International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

A Multimodal Deep Learning Framework for Explainable and Calibrated AI-Assisted Disease Prediction in Clinical Decision Support

Kishor Mugale, Milind Lanje, Mahesh Dhote, Deepak Deshmukh, Nita Gore, Vaibhav Rewaskar
Department of Computer Engineering
MIT Academy of Engineering, Alandi, Pune, India

***Abstract*—Early and consistent screening for multiple disease categories remains a bottleneck in routine clinical workflows, particularly where radiology and laboratory expertise is unevenly distributed. This paper presents a multimodal clinical decision support system (CDSS) that unifies medical image analysis and structured clinical-data analysis under a single prediction and reporting pipeline. The system screens for ten disease categories drawn from oncology, pulmonology, ophthalmology, orthopaedics, and metabolic medicine, using six image-based deep learning classifiers and four tabular ensemble learning models. Image classification is performed using an EfficientNet-B4 backbone augmented with a Convolutional Block Attention Module (CBAM) for five disease categories, a DenseNet-169 architecture for musculoskeletal fracture detection, and a domain-pretrained Vision Transformer (RETFound) for retinal fundus screening. Structured clinical data is processed through voting and stacking ensembles built from gradient-boosted trees. To address the opacity of deep models in a clinical setting, the system produces per-prediction explanations using Gradient-weighted Class Activation Mapping (Grad-CAM) for image-based predictions and SHapley Additive exPlanations (SHAP) for tabular predictions. Raw softmax confidence scores from the image models are further corrected through post-hoc temperature scaling to reduce overconfidence, and all predictions are translated into a three-tier clinical risk label. The complete platform is implemented as an asynchronous FastAPI backend with a React-based frontend, JWT-based authentication, rate limiting, Prometheus-based monitoring, and automated PDF report generation. On held-out validation data, the constituent models achieve accuracies ranging from 88.5% (diabetes) to 99.2% (brain tumour detection). We describe the system architecture, model design choices, explainability and calibration methodology, and discuss the limitations of the current design together with directions for future work.**

***Index Terms*—clinical decision support, multimodal deep learning, explainable AI, Grad-CAM, SHAP, model calibration, medical image classification, ensemble learning**

I. Introduction

Artificial intelligence has demonstrated strong empirical performance across individual medical imaging and clinical-risk-prediction tasks, from radiograph interpretation to structured-data-based risk scoring. However, most reported systems specialise in a single disease and a single input modality, and relatively few integrate multiple disease categories, multiple data modalities, and clinician-facing explainability and calibration into one deployable platform. Clinicians are understandably reluctant to act on an opaque probability score, particularly one that is poorly calibrated: a model that reports 95% confidence but is correct only 70% of the time at that confidence level can mislead triage decisions rather than assist them.

This work presents the design and implementation of a multimodal AI disease-screening platform intended to function as a decision-support layer rather than an autonomous diagnostic authority. The system accepts either a medical image (magnetic resonance imaging, radiographs, dermoscopic images, or fundus photographs) or structured tabular patient data (vitals and laboratory biomarkers) and returns a predicted class, a calibrated confidence score, a derived clinical risk level, and a modality-appropriate explanation of the prediction. The primary contributions of this paper are:

- A unified multimodal architecture spanning six image classifiers (built on EfficientNet-B4 with attention, DenseNet-169, and a retinal foundation Vision Transformer) and four tabular ensemble classifiers covering ten distinct disease-prediction tasks.
- An explainability layer that produces Grad-CAM saliency maps for image predictions and Shapley-value feature attributions for tabular predictions, generated per request rather than offline.
- A post-hoc temperature-scaling calibration stage applied to every image model, together with a risk-stratification scheme that converts calibrated probabilities into clinician-facing risk labels.
- A description of the production system architecture — asynchronous API design, authentication, rate limiting, and monitoring — required to operate such a model suite reliably.

