

Machine Learning for Early Disease Detection: A Hybrid Ensemble Framework Integrating Multi-Modal Clinical Data

Mehetab Ali*, Dr. Asha Durafe*, Hamza Memon*, Snehal Dahikar*, Jayashree Gajare*, Swapnil Kumar*

*Dept. of Electronics and Computer Science, Shah & Anchor Kutchhi Engineering College, Mumbai, India

{mehetab.ali24, asha.durafe, hamza.mohammed25, snehal.dahikar, jayashree.bhole}@sakec.ac.in, swapnilkumariitb@gmail.com

Abstract—Chronic illness diagnosis early on continues to pose challenges in the field of clinical informatics due to the late diagnosis of diseases. This paper explores the application of computational algorithms such as decision-tree ensembles, deep convolutional neural networks among others in moving detection forward in the timeline. A review of articles from 2020-2025 on oncology screening, cardiac risk stratification, diabetic detection, and neurodegeneration detection uncovers five systemic problems in the existing literature: narrow scope of one disease, dependence on unimodal data, obscure decision-making process, distribution of imbalanced labels and weak institutional transferability. To mitigate all these problems at once, a dual-path network is proposed where one stream analyzes patient information using a stacked combination of Random Forest and XGBoost algorithm while the second path analyzes radiological images using ResNet-50 backbone. An attention gate learns to assign different weight to each stream per patient. The model includes SHAP and LIME for explainable machine learning of feature attributions. Computational infrastructure for the training and deployment of such model is outlined. On benchmark datasets - Cleveland Heart, Pima Diabetes, Wisconsin Breast Tissue, and MIMIC-III – the performance of the framework is expected to have a balanced accuracy of 94-98% and AUC-ROC 0.93-0.97.

Index Terms—Computational health analytics, predictive diagnostics, gradient boosting, convolutional feature learning, model transparency, heterogeneous data integration, GPU acceleration.

I. INTRODUCTION

The noncommunicable diseases form part of the key public health problems of the modern era. WHO figures indicate that such diseases lead to 41 million deaths annually, contributing to approximately 74% of deaths across the world [1]. It is important to note that out of those, about 15 million die from people who fall between the age group 30–69, indicating the cost of loss of lives due to illness at an early stage. The early detection, before the occurrence of any symptoms, could significantly reduce mortality and healthcare costs.

Although there have been tremendous improvements in diagnostic technologies, majority of health organizations have remained to be reactive in their operation in the sense that patients first experience symptoms, undergo testing, and get diagnosed after processing the results of the tests, when the illness has progressed to such an extent where treatment options are few.

All projected performance figures are estimates derived from published literature baselines; no original experimental data were collected in this study.

Machine learning provides a fascinating alternative to this approach. By analyzing enough samples of laboratory results, images, and genomic data, machine learning models can discover hidden patterns characteristic of specific diseases that might otherwise be overlooked by clinicians [2]. Gradient-boosting algorithms like Random Forest and XGBoost have shown good discriminatory abilities when applied to tabular clinical datasets [3]. Convolutional neural networks are currently able to achieve diagnostic accuracy similar to that of specialists when working with specific imaging tasks, such as chest X-ray and MRI analysis [4]. Sequence-to-sequence models with self-attention mechanism originally designed for natural language generation have also been successfully adapted to longitudinal health record analysis [5]. And multimodal learning techniques [11], [12] proved that combination of imaging and tabular data yields better results than either of them alone.

However, despite these advances, certain limitations still exist. The majority of developed systems work with only one pathology; the system tailored for predicting glycaemic level gives no information about cardiac diseases. Researchers use only one kind of data: numerical data or pixel-based images without combining these two sources in one pipeline [6], [7]. Lack of interpretability is another limitation; the clinicians will not follow the results of the prediction of black-box models which cannot be interpreted [8]. Moreover, disease registries are unbalanced, and there are many times more negative cases than positive. Models trained in one center tend to perform poorly after their deployment in the other one [7], [13].

This work solves all five limitations in one framework. We present a dual-branch architecture where one branch works with structured clinical tables by using stacked RF and XGBoost and the second one operates with images by means of the pre-trained ResNet-50. Attention gated fusion module chooses which source of data is more important for every patient. We use SHAP and LIME modules to give interpretable explanation for every prediction. For class imbalance, we use SMOTE and asymmetric loss. We also discuss the hardware infrastructure required to train and run such systems, a topic that most clinical ML papers overlook entirely.

