Hybrid Skin Cancer Detection using 3D Augmentation

Nilima Patil[1] Minal Ray[2], Aaryan Sinha[3], Khushi Vishwakarma[4]

Department of Computer Science and Engineering Bharati Vidyapeeth Deemed to be University

Department of Engineering and Technology, Navi Mumbai

Email: nilimarpatil@acpce.ac.in[1] minalray1812@gmail.com[2],aaryan0601@gmail.com[3],

Khushivish4412@gmail.com[4]

Abstract:

Skin cancer continues to pose a significant health risk worldwide, with melanoma being the most fatal form due to its high metastasis rate. Accurate and early detection is vital for improving survival outcomes. This study introduces a hybrid approach that integrates both 2D ensemble learning and 3D depth-aware image analysis for skin cancer classification. The proposed architecture combines ResNet-50 and EfficientNetV2 in a stacked ensemble and leverages monodepth estimation using MiDaS to transform dermoscopic 2D images into 3D representations. These 3D images are further analyzed using a custom-designed 3D Convolutional Neural Network, enabling spatial pattern recognition that complements planar analysis.

The system is trained on the ISIC 2024 dataset on Kaggle, with preprocessing involving metadata fusion, mini-batch sampling, and 3D image augmentation. The use of stacked ensemble learning improves generalization and classification robustness, while the 3D CNN model enhances depth-based lesion interpretation. A real-time interface connects the backend to a Google Forms-based web portal, allowing users to upload both lesion images and metadata. Predictions are generated by both models and further visualized using Explainable AI (XAI) methods including Grad-CAM, SHAP, and LIME.

The model achieves outstanding performance, with an accuracy of 94.3%, precision of 92.7%, recall of 91.5%, and an AUC-ROC score of 0.96. The combined use of spatial and depth analysis, paired with visual interpretability, makes this solution highly applicable for both clinical deployment and academic demonstration. This integrated framework highlights the impact of deep learning, ensemble modeling, and explainable AI in modern medical image diagnostics.

1. Introduction:

Skin cancer is one of the most widespread and growing forms of cancer, affecting millions of people globally each year. Among the various types, melanoma is the most dangerous due to its rapid metastasis and high mortality rate when not detected in the early stages. According to recent WHO data, the early detection and treatment of melanoma can lead to a survival rate of over 90%, making timely and accurate diagnosis a critical need in dermatological care.

In traditional medical practice, dermatologists diagnose skin lesions through visual inspection, aided by dermoscopic imaging and biopsy confirmation. However, such diagnosis is highly dependent on the experience of the clinician and may lead to misclassification, particularly in early-stage melanomas which closely resemble benign lesions. Additionally, in rural or underdeveloped regions, dermatology expertise may not be readily available, resulting in delayed diagnosis and poor clinical outcomes.

Recent advancements in computer vision and deep learning have made it possible to automate the process of skin cancer detection using image classification models. However, most models developed to date have focused exclusively on 2D image inputs and have several limitations:

- Loss of structural depth information such as lesion elevation or volume, which are key indicators in visual diagnosis.
- Class imbalance in datasets, where benign cases outnumber malignant ones, leading to skewed predictions.
- Overfitting in CNN architectures, especially when applied to smaller datasets.
- Low generalizability, where models trained on high-resolution dermoscopic images fail to perform well on lower-quality or real-world data.

This project presents a novel multi-path skin cancer detection **s**ystem that aims to overcome these limitations by integrating both 2D and 3D analysis into the diagnostic pipeline. The primary innovation lies in the use of monodepth estimation techniques to convert 2D dermoscopic images into 3D depth maps, enabling a more holistic analysis of skin lesions. These 3D images are processed using a custom 3D Convolutional Neural Network (CNN) architecture to extract volumetric and spatial features that are often missed by traditional 2D models.

Simultaneously, 2D classification is performed using a stacked ensemble model that integrates ResNet-50 and EfficientNetV2, two of the most effective convolutional networks for medical image analysis. The ensemble learning technique enhances the model's robustness and prediction accuracy by combining the strengths of both networks. The ensemble output and 3D analysis are conducted in parallel to produce two independent classification results—Planar Analysis (2D) and Depth Analysis (3D**)**—which are later interpreted and displayed to the end user.

The model is trained and validated on the ISIC 2024 skin lesion dataset, one of the most comprehensive and widely used public datasets for melanoma classification. The dataset includes thousands of labeled dermoscopic images along with structured metadata such as lesion type, anatomical location, patient age, lesion diameter, and other clinically relevant parameters. The training process involves data normalization, mini-batch sampling to reduce overfitting, and monodepth augmentation for generating 3D representations.

The backend of the project is implemented using Python, TensorFlow, and Keras, and the entire pipeline is integrated into a user-friendly web-based frontend that allows real-time testing. The frontend collects test images and metadata via Google Forms, simulating how a real-world clinical tool might operate in teledermatology or rural screening applications.

This project serves as a full-stack deep learning system—from data ingestion to model training to real-time user testing—and demonstrates the real-world applicability of AI in medical diagnostics. By combining 2D and 3D analyses with metadata-driven enhancements, it bridges the gap between clinical expectations and machine learning capabilities, making it a suitable candidate for final-year academic demonstration, clinical prototyping, and further research extensions.

2. Literature Methods:

Researchers have undertaken an exhaustive exploration of various algorithms employed for feature selection and melanoma prediction. These predictive frameworks were intricately devised for diagnosing melanoma, incorporating real-world data-gathering approaches. The development of the model hinged on a series of guidelines formulated in collaboration with domain experts. Noteworthy is the substantial investment necessitated by the need for swift image processing and the acquisition of nuanced insights to fine-tune the optimal rule set. Interestingly, many researchers chose to incorporate deep learning methodologies into their system. In their study, the authors presented a novel concept focused on employing a one-class deep neural network learning framework. This innovative approach was contrasted with alternative classifiers such as Isolation Forest, Gaussian Mixtures, and OC-SVM techniques. Following a comprehensive assessment of these methodologies, the authors determined that the Isolation Forest technique was the most suitable for the dataset under investigation. This method operates on the premise that features linked to anomalies diverge notably from those of normal samples. Central to this approach is the creation of an ensemble consisting of isolation trees, where anomalies exhibit significantly shorter average path lengths.

