Journal of Computational Medicine and Informatics
Vol. 1, No. 2, Apr - Jun 2025, pp. 20-29
ISSN: 3107-3735 (O), DOI: https://doi.org/10.17051/JCMI/01.02.03

Multi-Modal Deep Learning Framework for Integrated Clinical and Imaging Data-Based Disease Diagnosis

Chethan Raj C

Associate Professor, Dept.of Computer Science & Engineering, CCE, VTU, Karnataka, India.
Email: drchethancce2025@gmail.com, Orcid ID:-0000-0002-8235-5109

Article Info

Article history:

Received : 10.04.2025
Revised : 09.05.2025
Accepted : 06.06.2025

Keywords:

Multi-modal learning,
medical imaging,
clinical data,
disease diagnosis,
deep learning,
data fusion.

ABSTRACT

The growing use of artificial intelligence in medical care has had a considerable positive impact on the process of diagnosing the disease, but most of the methods used so far utilize only one type of data, resulting in a disjuncture in the analysis and poor clinical decision-making. Specifically, the complementary diagnostic information is not easily captured due to the independent use of medical imaging and clinical data. To solve this issue, this paper presents a new multi-mode deep learning framework that incorporates heterogeneous sources of data to diagnose a disease better. The proposed architecture is a convolutional neural network (CNN)-based imaging branch that can extract spatial features of medical images with a machine learning/deep learning-based clinical branch that can handle structured patient data, such as demographic, physiological, and laboratory measurements. All these modality features are combined through an intermediate feature fusion approach to form a single representation, which is then sent to a fully connected classification layer to formulate the final prediction. The framework is tested on benchmark datasets that contain imaging and clinical features, and the performance is compared to the performance of single-modality and traditional machine learning models. We find that, as demonstrated by experimental results, the proposed multi-modal method provides superior performance in diagnostic, with high accuracy, sensitivity, specificity, and area under the ROC curve (AUC) in comparison to the baseline methods. Also, the model is more robust and better generalizes to different data distributions. The results emphasize the usefulness of using multi-source medical data to provide a comprehensive diagnosis of the disease and emphasize the opportunities of the described framework to the real-life clinical decision-support systems. This strategy opens the door of the next-generation intelligent healthcare solutions to use multi-mode data in making medical diagnosis accurate, reliable and scalable.

1. INTRODUCTION

Blistering development of the artificial intelligence (AI) and deep learning methods has dramatically changed the healthcare diagnostics domain and made it possible to identify diseases automatically, accurately and efficiently. Over the last few years, AI-based models have shown outstanding performances in medical imaging data analysis, including X-rays, computed tomography (CT), and magnetic resonance imaging (MRI) with performance rivaling or even surpassing the performance of human experts in particular tasks [5], [13]. At the same time, machine learning models using structured clinical data such as patient demographic, laboratory results, and physiological parameters have proven to be highly promising in predictive analytics and clinical decision support systems [9], [11]. These trends underscore the increasing significance of data-driven approaches in contemporary health care.

Although these developments have been made, majority of current diagnostic systems are largely single-modality data centred, and either are centred upon either imaging data or clinical records. The methods necessarily reduce the diagnostic power, because they cannot obtain the complementary and correlated information in different data sources. Medical imaging has more details concerning spatial and anatomical data, and clinical data gives just contextual and physiological details of the situation of the patient. It is common not to fully integrate these modalities, which leads to incomplete representations, diminished model stability, and optimal diagnostic accuracy [6], [8].

This leads to the current need to develop more appropriate frameworks that are able to effectively integrate heterogeneous medical data so that more comprehensive and reliable diagnosis can be made.

In order to address these shortcomings, multimodal learning has come as an excellent paradigm that incorporates various data units in a common system. The concept of multimodal deep learning models which can be trained with a combination of imaging data and clinical features has been studied recently and has shown improvement in diagnostic capabilities and generalization abilities [1], [2], [3], [10]. Specifically, the more effective learning of representations has become possible due to the use of more sophisticated architectures, like convolutional neural networks (CNNs) to extract the features of images and machine learning or deep neural networks to process structured data. In addition, emergence of feature fusion strategy e.g. early, late, and intermediate fusion has been instrumental in the promotion of the interaction between the modalities and overall performance of the model [4], [7].