II. Related Work

Convolutional architectures with compound model scaling, such as the EfficientNet family, have become a standard backbone for medical image classification tasks because they offer a favourable accuracy-to-parameter trade-off relative to earlier hand-designed networks [1]. Attention modules such

as the Convolutional Block Attention Module (CBAM) have been shown to improve localisation of diagnostically relevant regions by sequentially applying channel- and spatial-attention gates to intermediate feature maps [2], which is particularly useful when pathology occupies a small, spatially concentrated region of an otherwise large image. Densely connected networks, in which each layer receives the concatenated feature maps of all preceding layers, have separately been shown to preserve fine-grained texture information relevant to tasks such as fracture-line detection [3], and were the reference architecture used in the large-scale MURA musculoskeletal radiograph benchmark [10].

For domains where large labelled datasets are scarce relative to natural images, self-supervised foundation models pretrained on domain-specific data have begun to outperform ImageNet-pretrained convolutional networks. RETFound is one such example: a Vision Transformer pretrained with masked autoencoding on a large corpus of retinal fundus photographs, which has been shown to generalise well across several ophthalmic disease-detection tasks after light fine-tuning [4]. The underlying Vision Transformer and masked-autoencoding techniques were introduced, respectively, by Dosovitskiy et al. [5] and He et al. [6].

Model interpretability for clinical use has largely converged on two complementary techniques. Gradient-weighted Class Activation Mapping (Grad-CAM) produces a coarse localisation map for convolutional and, with adaptation, transformer-based classifiers by back-propagating the gradient of the target class score to a chosen feature layer [7]. For tabular models, SHapley Additive exPlanations (SHAP) provide a game-theoretic feature attribution grounded in Shapley values, offering a consistent way to rank the contribution of each clinical variable to an individual prediction [8]. Separately, Guo et al. demonstrated that modern deep networks are frequently miscalibrated and proposed temperature scaling, a lightweight post-hoc correction that rescales the logits of a trained network using a single learned scalar without altering its decision boundary [9].

On the tabular side, gradient-boosted decision tree implementations such as XGBoost [11] and LightGBM [12] remain strong baselines for structured clinical data with a moderate number of features, and are frequently combined into voting or stacking ensembles to reduce variance. Class imbalance in clinical tabular datasets, such as the skew present in the Pima Indians Diabetes dataset, is commonly addressed with synthetic oversampling techniques such as SMOTE [13]. For dermatological screening and thoracic radiography, respectively, the ISIC/HAM10000 skin lesion archive [14] and enhancement studies on COVID-19 chest radiography [15] illustrate the dataset and preprocessing choices adopted for those tasks in this work.

III. System Overview

A. Design Objective

The system is designed as a screening and triage aid rather than a replacement for clinical judgement. Its intended role is to reduce time-to-first-assessment and to improve consistency across cases by surfacing a probability-ranked prediction, an interpretable explanation, and a derived risk tier for the treating clinician to review before any decision is made.

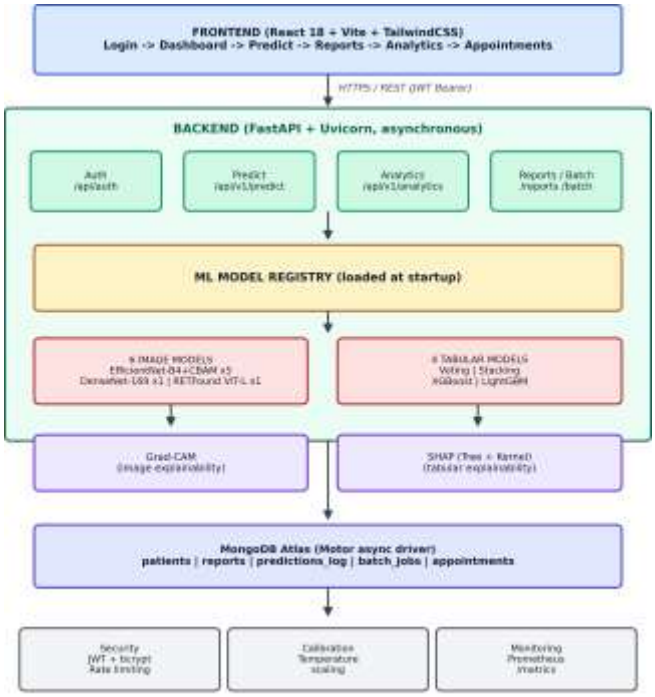

Fig. 1. System architecture of the multimodal disease-prediction platform.