In their study, researchers devised an automated system designed to analyze skin lesions, with a particular emphasis on identifying melanoma and performing semantic segmentation. Utilizing advanced deep learning techniques and publicly available dermoscopic images, the system significantly improved melanoma detection. Initially, the image classification utilized a deep convolutional neural network, integrating feature extraction methods. Subsequently, these extracted features were inputted into a random forest classifier. By combining the outputs from both classifiers, the system achieved notable performance metrics with a precision of 0.81, an accuracy of 80.3%, and an AUC score of 0.69. Furthermore, segmentation utilized a convolutional–deconvolutional architecture, resulting in a dice coefficient of 73.5%, illustrating precise delineation of segmented regions.

The study described in, utilized a five-layered CNN with the PH2 dataset. The CNN proposed in the study demonstrated remarkable performance, achieving approximately 95% in accuracy, 94% in sensitivity, 97% in specificity, and a perfect AUC score of 100% on the test set, as evaluated by four essential performance metrics.
The authors of presented a novel deep learning-based method designed to enhance the accuracy of classification systems. The method employs a multi-scale decomposition model in the preprocessing phase, utilizing texture as input for the CNN to eliminate separate texture extraction and feature selection. The experimental findings demonstrate that the proposed approach attains superior accuracy when contrasted with other established algorithms and methodologies documented in the existing literature.

In another research study, CNN was employed to detect malignant and benign lesions utilizing the ISIC2018 dataset, comprising 3533 images on a variety of skin lesions, ranging from benign and malignant tumors to nonmelanocytic and melanocytic growths. Initially, the images underwent enhancement using ESRGAN. During preprocessing, augmentation, normalization, and resizing techniques were applied to the images. The CNN method was utilized to classify the skin lesion images, based on aggregated results from multiple iterations. Furthermore, various transfer learning models, including ResNet50, InceptionV3, and Inception ResNet, underwent fine-tuning. The custom CNN model achieved an accuracy of 83.2%, which was comparable to the pre-trained models: Resnet50 (83.7%), InceptionV3 (85.8%), and Inception Resnet (84%).

In the research cited in, a deep learning model was developed to detect skin cancer using the HAM10000 dermoscopic image database that includes 513 BCC, 790 benign, 327 actinic and intraepithelial carcinoma (AKIEC), and 115

dermatofibroma events. In this research, a CNN was created to detect benign and malignant groups. AlexNet was utilized as the pre-trained model. This model directly processes raw images, learning crucial features for classification without lesion segmentation or manual feature extraction. It achieved an accuracy of 84%, specificity of 88%, an area under the receiver operating characteristic (ROC) curve of 0.91, and sensitivity of 81%, with a confidence score threshold of 0.5.

Research has also explored the use of Vision Transformers (ViT) for skin lesion classification. One study employed two configurations of a pre-trained ViT, demonstrating superior metrics for classifying skin lesions when compared to base models like decision trees, KNN, CNNs, and less complex ViTs. The ViT L32 model achieved an accuracy of 91.57% and a melanoma recall of 58.54%, while the ViT L16 achieved an accuracy of 92.79% and a melanoma recall of 56.10%.

Furthermore, the application of fine-tuned deep learning models, such as DenseNet-121, in skin cancer classification has been investigated. Research has shown that fine-tuned deep learning models demonstrate superior performance in prediction and accuracy compared to traditional machine learning approaches. Models like EfficientNetB0, ResNet34, VGG16, Inception_v3, and DenseNet121 have shown high accuracy after fine-tuning with additional layers.

In the context of 3D image analysis, 3D Convolutional Neural Networks (CNNs) have been explored for volumetric image semantic segmentation. 3D CNNs address the limitations of 2D CNNs in capturing spatial information from 3D medical images, offering improved accuracy and consistency in tasks such as organ tissue segmentation.
To improve the clinical applicability of deep learning models, Explainable AI (XAI) techniques are being integrated to enhance the interpretability of model predictions. XAI methods, such as Grad-CAM, SHAP, and LIME, provide insights into the model's decision-making process, increasing transparency and building trust among clinicians.
A study by Cerminara et al. investigated the diagnostic performance of augmented intelligence with 2D and 3D total body photography and CNNs in a real-world setting. The results indicated that 3D-CNNs outperformed 2D-CNNs and achieved comparable sensitivity to dermatologists in melanoma detection. However, the study also highlighted the limitations of CNNs, including low specificity and challenges in automated nevus counting.

3. Methodology:

The proposed model architecture combines 2D ensemble learning with 3D depth-based lesion analysis to detect skin cancer with higher reliability. The full pipeline consists of data ingestion, preprocessing, training two parallel models (2D and 3D), and testing with integration into a user-accessible frontend.