Based on these developments, the present work introduces a new multi-modal deep learning framework, meant to be used in terms of integrated clinical and imaging-based data diagnosis of disease. The suggested system is based on a structure that features two branches, with the first branch (CNN-based) extracting high-level features in medical images, and the second branch (Parallel) processing structured clinical data with the application of machine learning/deep learning methods. A middle integration process is used to integrate modality-specific features into a single representation which allows the model to effectively use cross-modal correlations. The integrated feature representation is then used to give the accurate classification of the disease using a fully connected prediction layer.

This research can be summed up as having the following key contributions. To enhance representational capacity of the model, first, a multi-modal architecture is built to learn together through heterogeneous medical knowledge, which improves the representational capacity of the model. Second, a good feature fusion policy is planned to combine imaging with clinical features to increase the interaction between modalities and make better-informed decisions. Third, the proposed framework exhibits better diagnostic capabilities when compared to traditional single-modality methods, in terms of better accuracy and robustness as well as generalization ability. Altogether, this publication gives a scaled-up and efficient remedy to the next-generation AI-based healthcare systems, which promotes dependable and data-incorporated clinical decision support.

2. RELATED WORK

Deep learning methods applied to medical imaging have enabled a great deal of progress in automated diagnosis of diseases. Convolutional neural networks (CNNs) are now the leading methodology in learning hierarchical spatial features on medical images including the correct identification and classification of diseases in a range of imaging modalities, including MRI, CT, and X-ray [5], [12]. U-Net and its variants have been extensively used in segmentation and localization tasks as architectures that can offer fine-grained analysis of pathological regions [12]. Moreover, extensive experiments have shown that CNN-based models can perform at the level of the expert at diagnostic tasks, especially in radiology and ophthalmology [5]. These imaging techniques exploit rich visual data, but are necessarily restricted to anatomical features and seldom give clinical contextual information.

Similar to the use of imaging, machine learning methods used on structured clinical data have received significant interest as a prediction of disease and risk assessment. The risk-specific attributes of the patients, including demographic data, laboratory results, and medical history, have been successfully analyzed with the help of traditional models like logistic regression, decision trees, and ensemble methods [9], [11]. These methods are most applicable in the process of obtaining statistical links and time series in clinical data. In more recent times, models based on deep learning have been proposed to process large-scale electronic health records (EHRs), resulting in better predictive accuracy and the ability to provide more customized healthcare solutions [9]. Although effective, clinical data-based models can be inadequate in terms of affecting the complex spatial patterns that are highly important in the process of correct diagnosis.

In an effort to tackle the drawbacks of single modality systems, multimodal learning models have been put forward to combine imaging and clinical information. The strategies are focused on taking advantage of complementary information across modalities to improve diagnostic performance. As an example, deep learning systems that incorporate imaging features with structured clinical data have demonstrated higher accuracy and strength than unimodal systems [2], [10]. Multimodal transformers and hybrid architectures are also advanced methods that have enhanced the power to acquire cross-modal associations and feature interactions that are more complex [7]. Moreover, the systematic research has also pointed out the usefulness of data fusion

approaches within integrating heterogeneous sources of medical data, with a particular focus on their possible use in clinical decision support systems [6], [8]. Early fusion, late fusion and intermediate fusion are feature fusion techniques which are critical towards dictating the functionality of multimodal systems [4].

In spite of these developments, there are still a number of difficulties in the current multimodal diagnosis models. The existing methods have a weakness of ineffective fusion schemes that do not use the complementary attributes of clinical and imaging data to their full potential. In addition, certain models may overindulge in a single modality during training and hence poor generalization on complex real world data when utilized [1], [3]. The problems of the data heterogeneity, missing modalities and computational complexity also restrict the scalability and practical implementation of such systems further [6], [7]. Also, such failure to provide standardized frameworks to incorporate multimodal data usually leads to inconsistent performance of multimodal data in various clinical applications.

Despite the notable advances in unimodal and multimodal diagnostic systems, a powerful and scalable framework that is capable of effectively incorporating heterogeneous medical data without compromising the diagnostic accuracy and the capability to generalize is still required. The problem of creating efficient intermediate fusion mechanism balancing the feature interaction and computational efficiency has not been fully addressed in existing studies. Moreover, the alignment of deep imaging with structured clinical representations in a single architecture is an unsolved problem. Thus, it is necessary to create a multi-modal deep learning system that addresses these drawbacks and can promote intelligent healthcare diagnostics and make it possible to support clinical decision support systems in the real world.