Fig. 1. End-to-end system architecture, from the frontend through the FastAPI backend, model registry, explainability layer, and persistence tier.

B. Architecture

The platform follows a three-tier architecture. A React-based single-page frontend collects either an uploaded image or a structured clinical-data form and communicates with the backend over HTTPS. A FastAPI backend exposes versioned REST endpoints for authentication, prediction, batch prediction, reporting, and analytics, and hosts an in-process model registry that loads all ten trained models at start-up. A MongoDB Atlas document store persists patient records, prediction reports, monitoring logs, batch job state, and appointment data. Fig. 1 summarises the request path from image or form submission through inference, explanation generation, calibration, and persistence.

C. Core Capabilities

- **Multimodal input**: raster medical images (PNG/JPEG) or JSON-encoded structured clinical variables.
- **Ten-disease coverage** spanning oncology, pulmonology, ophthalmology, orthopaedics, and metabolic disease.
- **Per-prediction explainability**: a Grad-CAM heatmap for image-based predictions or a SHAP feature-attribution chart for tabular predictions.
- **Calibrated confidence** via post-hoc temperature scaling applied uniformly to all image classifiers.

- **Automated reporting**: server-rendered PDF clinical reports embedding the explanation artefact and a standard disclaimer.
- **Batch and single-case inference**, asynchronous job tracking, and a live analytics dashboard refreshed on a fixed polling interval.

IV. Disease Coverage and Risk Stratification

Table I summarises the ten supported prediction tasks, the associated input modality, and the model family used for each.

Each prediction is mapped to one of three risk tiers rather than a bare probability. Benign or negative outcomes (e.g. "No Tumour", "Normal", "Negative", and the corresponding benign dermatological classes) are always reported as low risk regardless of confidence, since a low-severity label with high model confidence carries different clinical weight than a malignant label with the same confidence. Positive or pathological outcomes are stratified further: calibrated confidence at or above 85% is reported as high risk, confidence between 60% and 85% as medium risk, and confidence below 60% as low risk. This mapping is intended to make the model's calibrated uncertainty directly actionable rather than requiring the clinician to interpret a raw probability.

V. Model Architectures

A. EfficientNet-B4 with Convolutional Block Attention

Five of the six image-classification tasks (brain tumour, pneumonia, skin cancer, the base eye-disease pathway, and COVID-19) use an EfficientNet-B4 backbone [1]. EfficientNet scales network depth, width, and input resolution jointly according to a single compound coefficient ϕ, subject to a constraint on the product of the three scaling factors, which was found through architecture search to give a better accuracy-to-compute ratio than scaling any one dimension in isolation. The B4 variant used here has approximately 19 million parameters and produces a 1792-channel feature map prior to global pooling.

Before global average pooling, a CBAM block [2] is inserted to re-weight the feature map along two axes in sequence. Channel attention first pools the feature map along the spatial dimensions using both average and max pooling, passes each result through a shared bottleneck multilayer perceptron, and combines them through a sigmoid gate to obtain a per-channel weight. Spatial attention then pools the channel-attended feature map along the channel dimension and applies a 7×7 convolution to obtain a single-channel spatial weight map. The two gates are applied multiplicatively in sequence:

$$X' = X \otimes \sigma\ \mathrm{MLP}(\mathrm{AvgPool}(X)) + \mathrm{MLP}(\mathrm{MaxPool}(X)) \quad (1)$$

$$X'' = X' \otimes \sigma\ \mathrm{Conv}_{7\times7}([\mathrm{AvgPool}(X'); \mathrm{MaxPool}(X')]) \quad (2)$$

The attended feature map is pooled and passed through a two-layer classification head with intermediate dropout. Because medical pathology is typically confined to a small image region, this attention mechanism both improves classification accuracy and yields sharper, more clinically interpretable Grad-CAM localisations than the unattended backbone.

