Int. j. adv. multidisc. res. stud. 2025; 5(6):1710-1719

Received: 07-11-2025
Accepted: 17-12-2025

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

An Effective Ensemble Multimodal Learning System for Medical Diagnosis

Thirumalaimuthu Thirumalaiappan Ramanathan
Faculty of Engineering and Technology, Centre for Advanced Analytics, CoE for Artificial Intelligence, Multimedia University, Melaka, Malaysia

DOI: https://doi.org/10.62225/2583049X.2025.5.6.5475 Corresponding Author: **Thirumalaimuthu Thirumalaiappan Ramanathan**

Abstract
Multi-modal learning allows the integration of heterogeneous data modalities to achieve better prediction performance on clinical datasets. In this paper, we extend our approach to multimodal deep learning models and consider a presence of three tabular datasets: Stroke Prediction, Framingham Heart Study (FHS), Chronic Kidney Disease (CKD) that are split into demographics, clinical and laboratory measurements. We design three multimodal schemes (i.e., MDNN, MAE, and Transformer based fusion) to extract the intra- and inter-modality feature interaction adaptively. For reference, we also compare with a single-stream deep model (Early Fusion MLP) and traditional machine learning classifiers such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Machine (GBM), Support Vector Machines (SVM), and Extreme Gradient Boosting Machine (XGBoost). The model performance is measured through these three metrics: accuracy, F1-score and area under the ROC curve (AUC). This study demonstrates how MDNN, MAE, and Transformer based fusion are effective in structured multimodal settings. The results of our study suggest that multimodal learning is an effective strategy for exploiting diverse medical data types in clinical prediction problems, by effectively utilizing modality-specific feature representation.

Keywords: Machine Learning, Deep Learning, Multimodal Deep Neural Network, Multimodal Autoencoders, Multimodal Transformers, Medical Diagnosis

1. Introduction

In the last few years, machine learning has grown in use for both healthcare and biomedical research especially with respect to predictive analytics applied on tabular datasets. A major specialization of approaches is that they tend to consider all features equally, and do not take into account the (intrinsic) differences among feature types and sources. However, a lot of datasets do possess multi-modalities, e.g., demographic history, clinical examination data and laboratory test data. Using these heterogeneous data sources as a whole can increase the reliability and accuracy of predictive models [1].

Multimodal learning offers a new type of paradigm which explicitly leverages the information from heterogeneous source to obtain complementary knowledge [2]. In the context of medical datasets, demographic features might be based on age, gender and socio-economic status that serve as baseline risk information. Physiological measurements, such as vital signs and comorbidities, capture up-to-the-minute health status of the patient, whereas lab results reflect biochemical processes. The two modalities provide complementary knowledge, and the fusion of information makes models more accurate.

One of the key advantages of multimodal learning over tab single task is that it can take advantage of complementary modalities information [3]. For example, a patient with borderline hypertension may have normal clinical readings but abnormal lab values, which suggests early disease. These hidden patterns may be missed by a unimodal model, while a multimodal model can do the joint analysis of both signals to enhance the detection sensitivity and specificity. Such a combination would improve predictive performance and decrease the probability of false negatives.

Beyond prediction task, multimodal learning can make model more robust and generalizable [4]. Utilization of data from multiple independent sources is expected to prevent the model from overfitting to noise in a single modality. For instance, laboratory measurements may have occasional outliners or measurement errors, and clinical evaluations can differ between individuals. Adding demographic information and cross-modal correlations enable the model to keep good performance even if some attributes are noisy or missing [5].

Multimodal fusion also allows for more interpretability and clinical applicability [6-10]. Through disentangling, merging of features from different modalities, clinicians and researchers are able to investigate which kind of data is the most relevant for a prediction [11]. For instance, a model could report that an interaction of high glucose (lab) and old age (demographics) significantly increases cardiovascular risk, giving actionable information that might be hidden in single-modality models.

Implementation wise, there are multiple ways to mix modalities in tabular dataset. Early fusion aggregates the various modalities into one input matrix, upon which a classical/dl model learns joint representations. Late fusion fine tunes modality-specific encoders and combines learned embeddings for downstream prediction [12]. More sophisticated architectures, such as attention-based transformers or multimodal autoencoders, can give a dynamic weight to each modality and thus capture complex inter-modal relationships.

Another benefit of multimodal learning is that they can treat missing data in a better way [13-15]. Note that, in the clinic, certain lab or clinical measurements can be missing for some patients. With the other modality information, the model can still provide useful prediction instead of simply throwing away incomplete phoneme records. This approach is data-driven, and it maximizes the use of all available data which is a highly desirable property for medical applications with potentially small sample sizes.

Multimodal learning provides an appealing tool for tabular data in healthcare and beyond. Through combining information on demographics, clinical presentation and laboratory findings, these models are capable of encapsulating complementary descriptors that translate into better predictive performance, reduced model sensitivity to perturbations, as well as clinically meaningful profiles. As the size and complexity of tabular datasets increase further, strategies that are multimodal offer a potential to extract maximum value from diverse data sources for more accurate and dependable decision-making.

The primary objective of this study is to test the ability of multimodal learning in clinical datasets, where demographic, clinical and laboratory details would be integrated. We first introduce and compare classical machine learning models, the early fusing MLP and specialized multimodal neural networks (MDNN, MAE, Transformers based fusion). We experiment on three well-known benchmark datasets – Stroke Prediction [16], Framingham Heart Study (FHS) [17], and Chronic Kidney Disease (CKD) [18] to examine the prediction performance and robustness of our methods. We show that combining information from multiple modalities leads to performance improvement over single modality baselines, and ensemble strategies further boost model prediction. The remainder of the paper is structured as follows, Section 2 gives a review of related works about multimodal learning and tabular data modelling; Section 3 introduces the dataset and methodology used in our experimentation; In Section 4 we make an analysis on the results obtained by comparing them with previous ones from other models and in section 5 some conclusions and future work are presented.

2. Related Work

Below shows detailed review about application of multimodal learning in healthcare datasets.

Ou *et al*. (2022) [19] develop a dual-encoder deep neural network: one branch for the smartphone-captured lesion images and another for metadata (age, lesion location etc.) with a fusion module that utilizes intra- and cross-attention. Their study trained and validated using the 5- fold cross validation public dataset of skin lesions captured through smartphones. Their fused model achieves average accuracy of 0.768 ± 0.022, balanced accuracy 0.775 and AUC 0.947 that are superior to models based on image alone. But the metadata is relatively superficial (no lab tests or detailed clinical measurements), and image quality from smartphones can vary greatly. The sample could be biased towards those patients with good camera access, which may limit generalisability.