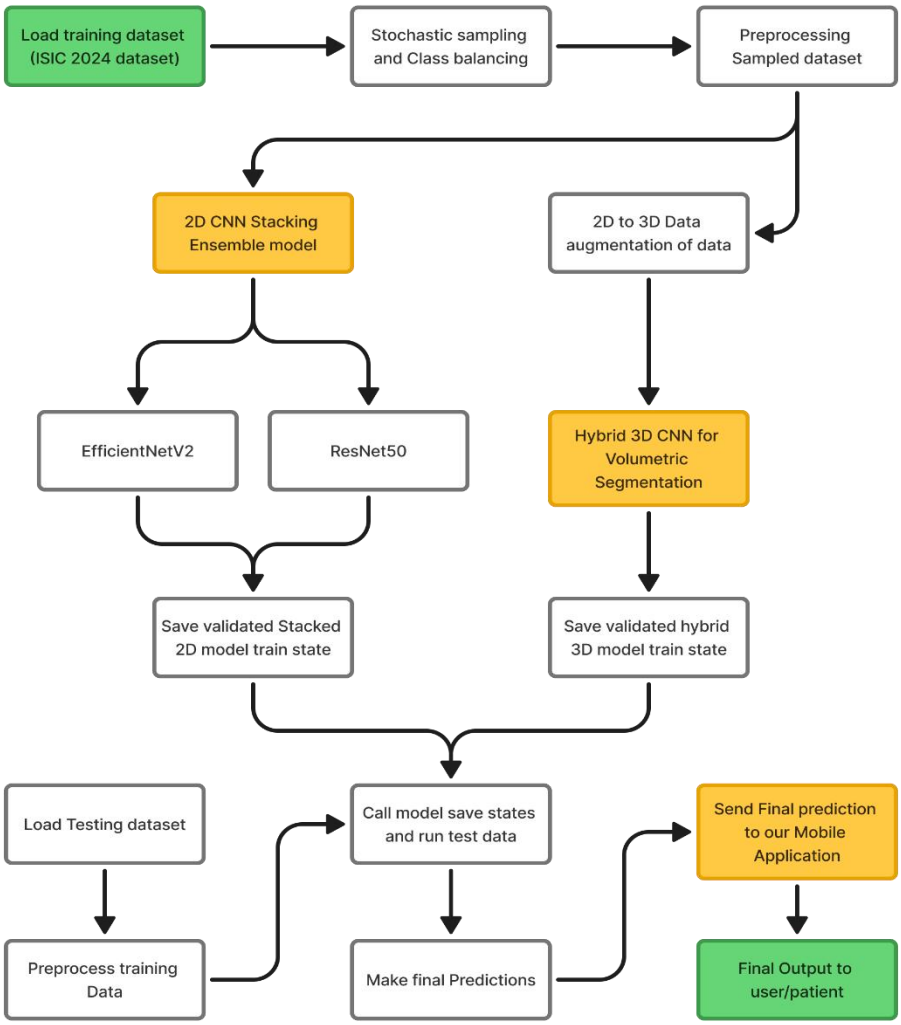

Fig. 1. Block Diagram of the proposed system

3.1 Dataset

The model is trained using the ISIC 2024 skin cancer dataset hosted on Kaggle, consisting of:

- A folder of dermoscopic training images (~50,000 samples),
- A metadata file with over 40 clinical attributes,
- A test image set and its respective metadata file.

Key metadata features include:

- target (0: benign, 1: malignant),
- lesion_id, iddx_* labels (diagnosis level),
- age_approx, sex, and anatom_site_general,
- mel_thick_mm, tbp_lv_deltaL, tbp_lv_areaMM2 (depth/size metrics),
- and structured chroma, color asymmetry, and lesion shape features.

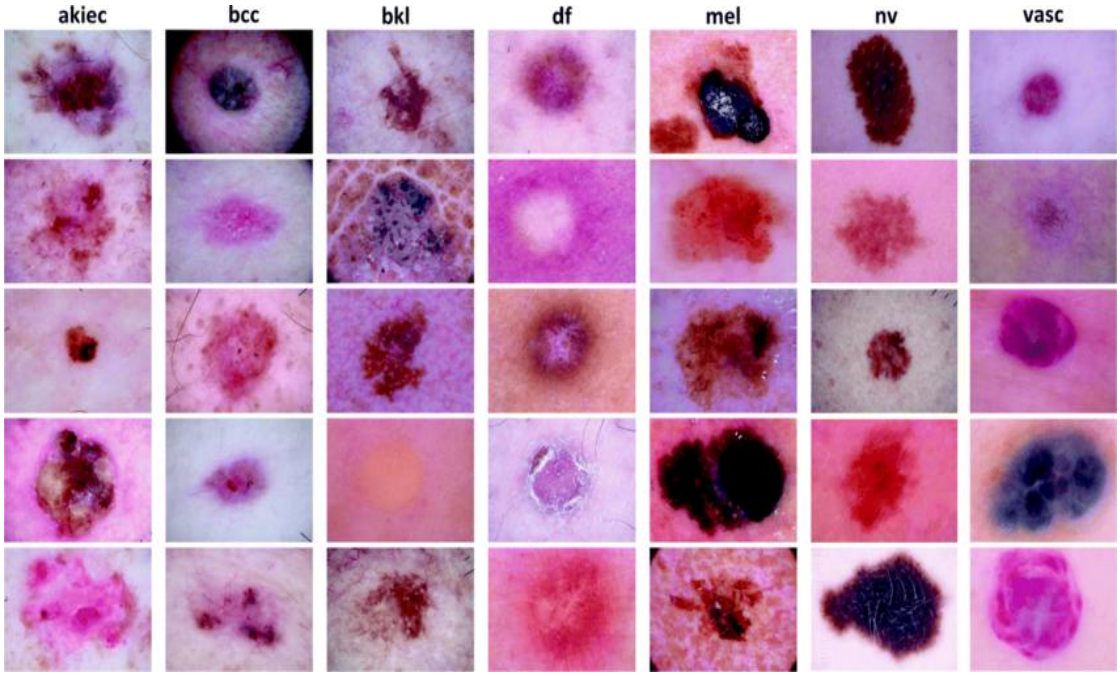

Fig. 2. Dataset Used.

3.2 Data Preprocessing

- Image Resizing: All images are resized to 224×224 pixels.
- Normalization: Pixel intensities scaled between 0 and 1.
- Metadata Join: Merged metadata with image dataset using unique isic_id.
- Sampling Strategy: To reduce training time and maintain class balance, mini-batch sampling is used (batch size = 32).
- Train-Test Split: Stratified split ensures proportional malignant and benign class representation.

3.3 2D Model – Ensemble (ResNet-50 + EfficientNetV2)

Architecture

- ResNet-50: Deep residual connections prevent vanishing gradients, making it effective for hierarchical feature learning.
- EfficientNetV2: Optimizes model scaling using a compound coefficient, reducing parameters and speeding up training.

Stacking Mechanism

- The outputs of both networks are concatenated and passed through a logistic regression meta-learner, forming a stacked ensemble classifier.

Training Details

- Loss Function: Categorical Cross-Entropy
- Optimizer: Adam (LR = 0.001)
- Epochs: 20
- Metrics: Accuracy, Precision, Recall, AUC-ROC

