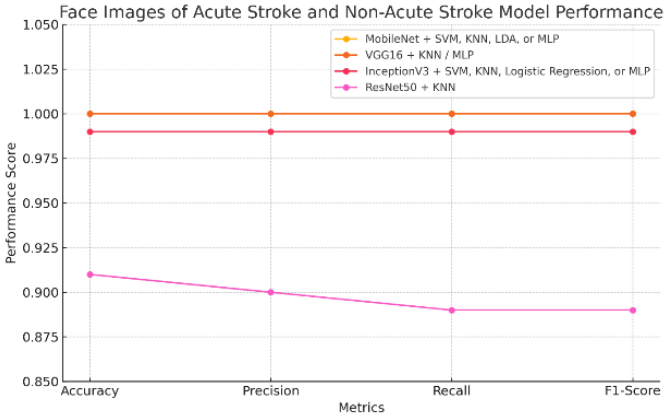

Fig. 8. Acute stroke model performance.

TABLE XXI. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH RESNET50 FOR DIFFERENT EYE DISEASE DATASET

Model	Accuracy	Precision	Recall	F1-Score
ResNet50+SVM	0.74	0.12	0.17	0.14
ResNet50+Random Forest	0.92	0.68	0.61	0.62
ResNet50+Gradient Boosting	0.90	0.54	0.53	0.53
ResNet50+AdaBoost	0.94	0.30	0.33	0.31
ResNet50+KNN	**0.93**	0.65	0.65	0.63
ResNet50+Logistic Regression	0.91	0.43	0.46	0.45
ResNet50+Decision Tree	0.88	0.39	0.32	0.34
ResNet50+Naïve Bayes	**0.38**	0.29	0.58	0.37
ResNet50+LDA	**0.93**	0.69	0.73	0.71
ResNet50+MLP	0.92	0.60	0.59	0.56

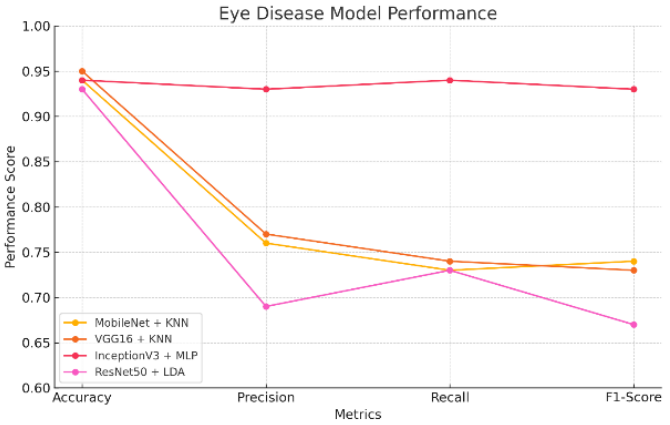

Fig. 9. Eye disease model performance.

TABLE XXII. PERFORMANCE COMPARISON OF ML CLASSIFIERS WITH INCEPTIONV3 FOR DIFFERENT EYE DISEASE DATASET

Model	Accuracy	Precision	Recall	F1-Score
InceptionV3+SVM	0.94	0.90	0.94	0.92
InceptionV3+Random Forest	0.91	0.88	0.91	0.89
InceptionV3+Gradient Boosting	0.91	0.89	0.90	0.90
InceptionV3+AdaBoost	0.85	0.85	0.78	0.79
InceptionV3+KNN	0.93	0.92	0.93	0.92
InceptionV3+Logistic Regression	0.93	0.93	0.93	0.93
InceptionV3+Decision Tree	0.79	0.86	0.76	0.83
InceptionV3+Naïve Bayes	0.70	0.73	0.75	0.74
InceptionV3+LDA	0.93	0.93	0.94	0.93
InceptionV3+MLP	0.94	0.93	0.94	0.93

- If *ResNet50* is used, *KNN* is best (Accuracy: **0.91**).

If computational efficiency and balanced performance are priorities, **MobileNet**-based models (especially with SVM, KNN, LDA, or MLP) are preferable due to MobileNet's lightweight nature. If interpretability and robustness are crucial, **VGG16** with **KNN** or **MLP** is an excellent choice (see Fig. 8 and Table XX).

h) ResNet50 performance: Table XXI shows that *ResNet50+KNN* and *ResNet50+LDA* achieved the highest Accuracy (0.93), with *ResNet50+LDA* delivering consistently strong results across all metrics. For the Precision metric, *ResNet50+Random Forest* obtained the best result (0.68). In contrast, *ResNet50+Naïve Bayes* was the weakest model, with the lowest Accuracy (0.38) and overall poor performance compared to the others.