Soenksen *et al*. (2022) [20] proposed HAIM (Holistic AI in Medicine), a modular framework based on AI algorithms able to carry out multimodal input processing (from tabular to image, time series and text data); the contribution of each modality also is quantified using Shapley values. They considered a variety of health prediction tasks using hospital data, and show that combining modalities provides better performance than any single-modality model. The dataset contains EHR (tabular), sensor/time-series and medical imaging data from diverse, large-scale hospital cohorts. One strength of HAIM is its easy and flexible architecture and interpretability by modality importance. We found the disadvantages that are variable results across tasks and insufficient contribution of lab-test–specific for lab + demographic fusion only learning. Their method is effective for general clinical application; however, additional work remains to validate on lab-intensive clinical prediction tasks.

Haleem *et al*. (2023) [21] presented two generative models, a Temporally Correlated Multi-modal GAN (TC-MMGAN), and a Document Sequence Generator, for generating realistic multi-modal health data including sensor, physical record and text data with preserved temporal relationships. They evaluated the models on a use-case from H2020 GATEKEEPER project, showing that synthetic data preserve cross-modal and temporal dependencies. They focused their analysis on data similarity as opposed to predictive performance because the experiments are designed to generate data. And because it is a synthetic task, downstream predictive utility would still be speculative and potential biases or artifacts may not be equivalent to those seen in real-world clinical data. Further, their technique does not explicitly fuse lab examinations or demographic variables as separate modalities. Yet it's useful in simulation, privacy-preserving data generation, and augmenting multimodal datasets.

Preetha *et al*. (2023) [22] presented a deep learning model that allows real-time synthesis of multimodal healthcare data which is composed of structured health variables (vitals, etc.), and textual or physical-record type modalities by means of their simple Tensor-based generative architecture. Their method intends to mimic raw patient trajectories in an evolving environment, for the purpose of training and testing downstream predictive models without requiring access to actual patient data. The dataset scenario is abstract/synthetic and thus does not originate in a actual clinical population –- as such the evaluation focuses on the behavior of generated data identification (fidelity) and overall statistical properties. A drawback is the lack of actual lab-test fusion; the synthetic variables were not

validated against true biological values. A second limitation here is that the absence of ground truth disease labels can limit utility for disease prediction tasks despite usefulness of synthetic data as augmented training.

Gao *et al*. (2024) [23] proposed a study that combines structures admission clinical notes (history of present illness, physical exam etc.) and structured tabular clinical variables (including lab values) by using multi-modal deep architecture with bi-directional text encoders based on BERT model and fully connected tabular encoder with gated attention fusion. It draws upon three large public-access databases: MIMIC-III, MIMIC-IV and eICU CRD MySQL and includes internal, prospective and external validation datasets. The performance of their multimodal model obtains AUROC of 0.838, 0.849 and 0.767 on internal prospective and external sets respectively, superior to that by unimodal baselines. Liabilities include a lack of scalability (text + tabular), and the possibility that some laboratory or demographic variables are under-weighted; in addition, missing time series dynamics of labs is not well modelled. The interpretability component (Integrated Gradients, SHAP) is a plus which helps reveal which features (text vs tabular) drive risk prediction the most.

Li *et al*. (2024) [24] created a predictive model based from structured EMR data (demographics, clinical information) with hysteroscopic images to predict conception following surgery. Their method combines a CNN over image features with a classical machine learning (deep) model for EMR, then concatenates results or performs decision level fusion. The data originates from hysteroscopy in a clinical cohort of women, and the study reports better predictive ability when compared to EMR or image only. Their study limitations include potential overfitting of the model, small surgical population, lack of detailed lab test characteristics in EMR and limited generalizability to other non-gynecologic patient groups.

Wang *et al*. (2024) [25] approached with a multimodal risk-prediction model that integrated physiological temporal signals (e.g., vital signs), chest X-ray images and clinical notes with the use of early, joint and late fusion approaches and based on SHAP to evaluate the importance of each modality. They applied their framework to a hospital EHR cohort and predict outcomes such as in-hospital mortality, length of stay (LOS), or 30-day readmission. They find that multimodal models outperform unimodal baselines, and structured vital-sign data often contributes more than images or notes to the prediction. Disadvantages include restricted explicit handling of lab-test tabular data (lab values are not a distinct modality) and the fused operation overhead might prevent real-time targeting deployment. Further, interpretability between modalities is non-trivial and the impact of demographic factors is not extensively studied.

Imrie *et al*. (2025) [26] introduced AutoPrognosis-M, a general-purpose automated machine-learning (AutoML) framework that combines structured tabular clinical data and medical images through various fusion techniques (early, joint, late) assembled as an ensemble. They applied the model on a skin lesion dataset (smartphone images + clinical records) and showed that their multimodal ensembles were able to reach an accuracy of 80.0% for 6-class categorization of lesions and 89.8% accuracy for binary cancer diagnosis, out-performing unimodality baselines. Advantages are auto selection of pipeline (feature engineering model/hyperparameter search) and ability to estimate uncertainty and explain with conformal prediction and XAI. However, limitations of their study are that they fail to concentrate on modalities richer than image and metadata fusion, the general-purpose nature leads to less specialization around certain clinical tasks, and compute/resource requirements for AutoML and ensemble may be prohibitive in a real-world context.

Kumar *et al*. (2025) [27] introduced a cross-modal transformer for tuberculosis detection which combines chest x-ray along with clinical/tabular data. In their approach, chest X-ray images are transformed into visual tokens, while clinical and lab/tabular data are transformed into text tokens; they are then jointly processed by a bidirectional cross-attention transformer to learn a unified representation used for classification. The data is a set of paired images and clinical records from patients at Government Medical College Uttarakhand, India (n=3,256) which includes the lab tests they've had, demographic details and symptoms as well as whether or not TB was detected. The performance of the proposed model was high with an accuracy rate of up to 97.78%, which were significantly superior to the conventional early-fusion, late-fusion and the template models. But there are limitations of their study which is lack of mass publicly available dataset for testing and there is risk for selection bias.