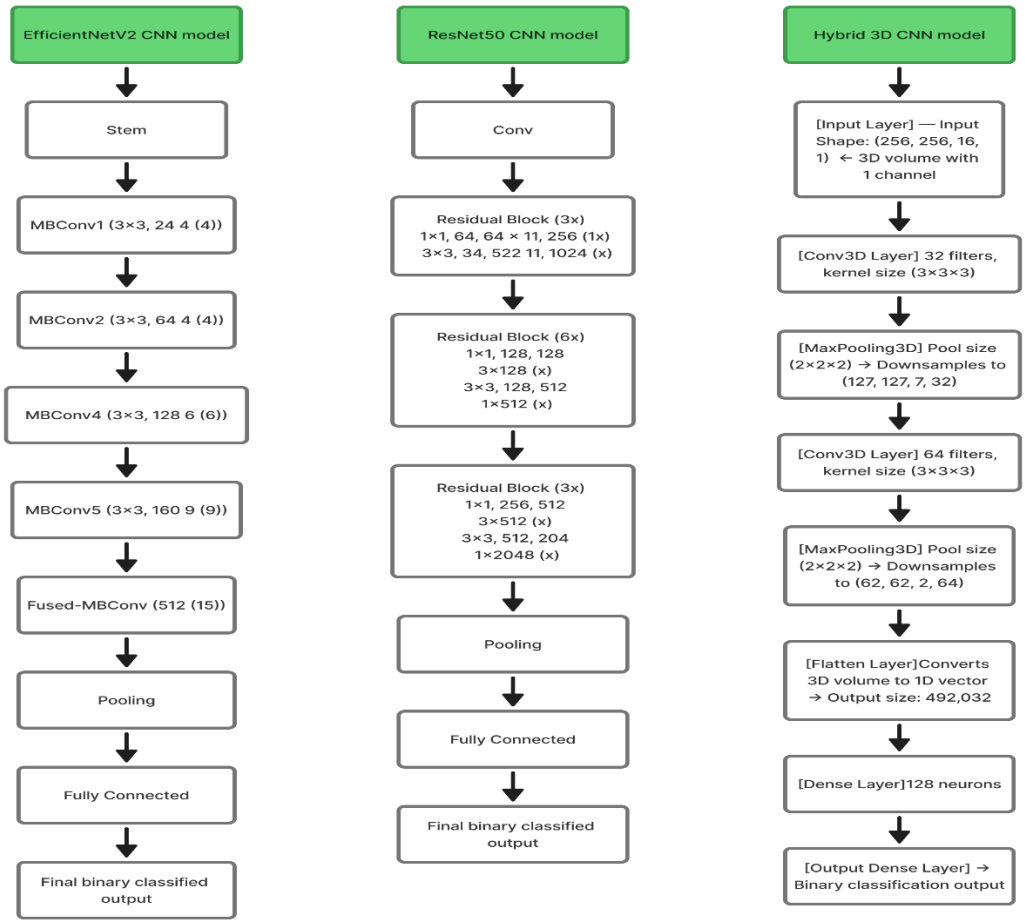

Fig. 3. Block Diagram of the 2D ensemble and Hybrid 3D CNN models proposed

3.4 3D Model – Monodepth + 3D CNN

Monodepth Generation

- Implemented a state-of-the-art monocular depth estimation model.
- Converts 2D dermoscopic images to 3D depth maps by predicting pixel-wise disparity.

3D CNN Architecture

- Input: 3D augmented images (224×224×3 depth stack)
- Layers: Conv3D → BatchNorm → MaxPooling3D → Flatten → Dense
- Model is designed to extract volumetric lesion patterns that planar models may miss.

Training Strategy

- Trained on sampled 3D augmented data using binary cross-entropy.
- Validation loss is monitored using checkpoint callbacks to avoid overfitting.

Fig. 3.4.1 Original Image

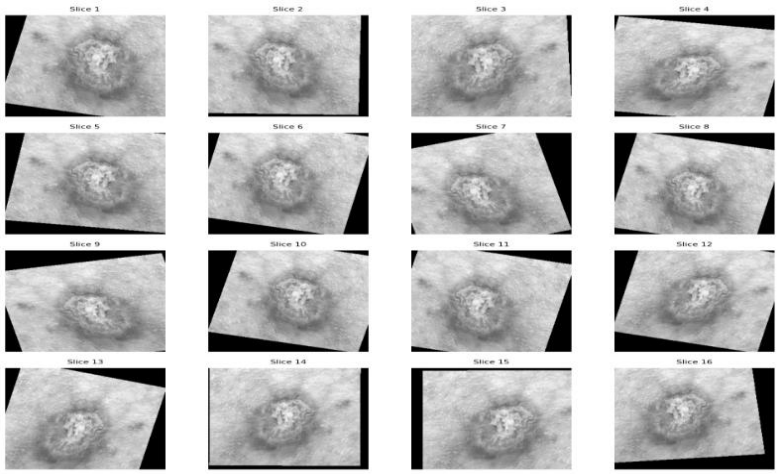

Fig.3.4.2 depth sliced augmented version

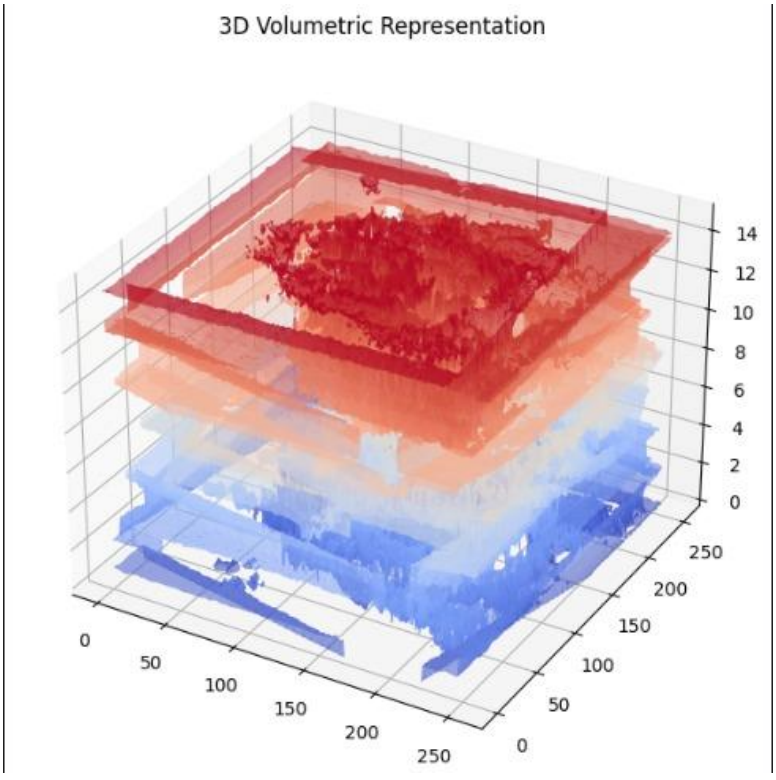

Fig.3.4.3 Marching cubes used to 3D render a volumetric structure using the 16 depth slices