B. DenseNet-169 for Fracture Detection

Bone-fracture detection uses a DenseNet-169 backbone [3], in which each layer l receives the concatenation of the outputs of all preceding layers, $x_l = H_l([x_0, x_1, \ldots, x_{l-1}])$. This dense connectivity pattern encourages feature reuse and preserves the fine edge and texture information that is diagnostically relevant to identifying a fracture line, which motivated its earlier adoption as the reference model for the large-scale MURA musculoskeletal radiograph benchmark [10]. The network produces a 1664-dimensional pooled feature vector, which is passed through dropout and a single linear layer to a binary fracture/normal output.

C. RETFound Vision Transformer for Retinal Screening

Fundus-photograph screening uses RETFound [4], a ViT-Large/16 Vision Transformer [5] pretrained with a masked-autoencoding objective [6] on approximately 1.6 million retinal fundus images. The input image is partitioned into a 14×14 grid of 16×16 patches, linearly embedded to a 1024-dimensional token sequence with a prepended classification token, and processed by 24 transformer encoder blocks with 16-head self-attention. Because self-attention operates globally across all patch tokens from the first layer, the model can relate distant anatomical structures — the optic disc, macula, and peripheral vasculature — in a way that a convolutional network's local receptive field cannot until much deeper layers, which is particularly relevant for retinal pathologies whose diagnostic signature depends on this global structure. Domain-specific masked-autoencoding pretraining on fundus images further gives the model a substantially better initialisation for this task than a generic ImageNet-pretrained network. A lightweight head (layer normalisation, a linear projection, GELU activation, and dropout) maps the classification-token representation to the five output classes.

D. Tabular Ensemble Models

All four tabular tasks share a common preprocessing pipeline consisting of feature standardisation followed by a classifier. Heart-disease prediction uses a soft-voting ensemble of gradient-boosted trees, a random forest, and a second gradient-boosting variant, in which the final class probability is the average of the three constituent estimators' probabilities. Diabetes prediction uses a LightGBM classifier [12] trained with class-imbalance correction and SMOTE oversampling [13] to address the skewed positive-class prevalence in the underlying Pima Indians dataset. Chronic kidney disease prediction uses an XGBoost classifier [11] on 24 clinical and laboratory features. Liver-disease prediction uses a stacking ensemble in which out-of-fold predictions from XGBoost and LightGBM base learners are combined by a logistic-regression meta-learner trained under five-fold cross-validation. Ensemble estimators that are not natively tree-based at the top level (the voting and stacking ensembles) are explained using a

TABLE I
DISEASE PREDICTION MATRIX

#	Disease	Modality	Model Family	Output Classes
1	Brain Tumour	MRI	EfficientNet-B4 + CBAM	4
2	Pneumonia	Chest X-Ray	EfficientNet-B4 + CBAM (+CLAHE)	2
3	Skin Cancer	Dermoscopy	EfficientNet-B4 + CBAM	7
4	Eye Disease	Fundus Photograph	RETFound ViT-Large	5
5	COVID-19	Chest X-Ray	EfficientNet-B4 + CBAM (+CLAHE)	4
6	Bone Fracture	Musculoskeletal X-Ray	DenseNet-169	2
7	Heart Disease	Tabular (13 features)	Soft-Voting Ensemble	2
8	Diabetes	Tabular (8 features)	LightGBM + SMOTE	2
9	Chronic Kidney Disease	Tabular (24 features)	XGBoost	2
10	Liver Disease	Tabular (10 features)	Stacking Ensemble	2

model-agnostic Kernel SHAP explainer, while the remaining tree-based models use the exact and substantially faster Tree SHAP explainer.