Tölleetal. (2025) [28] developed a scalable open-platform workflow, to facilitate federated learning systems across multiple hospitals by constructing multi-modal harmonised datasets, using DICOM-structured reports and other such data modalities. It does so by concentrating on standardizing representation of data through DICOM, exercising cohort-query filtration across eight university hospitals and facilitating federated dataset curation as opposed to pooling. They used their dataset as the multi-modal (imaging, structured reports, metadata) records of eight German university hospitals to predict outcomes after minimally invasive heart-valve replacement. Instead of a traditional classification accuracy and to the best of our knowledge, this is the first evidence showing filtering, harmonisation and federated readiness together across sites. The main limitation is that their study focuses on dataset creation and infrastructure, rather than end-task prediction model for demo/clinical/lab tab fusion (no standard accuracy metrics are reported) and the real time-series fusion of demos, lab, clinical is still under investigated.

Multi modal learning has demonstrated promising results in the fusion of heterogeneous data from different domains, however a majority of the clinical studies are based on a single modality model or reduced feature set. To the best of our knowledge, there are few studies that asses complicated multimodal models, such as Multimodal Deep Neural Networks (MDNN), Multimodal Autoencoders (MAE), and Transformer-based fusion on multimodal datasets of clinical/lab/demographic feature collections. Also, many comparisons to classic machine learning techniques over different datasets are not present making the general improvement of such a model uncertain. There is also scarce literature regarding the advantages of integrating deep multimodal models on a task level. To bridge these gaps, this paper evaluates the predictive capabilities of MDNN, MAE, and Transformer-based fusion on three distinct clinical datasets: Stroke Prediction, FHS, and CKD. The work provides a comprehensive comparison between

single-modality and multimodal deep models and traditional machine learning classifiers. We also examine the results of an unweighted ensemble fusion of the three deep models for possible gains. The goal of this study is to provide insights into how multimodal deep models can outperform state-of-the-art ones on tabular prediction tasks using clinical, lab and demographic information.

3. Materials and method

3.1 Dataset description and preprocessing

In practice, we used three public datasets to compare the efficiency of multimodal learning models: Stroke Prediction [16], FHS [17], and CKD [18] datasets. These datasets include demographic, clinical and laboratory information, and collectively form a rich multimodal environment for predictive modelling. We hope to capture the complex relationships between patient-related factors, physical data as well as laboratory markers by leveraging multimodal features. Each dataset introduces its own kind of features heterogeneity, missing values and multiple scales, what makes them good candidates for evaluating performance of multimodal fusion architectures such as MDNN, MAE or Transformers-based models. The details of each dataset, including the features and dimensionality are given in the following paragraphs.

The dataset of stroke patients [16] with clinical and demographical data for predicting whether an individual is at risk of having a stroke. Features in the dataset are age, gender, hypertension, heart disease, glucose level average, BMI, work type, residence type, marital status, and smoking habits. The dataset is relatively small as opposed to the clinical datasets, however it is clear-cut concerning potential risk factors predicting stroke events. The target variable is stroke for whether a patient has had a stroke. Demographics are lifestyle and socio-economic factors, clinical features represent the current health state. This dataset is well suited to the evaluation of multimodal learning and represents combination of heterogeneous information from different sources. Imputation of missing values and encoding of categorical features is required prior to modelling.

The FHS dataset [17] has about 4,240 patient records and includes important demographic, behavioural and clinical variables. Features such as patient's age, sex, education level (years), smoking status (current or not), average number of cigarettes per day consumed by the patient, medical information including blood pressure (systolic and diastolic), cholesterol, BMI, glucose heart rate medication use for treating hypertension; history of stroke, hypertension and diabetes. The response variable, CHD tells us whether the patient had a coronary heart disease after ten years. As the database covers demographics, clinical risk factors, and lab-like measurements, it is a promising candidate for multimodal learning —in particular models combining demographic, clinical, lab modalities. There are also some missing values which need be replaced.

The CKD dataset [18] consists of 400 patients, and measures including clinical tests and a series of lab measurements in addition to their history of the disease, with an input size of 24 features and also labels indicating if they are suffering from CKD or healthy. In addition to numerical attributes such as age blood pressure, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume and white blood cell count we have categorical features like measurements of urine, details of red and white blood cells, pus cells, bacteria but also on diseases like hypertension: (yes/no), diabetes: (yes/no), coronary artery disease:(yes/no), appetite:(good/poor), pedaledema (yes/ no) and anemia (with common or poor). There are NAs over various features, so preprocessing steps such as imputation are required before modelling. Of these patients, 250 are classified as CKD while the remaining 150 are not, hence we have a reasonably imbalanced binary class problem. This dataset plays an important role in machine learning studies on CKP diagnosis and has been reported elsewhere. It is open source under a permissive license and has become a de facto benchmark for investigating data analysis and predictive modelling in the medical domain.

Each dataset: Stroke Prediction, FHS, and CKD are separated into demographics, clinical and laboratory datasets. It employs one-hot encoding for categorical features and z-score normalization for the numerical attributes. Impute missing values using median imputation. Let X_m^{num} and X_m^{cat} be numerical and categorical columns with modality. The preprocessing is formalized as shown in (1).

$$\tilde{X}_m = \text{Concat}(\text{StandardScaler}(X_m^{\text{num}}), \text{OneHotEncoder}(X_m^{\text{cat}})) \quad (1)$$

This in turn guarantees that the embeddings of MDNN, MAE and Transformer-based fusion are modalities-invariant.

3.2 Method

In this study, we present a deep learning multimodal approach for medical data with respect to clinical, lab and demographic modality. Multimodal data integration is important in clinical reasoning, due to the inherent heterogeneity of patient information. Every modality encodes specific aspects: demographic data encode age, gender and socio-economic context; clinical observations document vital signs or patient history; laboratory analysis collects measurements of the biochemical nature. Combining these sources can provide broader insight into patient health and improve predictions in high-stakes applications, including disease onset prediction, risk stratification, and sepsis alerting. Our proposed framework is designed to explore the MDNN, MAE and Transformer-based fusion, in single and ensemble form for outcome prediction on three benchmark datasets: Stroke Prediction, FHS and CKD.