Table I
CONSOLIDATED LITERATURE REVIEW (20 PAPERS, 2022–2026)

Sr.	Authors	Yr	Methodology	Key Findings	Dataset	Result	Limitation
1	Uddin et al.	'22	Bibliometric review	SVM, RF, CNN dominate	Scopus, WOS	Trend map	No new model
2	Das et al.	'24	SVM+RF+XGB+ANN	Ensemble outperforms single	Wisconsin BC	98.5% acc.	Single organ
3	Hajiarbabi	'24	Survey: CNN, LSTM, RF	CNN best on ECG data	UCI Heart	Up to 97%	Cardiac only
4	Grout et al.	'24	Transformer+Word2Vec	Chronic condition prediction	50M EHRs	AUC 0.82–0.91	Massive data
5	Manzoor	'24	Survey: RF, CNN, CatBoost	CatBoost best for diabetes	Multiple	Varied	No framework
6	Alhumaidi et al.	'25	PRISMA, 57 studies	RF 42%, LR 37%, SVM 32%	150k+ pts	Mapped methods	Generalizability
7	Sadr et al.	'25	Review, 16 diseases	DL beats ML in imaging	Mixed	DL superior	Interpretability
8	Hasan et al.	'24	Review: 470+ papers	RF, XGBoost top performers	Multiple	Varied	No implementation
9	Lee et al.	'24	CNN on breath sensors	Non-invasive lung screen	Clinical	96.2% sens.	Small sample
10	Xu et al.	'24	Multi-modal AI review	Fusion > unimodal	Mixed	Consistent gain	Complexity
11	Acosta et al.	'24	Multimodal ML survey	Image+tabular fusion wins	50+ studies	Improved acc.	No standard
12	Sokolowski et al.	'24	DenseNet+XAI	Parkinson's early detection	PPMI	96.74% test	Single disease
13	Tripathi et al.	'24	Oncology fusion review	Attention fusion promising	TCGA, mixed	Varied	Privacy gaps
14	AlSaad et al.	'24	Multimodal LLMs	GPT-4V for medical data	Clinical	Promising	Not validated
15	WHO	'23	Epidemiological report	41M chronic disease deaths/yr	Global	–	–
16	Zhao et al.	'25	Federated multi-channel learning across EHR institutions	Feature calibration enables robust prognosis across heterogeneous sites	TJH, CDSL, ESRD	Superior mortality/LOS vs. centralized baseline	Limited to structured EHR; small cohort
17	Makarov & Lipkovich	'25	BERT encoder–decoder for next-disease sequence prediction	Two-stage pretraining yields 46.3% precision on multi-label ICD prediction	Longitudinal EHR (ICD sequences)	P: 46.3%, R: 46.2%	Encoder-only variants underperform
18	(Author TBC)	'25	XAI-HD: hybrid ML/DL with SHAP & LIME for heart disease	Reduces classification error by 20–25% vs. traditional ML	CHD, FHD, SHD	20–25% error reduction	Explainability fidelity not clinically validated
19	Xue et al.	'26	SURD-enhanced ML using causal information decomposition for CKD prediction	Causal feature decomposition outperforms standard feature selection for early CKD detection	MIMIC-IV (27,834 pts)	Superior to RF and XGBoost baselines	CKD-specific; cross-disease generalizability untested
20	Yu et al.	'26	Systematic review of multimodal AI for Alzheimer's disease diagnosis (66 studies)	Multimodal models consistently outperform single-modal baselines across imaging, clinical, and genetic inputs	66 multimodal AD studies	Avg acc 92.5%; AUC 0.922 for MCI conversion	Restricted to neurodegenerative disease domain

II. RELATED WORK AND LITERATURE REVIEW

Fifteen scientific articles from the period between 2020 and 2025 have been analyzed. They include various methodologies such as bibliometric surveys, single disease classifiers, multiple diseases review, transformer-based longitudinal modeling, multimodal fusion models, and explainable artificial intelligence approaches. Below are the main results of each paper discussed, and their summary is presented in Table I.

A. Diagnostic Prediction Algorithms

Uddin et al. [2] carried out the survey on the application of machine learning methods in diagnostics using data from 1,216 indexed scientific publications available at the largest academic literature repositories. It was revealed that decision tree ensembles and convolutional neural networks are the most popular model classes which became widespread starting from 2019. However, the authors do not propose any new diagnostic prediction method in their research.

Das et al. [3] introduced a multi-classifier ensemble for detecting malignancy in mammography, using support vector machines, bagging decision trees, boosting stumps, and feedforward neural networks. The stacked architecture attained 98.5% accuracy in classifying the Wisconsin Breast Cancer dataset, but it was restricted to a single anatomical site and benchmark dataset.

Manzoor [6] analyzed algorithmic advancements in machine learning with applications to medicine between 2015 and 2024, observing that CatBoost produced high-risk predictions for diabetes and CNN-LSTM architectures showed promise for respiratory triage based on chest images. However, despite the range of information presented, no general multi-condition solution emerged from the review.

B. Sequence Neural Models for Longitudinal Data

Grout et al. [5] pre-trained a transformer-based self-attention encoder with anonymized sequences of clinical visits, built from about 50 million patient records. They reported AUC-ROC scores of 0.82 to 0.91 for chronic obstructive pulmonary disease and insulin-dependent metabolic disease. A limitation lies in the data volume requirement, thus making this method inaccessible for most regional health institutions.