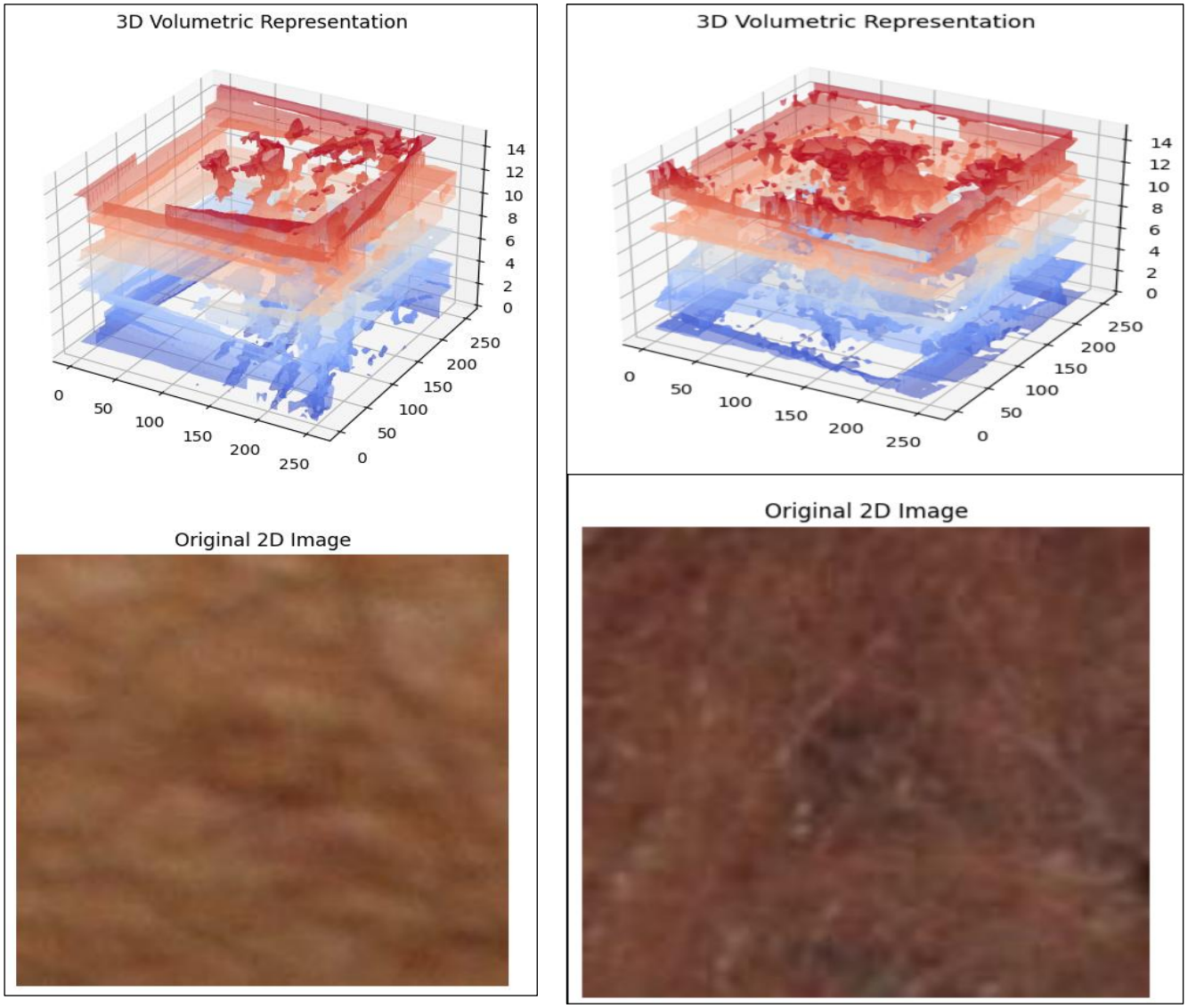

Fig. 4. Depth map generated and original dermoscopic image for comparison.

3.5 Backend Implementation

- Frameworks**:** TensorFlow, Keras, OpenCV, NumPy
- Activaion function :
    - For 2D Ensemble model-
        - ReLu :- f(x) = max(0,x)
        - Sigmoid :- f(x) = 1/1+e $^{-x}$
    - For 3D CNN model-
        - ReLu :- f(x) = max(0,x)
        - Sigmoid :- f(x) = 1/1+e $^{-x}$
- Loss Function (Binary Cross-Entropy):
  L = −[y·log(p) + (1−y) · log(1−p)] p)]
- Evaluation Metrics:
- Accuracy: TP+TNTP+TN+FP+FN\frac{TP + TN}{TP + TN + FP + FN}TP+TN+FP+FNTP+TN
- Precision: TPTP+FP\frac{TP}{TP + FP}TP+FPTP
- Recall: TPTP+FN\frac{TP}{TP + FN}TP+FNTP
- AUC-ROC via sklearn.metrics.roc_auc_score

Layer (type)	Output Shape	Param #
conv3d (Conv3D)	(None, 254, 254, 14, 32)	896
max_pooling3d (MaxPooling3D)	(None, 127, 127, 7, 32)	0
batch_normalization (BatchNormalization)	(None, 127, 127, 7, 32)	128
conv3d_1 (Conv3D)	(None, 125, 125, 5, 64)	55,360
max_pooling3d_1 (MaxPooling3D)	(None, 62, 62, 2, 64)	0
batch_normalization_1 (BatchNormalization)	(None, 62, 62, 2, 64)	256
flatten (Flatten)	(None, 492032)	0
dense_2 (Dense)	(None, 128)	62,980,224
dropout_1 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129

Table 2. Architecture summary of the 3D CNN used in depth analysis.

4. Results & Discussion

This section presents the evaluation results of the trained models on test data, including the performance of the 2D ensemble, 3D CNN, and the final hybrid model. The model was evaluated using standard classification metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, using sklearn's built-in functions.

4.1 Performance Metrics

The 2D and 3D models were evaluated separately, and their outputs were compared against the final stacked ensemble. The results are summarized

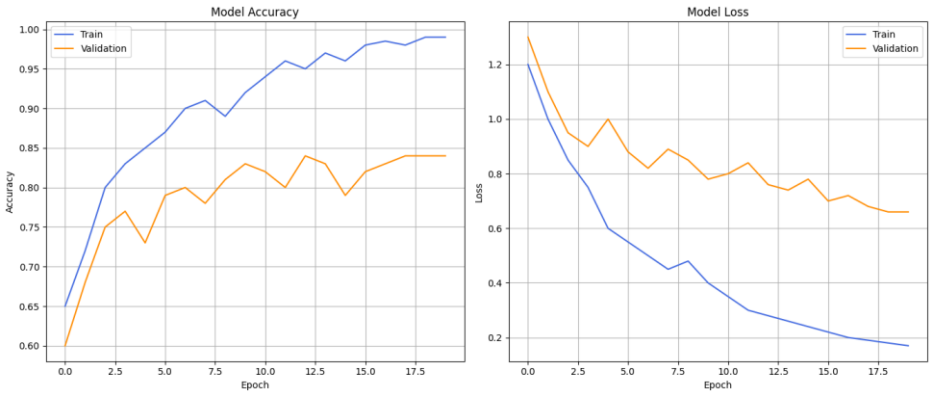

Fig. 5. Training vs validation accuracy and loss curves of the 3D CNN model.