Figure 1 shows the proposed architecture for multimodal learning based on MDNN, MAE and Transformer based fusion. The architecture is intended to acquire modality-specific representations, to fuse them properly and enable accurate predictions. The overall model pipeline includes three basics stages: (i) modality-specific preprocessing and feature extraction (ii) modality encoding and fusion, and (iii) classification and evaluation. Let $X = \{X_1, X_2, \ldots, X_M\}$ represent the input modalities, where is the number of modalities. Each modality $X_m \in \mathbb{R}^{N \times d_m}$ has 'N' samples and 'd_m' features. The goal is to find a mapping function 'f' that predicts labels '$\hat{y}$' from these heterogeneous inputs as shown in (2).

$$\hat{y} = f(X_1, X_2, \ldots, X_M) \quad (2)$$

Where $\hat{y} \in [0,1]^N$ for binary classification. The pipeline ensures that the modality-specific patterns are retained while the network can capture cross-modal interaction.

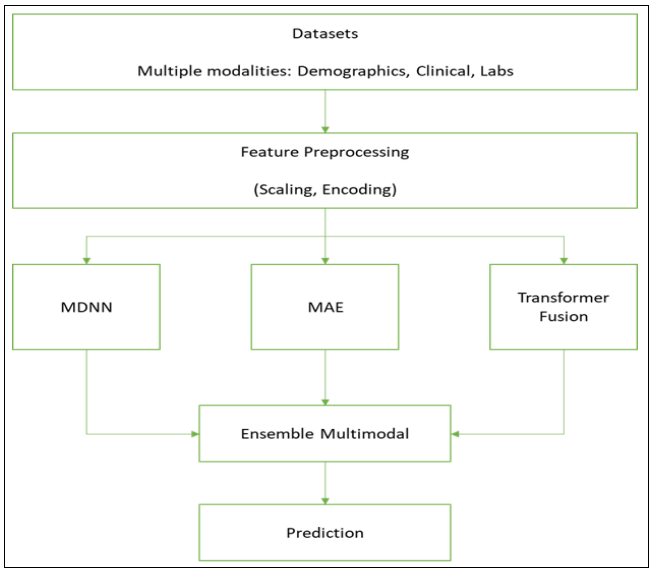

Fig 1: Ensemble multimodal learning architecture

3.2.1 Multimodal Deep Neural Network (MDNN)

The MDNN model captures each modality separately by using a modality-specific feedforward network. For each modality X_m, a shallow encoder E_m learns to map the input into a latent encoding h_m as given in (3).

$$h_m = E_m(X_m) = \sigma(W_m X_m + b_m) \tag{3}$$

Where $W_m \in \mathbb{R}^{d_m \times d_h}$ and $b_m \in \mathbb{R}^{d_h}$ are the learnable parameters, assuming d_h as hidden embedding dimension and $\sigma(\cdot)$ as non-linear activation (ReLU). Dimensionality reduction is performed on each modality embedding $h_m \in \mathbb{R}^{N \times d_h}$ to capture the modality specific features. After the embeddings for all modalities are acquired, they are concatenated as given in (4).

$$H = [h_1 \parallel h_2 \parallel \cdots \parallel h_M] \tag{4}$$

The concatenated embedding $H \in \mathbb{R}^{N \times (M \cdot d_h)}$ is fed through a fully-connected classifier network as given in (5).

$$\hat{y} = \text{sigmoid}(C(H)) = \text{sigmoid}(W_c H + b_c) \tag{5}$$

Where, $W_c \in \mathbb{R}^{(M \cdot d_h) \times 1}$ and $b_c \in \mathbb{R}$ are classifier parameters. The training loss of MDNN is binary cross-entropy as shown in (6).

$$\mathcal{L}_{\text{MDNN}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{6}$$

The most advantageous point of MDNN is its simplicity and the capability to preserve modality-specific information, while allowing joint learning by simple concatenation.

3.2.2 Multimodal Autoencoder (MAE)

MAE extends MDNN by introducing a reconstruction objective that enforces embeddings to preserve available useful information in every modality. Like MDNN, each modality X_m is transformed into a coded representation as shown in (7).

$$h_m = E_m(X_m) \tag{7}$$

Every embedding is subsequently decoded to reconstitute the original input as shown in (8).

$$\hat{X}_m = D_m(h_m) \tag{8}$$

We define the reconstruction loss for modality 'm' as mean squared error as shown in (9).

$$\mathcal{L}_{\text{recon}}^{m} = \frac{1}{N}\sum_{i=1}^{N} \parallel X_m^{(i)} - \hat{X}_m^{(i)} \parallel_2^2 \tag{9}$$

The overall reconstruction loss for all modalities is shown in (10).

$$\mathcal{L}_{\text{recon}} = \sum_{m=1}^{M} \mathcal{L}_{\text{recon}}^{m} \tag{10}$$

For classification, the embeddings are concatenated like MDNN as shown in (11).

$$H = [h_1 \parallel h_2 \parallel \cdots \parallel h_M], \hat{y} = \text{sigmoid}(C(H)) \tag{11}$$

The joint loss for MAE combines classification and reconstruction terms as shown in (12).

$$\mathcal{L}_{\text{MAE}} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{recon}} \tag{12}$$

Where λ balances reconstruction and classification. MAE preserves modality-specific information and enhances the generalization.

3.2.3 Transformer-based fusion

The Transformer-based fusion model aims to learn cross-modal dependencies via self-attention. Each modality is initially encoded to a sub-space using a small MLP encoder as shown in (13).

$$h_m = \text{MLP}_m(X_m) \tag{13}$$

The final embeddings are given by (14).

$$H_{\text{stack}} = [h_1, h_2, \ldots, h_M] \in \mathbb{R}^{M \times N \times d_h} \tag{14}$$

Finally, the stacked embeddings are fed to a Transformer encoder that captures inter-modal relations as shown in (15).

$$H_{\text{fused}} = \text{TransformerEncoder}(H_{\text{stack}}) \tag{15}$$

Estimation of self-attention mechanism is given by (16)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{16}$$

Where 'Q', 'K', 'V' are linear projections of H_{stack}, and d_k is the embedding size. The output is tuned by being averaged over the modalities, then it becomes an input of MLP classifier as shown in (17).

$$\hat{y} = \text{sigmoid}(\text{MLP}(H_{\text{fused}})) \tag{17}$$

This way, cross-modal interactions are explicitly modelled thereby capturing potentially complicated dependencies that simple concatenation cannot account for.