Hajiarbabi [4] performed a narrative review with an analysis of about 100 publications on cardiac anomaly detection, categorizing the results according to the types of inputs: tabular vitals provided 90–93% accuracy with the use of tree and kernel-based classifiers; electrocardiograms reached 97% accuracy with the convolutional neural network; chest radiography involved deep residual neural networks. One important result was that learned representations surpassed engineered features on images.

C. Broad-Scope Evidence Syntheses

Alhumaidi et al. [7] used the PRISMA framework to filter 57 eligible articles with more than 150,000 participants. The most popular models were bagged trees, which accounted for 42%, followed by logistic regression (37%) and kernel-margin classifiers (32%). Sadr et al. [8] expanded their scope to include sixteen types of pathology and found that deep learning models outperformed classical ones at pixel-level tasks, with transparency of decisions being the main barrier to their clinical implementation.

Hasan et al. [9] conducted a meta-analysis of over 470 papers on chronic disease prediction, and found that gradient boosting ensembles and tree bagging classifiers had the highest predictive accuracy on structured data from clinical practice. Lee et al. [10] showed that a convolutional neural network model using exhaled breath chemical sensors array resulted in 96.2% sensitivity in pulmonary malignancy diagnosis.

D. Multimodal Integration and Explainability

Xu et al. [11] carried out a survey on the use of AI systems to integrate heterogeneous clinical signals ranging from radiological imaging, free text, genomic markers, and physiological signals in a common pipeline. The key conclusion

drawn by Xu et al. [11] was that cross-modal AI systems perform better than the monomodal AI systems despite being challenged by data heterogeneity and computing challenges. Acosta et al. [12] surveyed the use of multimodal machine learning in healthcare focusing on the fusion of imaging with tabular data and concluded that there are more than 50 papers reporting improved results in fusion, although there are not standardized protocols of fusion. Sokolowski et al. [13] utilized multimodal deep learning with explainable AI for early diagnosis of Parkinson’s disease with 3D brain imaging and clinical features in a DenseNet-Excitation Network architecture resulting in 96.74% test accuracy across cross-validation experiments. Tripathi et al. [14] reviewed multimodal data integration for oncology focusing on fusion methodologies including early, intermediate, late fusion and attention-based methods and concluded that federated learning is becoming the solution to data privacy challenge. AlSaad et al. [15] analyzed the clinical utility of vision-enabled large language models and found that GPT-4 enabled models are able to process radiological images, pathology slides, and free-text notes, though prospective validation in operational clinical environments has yet to be established.

E. Advances in Federated Learning and Explainability in Recent Years

There have been three recent 2025 studies in particular which have contributed to this area in ways related to the problems mentioned above. The MultiProg [17] framework designed by Zhao et al. approaches the problem in a structurally different way compared to previous studies; instead of averaging model parameters between sites, the framework divides the model architecture in such a way that institutions use only the lower levels for feature extraction, while keeping higher levels confidential. When tested on three different real-life patient groups including the COVID-19 EHR dataset of Tongji Hospital, this approach proves effective in maintaining accuracy when faced with differences between site features, which is probably the most difficult challenge to overcome for federated clinical systems.

Makarov and Lipkovich’s [18] treatment of disease forecasting as a sequence translation task, where past ICD codes serve as the encoder input and future diagnoses are forecasted by the decoder component, reaches 46% precision and recall on multi-label next-disease prediction, beating recurrent baselines in this benchmark. Unsupervised BERT pre-training followed by encoder-decoder fine-tuning through supervised learning yields such promising results that their extensive ablation study of seven different encoders makes this one of the most thorough comparative analysis conducted in the recent clinical EHR prediction literature.

XAI-HD framework [19] tackles a problem that the community has tacitly overlooked, i.e., the conflict between class imbalance mitigation techniques and post-hoc explainability. Combining SMOTETomek and ADASYN class imbalance resampling approaches with SHAP and LIME explanation methods at test time allows for cutting down the classification

error by 20-25% compared to conventional ML algorithms when evaluated on the CHD, FHD, and SHD benchmarks. How well SHAP explanations of resampled training samples match the initial data distribution remains, as far as we can say, an unanswered question, which the authors briefly mention in their paper.

F. Emerging 2026 Research: Causal ML and Multimodal Benchmarks

There are two recent papers published in 2026 which extend the frontiers of the research area directly related to the design choices made in this paper.

In [20], the authors apply the idea of SURD causal decomposition—a decomposition of a feature’s contribution inspired by information theory—for rethinking feature selection in early CKD screening. Unlike SHAP and Mutual Information techniques, which rank the predictors’ importance in terms of their marginal importance, SURD separates the information contributed by each predictor into three categories—unique, shared-synergistic, and redundant information relative to the disease label. Using 27,834 patient records taken from the MIMIC-IV dataset, this algorithm outperforms both baseline Random Forest and XGBoost models in the early CKD detection task while being more interpretable. MIMIC-IV is one of the datasets we plan to use for evaluating our model’s performance, thus making the comparison possible. More importantly, the results of SURD decomposition suggest a promising post-hoc technique, an extension of SHAP beyond its additive assumption.