i) InceptionV3 performance: Table XXII shows that *InceptionV3+SVM* and *InceptionV3+MLP* achieved the highest Accuracy (**0.94**). The highest Precision was obtained by *Logistic Regression*, *Naïve Bayes*, *LDA*, and *MLP*. The Recall was highest for *SVM* and *MLP*. Regarding F1-score, *Logistic Regression*, *LDA*, and *MLP* achieved the best results (**0.93**). In contrast, *InceptionV3+Naïve Bayes* was the weakest model overall, with the lowest Accuracy (**0.70**).

j) Summary: Different eye diseases dataset: Based on the previous performance metrics, the most suitable ML classifier depends on the chosen DL feature extractor. The best combinations by *Accuracy* are:

- If *VGG16* is used, *KNN* is best (Accuracy: **0.95**).
- If *InceptionV3* is used, *MLP* is best (Accuracy: **0.94**).
- If *MobileNet* is used, *SVM*/*KNN* are best (Accuracy: **0.94**).
- If *ResNet50* is used, *LDA* is best (Accuracy: **0.93**).

Overall, **VGG16+KNN** and **InceptionV3+MLP** are strong choices, while **MobileNet+SVM/KNN** provides competitive performance with efficient inference (see Fig. 9).

k) Different eye diseases dataset: For the *Different Eye Diseases* dataset, **VGG16+KNN** is the most appropriate model based on its highest Accuracy and balanced performance across evaluation metrics. However, **InceptionV3+MLP** is also a strong contender. The choice between them depends on application requirements such as computational efficiency and real-time processing needs (see Fig. 9).

B. Discussion

The primary aim of this study was to create and assess a hybrid model architecture that combines deep learning (DL) feature extraction with machine learning (ML) classification techniques for medical image analysis. Across the five datasets, performance varied with the particular pairing of DL feature extractor and ML classifier. Key findings are summarized below:

- COVID-19 Dataset: *MobileNet+MLP* achieved the highest Accuracy (0.97).
- Melanoma Skin Cancer Dataset: *MobileNet+SVM* and *VGG16+MLP* both reached top accuracies (0.92), showing that different hybrid configurations can yield competitive results depending on computational or interpretability needs.
- Chest X-Ray Dataset: *MobileNet+MLP* again achieved the highest accuracy (0.98).
- Face Images of Acute Stroke Dataset: *MobileNet* with several ML classifiers (SVM, KNN, LDA, MLP) achieved perfect accuracy (1.00), indicating strong discriminative power of facial feature embeddings for neurological conditions.
- Different Eye Diseases Dataset: The top performer was *VGG16+KNN* (Accuracy: 0.95), while *InceptionV3+MLP* also delivered competitive results (Accuracy: 0.94).

Overall, *MobileNet* consistently appears as the most adaptable and high-performing DL feature extractor across datasets, especially when paired with *MLP* or *SVM*, depending on the classification challenge. These results confirm the potential of hybrid DL–ML models in medical image classification.

Model selection should not rely on accuracy alone; additional factors include:

- Computational efficiency: e.g., *MobileNet* is favorable for time-sensitive applications due to high speed and low resource use.
- Interpretability: e.g., *LDA* or *Logistic Regression* offer clearer clinical insight.
- Dataset specifications: image resolution, noise levels, and class imbalance must be considered.

The evaluation demonstrates that combining DL and ML can yield hybrid models that improve classification while providing flexible solutions for varied medical domains.

V. Conclusion and Future Work

This research introduced a hybrid architecture that couples Deep Learning (DL) feature extraction with Machine Learning (ML) classification for medical image analysis across multiple datasets. Four pre-trained CNNs (VGG16, ResNet50, MobileNet, InceptionV3) served as deep feature extractors; their features were fed to ten traditional ML classifiers, forming a comprehensive hybrid framework. We evaluated the framework on five datasets: COVID-19 chest X-rays, melanoma skin cancer, general chest X-rays, acute stroke facial images, and a multi-class eye disease dataset.

Results show that performance depends on both the DL extractor and ML classifier. In most cases, *MobileNet* proved the most consistent and effective extractor, especially with *MLP* or *SVM*. This pairing achieved the highest accuracies in several datasets, such as COVID-19 (*MobileNet+MLP*, 0.97), chest X-rays (*MobileNet+MLP*, 0.99), and stroke images (*MobileNet+MLP/KNN/LDA*, 1.00). By contrast, *ResNet50* showed relatively lower performance on stroke and eye-disease classification, whereas *VGG16* and *InceptionV3* achieved strong results elsewhere and are good alternatives when interpretability, computational complexity, or domain-specific constraints are important.

The study further demonstrates that well-optimized hybrid DL–ML pipelines can outperform standalone ML or DL approaches, offering better generalization across neurological, dermatological, and respiratory imaging tasks.