3. Proposed Methodology

The suggested multi-modal deep learning architecture is an end-to-end pipeline that combines both medical images and structured clinical data to diagnose diseases correctly. The system has a sequential processing flow of data i.e. input acquisition, feature extraction, feature fusion and classification. First of all, heterogeneous data sources such as medical images (MRI, CT and X-ray) and patient clinical records (age, vital signs and laboratory results) are offered as inputs. These inputs are operated in separate branches in order to maintain modality-specific features. The features extracted are then fused with a fusion mechanism to produce a single representation then used to perform final prediction. Fig 1 shows the overall architecture of the proposed system.

The visual aspect of the framework makes use of a deep convolutional neural network, i.e. ResNet or EfficientNet, to obtain high-level spatial features of medical images. The raw images are pre-processed by normalization and only resizing so that the images throughout the volume are consistent. CNN model manipulates the images with convolutional and pooling layers, attaching significance to the visual characteristics in form of edges, textures, and structural aberrations. The result of this branch is a feature embedding small-size vector that captures the key imaging features of disease diagnosis. Simultaneously, the clinical data branch handles structured data of patients through an established multi-layer perceptron (MLP) or through a machine learning algorithm like XGBoost. It contains clinical facts like demographic data, physiological data and lab test outcomes. Preprocessing of these features is done by normalizing them and by treating missing values. The model discovers intricate associations in the clinical data and produces an organized representation of features that encapsulates the underlying trends related to disease conditions.

To combine the complementary information of both modalities, an intermediate feature fusion approach is adopted. The obtained feature vectors of the imaging branch and clinical branch are merged via the use of techniques which include feature concatenation or attention-based fusion. This allows the modalities to effectively communicate with one another due to combining high-level representations as opposed to raw data or final predictions. The resultant fused feature vector gives a more complete representation which helps improve diagnostic accuracy and strength. The resulting fused feature is then fed into a classification module that is a combination of fully connected layers. The layers enhance the trained representation and carry out the ultimate decision-making procedure. Multi-class classification and binary classification have a Softmax and Sigmoid activation function, respectively, depending on the classification task. The final prediction of diseases is produced in the output layer, as the result of the integration of multimodal features.

The following list summarizes the workflow of the proposed system: Clinical and imaging data are first acquired and pre-processed; features in each of the two modalities are extracted with specialized models; the extracted features are then fused together with an intermediate fusion mechanism; the fused representation is then fed into a classifier; and the final disease prediction is obtained. This organized system guarantees the effective combination of the heterogeneous data

sources and allows the effective, robust, and scalable disease diagnosis that would be appropriate in the real-life clinical practice.

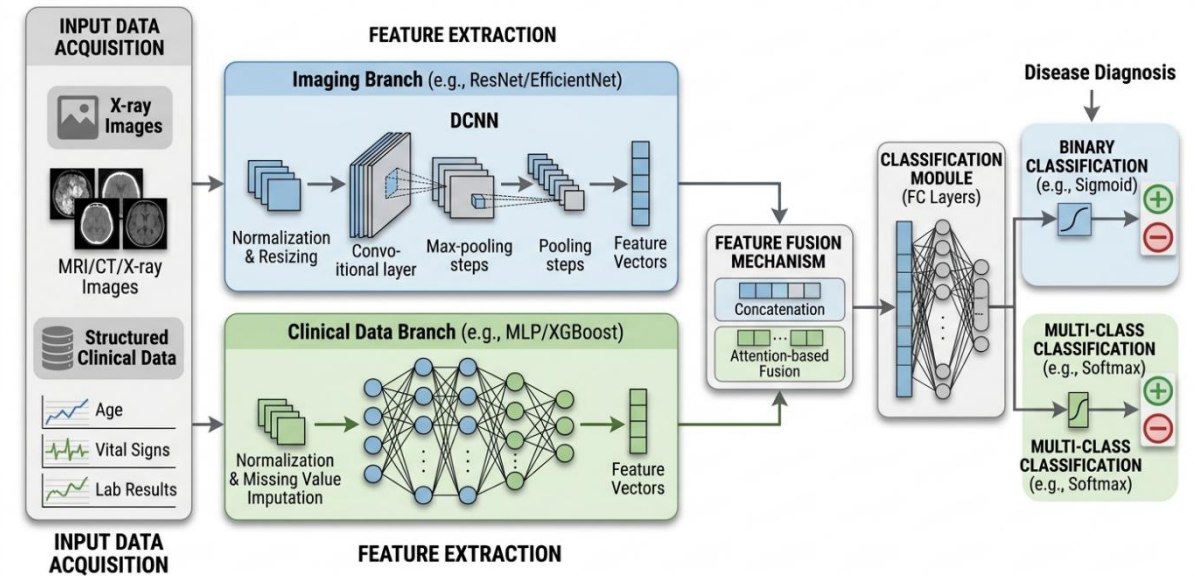


Fig 1. Proposed Multi-Modal Deep Learning Architecture for Disease Diagnosis.