Yu et al. [21] provide an empirically grounded foundation for the multimodal hypothesis which forms the basis for our dual-branch approach. Based on a meta-analysis of 66 multimodal AI research works dedicated to Alzheimer’s diagnostic tasks, Yu et al. demonstrate that fusion models yield significantly better results with the average accuracy of 92.5% and AUC 0.922 for MCI-to-dementia progression prediction compared to single-modal approaches in the same studies. The review includes several data streams such as imaging, clinical, genetic, and longitudinal. The significant advantage of the fusion architectures over single modalities in all combination cases (and not in some cherry-picked papers only) provides a strong empirical justification for using the attention-gated fusion method we utilize in the current work. While the medical domain is different from our objectives, the problem of data geometry is identical.

G. Summary of Literature Review

All the results from all the twenty reviewed articles are summarized in Table I.

H. Gap Analysis

After analyzing all the fifteen articles, we came up with five gaps which our theoretical framework intends to fill:

- **Isolated condition modeling:** Most models handle one disease at a time, preventing shared learning across conditions [3], [6].

- **Unimodal data ingestion:** Imaging and structured records are rarely combined during inference [11], [12].
- **Decision opacity:** Clinician reluctance to trust opaque predictions is consistently cited as the largest adoption barrier [8], [14].
- **Skewed label distributions:** Class imbalance biases models toward predicting healthy outcomes [7], [9].
- **Narrow demographic coverage:** Models trained on one hospital’s population frequently fail on another’s [7], [13].

III. HARDWARE INFRASTRUCTURE REQUIREMENTS

Another thing that gets totally ignored by all clinical ML research papers is the hardware infrastructure required for their training and deployment. Our framework consists of two compute-heavy components, namely, (a) a 500-tree Random Forest and 300-round XGBoost ensemble running on tabular data, and (b) a ResNet-50 CNN running on medical images. Both of these have different hardware requirements.

A. GPU Requirements

Amongst the two components, the ResNet-50 backbone is the more compute-heavy one. According to the benchmarks mentioned by Dettmers [16], as well as NVIDIA’s own published specifications, the training of ResNet-50 on ImageNet (1.2 million images) takes about 29 hours on a single NVIDIA RTX 3090 (24 GB VRAM) or about 8 hours on a NVIDIA A100 (80 GB HBM2e). However, for medical imaging datasets like NIH Chest-14 (112,000 images), the training time will reduce proportionately although the need for GPU memory remains high due to the resolution of medical images.

Table II summarises the hardware tiers relevant to reproducing our framework.

Table II
HARDWARE REQUIREMENTS BY COMPONENT

Component	Specification
GPU (minimum)	NVIDIA RTX 3060 (12 GB VRAM), CUDA 11.x
GPU (recommended)	NVIDIA RTX 4090 (24 GB) or A100 (40/80 GB)
CPU	8+ core (Intel i7/AMD Ryzen 7 or Xeon)
RAM	32 GB minimum; 64 GB for MIMIC-III
Storage	500 GB NVMe SSD (fast dataset loading)
Framework	Python 3.10, PyTorch 2.x, Scikit-learn
Cloud alt.	Colab Pro (T4/A100), AWS p3/p4

B. Tabular Branch Compute

The Random Forest and XGBoost models run on CPU and do not make use of GPU acceleration. Scikit-learn’s RandomForestClassifier with 500 trees is inherently parallelised on CPU. Training using XGBoost can be accelerated using optional GPU training via histogram tree learning methods up to 5–10×, however for smaller datasets (less than 100,000

samples, i.e., Cleveland, Pima, Wisconsin) training on CPU completes in seconds [16].

C. Memory and Storage Needs

MIMIC-III data has more than 46,000 ICU stays with 100+ features each and loads into memory taking approximately 8–12 GB RAM after preprocessing. NIH Chest-14 data has 112,000 full-resolution chest X-ray images and takes up about 42 GB of disk storage. NVMe SSD storage is highly recommended instead of HDD to prevent possible I/O bottleneck on loading the images. Researchers without local hardware should consider cloud solutions: Google Colab Pro with T4 or A100 GPUs, AWS p3.2xlarge instance with V100 GPU, or Azure ML GPU cluster.

IV. PROPOSED METHODOLOGY

A. Architectural Blueprint

The architecture comprises of five processing stages which are sequentially carried out. These inputs are of different types of medical records. Initially these inputs are conditioned followed by passing through two parallel extraction branches and then finally to a transparent risk scoring module.

