Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

# Dalikmata-Ultima: Revolutionizing Medical AI Diagnosis for Skin Cancer and Pneumonia with Hierarchical Classification, Model Hybridization and OOD Evaluation

Ricafrente, John Patrick M.
Pasig City Science High School
Pasig, Philippines

Taño, Sean Joshua P.
Pasig City Science High School
Pasig, Philippines

Samson, Kean Josiah S.
Pasig City Science High School
Pasig, Philippines

Ramos, Chanelle Joey S.
Pasig City Science High School
Pasig, Philippines

Cano, Gabrielle Simeone B.
Pasig City Science High School
Pasig, Philippines

***Abstract* - Global healthcare systems face critical workforce and diagnostic gaps, necessitating scalable AI solutions. This article presents Dalikmata-Ultima, a framework employing two distinct methodologies for dermatological and radiological diagnosis grounded in the Universal Approximation Theorem and VC bounds for structural reliability. For dermatological diagnosis, the system utilizes a hybrid CNN-MLP architecture to integrate EfficientNetV2 image features with patient metadata. Trained and evaluated using the ISIC 2019 dataset, the framework employs the Forced Dominance Transform (FDT) to synthesize these inputs while maintaining a clinical hierarchy that separates primary malignancy screening from subtype classification. For pneumonia detection, a vision-only CNN pipeline was implemented using the Mendeley Chest X-ray dataset. Due to the absence of structured metadata, hybridization was omitted to maintain a data-dependent design. This model underwent out-of-distribution (OOD) validation using clinical data from Pasig City Children's Hospital (PCCH), addressing a critical gap in Philippine medical AI research and proving the feasibility of deploying contextually relevant diagnostic support in real-world settings.**

***Keywords* - *EfficientNetV2; Hierarchical classification; Forced Dominance Transform (FDT); CNN MLP hybrid architecture; Medical artificial intelligence***

## I. INTRODUCTION

### *A. AI in Philippine Healthcare Systems*

Access to healthcare, which is a fundamental human need, is currently in crisis in the Philippines. The country is hindered by various health workforce challenges, despite the Philippines being one of the world's leading sources of medical practitioners [1]. One such challenge is a 45% local government budget cap for personnel, including medical workers, which limits healthcare capacity nationwide. Additionally, specialists and training hospitals are heavily concentrated in the National Capital Region (NCR), leaving other regions underserved. Many healthcare providers also raised concerns regarding low salaries, poor working conditions, and unprepared graduates. Evidence of budget mismanagement and flawed vaccine rollout during the COVID-19 era further highlights these systemic failures and the urgent necessity for a national healthcare overhaul [2].

These systemic and financial hurdles inevitably impact patient care outcomes, particularly regarding medical diagnosis. Accuracy is crucial to ensure the correct treatment for patients and is essential for patient safety, as failure to do so may lead to avoidable harm, increased healthcare costs, and morbidity. Research shows that approximately 20% of patients may experience diagnostic errors in emergency department settings [3]. Despite such importance, financially disadvantaged households in the Philippines avoided their need for medical diagnoses for other basic necessities amid rising costs. This puts them at risk for untreated morbidities [4].

One innovation and cost-effective approach in addressing these issues is the utilization of Artificial Intelligence (AI) for medical diagnosis. The field of AI research in healthcare institutions of the Philippines is dynamic, yet exploratory and in the development stage, having more emphasis on applied solutions rather than large-scale and rigorously benchmarked systems [5]. The Japan International Cooperation Agency tested their AI-powered tuberculosis detection technology from Japanese company Fujifilm in the Philippines under their

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

cooperation with the Department of Health (DOH) in Muntinlupa City, with the diagnosis of 382 patients as tuberculosis suspects, although only 32 were validated by a radiologist to have actual tuberculosis [6]. Results indicate low precision leading to a large quantity of false positives while false negatives were not considered in the evaluation, which can be fatal for those who actually had tuberculosis. While innovative, the AI requires further refinement before full integration. In addition to the AI projects of the DOH, the Department of Science and Technology (DOST) supports several local AI projects designed for healthcare. These include: Intelligent Stroke Utilization, Learning, Assessment and Testing (i-SULAT), a diagnostic AI tool for neurological disorders; and Immersive Gamification Technology Systems (ImGTS), an AI-based system for rehabilitation in cerebral palsy and Alzheimer's disease [7]. Universities, government agencies, and small-scale collaborations between the private sector dominate the research activities, which leads to the situation when innovation is there, yet not methodologically standardized and clinically validated in the long term.

When such representations are misaligned with clinical reasoning, particularly in diagnostic settings where certain decisions logically precede or dominate others, high-confidence false positives and underestimated false negatives can pose significant clinical risk. Beyond model accuracy, such outcomes reveal a deeper challenge in how AI systems represent diagnostic confidence. Issues of transparency, interpretability, and explainability are still relevant, undermining clinical decision-making by providing misleading confidence in AI output [8]. Such perceptions indicate an optimistic pessimism that is more inclined towards integrating under the watch of physicians than complete automation [9]. Limited training exposure, formal AI skills, infrastructural barriers and risk perception among healthcare workers are some of the constraints that could impede the fast translation of prototypes to practice in a rapid research [10]. NCR hospitals and universities that are more equipped to initiate AI projects remain equally susceptible to the issues of data access, standardized assessment, and clinician cooperation. This highlights the need to examine not only how medical predictions are generated, but also how their confidence is mathematically structured and interpreted.

Traditionally, medical practitioners regard AI as an assistive tool rather than a diagnostic agent, especially in the area of diagnostic work that requires pattern recognition and supports image-intensive specialties such as radiology and dermatology [11]. The potential of AI to enhance efficiency, cut the workload of the diagnostics, and help with the early detection of a disease is widely recognized. However, utilization of these tools in diagnosis comes under the terms of transparency, reliability, and clinician control. There is still some debate over the excessive use of automated systems, the risk of errors in diagnosis, and the question of ethical responsibility, particularly in life or death medical choices.

These views suggest a human-in-the-loop diagnostic paradigm, in which AI is used as an assistive system that is part of clinical workflows. This places AI as an augmentation tool which improves physician decision-making by providing a hierarchical output and confidence-sensitive evaluation [11]. The use of the hierarchical classification and OOD-conscious assessment approaches of a system leads to the provision of responses to the typical aspects of trust, reliability, and clinical control in the mind of clinicians. Such views are particularly applicable to the healthcare systems such as the Philippines where acceptance of the physicians is critical to the successful adoption of AI [9].

### *B. Hierarchal Classification*

One of the clinical areas central to this study is dermatology, particularly the detection of three major skin cancers: melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC). In the Philippines, BCC and SCC are the most common malignant skin tumors, accounting for 39.13% and 34.78% of cases respectively according to a five-year Philippine General Hospital (PGH) histopathology review [12], while melanoma represents only 4.55% of cases but has a higher fatality rate due to aggressive metastasis and late detection. Furthermore, over 2,700 cases of BCC and SCC were documented from 2011 to 2021, a figure likely underestimated due to the absence of a national cancer registry and the fact that there are only 1,275 board-certified dermatologists for 109 million people [13]. Although melanoma is rare among Filipinos, its exploration is vital as the most lethal form of skin cancer; it spreads aggressively to distant sites like the lungs and brain, leading to a survival rate of only 32% once metastatic. Since early detection yields a high survival rate of 94%, screening for even rare cases is essential to reduce the significant mortality associated with this rapidly metastasizing cancer [14].

In many medical domains, diagnostic outcomes are not only structured but also highly imbalanced, further complicating probability interpretation. Conditions such as melanoma are relatively rare compared to benign skin lesions, yet they carry significantly higher clinical risk. As a result, diagnostic workflows are often hierarchical in nature, where an initial decision such as determining malignancy precedes the identification of a specific subtype. However, standard classification models typically output flat probability distributions over all classes, treating each outcome as independent and equally comparable. A hierarchical loss function tailored for multi-class skin lesion classification, acknowledging that misclassification among major lesion types constitutes a significant real-world problem. It illustrates current hierarchical classifier development that goes beyond flat classification by embedding hierarchy into the learning objective [15]. A deep architecture (HFFNet-2d) for hierarchical multi-label classification, achieving strong performance on level-specific metrics (e.g., hierarchical precision and hierarchical recall) showed that incorporating hierarchical structure into network design enables more accurate identification, exemplifying how hierarchical models are optimized through specialized architectures [16].