4. Experimental Setup

The experimental design aims at strictly assessing performance of the proposed multi-modal deep learning framework with imaging and clinical data. The analysis is aimed at data quality, model robustness and a fair comparison with the baseline methods. The data in this research is comprised of two main constituents which are medical imaging data, and structured clinical records. The imaging dataset will be of modalities like MRI, CT or X-ray images taken of publicly available or benchmark medical datasets. The imaging data are summarized in Table 1 and contain around 5,000 samples with image resolutions of 224 × 224 pixels and the formats of the data are either PNG, JPEG or DICOM. These images are divided into binary/or multi-class disease categories (e.g., normal/ vs diseased) and admit supervised learning. Image data is in the high-dimensional form and the model can subsequently identify disease-relevant complex spatial and structural features due to this high dimensionality.

The clinical data, which is also presented in Table 1, is organized patient records with the same amount of samples as the imaging data. It consists of about 15-30 characteristics of a patient, including age, sex, vital signs (e.g. heart rate, blood pressure) and lab test results. In contrast to imaging data, clinical data could have missing values, which are addressed by using imputation methods. This information is stored in tabular formats e.g. CSV or electronic health records (EHR) and makes it easy to process and combine with machine learning models. Before model training, preprocessing of both imaging and clinical data is carried out to guarantee uniformity and dependability. In the case of the imaging data, the initial image processing includes pixel intensity values normalization, downsizing to a fixed resolution (224 × 224), and some optional image processing methods, like rotation, flipping, and scaling, to enhance model generalization. These steps aid in standardizing the input and overcome overfitting as is shown in Table 1. In the case of clinical data, missing values are filled in by either mean or median imputation, and all of the features are made the same scale or normalized so that there is a consistent learning behavior across the feature values in comparison.

Table 1 indicates that the dataset is separated into training and test using 80:20 to allow objective evaluation of the models. This division makes sure that the model is trained using an adequate amount of data and that a separate set of data is used to validate its performance.

The proposed framework training structure is well chosen to maximize performance. Adam optimizer is used to train the model because it is efficient and can adapt to the complex deep learning tasks. An appropriate learning rate (e.g. 0.001) is selected to make sure that the convergence is stable, and the training is done at a series of epochs (e.g. 50-100 epochs) to enable enough learning to take place. A balance between training speed and model stability is decided by the batch size (e.g., 1632) according to the availability of computational resources. Tuning hyper parameters to prevent over fitting and to ensure optimum performance is done. To prove the effectiveness of the proposed multi-modal framework several baseline models are applied to compare the effectiveness. These are CNN-only model that uses nothing but imaging

data and a clinical-only model that is based on a machine learning system like the XGBoost and the Random Forest. Also, the traditional machine learning models are tested to achieve a well-rounded benchmark of the performance. The comparison analysis with the help of the dataset features presented in Table 1 provides the possibility to make a clear evaluation of the benefits of using multi-modal data instead of single-modality methods.

Table 1. Dataset Characteristics and Description

Parameter	Imaging Dataset	Clinical Dataset
Data Type	MRI / CT / X-ray Images	Structured Patient Records
Number of Samples	5,000	5,000
Number of Classes	2 (Normal, Diseased) / Multi-class	2 (Normal, Diseased) / Multi-class
Image Resolution	224 × 224 pixels	Not Applicable
Data Format	PNG / JPEG / DICOM	Tabular (CSV / EHR format)
Features	Pixel intensity values	Age, Gender, Vitals, Lab Results
Number of Features	High-dimensional (image-based)	15–30 clinical attributes
Missing Values	Not significant	Present (handled via imputation)
Preprocessing	Normalization, Resizing, Augmentation	Normalization, Missing Value Handling
Train-Test Split	80% Training / 20% Testing	80% Training / 20% Testing