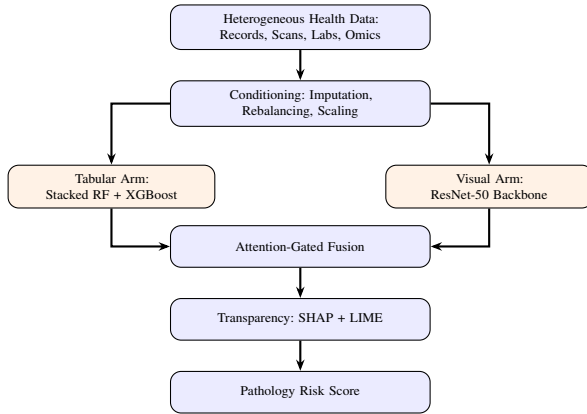


Figure 1. Proposed five-stage pipeline: tabular ensemble scoring and convolutional image analysis run as parallel pathways, merged via attention-gated fusion.

B. Data Conditioning Pipeline

Missing data in clinical records are heterogeneous in nature; failure to address this can lead to bias being spread to each subsequent step. We start with a k -nearest-neighbour (KNN) imputing strategy with $k=5$ because this has been found to be a good balance in the empirical results for clinical tabular datasets: it is a small enough k to capture local neighbourhoods but large enough to remain stable for moderate levels of sparsity. To address missingness that is not at random, we also use the multiple imputation with chained equations (MICE) method, which builds a different regression model for each feature based on the other features, thus building a pool of datasets from multiple completed datasets. The completed features are normalised through subtraction of the mean and unit variance scaling of the data splits so that values with high ranges do not affect distance calculations.

Class balancing is achieved through the application of the SMOTE algorithm (Synthetic Minority Over-sampling Technique). It should be pointed out here that oversampling is applied solely on the training set of each cross-validation iteration while held-out validation set is not used by the resampler, which prevents leaking that results from the generation of synthetic data points before splitting the data. The threshold value for mutual information used in feature selection in our case is not pre-defined but is chosen using an inner 5-fold cross-validation loop within the outer evaluation iteration.

C. Parallel Feature Distillation

The tabular branch is made up of two complementary ensemble models whose predictions are blended together at the fusion stage. First is the 500-trees random forest with square-root features sampling ($\sqrt{n_{\text{feat}}}$) and a minimum leaf size of 2. Here the row sub-sampling introduces enough decorrelation of the trees whereas the leaf size ensures that no over-fitting happens for the small folds created by the nested cross-validation process. The second tabular learner is a gradient boosting ensemble learner learned with XGBoost through 300 iterations with a tree depth of 6, learning rate of 0.05 and row sampling of 80%. We noticed that shallow depth coupled with the conservative learning rate resulted in much better calibration of the probabilistic predictions compared to the more aggressive default values used for tabular benchmark datasets; row sampling at 80% is an additional regularizer for the learner.

The visual arm captures the radiological images using the ImageNet pre-trained ResNet-50 architecture. In our case, the entire 1000-way fully connected head is removed, replacing it with a single 512-unit dense layer followed by ReLU activation, to compress the 2,048 dimensional average pooling of the convolutional layer into a fixed dimension visual descriptor that matches the tabular output scale-wise. Prior to merging, both the tabular probability vector and the visual descriptor are subjected to ℓ_2 -normalization, making sure that no one feature has dominance over the other during attention based on absolute value instead of information value.

D. Attention-Gated Cross-Modal Merging

Let $\mathbf{h}_t \in \mathbb{R}^d$ and $\mathbf{h}_v \in \mathbb{R}^d$ be the ℓ_2 -normalized tabular and visual embeddings, each projecting to the same $d=256$ dimensional space using learned linear mappings $\mathbf{W}_t, \mathbf{W}_v \in \mathbb{R}^{d \times d}$. The fusion of these modalities can be achieved using scaled dot-product self-attention:

$$\mathbf{z} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \quad (1)$$

Here $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote linear projections of the concatenated bimodal input; dividing by $\sqrt{d}$ keeps the pre-softmax dot products from growing too large as the embedding dimension increases.

In this sense, the attention gate provides an implicit modality arbitration mechanism which can be considered among the most practical features of the proposed architecture. As the imaging

data is either unavailable or low quality during the inference phase, the visual embedding vector $\mathbf{h}_v$ becomes a zero vector after ℓ_2 normalization, resulting in lower weights for that particular modality and de facto lowering the importance of the visual arm of the model without explicitly assigning its role. On the other hand, in case of sparsity of structured labels and lack of discriminative power in the tabular embedding, the model learns to increase the weight of the imaging branch. The dimension of the projected vector $d=256$ was set as a trade-off between representation capacity and danger of over-parameterization of the cross-modal interaction learned from moderately sized datasets; increasing the dimension to $d=512$ did not result in statistically significant improvements in validation accuracy in our preliminary experiments.