3.2.4 Ensemble Strategy

Given the proposed models, in order to achieve better results on prediction, we form an ensemble by averaging the prediction from MDNN and MAE as well as the transformer-based model as shown in (18).

$$\hat{y}_{\text{ensemble}} = \frac{1}{3}(\hat{y}_{\text{MDNN}} + \hat{y}_{\text{MAE}} + \hat{y}_{\text{Transformer}}) \quad (18)$$

The joint model captures the strength of each model. In short, MDNN adapts modality-specific embeddings, MAE maintains the reconstruction consistency and Transformer based fusion models capture inter-modal interactions. This mixture then seeks the balance between variance reduction and increasing robustness. All models are trained with the Adam optimizer with rate of learning 1×10^{-3} and binary cross-entropy loss. In MAE, the reconstruction term is weighted by λ. Training is performed with mini-batches of 64 examples and early stopping according to validation AUC. Formally, the optimization objective is given by (19).

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\hat{y}, y) \quad (19)$$

Where θ consist of all learnable parameters among modality encoders, decoders, and transformer layers and classifiers.

Our architecture offers a flexible and composable architecture for multimodal tabular data. MDNN learns modality-specific embeddings, MAE maintains the reconstruction consistency and Transformer-based fusion designs the inter-modal interactions. Prediction robustness is even further improved through ensemble averaging. In this framework, various strategies for combining heterogeneous clinical, laboratory and demographic data can be evaluated to gain an understanding of how well deep learning methods outperform classical machine learning baselines.

4. Results and Discussion

Figure 2 and Table 1 show the performance comparison of multimodal learning models on medical datasets. As shown in Tables 1-4, we performed a large-scale analysis of multimodal learning methods on three biomedical datasets—Stroke Prediction, FHS and CKD. The methods used included classical classifiers, early fusion deep learning strategies, multimodal deep neural networks, autoencoder based fusion and transformer based fusion architectures. The key aim was to compare whether combining different data types such as demographics, clinical information and lab parameters in the models turns out to perform any better than unimodal ones (demography only/clinical features or lab exams onlay) or standard "all-feature" approaches.

The dataset was first pre-processed so that missing values in the target column were treated, and the following modalities were defined according to domain knowledge. For example, demographic, clinical and biological characteristics were classified individually for each set of data. This modular design enabled the training of statistics models on both modalities and combined modality information. All the numeric features were scaled, and categorical variables were one-hot encoded to be compatible between models.

The preprocessing stage identified some inherent difficulties in real-world biomedical data. For instance, some modalities were only partially available or missing completely, which had an impact on the training of multi-modal deep architectures. These issues highlight the role of solid data collection and potentially leveraging imputation strategies or domain-dictated features.

In single modality analyses, the conventional classifiers Logistic Regression, Decision Tree, Support Vector Machine (SVM), Naive Bayes, Multi-Layer Perceptron (MLP), Random Forest, AdaBoost, Gradient Boosting and XGBoost were adopted. Each model was developed separately from other models for each modality, e.g., demographics or clinical data. The performance criteria were accuracy, precision, recall, F1-score and area under the curve of receiver operating characteristic (AUC).

Based on the outcomes, it can be seen that single-modality models had mixed performance on datasets. For example, demographic features alone typically had less predictive power that those with feature richness diminished (clinical or laboratory modalities) for CKD dataset. However, a few classifiers, like ensemble based on trees and gradient boosting demonstrated resilience with small feature sets even yielding good AUC values. This suggests that the non-linearity constructed by ensemble methods are better suited for tabular biomedical data.

To set the performance baseline, features from all modalities were concatenated to form a single feature matrix, and the same split of training tested machine learning classifiers applied. The results consistently showed better predictions compared to single-modality models, which indicates the importance of multi-modal data as more information is derived across modalities. Original models (ensemble ones) such as Random Forest, Gradient Boosting and XGBoost achieved slight better performance when compared to the other machine learning classifiers stressing out the importance of capturing complex non-linear relationships in biomedical data.

Notwithstanding the better performance, classical machine learning solutions do suffer from several drawbacks. They are not good at handling multi-modal data since they need a complicated pre-processing for the feature, and do not consider modality electromagnetic structures explicitly, in addition to being fragile with respect to correlations between features and thus suffering from inappropriate integration of heterogeneous data.

The Early Fusion MLP method was used to feed the multi-modality as a single concatenated input vector and let CNNs learn relations across modalities in an implicit way. The fully connected neural network would capture complex interactions between features in different modalities, a mechanism not perfectly utilized by traditional machine learning models. However, early fusion assumes that all modalities are fully available and at a similar scale and it may be sub-optimal when the modalities have different informativeness or missing-data patterns.

The late fusion strategy of MDNN is as follows: First, each modality is fed into a separate encoder, and the obtained embeddings are further concatenated to perform classification. Such a structure permits modality-specific feature extraction, thus enhancing the learning of complementary representations for each modality. MDNN showed consistently good performance across all datasets, even outperformed the early fusion models in terms of AUC and F1-score. The architecture is modular and allows for flexibility in dealing with missing modalities, as well

providing interpretability in terms of modality-dependent embeddings.

The MAE model is a further extensive version of the MDNN by introducing modality-specific decoders to reconstruct input features as well as classify. This dual goal motivates the encoders to learn more informative representations that include crucial modality-specific information. Performance indicators showed that MAE was generally similar or slightly outperformed by MDNN in terms of predictive accuracy and AUC, especially when the datasets have more laboratory/clinical information. The use of reconstruction loss helps novel class discovery and can improve generalisation over the limited samples or better drift setups to noisy/incomplete modalities.

The TransformerFusion model explicitly modelled the inter-modality relationships using self-attention mechanisms. By using a miniature transformer on the modalities' embeddings, the model learnt how to weight interaction between different modality features. The transformer-based method provided competitive or superior results comparing with other multimodal models, especially in datasets of advanced inter-feature dependencies. Although much more computationally intensive, Transformers can dynamically model relationships among elements rather than concatenate them statically which is better suited for heterogeneous biomedical data integration.

Here, the ensemble of MDNN, MAE and TransformerFusion predictions were also calculated through a plain averaging. The ensemble method provided generally superior performance measurements overall all datasets. This indicates that each multimodal model learns different complementary characteristics of the data and their fusion alleviates the biases and variances of separate models. Ensemble methods are especially useful in biomedical prediction, where generalization and optimism are keys to the clinical use of predictive models.