Building on the importance of imaging for early and accurate diagnosis in dermatology, radiology provides another critical avenue for AI-assisted evaluation particularly in the detection of pneumonia. It remains a leading cause of death across all ages in the Philippines and is the most common cause of death among children under five, highlighting its critical public health impact. Its substantial economic burden, especially for severe cases, combined with risk factors such as undernutrition, comorbidities, and limited access to healthcare, underscores the need for timely and accurate diagnosis [17].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

### C. Framework of Medical AI

The development and scrutiny of these diagnostic tools, highlight the vital role of specific machine learning architectures in determining accuracy and reliability. Central to many medical AI applications, particularly those involving image and signal processing like X-rays, computed tomography (CT) scans, and handwriting analysis, are Convolutional Neural Networks (CNNs). CNNs are multi-layered neural networks that can identify and classify objects within an image. Such networks are frequently used in various studies around medical applications as it is specifically designed to deal with a range of 2D images. These are more effective than regular networks as CNNs have the ability to automatically recognize different elements such as lines, shape and even more intricate features without the need for human intervention [18]. However unlike medical practitioners, CNNs do not make use of patient symptoms and medical history when diagnosing as the model architecture is not suited for analyzing nonimaging data [19]. CNNs alone are not sufficient for handling non-image metadata as it is designed for extracting features from image data [20]. Meanwhile, other clinical variables such as patient and demographic information require integration with CNN outputs in a multimodal framework. This unimodal approach results in an incomplete or fragmented view of patient health, requiring a fundamental shift toward multimodal architectures to achieve true clinical utility.

To effectively integrate the structured and tabular clinical data describing patient symptoms and history, a distinct neural network architecture is necessary. A known architecture for analyzing structured data is the Multilayer Perceptron (MLP). Unlike CNNs, MLPs use flattened vectors which discard spatial structure and fully connect all input features to allow the analyses of their complex and non-linear interactions. While this is unreliable for image data, it becomes more efficient when analyzing pure tabular metadata. The performance of MLPs for heart failure classification resulted in an accuracy of 93.9% within the 299-patient dataset [21]. This showcases the potential of MLPs in the field of medical AI. Careful selection of an appropriate activation function is a key factor in the effective performance of MLP architecture, being considered as a design strategy that can significantly influence overall performance, learning behavior, and classification accuracy [22]. Having control over the optimization process in generalization performance and architectural configurations is important in balancing model capacity particularly for medical AI that works with structured or low-dimensional clinical features to improve predictive reliability [23].

Through the combination of these neural networks into a hybrid CNN-MLP dual-branch architecture, diagnostic accuracy can be improved with the combination or the extraction of deep features from medical image analysis and utilization of MLP as a final classifier. The integration of multiple data modalities within a learning framework enhances overall model effectiveness [24]. Programmers integrate these into a single model through transforming the complex patterns learned using CNN into structured inputs suitable for the MLP, improving accuracy in diagnosing medical images [25].

### D. AI Utilization in Skin Lesion Classification and Pneumonia Detection

This study aims to develop a hierarchical hybrid CNN-MLP model to diagnose three major skin cancers: melanoma, basal cell carcinoma, and squamous cell carcinoma, using skin lesion images and patient metadata. Performance is evaluated using tenfold cross-validation balanced accuracy to select the best median performance across CNN-only and CNN-MLP configurations of EfficientNetV2-RW-T and EfficientNetV2-RW-S architectures. The study also examines the diagnostic stability of a CNN-only radiology model for pneumonia through higher median F1, which is then validated using an out-of-distribution dataset provided by a local hospital in Pasig City.

By providing an affordable diagnostic option that maintains high accuracy, the hierarchical CNN-MLP model supports greater healthcare accessibility for low-income individuals who may struggle to afford regular check-ups. It equips medical professionals with an AI-assisted tool to improve diagnostic efficiency and enhance patient care quality. The findings can inform policymakers in strengthening healthcare systems and expanding equitable access to diagnostic technologies. Beyond practical applications, the research contributes to the growing body of AI-driven medical studies, offering a foundation for future innovation and development in the field.

## II. NEURAL NETWORK THEORY

### A. Introducing Neural Networks

Neural networks are mathematical models which have adaptive parameters. Such functionality is utilized in optimization problems aimed at minimizing error. Thus, this makes neural networks function as universal approximators. This concept was inspired by biological neural networks where the brain learns to adapt to situations. However, unlike real brains, mathematical neural networks are not exact replicas. This study utilized neural networks to model a prediction system where an input of limited medical information interpretable by numbers may predict the presence of a disease within a patient. Here is the main backbone in modeling a neural network:

*Theorem 1. Universal Approximation Theorem*

Let $C(X)$ be the set of continuous functions on a compact subset $X \subset \mathbb{R}^n$. Let $\sigma$ be a non-constant, bounded, and continuous activation function. For any function $f \in C(X)$ and any $\epsilon > 0$, there exists a single-hidden-layer neural network $F(x)$ such that:

$$\max_{x \in X} |F(x) - f(x)| < \epsilon$$

This implies that a finite neural network can approximate any continuous mapping to arbitrary precision, provided it has a sufficient number of hidden units [26]. The aforementioned concept is useful for making medical AI models as even limited medical information is enough to make meaningful medical predictions. In the modern era of modeling neural networks, particularly high-end AIs, there are various model architectures for various purposes. Here justifies the exploration of certain architectures for this study:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

*Theorem 2. No Free Lunch Theorem*

*There is no machine learning algorithm that is universally superior for every task.*

Across all possible problem distributions, the average performance of any two algorithms is equivalent [27]. This principle implies a fundamental conservation law in mathematical optimization where any gain in performance on one class of problems is strictly offset by a loss on another. In the context of a medical prediction system, this means that a neural network does not possess an inherent or universal superiority over simpler models like linear regression. Instead, its effectiveness is entirely dependent on how well its architectural assumptions align with the specific statistical structure of the medical data provided.

The theorem suggests that for an algorithm to succeed, it must incorporate prior knowledge about the problem domain to narrow the search space. Because the set of all possible mathematical functions is infinite and mostly consists of random noise, an algorithm that performs well on every imaginable task is a mathematical impossibility. Therefore, the choice to use a neural network in this study is a strategic bet that the relationship between medical inputs and disease presence follows a non-linear and hierarchical pattern which the network is specifically designed to capture. The following showcases the limitation of neural networks:

*Theorem 3. Vapnik-Chervonenkis (VC) Generalization Bound*

For a binary classification model with a VC dimension $d$ (a measure of model complexity), trained on $N$ samples, the generalization error $R(f)$ is bounded by the empirical training error $r_N(f)$ with a probability of at least $1 - \delta$:

$$R(f) \leq r_N(f) + \sqrt{\frac{d\left(\ln \frac{2N}{d} + 1\right) + \ln \frac{4}{\delta}}{N}}$$

In practical terms, this bound proves that for a prediction system to be medically reliable, the architect must minimize the structural risk by ensuring the network is not overly complex for the size of the available clinical dataset. This mathematical relationship justifies the use of regularization techniques to keep the VC dimension $d$ under control, ensuring that the patterns learned from limited medical information remain valid when applied to the general population [28].

### *B. Convolutional Neural Networks*

Convolutional Neural Networks (CNNs) represent a foundational deep learning architecture that has revolutionized the capacity of machines to learn from data, particularly in the domain of computer vision [29] Unlike traditional neural networks that may struggle with the high dimensionality of raw input, CNNs leverage specialized convolutional layers to automatically and adaptively learn spatial hierarchies of features. By applying a series of learnable filters or kernels, the network systematically captures patterns ranging from low-level edges and textures to complex high-level shapes. This architectural design, which typically integrates convolutional, pooling, and fully connected layers, allows for the robust extraction of significant features without the need for manual human supervision or extensive preprocessing.