Future work will expand to broader medical conditions and explore stronger denoising methods to enhance image quality. Another promising direction is multi-modal integration—combining images with clinical data—to improve accuracy and provide a more holistic diagnostic aid.

Acknowledgment

Funding: The generous support from Princess Nourah bint Abdulrahman University in Riyadh, Saudi Arabia, through the Researchers Supporting Project number (PNURSP2025R909) is greatly appreciated by the authors. This support enabled key activities such as data collection, analysis, and dissemination. The authors also express sincere gratitude to King Saud University for invaluable assistance and continuous support during the preparation of this work. The collective support of these institutions significantly enhanced the quality and impact of the research.

REFERENCES

[1] S. Zhang and D. Metaxas, "On the challenges and perspectives of foundation models for medical image analysis," *Medical Image Analysis*, vol. 91, p. 102996, 2024.

[2] P. Rajpurkar and M. P. Lungren, "The current and future state of AI interpretation of medical images," *New England Journal of Medicine*, vol. 388, no. 21, pp. 1981–1990, 2023.

[3] S. Hussain *et al.*, "Modern diagnostic imaging technique applications and risk factors in the medical field: A review," *BioMed Research International*, vol. 2022, no. 1, p. 5164970, 2022.

[4] R. Najjar, "Redefining radiology: A review of artificial intelligence integration in medical imaging," *Diagnostics*, vol. 13, no. 17, p. 2760, 2023.

[5] S. V. M. Sagheer and S. N. George, "A review on medical image denoising algorithms," *Biomedical Signal Processing and Control*, vol. 61, p. 102036, 2020.

[6] O. Tarawneh, Q. Saber, A. Almaghthawi, H. A. Owida, A. Issa, N. Alshdaifat, G. Jaradat, S. Abuowaida, and M. Arabiat, "The effect of pre-processing on a convolutional neural network model for dorsal hand vein recognition." *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 3, 2024.

[7] H. M. Turki, E. Al Daoud, G. Samara, R. Alazaidah, M. H. Qasem, M. Aljaidi, S. Abuowaida, and N. Alshdaifat, "Arabic fake news detection using hybrid contextual features." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 15, no. 1, 2025.

[8] M. Radak, H. Y. Lafta, and H. Fallahi, "Machine learning and deep learning techniques for breast cancer diagnosis and classification: A comprehensive review of medical imaging studies," *Journal of Cancer Research and Clinical Oncology*, vol. 149, no. 12, pp. 10 473–10 491, 2023.

[9] A. Ehsan, Z. Iqbal, S. Abuowaida, M. Aljaidi, H. U. Zia, N. Alshdaifat, and N. K. Alshammry, "Enhanced anomaly detection in ethereum: Unveiling and classifying threats with machine learning," *IEEE Access*, 2024.

[10] N. A. Mahoto, A. Shaikh, A. Sulaiman, M. S. Al Reshan, A. Rajab, and K. Rajab, "A machine learning based data modeling for medical diagnosis," *Biomedical Signal Processing and Control*, vol. 81, p. 104481, 2023.

[11] M. Kamran and A. A.-A. Almaghthawi, "Case-based reasoning diagnostic system for antenatal research database." *International Journal of Online & Biomedical Engineering*, vol. 18, no. 7, 2022.

[12] A. Almaghthawi, F. Bourennani, and I. R. Khan, "Differential evolution-based approach for tone-mapping of high dynamic range images," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020.

[13] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[14] S. Abuowaida, H. A. Owida, D. M. Alsekait, N. Alshdaifat, D. S. AbdElminaam, and M. Alshinwan, "Ultrasegnet: A hybrid deep learning framework for enhanced breast cancer segmentation and classification on ultrasound images." *Computers, Materials & Continua*, vol. 83, no. 2, 2025.

[15] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons. B*, vol. 4, pp. 51–62, 2017.

[16] S. J. MacEachern and N. D. Forkert, "Machine learning for precision medicine," *Genome*, vol. 64, no. 4, pp. 416–425, 2021.

[17] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[18] H. A. Owida, I. A. El-Fattah, S. Abuowaida, N. Alshdaifat, H. A. Mashagba, A. B. Abd Aziz, A. Alzoubi, S. Larguech, and S. S. Al-Bawri, "A deep learning-based dual-branch framework for automated skin lesion segmentation and classification via dermoscopic images," *Scientific Reports*, vol. 15, no. 1, p. 37823, 2025.

[19] H. A. Owida, B. A.-h. Moh'd, and M. Al Takrouri, "Designing an integrated low-cost electrospinning device for nanofibrous scaffold fabrication," *HardwareX*, vol. 11, p. e00250, 2022.