VI. EXPLAINABLE AI

A. Grad-CAM for Image Predictions

For image-based predictions, an explanation is produced using Grad-CAM [7]. Given the logit score S_c for the predicted class c, the gradient of S_c with respect to the activations A^k of a chosen convolutional (or, for the transformer model, patch-token) layer is global-average-pooled over the spatial dimensions to obtain an importance weight α_c^k for each feature channel:

$$\alpha_c^k = \frac{1}{Z} \sum_{i,j} \frac{\partial S_c}{\partial A_{ij}^k} \tag{3}$$

The class-discriminative localisation map is then the ReLU-rectified weighted combination of the feature maps,

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_c^k A^k \right), \tag{4}$$

which is resized to the input resolution and overlaid on the original image as a colour heatmap. The target layer differs by architecture: the first convolution of the final mobile inverted-bottleneck block for EfficientNet-B4, the final dense block for DenseNet-169, and the final transformer block (with patch-token gradients reshaped back to a spatial grid) for the RETFound Vision Transformer.

B. SHAP for Tabular Predictions

For tabular predictions, each feature's contribution is estimated using the Shapley value from cooperative game theory [8]. For a model f over a feature set N, the attribution of feature i on input x is its average marginal contribution over all possible feature subsets $S \subseteq N \setminus \{i\}$:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{5}$$

A positive ϕ_i indicates that the feature increased the predicted disease probability, while a negative value indicates the opposite. The top-ranked features and their signed contributions are returned alongside a short natural-language explanation and rendered as a horizontal bar chart.

VII. PREPROCESSING AND CALIBRATION

A. Image Preprocessing

All input images are converted to RGB, resized with bicubic interpolation, converted to a normalised tensor, and standardised using ImageNet channel statistics, consistent with the pretraining distribution of the underlying backbones. For chest radiography tasks (pneumonia and COVID-19), Contrast Limited Adaptive Histogram Equalisation (CLAHE) is applied prior to the standard pipeline. CLAHE divides the image into an 8×8 tile grid, equalises the histogram of each tile independently, and clips the contrast gain at a fixed limit before interpolating between tiles, which improves the visibility of subtle parenchymal findings such as infiltrates and ground-glass opacities without amplifying noise in otherwise uniform lung fields — a known failure mode of global histogram equalisation.

B. Confidence Calibration

Deep classifiers are known to produce systematically over-confident probability estimates [9]. To correct this, each image model is calibrated post-hoc using temperature scaling: the trained network's weights are frozen, and a single scalar temperature T is fit on a held-out validation set by rescaling the logits before the softmax,

$$\hat{p}_i = \text{softmax}(z_i / T), \tag{6}$$

with T chosen to minimise the negative log-likelihood of the validation labels and clamped to $[0.5, 5.0]$ to avoid degenerate solutions. Calibration does not change the model's predicted class, only the sharpness of its confidence distribution. In the present system, all calibrated temperatures except brain-tumour detection exceed 1, indicating that the uncalibrated models were overconfident and that softening their probability estimates improves the correspondence between stated confidence and empirical accuracy.

VIII. SYSTEM IMPLEMENTATION

A. Backend and Inference Pipeline

The backend is implemented in FastAPI with an asynchronous request lifecycle. On start-up, a model registry loads all ten trained models into memory; if a given model file is missing or fails to load, the registry falls back to a mock

predictor that preserves the expected response schema and clinical labels, ensuring that the API remains available for development and demonstration even without a full set of trained weights. An incoming prediction request is authenticated via JSON Web Token, checked against a per-IP rate limit, and routed to either the image-inference path (preprocessing, forward pass, calibration, Grad-CAM generation) or the tabular-inference path (feature mapping, ensemble prediction, SHAP attribution), after which the result is persisted to the report store, logged for monitoring, and returned to the client.

B. Data Storage

Patient records, prediction reports, monitoring logs, batch-job state, and appointment bookings are stored in MongoDB. The variable structure of a prediction record — an image prediction carries a Grad-CAM heatmap, while a tabular prediction carries a SHAP chart and a ranked feature list — maps naturally onto a schema-flexible document store without resorting to nullable columns or an entity-attribute-value layout. Monitoring and batch-job collections use time-to-live indexes to expire transient data automatically, and the analytics dashboard is served from aggregation pipelines computed over the report collection.