The CNN model used in this study was from the EfficientNetV2 family. This improves upon its predecessor by utilizing Fused-MBConv layers and a revised scaling strategy that optimizes training speed and parameter efficiency. Unlike standard CNNs, EfficientNetV2 employs progressive learning, where the image size and regularization strength are adjusted dynamically during training. This approach is particularly beneficial for medical datasets with limited samples, as it enhances the model's ability to generalize without the massive computational overhead typically required by deep architectures [30]. Complementing this efficiency, the study incorporates the RW (Ross Wightman) variant, which demonstrates that refined training procedures can elevate classic architectures to state-of-the-art performance levels. This model ensures high robustness and stable convergence, allowing the predictive system to maintain reliability when navigating the noisy, high-dimensional boundaries inherent in patient health data [31]. Here are the two CNN models chosen for this study:

| Model | Training Image Size | Testing Image Size |
|---|---|---|
| EfficientNetV2-RWT | 224x224 | 288x288 |
| EfficientNetV2-RWS | 288x288 | 384x384 |

*Table 1. Selected CNN Models*

While CNNs are optimized for spatial feature extraction, their utility is fundamentally limited when applied to the discrete and heterogeneous nature of patient metadata. This constraint presents that for clinical variables such as demographics, laboratory results, and BMI, standard convolutional architectures struggle to model the non-spatial relationships found in structured electronic health records [32]. Because CNNs rely on spatial proximity and translation invariance, the convolutional kernel fails to distinguish between independent clinical markers, often treating discrete medical values as localized "textures" rather than individual diagnostic predictors. Because of these limitations, there is a need for another neural network architecture to support patient metadata.

### *C. Multilayer Perceptrons*

The Multilayer Perceptron (MLP) offers a distinct mathematical advantage over convolutional architectures when processing structured medical records. This model possesses a rotationally invariant inductive bias that treats each clinical feature as an independent coordinate in a high-dimensional space [33]. Unlike CNNs, which assume neighboring features are physically related, the MLP utilizes a globally connected architecture to weight variables like age and BMI against the entire feature set simultaneously. This design allows the model to map non-linear diagnostic risks without introducing the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

informational noise caused by the spatial assumptions of a CNN.

In this study, a custom MLP was made for the dermatological model. The MLP takes into account the patient's age, gender and skin lesion location to support the classification of the lesion. Here is the activation function used for the MLP in accordance with Theorem 1:

*Definition 1. Rectified Linear Unit (ReLU)*

$$ReLU(x) = \frac{x + |x|}{2}$$

The integration of ReLU in the MLP hidden layers is justified by its ability to resolve the vanishing gradient problem, which frequently stalls the training of models using discrete clinical variables. This activation function is the optimal choice for multi-layer architectures because its piecewise linear nature prevents gradient saturation, a common issue when using sigmoidal functions on demographic data. By outputting a direct linear mapping for positive values and zero for negative values, ReLU encourages sparse activation while ensuring the model remains computationally efficient. Here is our full MLP architecture:

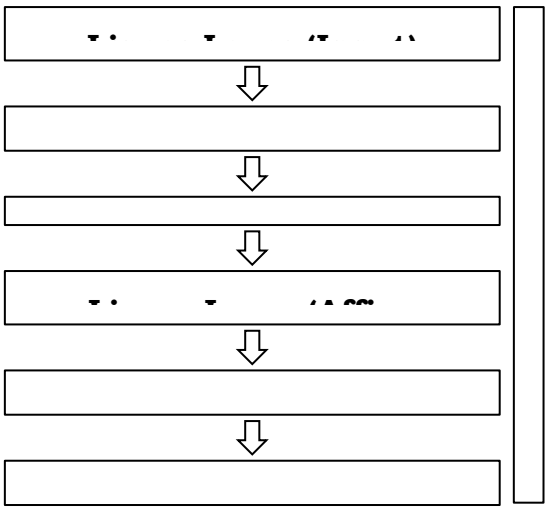

*Figure 1. MLP Architecture for Dermatological Classification*

### D. Hybridization of Architectures

For the dermatological model in this study, a hybrid CNN-MLP architecture was implemented to process multi-source heterogeneous data. This dual-stream approach is justified by its ability to simultaneously analyze visual features and textual metadata, effectively overcoming the bottlenecks associated with single-source models. By feeding image data into a CNN and clinical metadata into a MLP, the fusion model captures high-dimensional correlations across different data types [24]. This integration ensures the diagnostic prediction benefits from both the spatial patterns of the lesion and discrete patient variables, leading to higher predictive accuracy and more robust performance. Given the two architectures they are hybridized as:

*Definition 2. Feature Concatenation*

The fusion of image features and patient metadata is mathematically represented as the concatenation of two distinct feature vectors. Let $x_{\text{img}} \in \mathbb{R}^d$ be the feature vector extracted by the EfficientNetV2 backbone and $x_{\text{meta}} \in \mathbb{R}^k$ be the feature vector produced by the MLP. The fused representation $z$ is defined as the concatenation of these vectors:

$$\mathbf{z} = \mathbf{x}_{img} \oplus \mathbf{x}_{meta} = [\mathbf{x}_1, \ldots, \mathbf{x}_d, \mathbf{x}_{d+1}, \ldots, \mathbf{x}_{d+k}]^T$$

The final classification layer then learns a mapping $f: \mathbb{R}^{d+k} \rightarrow \mathbb{R}^C$ where $C$ is the number of diagnostic classes. This operation ensures that the high-dimensional spatial information from the CNN is preserved alongside the discrete clinical variables from the MLP, allowing the final linear layer to find an optimal separating hyperplane in the joint feature space.

## III. CLASSIFIER MODELS

### A. Solo Classifiers and Ensemble Classifiers

To utilize a neural network architecture for practical purposes, it must be trained with training data for a set of epochs until optimal performance is reached. For solo classifiers, it is simply using a single architecture and doing classification on that architecture alone. In fact, the radiological model for Pneumonia is simply a solo classifier of EfficientNetV2. However, ensemble classifiers use multiple trained instances of architecture/s. While ensemble classifiers often provide marginal performance improvements, the increased complexity of such architectures does not inherently guarantee superior results. Although ensemble methods like bagging and boosting generally enhance performance over individual base learners, advanced structures such as stacking can introduce instability and offer little additional gain despite their higher computational cost [34]. This suggests that the added layers of complexity may not always justify the trade-off in model interpretability and stability. Furthermore, a single, well-optimized model can actually outperform an ensemble approach, finding that a Gradient Boosting model achieved higher accuracy (88.8%) than a majority voting ensemble (85.6%) [35]. Together, these studies reaffirm Theorem 2, where no single model classifier dominates every field and choosing the right one is dependent and situational.

### B. Hierarchical Classifiers

In practice, dermatological diagnosis comes in stages. A skin lesion may be benign or malignant and from there, malignancy has various lesion types such as melanoma, basal cell carcinoma and squamous cell carcinoma. However, machine learning research currently does not reflect this hierarchy as classification there is done by evaluation of every lesion type at the same time instead of considering major categories. This study proposes a new type of classifier named the hierarchical classifier where such staging is considered during classification.

The hierarchical classifier works by using two model classifiers. The first classifier evaluates a sample based on two

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

top categories (benign and malignant) and then the second classifier evaluates the same sample on subcategories (cancer types under malignancy). From there, the two classifiers will produce probability distributions and then the two are combined to produce an overall probability distribution. However, there is limited prior work that explicitly combines hierarchical classification, multimodal fusion, and consistency enforcement in this manner. Which is why this study proposes the Forced Dominance Transform for the hierarchical classifier:

*Definition 3. Forced Dominance Transform*

Let $P_b$ and $P_m$ be be the probabilities for benign and malignant from the first classifier, and let $P_B \in \mathbb{R}^K$ be the vector of subtype probabilities from the second classifier. The final probability distribution is obtained by:

$$\delta = P_m - P_b$$

$$\gamma = \begin{cases} \frac{\delta^2}{\delta+\epsilon}, & \text{if } P_m > P_b \\ P_m, & \text{otherwise} \end{cases}$$

$$D = \begin{cases} P_b + \gamma, & \text{if } P_m > P_b \\ 1, & \text{otherwise} \end{cases}$$

$$P_b' = \frac{P_b}{D}, \quad P_m' = \frac{\gamma \cdot P_B}{D}$$

$$P_{final} = [P_b', P_m']$$

This transform serves as a decision-weighting mechanism that prioritizes the malignancy branch by applying a quadratic penalty to the probability gap whenever the malignant score exceeds the benign score. It dynamically redistributes the primary triage confidence across the specific subtype probabilities generated by the secondary classifier, ensuring the sub-classification is weighted by the initial triage strength. Finally, the transform utilizes a normalization factor to preserve a valid probability distribution while mathematically enforcing the hierarchical staging of the dermatological diagnosis.