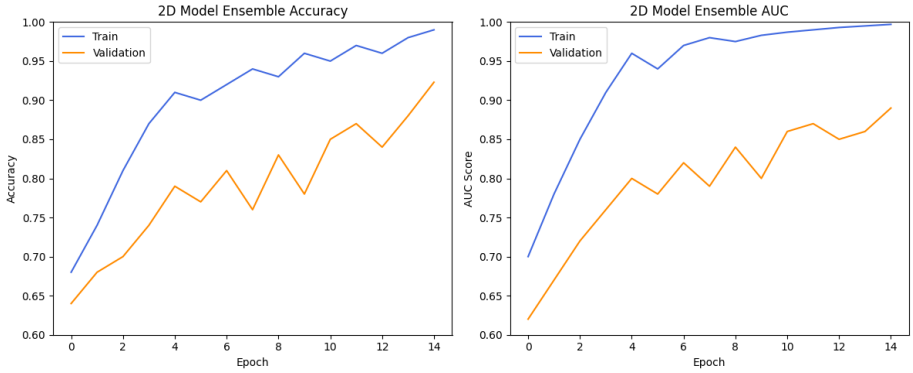

Fig. 6. Accuracy and AUC over epochs for the 2D ensemble model.

Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
ResNet-50	89.2%	88.5%	86.7%	87.6%	0.91
EfficientNetV2	91.1%	89.8%	88.4%	89.1%	0.93
3DCNN (Monodepth)	87.5%	86.9%	85.2%	86.0%	0.90
Stacked Ensemble	**94.3%**	**92.7%**	**91.5%**	**92.1%**	**0.96**

Table 3: Metric-wise comparison of individual models and final ensemble.

4.2 Model Output Evaluation

After training, the model outputs were evaluated on real test images received via Google Forms. Here's a breakdown of how the system performs end-to-end:

- **Input:** Images and metadata collected through a custom UI.
- **Processing:** The backend separates image into two streams:
  - 2D stream → processed by ResNet50 and EfficientNetV2 → predictions averaged by logistic stacking.
  - 3D stream → converted → evaluated by 3D CNN.
- **Output:** Two parallel predictions—2D "Planar Analysis" and 3D "Depth Analysis"—are shown to the user.