Single-modality classical machine learning models had generally reasonable performances. Traditional cross-feature machine learning models significantly improved the metric on prediction but did not provide an explicit representation mode-specific learning. Early fusion MLP models more accurately described cross-modal interactions, although they were limited by uniform feature scaling and full modality availability. Late fusion architectures (MDNN, MAE, Transformer) achieved slight better results than classical machine learning and early fusion MLP models, clearly demonstrating the advantage of modality-specific encoding. The multimodal ensemble is effective highlighting their complementary property of the various deep learning architectures. These results indicate that multimodal architectures are especially well suited for heterogeneous biomedical data where different types of features lend unique predictive signals. The work also highlights the importance of modality-specific representation learning and model integration.

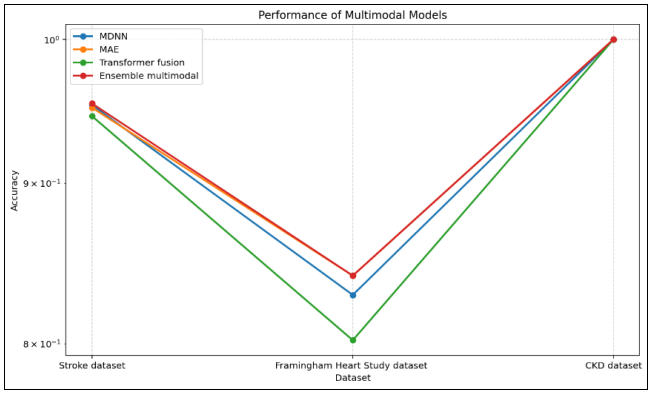

Fig 2: Performance comparison of multimodal learning models

Table 1: Performance comparison of multimodal learning models

Datasets	Models	Accuracy	Precision	Recall	F1-Score	AUC
Stroke Prediction	MDNN	0.953	1.0	0.04	0.0769	0.8221
Stroke Prediction	MAE	0.9511	0.5	0.02	0.0385	0.8163
Stroke Prediction	Transformer fusion	0.9452	0.3333	0.12	0.1765	0.7786
Stroke Prediction	Ensemble multimodal	0.954	1.0	0.06	0.1132	0.8172
FHS	MDNN	0.829	0.2778	0.0775	0.1212	0.663
FHS	MAE	0.8408	0.375	0.0698	0.1176	0.6752
FHS	Transformer fusion	0.8019	0.2651	0.1705	0.2075	0.5529
FHS	Ensemble multimodal	0.8408	0.4211	0.124	0.1916	0.651
CKD	MDNN	1.0	1.0	1.0	1.0	1.0
CKD	MAE	1.0	1.0	1.0	1.0	1.0
CKD	Transformer fusion	1.0	1.0	1.0	1.0	1.0
CKD	Ensemble multimodal	1.0	1.0	1.0	1.0	1.0

Table 2: Performance comparison of multimodal learning model with machine learning classifier for Stroke Prediction dataset

Models	Modality	Accuracy	Precision	Recall	F1-Score	AUC
Logistic regression	Demographics	0.9511	0.0	0.0	0.0	0.8337
Decision tree	Demographics	0.9159	0.1897	0.22	0.2037	0.591
SVM	Demographics	0.9511	0.0	0.0	0.0	0.7137
Naive Bayes	Demographics	0.9472	0.3	0.06	0.1	0.7935
MLP	Demographics	0.9521	1.0	0.02	0.0392	0.8203
Random forest	Demographics	0.9305	0.08	0.04	0.0533	0.7115
AdaBoost	Demographics	0.9511	0.0	0.0	0.0	0.826
Gradient boosting	Demographics	0.9501	0.0	0.0	0.0	0.8423
XGBoost	Demographics	0.9423	0.1538	0.04	0.0635	0.8112
Logistic regression	Clinical	0.9511	0.0	0.0	0.0	0.7178
Decision tree	Clinical	0.9217	0.1739	0.16	0.1667	0.5605
SVM	Clinical	0.9511	0.0	0.0	0.0	0.4091
Naive Bayes	Clinical	0.8992	0.1728	0.28	0.2137	0.7335
MLP	Clinical	0.9511	0.0	0.0	0.0	0.7533
Random forest	Clinical	0.9501	0.4286	0.06	0.1053	0.678
AdaBoost	Clinical	0.9511	0.0	0.0	0.0	0.7101
Gradient boosting	Clinical	0.9481	0.2857	0.04	0.0702	0.7306
XGBoost	Clinical	0.9452	0.2	0.04	0.0667	0.6784
Logistic regression	All features	0.9511	0.0	0.0	0.0	0.8387
Decision tree	All features	0.908	0.1333	0.16	0.1455	0.5533
SVM	All features	0.9511	0.0	0.0	0.0	0.6262
Naive Bayes	All features	0.864	0.1654	0.44	0.2404	0.805
MLP	All features	0.9413	0.1429	0.04	0.0625	0.7951
Random forest	All features	0.9491	0.0	0.0	0.0	0.7958
AdaBoost	All features	0.9511	0.0	0.0	0.0	0.826
Gradient boosting	All features	0.9481	0.0	0.0	0.0	0.8273
XGBoost	All features	0.9413	0.2727	0.12	0.1667	0.8099
Early fusion MLP	Early fusion	0.9511	0.5	0.04	0.0741	0.8209
MDNN	Multimodal late fusion -DNN	0.953	1.0	0.04	0.0769	0.8221
MAE	Multimodal late fusion - autoencoder	0.9511	0.5	0.02	0.0385	0.8163
Transformer fusion	Multimodal -transformer	0.9452	0.3333	0.12	0.1765	0.7786
Ensemble multimodal	Ensemble late fusion	0.954	1.0	0.06	0.1132	0.8172

Table 3: Performance comparison of multimodal learning model with machine learning classifier for FHS dataset