Furthermore, the VC bound from Theorem 3 suggests that fragmenting the classification process lowers error rates. Instead of using a solo classifier to evaluate every lesion type, two classifiers operating on simpler tasks may achieve lower error rates in practice. This advantage is then preserved during synthesis through FDT, as motivated by the following formulation:

*Theorem 4. Margin-Based VC Bound Theorem*

Let $S$ be a sample set from an input space with radius $r$ such that $||x_i||_2 \leq r$ for all $x_i \in S$. For a hypothesis class $H_{S,\Lambda}$ of linear classifiers with a functional margin determined by the weight norm $||w||_2 \leq \Lambda$, the VC-dimension $d$ is bounded by:

$$d \leq r^2 \Lambda^2$$

This theorem provides the mathematical foundation for why the high accuracy of fragmented sub-classifiers is not lost during synthesis. In a solo classifier, the model must encompass a large radius r of disparate features (e.g., both benign textures and malignant vascularity) within a single hypothesis space, which inflates the VC-dimension and increases the risk of misclassification. By contrast, fragmenting the process allows each sub-classifier to operate within a smaller, task-specific radius $r_{sub}$ [36].

When these models are combined via the Forced Dominance Transform, the quadratic penalty on the decision gap maximizes the margin $\gamma$ (where $\gamma = 1/\Lambda$). Since the VC-dimension d is directly proportional to the square of the inverse margin $\Lambda$, the FDT may help maintain a lower effective complexity for the integrated system. This suggests that the synthesis stage does not significantly increase structural risk, but instead helps retain the performance advantages of the individual sub-classifiers while remaining consistent with generalization bound intuition.

## IV. METHODOLOGY

### A. Skin Lesion Dataset

The skin lesion dataset used in this study is the International Skin Imaging Collaboration (ISIC) 2019 Challenge Dataset. These image samples with patient metadata comprise a total of 25,331 dermoscopic images categorized into eight distinct diagnostic classes: Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), and Squamous Cell Carcinoma (SCC). The dataset is characterized by a significant class imbalance, with Melanocytic Nevus representing the majority class (12,875 images), while rare malignant types like SCC and DF contain fewer than 700 samples each. Beyond the visual data, each entry is paired with clinical metadata including the patient's age, sex, and the anatomical site of the lesion, which provides the high-dimensional heterogeneous markers necessary for the hybrid CNN-MLP architecture utilized in this research.

This dataset is medically validated by a local dermatologist in Pasig City with some suggested changes for practicality in the clinical field. While Melanoma, BCC and SCC are retained as unique critical classes under malignancy, every other class is synthesized under the category “Benign Lesions”. This is the main idea that inspired the researchers to formulate the FDT in the first place. Furthermore, images with missing metadata were omitted to ensure complete testing. This resulted in 22,426 as the actual amount of skin lesion samples used in this study.

### B. Pneumonia In-Distribution Dataset

The Pneumonia Dataset used in this study is the Chest X-ray Dataset (Normal and Pneumonia Cases), a comprehensive collection of anterior-posterior chest X-ray images curated and hosted on the Mendeley Data repository. This dataset contains a total of 5,856 grayscale images with anteroposterior (AP) view obtained from pediatric patients which were collected in Guangzhou, China. The samples are categorized into two primary diagnostic classes: Normal and Pneumonia. Specifically, the dataset is composed of 1,583 normal cases and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

4,273 pneumonia cases, with the latter including both bacterial and viral manifestations of the disease [37].

This significant sample size provides the high-dimensional visual features required to train the EfficientNetV2 solo classifier used in this study, ensuring the model can effectively distinguish between healthy pulmonary tissue and the increased opacities characteristic of pneumonia.

Medical validation was done for this dataset by a radiologist from Pasig City Children's Hospital. It was the only x-ray dataset approved by the radiologist after the researchers presented various well-known x-ray datasets such as ChestX-Ray14 and SIIM-Pneumothorax which had apparent inconsistencies and errors.

C. Pneumonia Out-of-Distribution Dataset

To truly create an effective model for clinical trials, the model must be evaluated with a dataset geographically and demographically independent from the original dataset. This is why the researchers coordinated with Pasig City Children's Hospital to collect patient chest x-ray images to locally evaluate the radiology model. This was done in accordance with the Data Privacy Act of 2012 where images are deidentified and the researchers were monitored and guided by hospital staff to ensure accuracy of data collection.

The researchers gathered 800 x-ray images however the dataset was heavily omitted due to uncertainty of diagnosis by radiologists. To truly create a viable dataset for machine learning, the labels must be confidently verified. As a result, for AP views: 60 images were collected for Pneumonia-Positive and 34 images were collected for Pneumonia-Negative.

D. Training Process

The primary goal of the training process is Empirical Risk Minimization (ERM), where the model iteratively adjusts its internal weights to minimize the discrepancy between its predictions and the ground-truth clinical labels. By optimizing a loss function across the training dataset, the model learns to extract high-dimensional diagnostic features that generalize to unseen patient data. In this study, the training objective is specifically designed to handle the severe class imbalance inherent within datasets, ensuring that rare but critical cases are not overshadowed by prevalent classes. Here is the primary training metric:

*Definition 4. Focal Loss*

To address extreme data imbalance, the loss function *FL* for a predicted probability $p_t$ is defined as:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

This loss function extends Cross-Entropy by incorporating a modulating factor to down-weight the loss contribution from easy, majority-class samples. By adjusting the focusing parameter, the model is forced to concentrate on hard examples where the predicted probability is low. The alpha parameter further balances the influence of classes based on their frequency, ensuring the resulting probability distributions are sensitive to subtle unique features of certain classes [38].

*E. Regularization*

In medical machine learning, overfitting remains a primary obstacle to clinical implementation. This occurs when a model, often due to high parametric complexity or limited dataset diversity, learns the noise or specific artifacts of the training images (such as surgical markings or lighting conditions) rather than generalizable diagnostic features [39]. Recent studies have shown that while models pretrained on natural images (like ImageNet) converge quickly, they are significantly more susceptible to overfitting on non-clinically relevant features in dermatological and radiological tasks [40]. Here are the various techniques to conduct regularization:

**Definition 5.** Early Stopping

Early stopping is a form of regularization that halts the training process once performance on a validation set begins to decline, despite continued improvement on the training set. If *E(θ, t)* represents the error at epoch *t*, training is terminated at $t_{opt}$ such that:

$$t_{opt} = \arg\min_t E_{val}(\theta, t)$$

This prevents the model from entering the overfitting regime where it begins to memorize the training data rather than learning generalizable features.

*Definition 6. Dropout*

Dropout is a stochastic regularization technique where, during each training iteration, each neuron $h_i$ in a layer is set to zero with a probability *p*. The modified activation $h'_i$ is given by:

$$h'_i = r_i \cdot h_i, \quad r_i \sim \text{Bernoulli}(1-p)$$

By preventing neurons from co-adapting, Dropout forces the network to learn robust, redundant representations of medical data, ensuring the model does not rely on specific noisy pixels.

*Definition 7. $L_2$ Regularization (Weight Decay)*

$L_2$ regularization constrains model complexity by adding a penalty term to the original loss function *J(θ)* based on the Euclidean norm of the weights. The regularized objective function $J_2$ is defined as:

$$J_2(\theta) = J(\theta) + \frac{\lambda}{2}\|\theta\|_2^2$$

Where λ is the weight decay coefficient. This encourages smaller weight values, effectively smoothing the decision boundary and preventing the model from fitting high-frequency noise in the dataset.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

*Definition 8. Data Augmentation*

Data augmentation is the process of expanding the training space by applying a set of transformation functions $T$ that preserve the semantic label $y$. For an input image $x$, the augmented sample $x'$ is generated as:

$$x' = T(x), \quad T \in \{\text{Rotation, Flip, Color Jitter}\}$$

This ensures the model develops translation and rotation invariance, which is vital for medical imaging where the orientation of a lesion or the lighting of an x-ray can vary across different clinics.

### F. Learning Rate Parameters

The learning rate is a critical hyperparameter that dictates the step size the optimizer takes during each iteration toward the global minimum of the loss function. If the learning rate is too high, the optimization may oscillate or even diverge; conversely, if it is too low, the training process becomes computationally inefficient and may get trapped in suboptimal local minima. To navigate this optimization surface effectively, this study utilizes the AdamW optimizer in conjunction with a ReduceLROnPlateau scheduler.

*Definition 9. AdamW Optimizer*

AdamW is an extension of the Adam (Adaptive Moment Estimation) optimizer that decouples the weight decay penalty from the gradient update. In standard Adam, $L_2$ regularization is often less effective because the moving averages of the gradients interfere with the weight decay. AdamW corrects this by applying the weight decay directly to the weights:

$$\theta_{t+1} = \theta_t - \eta \left( \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right)$$

This ensures that the regularization effectively constrains the model's complexity, satisfying the structural risk requirements established in Theorem 3 while maintaining the adaptive benefits of the Adam algorithm.