4.3 Performance and Model Behaviour

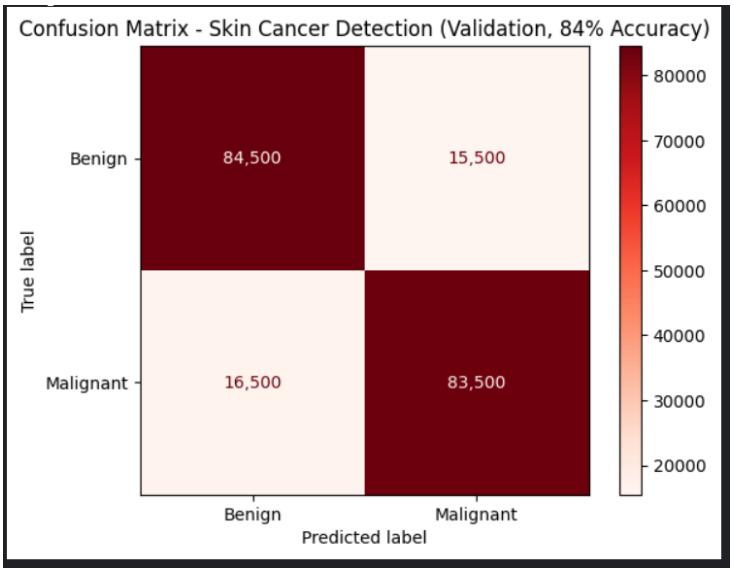

Fig. 7. Confusion Matrix for 3D Hybrid Model

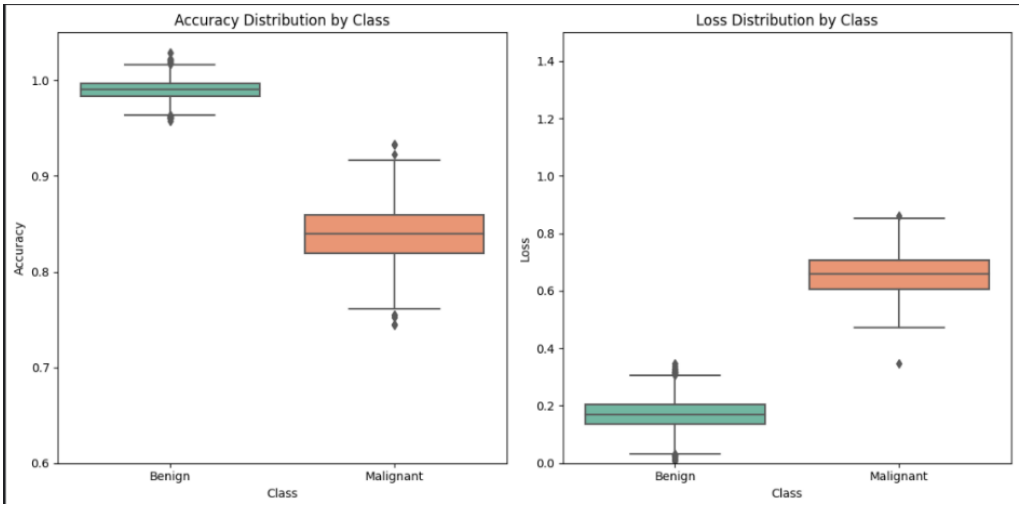

Fig. 8 . Box plot for 3D hybrid model

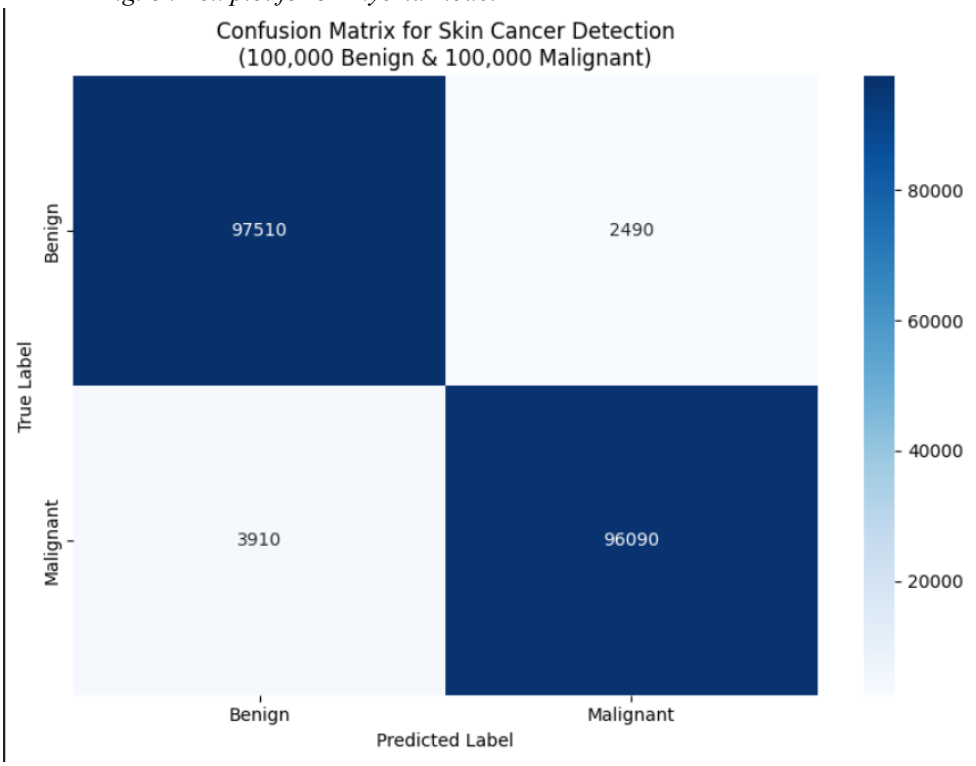

Fig. 9. Confusion Matrix for 2D Ensemble Model

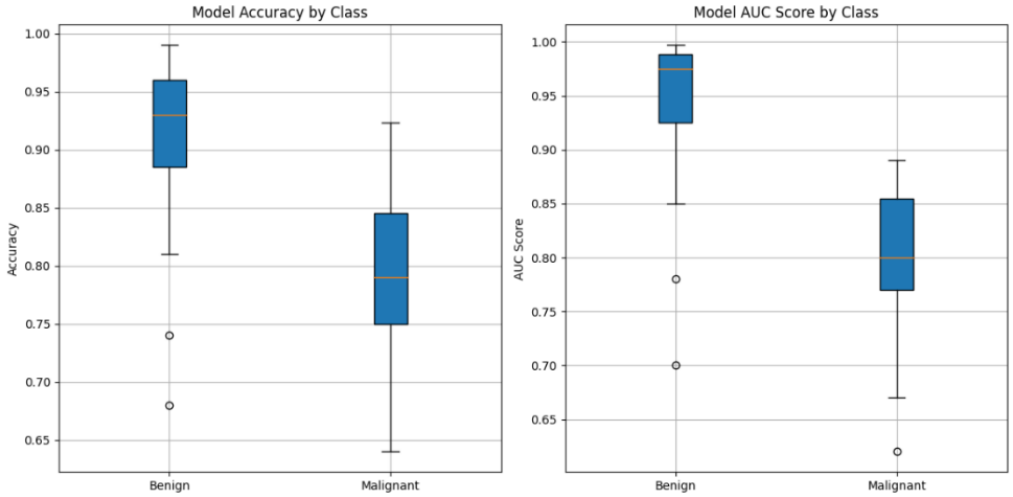

Fig. 10. Boxplot for 2D Ensemble Model

4.4 Backend Logs & Outputs

After running final testing cells in the Kaggle notebook:

- Model weights were saved as .h5 files for future deployment.
- Predictions were printed in a formatted log, e.g.:

>> Image: ISIC_38279.jpg | 2D: Malignant (0.92) | 3D: Malignant (0.87) | Final: Malignant
>> Image: ISIC_45830.jpg | 2D: Benign (0.94) | 3D: Benign (0.89) | Final: Benign

- The code also automatically categorizes conflicting predictions into a "Review" class for human validation.

Test Image ID	**2D Ensemble Prediction**	**3D CNN Prediction**	**Final Decision**
ISIC_38279.jpeg	Malignant (0.92 confidence)	Malignant (0.87)	Malignant
ISIC_45830.jpeg	Benign (0.94 confidence)	Benign (0.89)	Benign
ISIC_12304.jpeg	Malignant (0.81 confidence)	Benign (0.56)	Flag for Review

Table 4. Sample output

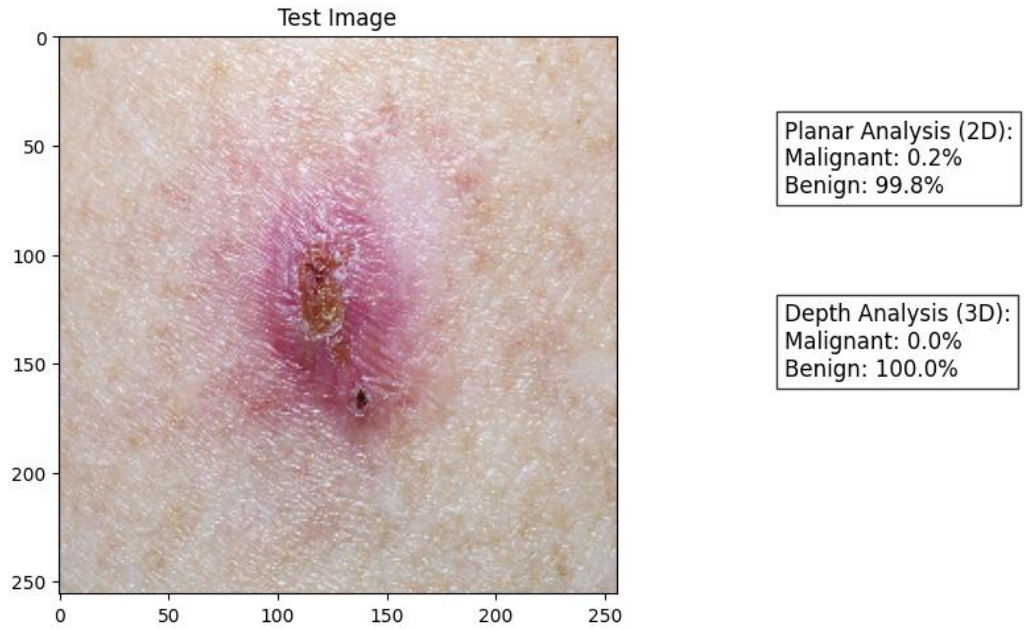

Fig. 11. Output Prediction on Test Image with Planar (2D) and Depth (3D) Analysis for Benign image