C. Security

Authentication uses HMAC-signed JSON Web Tokens with a configurable expiry, transmitted as a bearer token on all protected endpoints. Passwords are hashed with bcrypt, and a constant-time comparison against a dummy hash is performed when a submitted email does not match a known account, to prevent user enumeration via response-timing analysis. Request throughput is limited per client IP address, and Transport Layer Security is enforced on all connections to the managed database, using a certificate bundle to avoid platform-specific trust-store failures.

IX. Results

Table II reports the validation accuracy achieved by each disease-specific model. Image-based classifiers trained on well-represented, single-institution-style datasets (brain tumour, pneumonia, COVID-19) achieve the highest accuracies, while tabular models trained on comparatively small classical datasets (the 303-row heart-disease dataset and 768-row diabetes dataset) achieve moderately lower, though still clinically informative, accuracy. This pattern is consistent with the expected effect of training-set size and feature dimensionality on model performance, rather than an artefact of any single architectural choice.

Fig. 2 shows representative screens from the deployed platform: the AI-services catalogue presented to a clinician before a prediction is requested, and the resulting diagnostic output for a dermoscopic image, including the calibrated confidence score, the derived risk label, the recommended clinical action, and the accompanying explanation text.

Calibration analysis showed that five of the six image models required temperature values above 1.0 (ranging from approximately 1.13 to 1.47), consistent with the general finding that modern deep networks are overconfident prior to calibration [9]; the brain-tumour model was the only case requiring a temperature below 1.0, suggesting comparative underconfidence on that particular validation split. Applying these per-model temperatures leaves the predicted class unchanged while bringing the reported confidence closer to the model's empirical accuracy, which is a prerequisite for the confidence-based risk stratification described in Section IV to be clinically meaningful.

TABLE II
Validation Accuracy by Disease and Model

Disease	Model	Accuracy
Brain Tumour	EfficientNet-B4 + CBAM	99.2%
Pneumonia	EfficientNet-B4 + CBAM	97.5%
COVID-19	EfficientNet-B4 + CBAM	97.1%
Eye Disease	RETFound ViT-Large	96.4%
Kidney Disease	XGBoost	96.0%
Bone Fracture	DenseNet-169	95.3%
Skin Cancer	EfficientNet-B4 + CBAM	95.8%
Heart Disease	Voting Ensemble	92.0%
Liver Disease	Stacking Ensemble	91.0%
Diabetes	LightGBM + SMOTE	88.5%

X. Discussion and Limitations

The current system has several limitations that are relevant to any deployment beyond a demonstration setting. First, all models are static and trained on fixed, publicly available datasets; performance may degrade under distribution shift caused by different imaging equipment, acquisition protocols, or patient population characteristics not represented in training. Second, the system currently accepts only PNG/JPEG image input, whereas clinical picture archiving systems predominantly use the DICOM format. Third, tabular feature sets are limited to the variables available in the underlying public datasets (for example, thirteen features for heart disease), which is substantially narrower than the variable set available in a typical electronic health record. Fourth, only the image models are currently temperature-calibrated; the tabular ensembles do not yet receive post-hoc probability calibration. Finally, authentication in the present implementation uses a fixed set of demonstration accounts rather than a production-grade, role-based user-management system.

These limitations motivate several directions for future work: native DICOM support; multi-disease panels predicted jointly from a single scan (for example, screening for pneumonia, COVID-19, and pleural effusion from one chest radiograph); active-learning workflows that route low-confidence predictions to clinician review and feed corrected labels back into retraining; integration with hospital systems via HL7/FHIR interfaces; federated training across multiple institutions without centralising patient data; uncertainty quantification through techniques such as Monte Carlo dropout or deep ensembles; pixel-level lesion segmentation in addition to classification; and extension of post-hoc calibration to the tabular ensemble models.