*Definition 10. ReduceLROnPlateau Scheduler*

To further refine the convergence process, a ReduceLROnPlateau scheduling strategy is implemented. This dynamic scheduler monitors the validation metric and reduces the learning rate by a factor when the metric ceases to improve for a predetermined number of epochs.

The logic behind this approach is to use a relatively large learning rate during the initial phase of training to quickly move toward the general vicinity of the minimum, and then cool down the learning rate to allow the model to settle into the finest possible local minima without overshooting. This adaptive behavior is particularly useful for medical datasets where the loss surface can be highly non-convex due to class imbalance and heterogeneous data features.

### G. Hardware and Software Usage

The computational demands of training deep convolutional neural networks, particularly the EfficientNetV2 architectures used in this study, necessitate the use of high-performance hardware and specialized software acceleration. Training was conducted using NVIDIA Graphics Processing Units (GPUs), which are uniquely suited for the parallel matrix multiplications required by deep learning.

Local training and model prototyping were performed on an HP VICTUS laptop equipped with an NVIDIA GeForce RTX 4050 GPU. This local environment allowed for immediate iterative testing of the training scripts and preprocessing pipelines. For more intensive training sessions, the study utilized the Kaggle cloud platform, which provides access to the NVIDIA Tesla P100 GPU. The P100 architecture is specifically designed for data center workloads, offering high memory bandwidth that is essential for processing the large-scale image batches required by the EfficientNetV2 architecture.

The implementation of the classification models was conducted using Python, leveraging its comprehensive ecosystem of data science and machine learning libraries. The core of the development was built upon the PyTorch framework, which provides the essential autograd engines for calculating the gradients of the Focal Loss function from Definition 5 To access the specific EfficientNetV2 architectures, the timm (PyTorch Image Models) library was utilized. This library served as the source for the RWT and RWS variants, providing the highly optimized, pretrained weights necessary for the transfer learning phase.

Data management and evaluation were supported by several auxiliary packages. NumPy and Pandas were used for metadata manipulation and dataset structuring, while Scikit-learn provided the utility for stratified k-fold partitioning and metric calculation. Finally, Matplotlib and Seaborn were employed to generate the training curves and diagnostic visualizations presented in the following chapters.

## V. MODEL SELECTION AND EVALUATION

### A. Cross Validation

To ensure the diagnostic reliability of the medical models in this study, k-fold cross-validation is implemented. Unlike a static three-way split, k-fold cross-validation provides a more robust estimate of the model's generalization performance by utilizing the entire dataset for both training and validation. This is particularly critical in medical imaging, where the limited diversity of rare malignant samples means a single random split could accidentally exclude key features from the training phase. By rotating the validation set through $k$ different folds, the study minimizes the variance in performance metrics and ensures that the reported accuracy is not a result of an advantageous data split [41].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

*Definition 11. K-Fold Cross-Validation*

The dataset is partitioned into *k* mutually exclusive subsets of approximately equal size. In each of the *k* iterations, one subset is reserved for validation while the remaining subsets are used for training. To determine the final model performance, the median value of a chosen metric is calculated across all *k* folds. Utilizing the median ensures that the final evaluation reflects the typical performance of the model and remains resilient against outlier results from any single, unrepresentative data split.

In this study, tenfold cross-validation was done to conduct model selection for the skin lesion classification and pneumonia detection. Stratified sampling was done for each fold to preserve class proportionalities.

### *B. Skin Lesion Model Selection and Evaluation*

The skin lesion classification system in this study was chosen from four models. Namely the four combinations from the two EfficientNetV2 variants and the presence of the MLP. Model selection was done for the top classifier (malignancy detection) and sub classifier (cancer type). The primary metric for the tenfold cross-validation is balanced accuracy:

*Definition 12. Balanced Accuracy*

Balanced Accuracy is a performance metric designed to evaluate models on imbalanced datasets by calculating the arithmetic mean of class-specific recall values. Unlike standard accuracy, which can be misleadingly high if a model simply predicts the majority class, Balanced Accuracy assigns equal weight to each diagnostic category. For a multiclass problem with *K* classes, it is defined as:

$$Balanced\ Accuracy = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FN_i}$$

In this formula, $TP_i$ represents the True Positives and $FN_i$ represents the False Negatives for each individual class *i*. This ensures that the model's ability to correctly identify rare malignant subtypes is represented as accurately as its ability to identify common benign lesions.

With this evaluation system, the four models are then cross-validated with a training session of patience 5 to only determine potential performance and then the best models are trained with patience 20 where the FDT from Definition 3 applies.

### *C. Pneumonia Model Selection and Evaluation*

The Pneumonia detection system in this study was chosen from two. Namely the two EfficientNetV2 variants (Tiny and Small). Here were the four metrics considered:

*Definition 13. Accuracy*

Accuracy represents the ratio of correctly predicted observations to the total observations in the dataset. While it provides a general overview of model performance, it is often supplemented by other metrics in medical contexts to account for potential data imbalances.

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

*Definition 14. Precision (Positive Predictive Value)*

Precision measures the proportion of positive identifications that were actually correct. In this study, high precision ensures that patients flagged for pneumonia truly have the condition, thereby reducing the rate of unnecessary clinical interventions or follow-up imaging.

$$Precision = \frac{TP}{TP + FP}$$

*Definition 15. Recall (Sensitivity)*

Recall measures the proportion of actual positives that were identified correctly. This is the most critical metric for pneumonia detection, as a high recall minimizes False Negatives, ensuring that infected patients are not incorrectly cleared and sent home without treatment.

$$Recall = \frac{TP}{TP + FN}$$

*Definition 16. F1-Score*

The F1-Score is the harmonic mean of Precision and Recall. It provides a single balanced metric that penalizes extreme values in either category. This is particularly useful for selecting a model that maintains a high diagnostic catch rate (Recall) without sacrificing too much diagnostic certainty (Precision).

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

For the model selection phase, the metric that determines the selected model was the F1-Score since any surge in false positives or false negatives would automatically lower F1 even with high accuracy.

## VI. RESULTS AND DISCUSSION

### *A. Skin Lesion Model*

The model selection phase was designed to identify the optimal architecture for skin lesion classification by evaluating the impact of MLP-based clinical metadata integration and specific EfficientNetV2 scaling variants. To ensure a fair and efficient comparison, each model was subjected to a tenfold cross-validation tournament with a strict early stopping patience

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

of 5 epochs. This prevented overfitting during the search phase and highlighted the architectures capable of rapid, stable convergence. The primary selection metric was the Median Balanced Accuracy across all folds to ensure robustness against dataset outliers, while the Interquartile Range (IQR) was utilized to measure diagnostic consistency.

For the initial stage of the hierarchy, models were trained to distinguish between malignant and benign lesions. This binary classification serves as the "gatekeeper" for the subsequent FDT.

| Model Architecture | Median Balanced Accuracy (%) | Interquartile Range (%) |
|---|---|---|
| EfficientNetV2-RWT-MLP | 88.29 | 0.91 |
| EfficientNetV2-RWT-ONLY | 87.85 | 0.77 |
| EfficientNetV2-RWS-MLP | 86.93 | 4.13 |
| EfficientNetV2-RWS-ONLY | 85.49 | 5.08 |

*Table 2. Top Classifier Benchmark (Malignancy vs. Benign)*

As demonstrated in Table 2, the EfficientNetV2-RWT-MLP emerged as the superior architecture for the Top Classifier, achieving a median accuracy of 88.29%. Notably, the integration of clinical metadata via MLP provided a measurable performance uplift for both RWT and RWS variants, confirming the hypothesis that non-image clinical markers (age, sex, and anatomical site) provide critical diagnostic context. The stability of these results is further analyzed through the distribution of validation accuracies across the ten folds.