[20] A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi, and R. Ahmad, "Machine learning and deep learning approaches for cybersecurity: A review," *IEEE Access*, vol. 10, pp. 19 572–19 585, 2022.

[21] H. Abu Owida, "Recent biomimetic approaches for articular cartilage tissue engineering and their clinical applications: narrative review of the literature," *Advances in Orthopedics*, vol. 2022, no. 1, p. 8670174, 2022.

[22] A. Y. Alhusenat, H. A. Owida, H. A. Rababah, J. I. Al-Nabulsi, and S. Abuowaida, "A secured multi-stages authentication protocol for iot devices." *Mathematical Modelling of Engineering Problems*, vol. 10, no. 4, 2023.

[23] R. Alazaidah, H. A. Owida, N. Alshdaifat, A. Issa, S. Abuowaida, and N. Yousef, "A comprehensive analysis of eye diseases and medical data classification," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 22, no. 6, pp. 1422–1430, 2024.

[24] H. Abu Owida, B. A.-H. Moh'd, N. Turab, J. Al-Nabulsi, and S. Abuowaida, "The evolution and reliability of machine learning techniques for oncology," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 08, pp. 110–129, Jun. 2023. [Online]. Available: https://online-journals.org/index.php/i-joe/article/view/39433

[25] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive bayes classification algorithm for traffic risk management," *EURASIP Journal on Advances in Signal Processing*, vol. 2021, no. 1, p. 30, 2021.

[26] A. Bhardwaj, A. Gupta, P. Jain, A. Rani, and J. Yadav, "Classification of human emotions from EEG signals using SVM and LDA classifiers," in *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2015, pp. 180–185.

[27] S. R. Sain, *The Nature of Statistical Learning Theory*. Taylor & Francis, 1996.

[28] D. Bertsimas and A. King, "Logistic regression: From art to science," *Statistical Science*, pp. 367–384, 2017.

[29] A. M. Mubarek and E. Adalı, "Multilayer perceptron neural network technique for fraud detection," in *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2017, pp. 383–387.

[30] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, "KNN based machine learning approach for text and document mining," *International Journal of Database Theory and Application*, vol. 7, no. 1, pp. 61–70, 2014.

[31] S. Munawwar and P. V. Gopi Krishna Rao, "An efficient deep learning based multi-level feature extraction network for multi-modal medical image fusion," *International Arab Journal of Information Technology (IAJIT)*, vol. 22, no. 3, 2025.

[32] B. Kégl, "Introduction to adaboost," Technical report, 2009.

[33] B. Tatiparti and V. Dharmar, "GNN-DQL-based chronic kidney disease classification using GFR in the internet of medical things environment," *International Arab Journal of Information Technology (IAJIT)*, vol. 22, no. 3, 2025.

[34] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, pp. 1063–1095, 2012.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[37] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[39] R. A. Salma, H. Kafajeh, R. Alazaidah, M. Assasfeh, A. Alsheriadeh, and N. Alshdaifat, "Leveraging machine learning for effective breast cancer diagnosis," *WSEAS Transactions on Computer Research*, vol. 13, pp. 34–46, 2025.

[40] M. Alzyoud *et al.*, "Diagnosing diabetes mellitus using machine learning techniques," *International Journal of Data & Network Science*, vol. 8, no. 1, 2024.

[41] M. Rudra Kumar, R. Pathak, and V. K. Gunjan, "Diagnosis and medicine prediction for COVID-19 using machine learning approach," in *Computational Intelligence in Machine Learning: Select Proceedings of ICCIML 2021*. Springer, 2022, pp. 123–133.

[42] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, "A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework," *Sensors*, vol. 21, no. 3, p. 748, 2021.

[43] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature Communications*, vol. 11, no. 1, p. 3923, 2020.

[44] R. R. Ali *et al.*, "Learning architecture for brain tumor classification based on deep convolutional neural network: Classic and resnet50," *Diagnostics*, vol. 15, no. 5, p. 624, 2025.

[45] N. Şengöz, T. Yiğit, Ö. Özmen, and A. H. Isık, "Importance of preprocessing in histopathology image classification using deep convolutional neural network," *Advances in Artificial Intelligence Research*, vol. 2, no. 1, pp. 1–6, 2022.

[46] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016, pp. 241–246.

[47] H. A. Vu, "Integrating preprocessing methods and convolutional neural networks for effective tumor detection in medical imaging," *arXiv preprint arXiv:2402.16221*, 2024.

[48] S. Cui, Y. L. Su, K. Duan, and Y. Liu, "Maize leaf disease classification using CBAM and lightweight autoencoder network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 7297–7307, 2023.

[49] D. Müller, I. Soto-Rey, and F. Kramer, "An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks," *IEEE Access*, vol. 10, pp. 66 467–66 480, 2022.