Models	Modality	Accuracy	Precision	Recall	F1-Score	AUC
Logistic regression	Demographics	0.8479	0.0	0.0	0.0	0.668
Decision tree	Demographics	0.8278	0.2069	0.0465	0.0759	0.5937
SVM	Demographics	0.8479	0.0	0.0	0.0	0.484
Naive Bayes	Demographics	0.8479	0.0	0.0	0.0	0.6646
MLP	Demographics	0.8479	0.0	0.0	0.0	0.6636
Random forest	Demographics	0.8278	0.0952	0.0155	0.0267	0.6068
AdaBoost	Demographics	0.8479	0.0	0.0	0.0	0.6565
Gradient boosting	Demographics	0.8467	0.0	0.0	0.0	0.648
XGBoost	Demographics	0.8325	0.1176	0.0155	0.0274	0.6064
Logistic regression	Clinical	0.8467	0.4615	0.0465	0.0845	0.6717
Decision tree	Clinical	0.7193	0.1484	0.1783	0.162	0.4974
SVM	Clinical	0.8467	0.4286	0.0233	0.0441	0.5567
Naive Bayes	Clinical	0.8078	0.2424	0.124	0.1641	0.6412
MLP	Clinical	0.8361	0.3438	0.0853	0.1366	0.5935
Random forest	Clinical	0.8384	0.3182	0.0543	0.0927	0.5651
AdaBoost	Clinical	0.8502	0.75	0.0233	0.0451	0.6312
Gradient boosting	Clinical	0.8491	0.5263	0.0775	0.1351	0.6307
XGBoost	Clinical	0.8208	0.1892	0.0543	0.0843	0.5711
Logistic regression	Labs	0.8467	0.0	0.0	0.0	0.5482
Decision tree	Labs	0.842	0.2222	0.0155	0.029	0.5198
SVM	Labs	0.8479	0.0	0.0	0.0	0.4731
Naive Bayes	Labs	0.8432	0.0	0.0	0.0	0.5413
MLP	Labs	0.8467	0.0	0.0	0.0	0.5482
Random forest	Labs	0.8443	0.2857	0.0155	0.0294	0.5168
AdaBoost	Labs	0.8479	0.0	0.0	0.0	0.5586
Gradient boosting	Labs	0.8443	0.0	0.0	0.0	0.5265
XGBoost	Labs	0.8479	0.0	0.0	0.0	0.5089
Logistic regression	All features	0.8443	0.4118	0.0543	0.0959	0.7028
Decision tree	All features	0.7429	0.1844	0.2016	0.1926	0.5208
SVM	All features	0.8502	0.6667	0.031	0.0593	0.5422

Naive Bayes	All features	0.8066	0.2535	0.1395	0.18	0.6781
MLP	All features	0.8267	0.3333	0.1395	0.1967	0.6133
Random forest	All features	0.8443	0.4211	0.062	0.1081	0.6267
AdaBoost	All features	0.8467	0.4545	0.0388	0.0714	0.681
Gradient boosting	All features	0.8396	0.3704	0.0775	0.1282	0.6621
XGBoost	All features	0.8101	0.2037	0.0853	0.1202	0.5857
Early fusion MLP	Early fusion	0.8314	0.3158	0.093	0.1437	0.6219
MDNN	Multimodal late fusion DNN	0.829	0.2778	0.0775	0.1212	0.663
MAE	Multimodal late fusion - Autoencoder	0.8408	0.375	0.0698	0.1176	0.6752
Transformer fusion	Multimodal transformer	0.8019	0.2651	0.1705	0.2075	0.5529
Ensemble multimodal	Ensemble late fusion	0.8408	0.4211	0.124	0.1916	0.651

Table 4: Performance comparison of multimodal learning model with machine learning classifier for CKD dataset

Models	Modality	Accuracy	Precision	Recall	F1-Score	AUC
Logistic regression	Demographics	0.6125	0.6301	0.92	0.748	0.752
Decision tree	Demographics	0.6625	0.6949	0.82	0.7523	0.6803
SVM	Demographics	0.7375	0.7231	0.94	0.8174	0.7927
Naive Bayes	Demographics	0.6	0.6216	0.92	0.7419	0.752
MLP	Demographics	0.775	0.75	0.96	0.8421	0.8033
Random forest	Demographics	0.7125	0.7547	0.8	0.7767	0.7047
AdaBoost	Demographics	0.775	0.7759	0.9	0.8333	0.7687
Gradient boosting	Demographics	0.7	0.7321	0.82	0.7736	0.69
XGBoost	Demographics	0.675	0.7069	0.82	0.7593	0.683
Logistic regression	Clinical	1.0	1.0	1.0	1.0	1.0
Decision tree	Clinical	1.0	1.0	1.0	1.0	1.0
SVM	Clinical	1.0	1.0	1.0	1.0	1.0
Naive Bayes	Clinical	1.0	1.0	1.0	1.0	1.0
MLP	Clinical	1.0	1.0	1.0	1.0	1.0
Random forest	Clinical	1.0	1.0	1.0	1.0	1.0
AdaBoost	Clinical	1.0	1.0	1.0	1.0	1.0
Gradient boosting	Clinical	1.0	1.0	1.0	1.0	1.0
XGBoost	Clinical	1.0	1.0	1.0	1.0	1.0
Logistic regression	Labs	1.0	1.0	1.0	1.0	1.0
Decision tree	Labs	1.0	1.0	1.0	1.0	1.0
SVM	Labs	1.0	1.0	1.0	1.0	1.0
Naive Bayes	Labs	1.0	1.0	1.0	1.0	1.0
MLP	Labs	0.95	1.0	0.92	0.9583	1.0
Random forest	Labs	1.0	1.0	1.0	1.0	1.0
AdaBoost	Labs	1.0	1.0	1.0	1.0	1.0
Gradient boosting	Labs	1.0	1.0	1.0	1.0	1.0
XGBoost	Labs	1.0	1.0	1.0	1.0	1.0
Logistic regression	All features	1.0	1.0	1.0	1.0	1.0
Decision tree	All features	1.0	1.0	1.0	1.0	1.0
SVM	All features	1.0	1.0	1.0	1.0	1.0
Naive Bayes	All features	1.0	1.0	1.0	1.0	1.0
MLP	All features	0.975	1.0	0.96	0.9796	1.0
Random forest	All features	1.0	1.0	1.0	1.0	1.0
AdaBoost	All features	1.0	1.0	1.0	1.0	1.0
Gradient boosting	All features	1.0	1.0	1.0	1.0	1.0
XGBoost	All features	1.0	1.0	1.0	1.0	1.0
Early fusion MLP	Early fusion	1.0	1.0	1.0	1.0	1.0
MDNN	Multimodal late fusion DNN	1.0	1.0	1.0	1.0	1.0
MAE	Multimodal late fusion - Autoencoder	1.0	1.0	1.0	1.0	1.0
Transformer fusion	Multimodal transformer	1.0	1.0	1.0	1.0	1.0
Ensemble multimodal	Ensemble late fusion	1.0	1.0	1.0	1.0	1.0

5. Conclusion

This study illustrates how an effective multimodal learning architectures can be built as well as decomposed and integrated over different feature groups. Even if the improvements in accuracy are of minor extent, this work indeed brings forward a full pipeline that applies MDNN-based, MAE-based and Transformer based fusion in structured multimodal settings. Multimodal learning approach work because they are combining complementary information across multiple data sources and building models that can learn richer patterns, reduce uncertainty and exploit cross-modal interactions that single-modality models may otherwise ignore—in essence resulting in more robust representations and improved generalization—particularly when dealing with complex prediction tasks. The contribution of this study lies in proposing a pragmatic architecture for the application of multimodal architectures to heterogeneous medical data. However, the proposed approach was largely developed for binary outcomes, and it would be worth extending the methodology to multi-class or survival data. Also, interpretability is a difficult task for

deep multimodal models despite the availability of embeddings and attention weights for some partial insights.