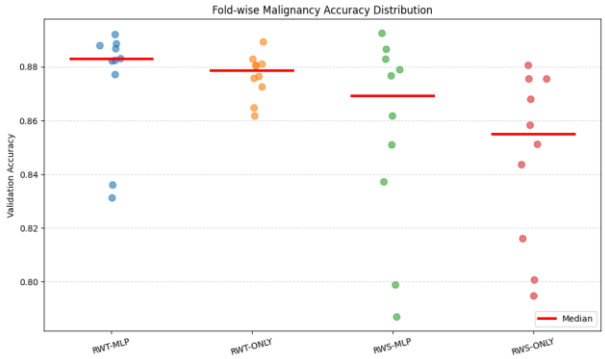


*Figure 2. Distribution and Stability Analysis of the Top Classifier*

As shown in Figure 2, the RWT-based models exhibited significantly lower IQR values (0.91%) compared to their RWS counterparts (4.13%). The RWT-MLP distribution displays a tight cluster of high-performing folds, whereas the RWS variants show significant dispersion and lower-bound outliers. This indicates that the RWT variant provides greater training stability and more consistent generalization. Due to this combination of high peak accuracy and low variance, the EfficientNetV2-RWT-MLP was selected as the foundational backbone for the subsequent full-training phase and the final 4-class FDT implementation.

Similarly, the model selection phase for the sub-classifier focused on differentiating between the three specific malignant pathologies: Melanoma (MEL), Basal Cell Carcinoma (BCC), and Squamous Cell Carcinoma (SCC). This stage utilized the same tenfold cross-validation framework and 5-epoch early stopping patience to identify which architecture best captured the fine-grained features necessary for intra-class cancer discrimination. The evaluation prioritized Median Balanced Accuracy and IQR to ensure the selected backbone could reliably categorize cancer subtypes across diverse patient metadata and lesion appearances.

| Model Architecture | Median Balanced Accuracy (%) | Interquartile Range (%) |
|---|---|---|
| EfficientNetV2-RWT-MLP | 91.77 | 2.76 |
| EfficientNetV2-RWT-ONLY | 88.63 | 2.49 |
| EfficientNetV2-RWS-MLP | 91.03 | 4.35 |
| EfficientNetV2-RWS-ONLY | 87.46 | 2.97 |

*Table 3. Comparative Performance of Sub-Classifier Variants*

The EfficientNetV2-RWT-MLP again outperformed all other variants with a median accuracy of 91.77%. Notably, the MLP variants for both RWT and RWS showed a significant performance jump (3-4) over their non-MLP counterparts, reinforcing that metadata fusion is essential for distinguishing between malignant subtypes that may appear visually similar but possess distinct clinical profiles. To evaluate the reliability of the sub-classifier across the different data folds, the accuracy distribution was plotted to visualize the spread and median performance.

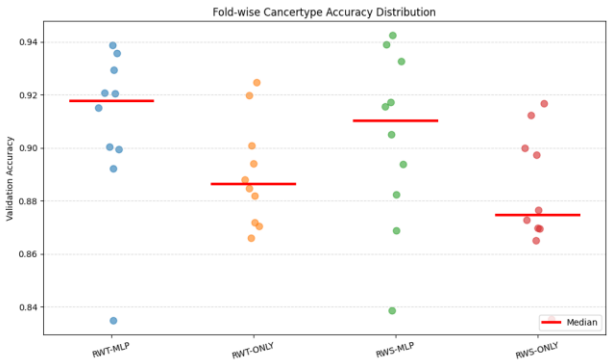


*Figure 3. Distribution and Stability Analysis of the Sub-Classifier*

As illustrated in Figure 3, while the RWT-MLP achieved the highest median, the RWS-MLP exhibited a much wider IQR (4.35%), indicating less predictable performance when training on smaller subsets of malignant data. The RWT-MLP maintained a more compact distribution, confirming its selection as the most stable architecture for the second level of the diagnostic hierarchy. Consequently, the EfficientNetV2-RWT-MLP was finalized as the architecture for both tiers of the system prior to the final full-training phase and FDT integration.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Following the model selection phase, the hierarchical classification system proceeded to the final training stage, where both the top classifier and the sub-classifier were trained using an extended early stopping patience of 20 epochs to ensure full convergence and maximum feature extraction. Once individual training was complete, the two separate models were synthesized into a unified diagnostic pipeline through the application of the Normalized FDT. This integration allows the system to transition from binary malignancy detection to a comprehensive four-class probability distribution, ensuring that the final output maintains the logical consistency required for clinical skin lesion diagnosis.

| Fold Index | Top-1 Balanced Accuracy | Top-2 Balanced Accuracy |
|---|---|---|
| Fold 1 | 0.8310 | 0.9910 |
| Fold 2 | 0.8607 | 0.9895 |
| Fold 3 | 0.8848 | 0.9943 |
| Fold 4 | 0.8411 | 0.9946 |
| Fold 5 | 0.8499 | 0.9778 |
| Fold 6 | 0.8838 | 0.9940 |
| Fold 7 | 0.8725 | 0.9893 |
| Fold 8 | 0.8410 | 0.9930 |
| Fold 9 | 0.8384 | 0.9911 |
| Fold 10 | 0.8838 | 0.9833 |
| Median | 0.8553 | 0.9911 |

*Table 4. Performance of the Integrated Hierarchical Pipeline across 10-Fold Cross-Validation*

The synthesis of the binary Top Classifier and the multi-class Sub-classifier via Normalized FDT yielded high diagnostic stability across all ten folds. The Top-1 Balanced Accuracy median of 0.8553 indicates that the system effectively handles the class imbalance inherent in the ISIC 2019 dataset, successfully differentiating between benign lesions and specific malignant subtypes (MEL, BCC, SCC).
Particularly noteworthy is the Top-2 Balanced Accuracy median of 0.9911. In a clinical context, this suggests that even when the primary diagnosis is incorrect, the true pathology is consistently identified as the secondary probability. This high Top-2 performance validates the "Forced Dominance" approach; by ensuring the sub-classifier's output is strictly gated by the top-level malignancy prediction, the system eliminates logically impossible predictions (e.g., a high-confidence MEL prediction for a lesion flagged as Benign). The low variance between folds further demonstrates that the extended 20-epoch patience allowed for a robust feature extraction that generalizes well across different data subsets.

To rigorously benchmark the hierarchical pipeline, it is essential to compare these results against the top-performing methodology of the ISIC 2019 Challenge, where they utilized a complex ensemble of EfficientNet (B0 through B6) and SENet154 architectures, incorporating advanced strategies such as loss balancing, color constancy, and multi-resolution cropping to address severe class imbalance [42]. Their final ensemble achieved a balanced accuracy (mean sensitivity) of 72.5% ± 1.7% across a five-fold cross-validation of the eight known diagnostic classes.

In comparison, the hierarchical system proposed in this study achieved a Top-1 Balanced Accuracy median of 0.8553 (85.53%) across ten folds. This represents a significant 13.03% improvement over the established baseline. This performance gap can be attributed to the Normalized FDT synthesis, which utilizes metadata-integrated sub-classifiers to refine diagnostic boundaries. While traditional ensembles rely on "flat" multi-resolution aggregation to handle the structural noise of an 8-class system, the hierarchical approach's use of a dedicated malignancy gate significantly reduces the probability of catastrophic classification errors between unrelated benign and malignant categories.

Furthermore, the near-perfect Top-2 accuracy of 0.9911 (99.11%) suggests a level of diagnostic stability that exceeds the reported metrics of competition-grade models. While previous optimization strategies focused on sensitivity through loss weighting, the hierarchical constraints in this pipeline ensure a "safety net" where the true pathology is virtually always captured within the top two predictions. This high performance validates the decision to move from a wide, flat 8-class model toward a structured diagnostic hierarchy, effectively outperforming one of the most robust ensemble methods developed for the ISIC 2019 dataset [43].