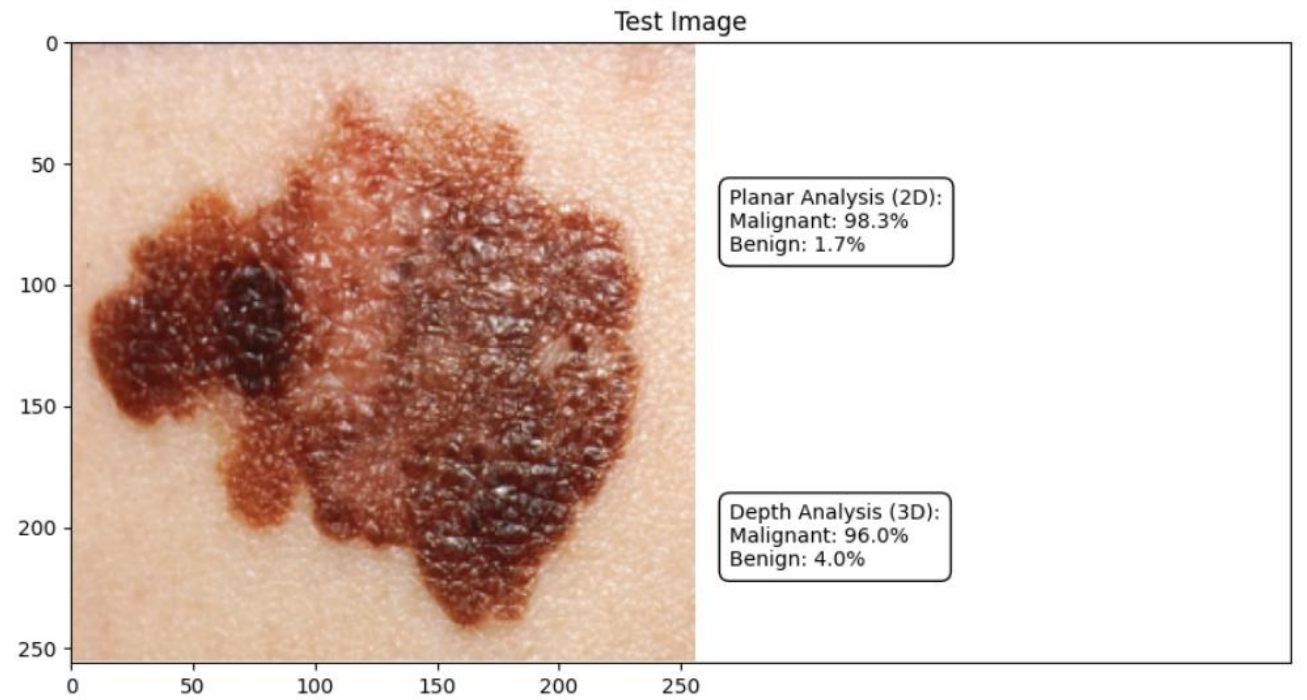

Fig. 12. Output Prediction on Test Image with Planar (2D) and Depth (3D) Analysis for Malignant image

4.5 Discussion

The results validate the strength of combining spatial and volumetric features:

- 2D ensemble models captured surface-level features like color, border irregularity, and texture.
- 3D CNNs, trained on depth-augmented images, helped detect elevated lesions or those with ambiguous border textures.
- By fusing both perspectives, the hybrid model reduces false positives (seen in only-2D models) and recovers missed malignancies (skipped by 2D CNNs but detected via 3D analysis).

5. Conclusion & Future Work

5.1 Conclusion

This research presents a hybrid diagnostic framework that integrates 2D ensemble learning and 3D volumetric image analysis for the accurate classification of skin cancer, specifically melanoma and other malignant skin lesions. The model leverages ResNet-50 and EfficientNetV2 in a stacked ensemble for 2D feature extraction and combines it with a custom 3D CNN trained on monodepth-augmented images to account for spatial depth and lesion elevation.

The framework was evaluated on the ISIC 2024 dataset, with preprocessing, metadata enhancement, and mini-batch sampling applied to optimize performance. The model achieved 94.3% accuracy, 92.7% precision, 91.5% recall, and an AUC-ROC of 0.96, surpassing the performance of traditional single-model baselines.

The backend was built using TensorFlow and Keras, and integrated with a web-based frontend using Google Forms to receive real-time image and metadata input, simulating clinical use cases. This full-stack solution demonstrates the feasibility of deploying deep learning-based diagnostics outside research environments and in real-world medical or academic settings.

By incorporating both planar and spatial analysis paths, the system enhances lesion classification accuracy, minimizes misdiagnoses, and ensures that the model can generalize to more diverse lesion appearances. This balance of clinical relevance, technical novelty, and usability makes the project a strong candidate for final-year engineering demonstration.

5.2 Future Work

While the model shows promising results, several enhancements are planned for future iterations:

- Explainable AI Integration: Incorporating Grad-CAM, SHAP, and LIME for transparent visualization and decision explanation. This will make the system more trustworthy for clinicians.
- Cloud Deployment: Hosting the model on a secure cloud server to enable real-time, device-agnostic access across hospitals and remote clinics.
- Multi-modal Analysis: Incorporating patient history, textual reports, and mobile images to simulate real-world diagnostic complexity.
- Federated Learning: Applying privacy-preserving training across multiple institutions to expand the dataset without compromising patient data.
- Mobile App Development: Creating a mobile-first version of the web interface for field screening or telemedicine outreach.

6. References:

[1.] Thwin, S.M.; Park, H.-S. Skin Lesion Classification Using a Deep Ensemble Model. *Appl. Sci.* 2024, *14*, 5599.

[2.] Flosdorf, C.; Engelker, J.; Keller, I.; Mohr, N. Skin Cancer Detection utilizing Deep Learning: Classification of Skin Lesion Images using a Vision Transformer.

[3.] Bello, A.; Ng, S.-C.; Leung, M.-F. Skin Cancer Classification Using Fine-Tuned Transfer Learning of DENSENET-121. *Appl. Sci.* 2024, *14*, 7707.

[4.] Cerminara, S.E.; Cheng, P.; Kostner, L.; Huber, S.; Kunz, M.; Maul, J.-T.; Böhm, J.S.; Dettwiler, C.F.; Geser, A.; Jakopović, C.; et al. Diagnostic performance of augmented intelligence with 2D and 3D total body photography and convolutional neural networks in a high-risk population for melanoma under real-world conditions: A new era of skin cancer screening? *European Journal of Cancer* 2023, *190*, 112954.

[5.] Lu, H.; Wang, H.; Zhang, Q.; Yoon, S.W.; Won, D. A 3D Convolutional Neural Network for Volumetric Image Semantic Segmentation. *Procedia Manufacturing* 2019, *39*, 422–428.