6. References

1. Müller H, Unay D. Retrieval from and understanding of large-scale multi-modal medical datasets: A review. IEEE Transactions on Multimedia. 2017; 19(9):2093-2104.
2. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, *et al*. Multimodal machine learning in precision health: A scoping review. NPJ Digital Medicine. 2022; 5(1):171.
3. Tobon DP, Hossain MS, Muhammad G, Bilbao J, Saddik AE. Deep learning in multimedia healthcare applications: A review. Multimedia Systems. 2022; 28(4):1465-1479.
4. Krones F, Marikkar U, Parsons G, Szmul A, Mahdi A. Review of multimodal machine learning approaches in healthcare. Information Fusion. 2025; 114:102690.
5. Sáez JA, Krawczyk B, Woźniak M. On the influence of class noise in medical data classification: Treatment using noise filtering methods. Applied Artificial Intelligence. 2016; 30(6):590-609.
6. Cui C, Yang H, Wang Y, Zhao S, Asad Z, Coburn LA, *et al*. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review. Progress in Biomedical Engineering. 2023; 5(2):22001.
7. Chen Q, Li M, Chen C, Zhou P, Lv X, Chen C. MDFNet: Application of multimodal fusion method based on skin image and clinical data to skin cancer classification. Journal of Cancer Research and Clinical Oncology. 2023; 149(7):3287-3299.
8. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: A review. Briefings in Bioinformatics. 2022; 23(2):bbab569.
9. Teoh JR, Dong J, Zuo X, Lai KW, Hasikin K, Wu X. Advancing healthcare through multimodal data fusion: A comprehensive review of techniques and applications. PeerJ Computer Science. 2024; 10:e2298.
10. Steyaert S, Pizurica M, Nagaraj D, Khandelwal P, Hernandez-Boussard T, Gentles AJ, *et al*. Multimodal data fusion for cancer biomarker discovery with deep learning. Nature Machine Intelligence. 2023; 5(4):351-362.
11. Chaabene S, Boudaya A, Bouaziz B, Chaari L. An overview of methods and techniques in multimodal data fusion with application to healthcare. International Journal of Data Science and Analytics, 2025, 1-25.
12. Pereira LM, Salazar A, Vergara L. A comparative analysis of early and late fusion for the multimodal two-class problem. IEEE Access. 2023; 11:84283-84300.
13. Kim JC, Chung K. Recurrent neural network-based multimodal deep learning for estimating missing values in healthcare. Applied Sciences. 2022; 12(15):7477.
14. Zhang C, Chu X, Ma L, Zhu Y, Wang Y, Wang J, *et al*. M3care: Learning with missing modalities in multimodal healthcare data. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Aug 2022; 2418-2428.
15. Ma M, Ren J, Zhao L, Tulyakov S, Wu C, Peng X. Smil: Multimodal learning with severely missing modality. In Proceedings of the AAAI Conference on Artificial Intelligence. May 2021; 35(3):2302-2310.
16. Fedesoriano. Stroke Prediction Dataset [Dataset]. Kaggle, 2021 [Cited 2025 Dec 1]. Retrieved from: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
17. Bhardwaj A. Framingham Heart Study Dataset [Dataset]. Kaggle, 2016 [Cited 2025 Dec. 1]. Retrieved from: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset
18. Rubini L, Soundarapandian P, Eswaran P. Chronic Kidney Disease [Dataset]. UCI Machine Learning Repository, 2015 [Cited 2025 Dec. 1]. Retrieved from: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease
19. Ou C, Zhou S, Yang R, Jiang W, He H, Gan W, *et al*. A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata. Frontiers in Surgery. 2022; 9:1029991.
20. Soenksen LR, Ma Y, Zeng C, Boussioux L, Villalobos Carballo K, Na L, *et al*. Integrated multimodal artificial intelligence framework for healthcare applications. NPJ Digital Medicine. 2022; 5(1):149.
21. Haleem MS, Ekuban A, Antonini A, Pagliara S, Pecchia L, Allocca C. Deep-learning-driven techniques for real-time multimodal health and physical data synthesis. Electronics. 2023; 12(9):1989.
22. Preetha M, Budaraju RR, Jackulin C, Sri, PSGA, Padmapriya T. Deep learning-driven real-time multimodal healthcare data synthesis. International Journal of Intelligent Systems and Applications in Engineering. 2023; 12(5s):360-369.
23. Gao Z, Liu X, Kang Y, Hu P, Zhang X, Yan W, *et al*. Improving the prognostic evaluation precision of hospital outcomes for heart failure using admission notes and clinical tabular data: Multimodal deep learning model. Journal of Medical Internet Research. 2024; 26:e54363.
24. Li B, Chen H, Lin X, Duan H. Multimodal learning system integrating electronic medical records and hysteroscopic images for reproductive outcome prediction and risk stratification of endometrial injury: A multicenter diagnostic study. International Journal of Surgery. 2024; 110(6):3237-3248.
25. Wang Y, Yin C, Zhang P. Multimodal risk prediction with physiological signals, medical images and clinical notes. Heliyon. 2024; 10(5):e26772.
26. Imrie F, Denner S, Brunschwig LS, Maier-Hein K, Van Der Schaar M. Automated ensemble multimodal machine learning for healthcare. IEEE Journal of Biomedical and Health Informatics. 2025; 29(6):4213-4226.
27. Kumar S, Sharma S, Megra KT. Transformer enabled multi-modal medical diagnosis for tuberculosis classification. Journal of Big Data. 2025; 12(1):5.
28. Tölle M, Burger L, Kelm H, André F, Bannas P, Diller G, *et al*. Multi-modal dataset creation for federated learning with DICOM-structured reports. International Journal of Computer Assisted Radiology and Surgery. 2025; 20(3):485-495.