### *B. Pneumonia Model*

The selection of the pneumonia detection system was finalized through a comparative analysis of two EfficientNetV2 variants. To determine the optimal architecture, tenfold cross-validation was conducted using the Mendeley Pneumonia Dataset, which contains 5,856 AP view chest X-ray images.

| Efficient NetV2 Model | Median Accuracy (%) | Median Precision (%) | Median Recall (%) | Median F1 (%) | Median Epochs (rounded) |
|---|---|---|---|---|---|
| RWT | 96.24 | 97.09 | 97.66 | 97.43 | 8 |
| RWS | 97.35 | 97.91 | 98.36 | 98.19 | 7 |

*Table 5. Mendeley Pneumonia Dataset Tenfold Cross-Validation Results*

As illustrated in Table 6, the RWS variant demonstrated superior diagnostic capability across all recorded metrics. Specifically, RWS achieved a median accuracy of 97.35%, precision of 97.91%, recall of 98.36%, and an F1-Score of 98.19%. These results indicate that the RWS architecture is more adept at capturing the subtle opacities characteristic of pneumonia while maintaining a lower rate of false positives compared to the RWT variant. Furthermore, the RWS model reached these peak metrics in fewer epochs (7), suggesting a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

more efficient convergence path. Following selection, the RWS model was trained on the full Mendeley dataset for 14 epochs (twice the cross-validation median) to maximize its parametric capacity. This finalized model was then subjected to Out-of-Distribution (OOD) testing using clinical images from Pasig City Children's Hospital (PCCH) to evaluate its performance on unseen hardware and patient populations.

| Testing Set | Accuracy (%) | Precision (%) | Recall (%) | F1 (%) |
|---|---|---|---|---|
| Mendeley (Median CV) | 97.35 | 97.91 | 98.36 | 98.19 |
| Anteroposterior PCCH Images | 72.92 | 72.92 | 100.00 | 84.34 |
| Posteroanterior PCCH Images | 19.21 | 18.75 | 100.00 | 31.58 |

*Table 6. Performance of the RWS Model on Mendeley and Out-of-Distribution (PCCH) Chest X-ray Datasets*

For the AP images from PCCH, the model maintained high clinical utility. While accuracy and precision both settled at 72.92%, the model achieved a Recall of 100.00%. In a triage context, this is a significant finding: the model successfully identified every single positive pneumonia case in the PCCH AP set. The resulting F1-Score of 84.34% represents a manageable 14% "generalization gap" from the Mendeley baseline, confirming that the RWS model is robust enough to handle different imaging environments provided the anatomical orientation remains constant.

The results for Posteroanterior (PA) images, while lower in accuracy (19.21%) and precision (18.75%), provide a highly valuable insight into the model's spatial sensitivity. The sharp decline in performance when switching from AP to PA orientation demonstrates that the model has learned the specific anatomical markers of the AP view (where the heart appears larger and the scapulae are often visible in the lung field). Rather than a failure, this result highlights the model's high degree of view-specificity. It confirms that the deep features extracted by EfficientNetV2 are sensitive to the orientation of the thoracic cavity, reinforcing the importance of standardized imaging protocols when deploying AI in a clinical setting.

The performance of the RWS model is further validated when compared against a landmark study, which utilized the same Mendeley dataset for the development of a diagnostic tool for pediatric pneumonia [37]. The study reported a classification accuracy of 92.8% using an InceptionV3-based architecture. In contrast, the EfficientNetV2-RWS variant in this study achieved a median accuracy of 97.35%, representing a 4.55% improvement over the previous benchmark.

This performance gain is particularly notable in the sensitivity of the model; while the landmark study achieved high diagnostic standards, the RWS model's median recall of 98.36% (and subsequent 100% recall on PCCH AP images) suggests that the Ross Wightman scaling optimizations and the use of Focal Loss are more effective at minimizing false negatives than standard transfer learning approaches. By exceeding the accuracy of a highly cited, clinically validated model, this study proves that modern, parameter-efficient architectures like EfficientNetV2 can enhance the reliability of AI-driven triage in pediatric respiratory care.

## VII. CONCLUSION

This study developed and evaluated Dalikmata-Ultima, a medical artificial intelligence framework designed to improve diagnostic assistance for dermatological and radiological conditions through hierarchical classification, model hybridization, and local clinical validation. The findings demonstrate that structuring diagnosis according to clinical reasoning, integrating heterogeneous data sources, and validating models within real-world healthcare environments collectively strengthen the reliability and applicability of medical AI systems.

For skin lesion model, the proposed hierarchical classifier using the Forced Dominance Transform (FDT) successfully modeled the staged nature of clinical decision-making by separating primary triage (benign versus malignant) from subtype classification (basal cell carcinoma, melanoma, and squamous cell carcinoma). Ten-fold cross-validation using balanced accuracy enabled objective model comparison, allowing selection of architectures that maintained stable median performance across folds. Findings indicate that fragmentation of classification tasks reduces structural risk while preserving diagnostic confidence when recombined through the semantic transform as shown in Theorem 3 and Theorem 4.

The hybrid CNN-MLP architecture further demonstrated the value of multimodal learning in medical AI. By combining EfficientNetV2 image feature extraction with a custom Multilayer Perceptron processing patient metadata, the system leveraged both spatial lesion characteristics and structured clinical variables. Current understanding of hybrid CNN-MLP systems suggests that feature concatenation enables complementary learning representations, and this study confirms its relevance in medical contexts where diagnosis depends on both visual and demographic information. The optimized MLP design, utilizing fully connected layers, ReLU activation, and dropout regularization, aligned with generalization principles necessary for limited clinical datasets, reinforcing the importance of controlled architectural complexity in healthcare applications.

Pneumonia detection models were evaluated using accuracy, precision, recall, and F1-score, with model selection based on median F1 performance. Unlike the skin lesion model, hybridization was not applied due to insufficient structured patient metadata, highlighting that architectural choices must remain data-dependent. Moreover, local validation using chest X-ray data collected from Pasig City Children's Hospital demonstrated the feasibility of evaluating AI systems beyond benchmark datasets, addressing a major gap in Philippine medical AI research where external clinical validation remains limited.

Overall, these developments in hierarchical staging, multimodal hybridization, and local clinical validation position AI not as a replacement for physicians but as a structured

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

decision-support system. By enhancing diagnostic transparency and contextual relevance, this approach facilitates earlier detection and improved triage within healthcare systems facing limited specialist access. This framework ultimately supports more accessible diagnostic assistance while ensuring that essential clinician oversight remains at the center of the process.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the teachers and administrators of Pasig City Science High School for providing the necessary resources, time, and support to carry out this study.

We extend our heartfelt thanks to our research adviser, Ms. Charito Corsiga, for her invaluable guidance, encouragement, and feedback throughout the course of this study. We also convey our special thanks to our consultant, Mr. Hernan Barrosa, for his expert guidance in statistical and mathematical validation, which ensured the accuracy and integrity of this study.

We would like to express our deep appreciation to Dr. Vanessa Anne Bernal, MD, FPDS, for validating the skin lesion datasets and for her professional advice in improving the medical application of this work. We likewise thank Dr. Anthony Alvarez, MD, of Pasig City Children's Hospital for validating the chest X-ray datasets, and Sir Rodrigo Roque for generously providing the local pneumonia dataset.

Above all, we give thanks to God, whose guidance and blessings made this work possible.

## REFERENCES

[1] Pepito, V. C., Loreche, A. M., Legaspi, R. S., Guinaran, R. C., Prudencio, T., Mae, M., & Dayrit, M. (2025). Health workforce issues and recommended practices in the implementation of Universal Health Coverage in the Philippines: a qualitative study. Archīum Ateneo. https://archium.ateneo.edu/asmph-pubs/314/

[2] Cordero, D. A. (2022). An Evaluation of the Philippine Healthcare System: Preparing for a Robust Public Health in the Future. Journal of Preventive Medicine and Public Health, 55(3), 310–311. https://doi.org/10.3961/jpmph.22.154

[3] Alharbi, A. F., Rababa, M., Alsuwayl, H., Alsubail, A., & Alenizi, W. S. (2025). Diagnostic challenges and patient safety: The critical role of accuracy – A systematic review. Journal of Multidisciplinary Healthcare, Volume 18, 3051–3064. https://doi.org/10.2147/jmdh.s512254

[4] Villanueva, W. M. D., & Aranas, P. J. B. (2023). Catastrophic Out-of-Pocket Expenditure on Health: Evidence from the Regions in the Philippines. Journal Healthcare Treatment Development(JHTD) ISSN : 2799-1148, 3(06), 18–28. https://doi.org/10.55529/jhtd.36.18.28

[5] Francis, R., Overgaard, S. M., Gai, C., Overgaard, J. D., & Ohde, J. W. (2025). Guiding responsible AI in healthcare in the Philippines. Npj Digital Medicine, 8(1). https://doi.org/10.1038/s41746-025-01755-3

[6] Japanese International Cooperation Agency. (2023). JICA-DOH cooperation takes TB diagnosis in PH to new heights using AI tech | Where We Work - JICA. Www.jica.go.jp. https://www.jica.go.jp/english/overseas/philippine/information/press/2023/1516623_16864.html

[7] Healthcare Asia. (2025). Philippine healthcare sees rise of AI tools. Healthcare Asia Daily News - Asia's Leading News and Information Source on Healthcare and Medical Industry, Medical Technology, Healthcare Business and R&D, Healthcare Events. Online since 2010. https://www.healthcareasia.org/2025/philippine-healthcare-sees-rise-of-ai-tools/