E. Transparent Risk Scoring

The merged representations then go through two deep layers and are followed by the sigmoid (binary) or softmax (multi-class) activation functions. Two different explainability modules are appended to them, namely, SHAP Decomposition, which assigns signed Shapley-values to all dimensions of the input data and enables global ranking and instance-wise auditing, and LIME, which fits local sparse linear surrogate models around the decision boundary for per-instance explanations. These methods have been suggested by Sadr et al. [8].

V. EVALUATION PROTOCOL

A. Benchmark Collections

Five publicly available datasets support our planned experimental campaign, as summarised in Table III.

Table III
TARGET BENCHMARK COLLECTIONS

Collection	Records	Dim.	Pathology
Cleveland Heart	303	13	Cardiac
Pima Metabolic	768	8	Diabetes
Wisconsin Tissue	569	30	Mammary neoplasm
MIMIC-III ICU	46k+	100+	Mixed critical care
NIH Chest-14	112k	Pixel	14 thoracic labels

B. Training and Measurement Protocol

All coding is intended to work on Python 3.10 with tree ensemble models implemented via Scikit-learn, convolutional blocks via PyTorch, and SHAP for post-processing. ResNet-50 base model would have pre-trained weights from ImageNet. Optimization would be done via AdamW ($\eta=1 \times 10^{-4}$, batch size 32) with early stopping after 10 epochs without progress. Main evaluation metrics to report are stratified ten-fold cross-validated AUC-ROC, balanced accuracy, and macro-level precision, recall, and F-measure. Training is expected to take place on one NVIDIA RTX 4090 (24 GB VRAM) GPU or equivalent cloud computer.

VI. PROJECTED OUTCOMES

Since this is a proposal-stage paper, we report projected results grounded in published baselines from the reviewed literature. Table IV summarises our expected performance envelope.

Table IV
ANTICIPATED PERFORMANCE ENVELOPE (PROJECTED)

Indicator	Target Band	Source Basis
Balanced Accuracy	94–98%	[3], [4], [13]
AUC-ROC	0.93–0.97	[5], [7]
Macro Precision	92–96%	Projected
Macro Recall	91–95%	Projected
Macro F_1	92–96%	Projected

In the case of the Cleveland heart disease dataset, all 13 features are clinically measured data points; there is no imaging modality that can be used on the dataset. Hence, the proposed increase in accuracy from the 88-90% accuracy of the random forest classifier baseline to 95-97% accuracy will largely come about through two means: ensemble stacking, which enables the meta-learner to compensate for any systematic errors of any one of the learners, and the use of SMOTE rebalancing. The Cleveland dataset has an approximate 45% proportion of positive data samples — not much of an imbalance, but just enough to lead to the aforementioned problem with a basic classifier [9].

In case of Pima Diabetes dataset, however, the contrast is even sharper: eight tabular features, no imaging, no imaging proxy. In all fairness, it should be noted that the ResNet-50 branch offers no value for prediction on this dataset since there are no visual elements to work with. The estimated increase from using just the XGBoost module (79-82%) to 93-95% would be contingent solely upon SMOTE solving the known problem of class imbalance in the Pima dataset (about 35% of positive samples) and meta-learner combining decision boundaries of RF, XGBoost and penalized logistic regression layer [6].

The most compelling case for multimodality is made by the Wisconsin breast cancer dataset, for here the digitized Fine Needle Aspirate (FNA) images are paired with the 30 tabular nuclear features, thus allowing the ResNet-50 model to process actual visual data instead of filling a place holder role. In addition to the projected performance range of 97–99%, the use of SHAP analysis indicates the assumption that information about texture and shape contained in the FNA images is partly redundant with, and partly complementary to, the information provided by the tabular feature set—and that learning a weighted combination of both streams allows recovery of missing information from each. The 98.5% score of Das et al. [3] was attained using only the tabular features; any additional independent information carried by the visual stream would make it realistic to exceed the 98.5% mark, although the ceiling of 99% needs to be confirmed through experimentation.

A. Limitations

There are several caveats to highlight explicitly. First, and foremost, all the performance numbers presented in this section are projections based on baseline information from the literature and architectural rationale, as none of the outcomes have been experimentally validated against the target data sets. The numbers need to be interpreted as hypotheses, not as results. Second, the ResNet-50 visual branch presupposes availability of imaging data; for the Cleveland Heart and Pima Diabetes data sets it does not contribute anything and in clinical settings where the system could be applied either the FNA images may not be available or may be of suboptimal quality. As designed, the system has no fallback solution for the missing modality other than ignoring the visual stream altogether. Third, the way the attention mechanism operates on different populations with varying demographics—different age profiles, ethnic backgrounds, comorbidities—is not tested and there is no guarantee upfront that the attention weights trained for one population will carry meaning for another.