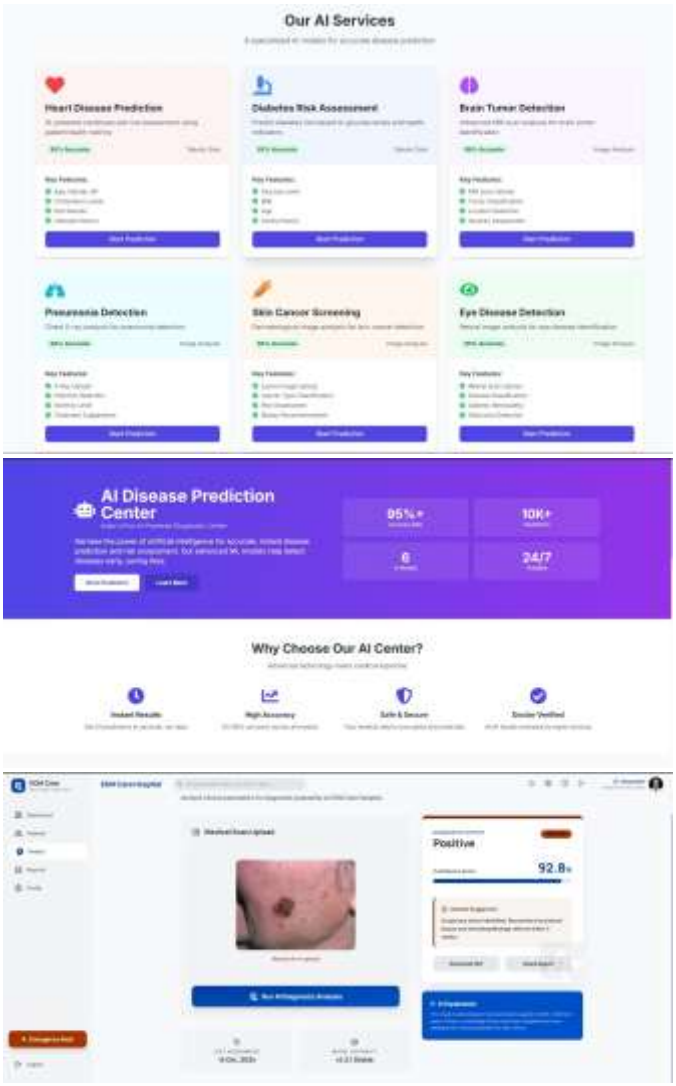

Fig. 2. Platform screens. Top: the AI-services selection view listing the ten disease-prediction tasks. Middle: the AI Disease Prediction Center overview. Bottom: a sample diagnostic output screen for a skin-cancer prediction, showing the calibrated confidence score, risk label, and clinical suggestion.

XI. Conclusion

This paper described a multimodal, multi-disease clinical decision support system that combines attention-augmented convolutional networks, a domain-pretrained Vision Transformer, and gradient-boosted ensemble models under a common prediction, explanation, and reporting pipeline. By pairing each prediction with a modality-appropriate explanation (Grad-CAM or SHAP) and a calibrated, risk-stratified confidence score, the system is designed to support rather than replace clinical judgement. Reported validation accuracies across the ten supported tasks range from 88.5% to 99.2%, and the accompanying production architecture — authentication, rate limiting, monitoring, and automated reporting — addresses several of the operational requirements for deploying such a system in a real clinical workflow. The identified limitations, particularly around DICOM support, dataset distribution shift, and tabular-model calibration, define concrete next steps for future development.

References

[1] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019.
[2] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.
[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.
[4] Y. Zhou, M. A. Chia, S. K. Wagner, *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, pp. 156–163, 2023.
[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.
[6] K. He, X. Chen, S. Xie, *et al.*, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.
[7] R. R. Selvaraju, M. Cogswell, A. Das, *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.
[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017.
[10] P. Rajpurkar, J. Irvin, R. L. Ball, *et al.*, "MURA: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, 2017.
[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 2016.
[12] G. Ke, Q. Meng, T. Finley, *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[14] N. C. F. Codella *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv preprint arXiv:1803.10417*, 2018.
[15] T. Rahman, A. Khandakar, Y. Qiblawey, *et al.*, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Comput. Biol. Med.*, vol. 132, 2021.