[8] Singh, Y., Hathaway, Q. A., Varekan Keishing, Salehi, S., Wei, Y., Horvat, N., Vera-Garcia, D. V., Choudhary, A., Almurtadha Mula Kh, Quaia, E., & Andersen, J. B. (2025). Beyond Post hoc Explanations: A Comprehensive Framework for Accountable AI in Medical Imaging Through Transparency, Interpretability, and Explainability. Bioengineering, 12(8), 879–879. https://doi.org/10.3390/bioengineering12080879

[9] Oh, S., Kim, J. H., Choi, S. W., Lee, H. J., Hong, J., & Kwon, S. H. (2019). Physician confidence in artificial intelligence: An online mobile survey. Journal of Medical Internet Research, 21(3), e12422. https://www.jmir.org/2019/3/e12422/

[10] Cusipag, M. N., Oluyinka, A. S., Jimenez, R. S., Gonzales, R. A., & Ferrer, R. L. (2025). Artificial intelligence: Its implementation in Philippine healthcare institutions. Environment and Social Psychology, 10(7). https://esp.as-pub.com/index.php/esp/article/view/3611

[11] Miguel, R., Alcazar, M. U., Babaran, H. G., Dominique, B., Corpuz, A. A., Victoria, M., Claire, A., & Thiele, I. (2025). Exploring Filipino Medical Students' Attitudes and Perceptions of Artificial Intelligence in Medical Education: A Mixed-Methods Study. MedEdPublish, 14, 282–282. https://doi.org/10.12688/mep.20590.2

[12] Villanueva, E. Q. (2022). Epidemiologic profile of skin tumors in the Philippine General Hospital: A descriptive cross-sectional study. Health Science Reports, 5(5), e796. https://doi.org/10.1002/hsr2.796

[13] Tan, N. M. G., Arevalo, Ma. V. P. N., Eala, M. A. B., & Siripunvarapon, A. H. (2022). Skin cancer in the Philippines: The Filipino narrative. JAAD International, 8, 163–164. https://doi.org/10.1016/j.jdin.2022.05.007

[14] Sundararajan, S., Thida, A. M., & Badri, T. (2024). Cancer, Metastatic Melanoma. PubMed; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK470358/

[15] Hsu, B. W., & Tseng, V. S. (2022). Hierarchy-aware contrastive learning with late fusion for skin lesion classification. Computer Methods and Programs in Biomedicine, 216, 106666. https://doi.org/10.1016/j.cmpb.2022.106666

[16] Han, R., Liu, C., Sun, W., Yu, S., Zheng, H., & Deng, L. (2025). Machine learning of automatic hierarchical multi-label classification method for identifying metal failure mechanisms. Scientific Reports, 15(1), 19904. https://doi.org/10.1038/s41598-025-05076-z

[17] Santos, J. (2021). A Review of Pneumonia in the Philippines A Review of Pneumonia in the Philippines. Santos JA. *Philippines Journal*, 22(2), 6–11. https://www.pidsphil.org/home/wp-content/uploads/2021/09/003_vol-22-no-2_SANTOS_PNEUMONIA.pdf

[18] Taye, M. M. (2023). Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. Computation, 11(3), 52. https://doi.org/10.3390/computation11030052

[19] Khader, F., Müller-Franzes, G., Wang, T. S., Han, T., Arasteh, S. T., Haarburger, C., Stegmaier, J., Bressem, K. K., Kuhl, C. K., Nebelung, S., Kather, J. N., & Truhn, D. (2023). Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters: A Case for Transformers. Radiology, 309(1). https://doi.org/10.1148/radiol.230806

[20] Jain, M., & Shah, A. (2020). A multi-modal CNN framework for integrating medical imaging for COVID-19 Diagnosis. World Journal of Advanced Research and Reviews, 8(3), 475–493. https://doi.org/10.30574/wjarr.2020.8.3.0418

[21] Kaya, M. O. (2021). Performance Evaluation of Multilayer Perceptron Artificial Neural Network Model in the Classification of Heart Failure. The Journal of Cognitive Systems. https://doi.org/10.52876/jcs.913671

[22] Feng, J., & Lu, S. (2019). Performance Analysis of Various Activation Functions in Artificial Neural Networks. Journal of Physics: Conference Series, 1237, 022030. https://doi.org/10.1088/1742-6596/1237/2/022030

[23] Sen, A. (2025). Optimization and Generalization Dynamics in Multi-Layer Perceptron Classifiers for Low-Dimensional Feature Embeddings. Advanced International Journal for Research, 6(6). https://doi.org/10.63363/aijfr.2025.v06i06.2011

[24] Li, M., Li, S., Tian, Y., Fu, Y., Pei, Y., Zhu, W., & Ke, Y. (2023). A deep learning convolutional neural network and multi-layer perceptron hybrid fusion model for predicting the mechanical properties of carbon fiber. Materials & Design, 227, 111760. https://doi.org/10.1016/j.matdes.2023.111760

[25] Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., & Atkinson, P. M. (2018). A hybrid MLP-CNN classifier for very fine resolution

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

remotely sensed image classification. ISPRS Journal of Photogrammetry and Remote Sensing, 140, 133–144. https://doi.org/10.1016/j.isprsjprs.2017.07.01

[26] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems, 2(4), 303–314. https://doi.org/10.1007/bf02551274

[27] Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1(1), 67–82. https://doi.org/10.1109/4235.585893

[28] Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE Transactions on Neural Networks, 10(5), 988–999. https://doi.org/10.1109/72.788640

[29] Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. Information, 15(9), 517. https://doi.org/10.3390/info15090517

[30] Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller Models and Faster Training. https://arxiv.org/pdf/2104.00298

[31] Wightman, R., Touvron, H., & Jégou, H. (2021). ResNet strikes back: An improved training procedure in timm. ArXiv.org. https://arxiv.org/abs/2110.00476

[32] Swinckels, L., Ziesemer, K. A., Janneke, Scheerman, F. M., & Bennis, C. (2024). The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review. Journal of Medical Internet Research, 26. https://www.researchgate.net/publication/384300019_The_Use_of_Deep_Learning_and_Machine_Learning_on_Longitudinal_Electronic_Health_Records_for_the_Early_Detection_and_Prevention_of_Diseases_Scoping_Review

[33] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? ArXiv:2207.08815 [Cs, Stat]. https://arxiv.org/abs/2207.08815

[34] Arévalo-Cordovilla, F. E., & Peña, M. (2025). Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data. Scientific Reports, 15(1), 29420. https://doi.org/10.1038/s41598-025-15388-9

[35] Chung, J., & Teo, J. (2023). Single classifier vs. ensemble machine learning approaches for mental health prediction. Brain Informatics, 10(1), 1. https://doi.org/10.1186/s40708-022-00180-6

[36] Abernethy, J. (2015). Margin theory (Lecture 11). EECS 598: Theoretical foundations of machine learning, University of Michigan. https://web.eecs.umich.edu/~jabernet/eecs598course/fall2015/web/notes/lec11_101315.pdf

[37] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., & Shi, A. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell, 172(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

[38] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. Openaccess.thecvf.com. https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html

[39] Diukarev, V., & Starukhin, Y. (2024). Proposed Methods for Preventing Overfitting in Machine Learning and Deep Learning. Asian Journal of Research in Computer Science, 17(10), 85–94. https://doi.org/10.9734/ajrcos/2024/v17i10511

[40] Matas, I., Serrano, C., Nogales, M., Moreno, D., Ferrándiz, L., Ojeda, T., & Acha, B. (2025). Mitigating Overfitting in Medical Imaging: Self-Supervised Pretraining vs. ImageNet Transfer Learning for Dermatological Diagnosis. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2505.16773

[41] Shao, Z., & Er, M. J. (2016). Efficient Leave-One-Out Cross-Validation-based Regularized Extreme Learning Machine. Neurocomputing, 194, 260–270. https://doi.org/10.1016/j.neucom.2016.02.058

[42] Alshahrani, M., Al-Jabbar, M., Ebrahim Mohammed Senan, Ibrahim Abdulrab Ahmed, & Saif, M. (2024). Analysis of dermoscopy images of multi-class for early detection of skin lesions by hybrid systems based on integrating features of CNN models. PloS One, 19(3), e0298305–e0298305. https://doi.org/10.1371/journal.pone.0298305

[43] Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin Lesion Classification Using Ensembles of Multi-Resolution EfficientNets with Meta Data. MethodsX, 100864. https://doi.org/10.1016/j.mex.2020.100864