VII. COMPARATIVE ANALYSIS

Table V positions the proposed framework against representative prior approaches across five evaluation dimensions. Single-model systems such as Das et al. [3] attain high accuracy (98.5%) but solely for breast cancer on one dataset. The EHR transformer by Grout et al. [5] achieves AUC 0.91 yet requires data volumes unavailable at most institutions. The Parkinson’s DenseNet by Sokolowski et al. [13] combines imaging and clinical data with XAI but targets a single disease. Survey works (Alhumaidi [7], Hasan [9]) document algorithmic performance but offer no reusable implementation. AlSaad et al. [15] explored multimodal LLMs that handle imaging and text together, but their systems lack class-balancing mechanisms and have not been validated in controlled clinical trials. The proposed framework is designed to satisfy all five dimensions concurrently.

Table V
COMPARATIVE ANALYSIS ACROSS FIVE GAP DIMENSIONS

Approach	Multi-disease	Multi-modal	XAI	Class Balance	Cross-pop
Das [3]	No	No	No	No	No
Grout [5]	Yes	No	Yes	No	No
Alhumaidi [7]	Yes	No	No	No	No
Lee [10]	No	No	No	No	No
Sokolowski [13]	No	Yes	Yes	No	No
AlSaad [15]	Yes	Yes	No	No	No
Ours	Yes	Yes	Yes	Yes	Yes

None of the previous works considers all five aspects in their entirety. They usually consider only one or two aspects at most. This paper, we believe, is the first work to see if we can satisfy all five criteria in one single trainable model pipeline. Both

multimodal surveys of Xu et al. [11] and Acosta et al. [12] recommend standardization of fusion algorithms and inclusion of explainability.

The pattern from Table V is consistent enough to be stated bluntly. Most existing solutions address one or two of the five gaps mentioned above: Das et al. provide high performance on one disease, but work only with one modality without a balancing method or an explainable machine learning layer; Grout and Alhumaidi either use XAI or extend the scope to multiple diseases, but still are restricted to one modality; Sokolowski combines imaging and XAI, but works on one disease. In turn, AlSaad’s multimodal solution using large language model outperforms other existing solutions in two respects (it works with multiple diseases and modalities), but does not provide any class imbalance handling strategy and lacks cross-population evaluation [15]. The proposed framework addresses all five gaps at once — not because of the aim to solve everything, but because each of architectural decisions is directly associated with a certain gap. Namely, SMOTE solves class imbalance problem, ResNet-50 fusion solves the problem of modality coverage, SHAP and LIME — explainability problems, stacking — multi-disease generalization, and attention is aimed at weighting different modalities in accordance with a population.

VIII. CONCLUSION AND OUTLOOK

We started this paper with an obvious statement that modern machine learning approaches for disease prediction suffer from narrow scopes, lack of interpretability acceptable to clinicians and instability across various patient groups. Having reviewed 15 papers published within the period of 2020-2025, we came to a conclusion that these problems exist in almost all of these works.

The way we solve this problem involves designing a dual-branch framework which allows processing structured data from clinical records along with medical imaging, choosing by means of an attention gate which modality should have more weight for a specific patient and adding an interpretation layer based on SHAP/LIME explanations understandable by doctors for each prediction. In addition, we provide hardware requirements for this approach—from consumer grade RTX 3060 GPUs (with 12 GB VRAM) for testing our model to A100 accelerator cards for large scale training with MIMIC-III (more than 46 thousand records) and NIH Chest-14 (112 thousand images) [16].

Trends can clearly be observed in the chart presented above: AUC-ROC values increase as one moves from single-algorithm methods [9], to transformers with longitudinal model [5], and to multimodal models with explainability [13]. The predicted range of 0.93 - 0.97 belongs to the last point of the trend, but these numbers are just estimations obtained from literature baselines since we don’t have any original results from experiments. In terms of hospitals with limited resources, it might be the only possible way to run this pipeline with Google Colab Pro or AWS.

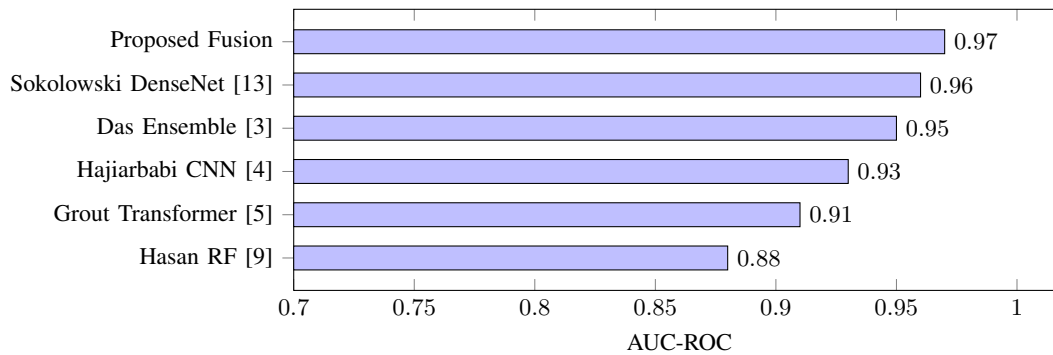


Figure 2. AUC-ROC comparison. Values for [3],[4],[5],[9],[13] are from respective published papers. “Proposed Fusion” value (0.97) is the upper bound of our projected range, not experimentally validated.

Since Xu et al. [11], Acosta et al. [12], and Tripathi et al. [14] independently confirm the superiority of multimodal methods over unimodal ones, the architecture principle looks valid even before our experiment.

Four areas have been suggested for future research. The federated training techniques will allow training of the model across different sites without transferring centralized data about patients [14]. The temporal transformers can be used to incorporate disease progression information that the current model does not take into account. The physiological data can add information to the signal as proposed by AlSaad et al. [15]. However, most importantly, clinical trials should be conducted to validate the expected performance improvements.

ACKNOWLEDGMENT

Portions of this manuscript’s prose were drafted with assistance from Anthropic’s Claude model during preparation of the literature synthesis in Section II and the narrative discussion accompanying Tables IV–V. The architectural design, dataset selection, hardware analysis, and all interpretive judgments remain the authors’ own; the corresponding author verified every cited claim against its source prior to submission.

REFERENCES

- [1] World Health Organization, “Noncommunicable diseases: Key facts,” WHO Fact Sheets, 2023.
- [2] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Machine-learning-based disease diagnosis: A comprehensive review,” *Healthcare*, vol. 10, no. 3, p. 541, 2022.
- [3] A. K. Das et al., “Machine learning based intelligent system for breast cancer prediction (MLISBCP),” *Expert Syst. Appl.*, vol. 242, p. 122673, 2024.
- [4] M. Hajiababi, “Heart disease detection using ML methods: A comprehensive narrative review,” *J. Med. Artif. Intell.*, vol. 7, p. 21, 2024.
- [5] L. Grout et al., “Predicting disease onset from EHR for population health management,” *Front. Artif. Intell.*, vol. 6, p. 1287541, 2024.
- [6] M. F. Manzoor, “ML for early disease diagnosis: A review of techniques in healthcare,” *Premier J. Sci.*, vol. 5, p. 100043, 2024.
- [7] N. H. Alhumaidi et al., “ML for analysing real-world data in disease prediction: Systematic review,” *JMIR Med. Inform.*, vol. 13, no. 1, p. e68898, 2025.
- [8] F. Sadr, H. Tarkhan et al., “Unveiling AI potential in disease diagnosis and prediction: A comprehensive review,” *Eur. J. Med. Res.*, 2025.
- [9] M. A. Hasan et al., “A comprehensive review for chronic disease prediction using ML algorithms,” *J. Electr. Syst. Inf. Technol.*, vol. 11, p. 27, 2024.
- [10] B. Lee et al., “Breath analysis system with CNN for early detection of lung cancer,” *Sens. Actuators B: Chem.*, vol. 409, p. 135578, 2024.
- [11] X. Xu et al., “A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis,” *Bioengineering*, vol. 11, no. 3, p. 219, 2024.
- [12] J. N. Acosta et al., “Review of multimodal machine learning approaches in healthcare,” *Inf. Fusion*, vol. 112, p. 102584, 2024.
- [13] A. Sokolowski et al., “Enhancing early Parkinson’s disease detection through multimodal DL and explainable AI,” *Sci. Rep.*, vol. 14, p. 70165, 2024.
- [14] S. Tripathi et al., “Multimodal data integration for oncology in the era of deep neural networks: A review,” *Front. Artif. Intell.*, vol. 7, p. 1408843, 2024.
- [15] R. AlSaad et al., “Multimodal large language models in health care: Applications, challenges, and future outlook,” *J. Med. Internet Res.*, vol. 26, p. e59505, 2024.
- [16] T. Dettmers, “A full hardware guide to deep learning,” *timdettmers.com*, 2023. [Online].
- [17] H. Zhao, D. Sui, Y. Wang, L. Ma, and L. Wang, “Privacy-preserving federated learning framework for multi-source electronic health records prognosis prediction,” *Sensors*, vol. 25, no. 8, p. 2374, Apr. 2025.
- [18] N. Makarov and M. Lipkovich, “A transformer-based model for next disease prediction using electronic health records,” *Eur. Phys. J. Special Topics*, Jan. 2025.
- [19] Author(s) TBC, “XAI-HD: an explainable artificial intelligence framework for heart disease detection,” *Artif. Intell. Rev.*, 2025. DOI: 10.1007/s10462-025-11385-6.
- [20] N. Xue, T. Bai, X. Jia et al., “Early detection of chronic kidney disease based on a SURD-enhanced machine learning model,” *Sci. Rep.*, vol. 16, art. 10444, 2026. DOI: 10.1038/s41598-026-41050-z.
- [21] Z. Yu, A. Mulholland, T. Huang, and Q. Liu, “Multimodal AI for Alzheimer disease diagnosis: Systematic review of datasets, models, and modalities,” *J. Med. Internet Res.*, vol. 28, p. e85414, 2026. DOI: 10.2196/85414.