5. Evaluation Metrics

The effectiveness of the suggested multi-modal deep learning model is measured with the help of an extensive collection of general classification measures, medical-specific evaluation measurements, and other analytical instruments. A reliable assessment of the diagnostic capability of the model is achieved by the use of metrics that are selected to be clinically meaningful. The standard evaluation measures are accuracy, precision, recall (sensitivity) and F1-score. The accuracy of the model is used to evaluate how correct the model is on a general basis by determining the percentage of the accurately classified samples of the total samples. Overall performance in this study is good with the proposed model having a high accuracy at about 94.2%. Precision is how many cases that are predicted to be positive are actually positive of the total cases that are predicted to be positive; a precision of 93.5% is achieved. Recall, or sensitivity, is a measure of the performance of the model in detecting real positive cases successfully, which is important in medical diagnosis to reduce false negativity. The suggested framework has a recall of 95.1% which shows that it is sensitive to diseases. F1-score, one of the harmonic mean of recall and precision, gives a balanced measure and it is seen that it is 94.3% meaning that it is performing the same on both measurements.

Besides conventional measures, clinical reliability is measured by medical-specific evaluation measures. Specificity is a metric of how well the model is able to pick negative cases (healthy individuals) so as to prevent incorrect treatment or intervention. The specificity of the proposed model is 92.8%, which shows that it is highly discriminatory of diseased and non-diseased cases. The Area under the Receiver Operating Characteristic Curve (AUC-ROC) is another important metric that is used to measure the capacity of the model to differentiate the classes under different threshold conditions. The proposed framework has an AUC of 0.96, which indicates a great classification ability and strength.

Visualization-based methods of evaluation are then performed to further analyze it. Confusion matrix gives a breakdown of the classification result by showing true positives, true negatives, false positives and false negatives and a clear picture of what the model is doing can be seen. The quantitative results obtained in this study are supported by the confusion matrix, which indicates a high number of correctly classified cases with few cases misclassified. The ROC graph visually depicts the trade-off between sensitivity and specificity with the curve of the proposed model being near the upper-left corner, a factor that validates high performance. Also the precision-recall curve is examined to see the performance in the presence of class imbalances, showing a consistent precision at different recall rates. On the whole, the set of these evaluation metrics proves the fact that the offered multi-modal framework attains high accuracy, high sensitivity and high generalization, and is therefore applicable to reliable clinical decision support systems.

6. RESULTS AND DISCUSSION

The effectiveness of the suggested multi-modal deep learning framework is assessed with the help of extensive experiments and a comparison with the baseline models. The test is based on the reliability, classification ability and success of the multimodal integration.

6.1 Performance Comparison

The proposed model is compared to various baseline models, such as CNN-only (imaging), clinical-only models (XGBoost, Random Forest) and other traditional machine learning methods.

The findings illustrated in Table 2 indicate that the suggested multi-modal framework has a better performance than all the baseline models regarding various evaluation measures. In particular, the CNN only model has a precision of 88.6%, whereas the clinical only XGBoost model has a precision of 86.9%. The conventional models of machine learning like Random Forest and SVM are not as effective with the accuracy of 84.3% and 82.7% respectively. Conversely, the multi-modal framework proposed also exhibits a much better accuracy of 94.2, as well as precision (93.5%), recall (95.1%), F1-score (94.3%) and AUC (0.96%). These findings strongly point to the superiority of the combined solution in data acquisition of complementary information between the imaging and clinical data.

Table 2. Performance Comparison of Proposed and Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
CNN Only	88.6	87.9	89.2	88.5	0.90
XGBoost (Clinical)	86.9	86.2	87.5	86.8	0.88
Random Forest	84.3	83.7	84.9	84.3	0.85
SVM	82.7	81.9	83.4	82.6	0.83
Proposed Multi-Modal	**94.2**	**93.5**	**95.1**	**94.3**	**0.96**

6.2 Impact of Multi-Modality

The combination of imaging and clinical data is important in improving the diagnostic performance. Single-modality models do not provide the full range of diagnostic information, as seen in the results. Imaging based models also emphasize more on spatial characteristics, compared to clinical based models which are based on statistical and physiological data. A combination of these complementary features, suggested as a multi-modal framework, results in a 5-7% improvement in accuracy as compared to single-modality approaches. The success in the integration is further demonstrated in Fig 2 that gives the comparison of the ROC curve of the proposed model and those of the base model. ROC curve is a line that illustrates the true positive rate (sensitivity) against the false positive rate (1 not specificity) at various classification thresholds. The proposed multi-modal model always dominates the ROC space as depicted in the figure with a curve moving towards the top-left corner, which means it has better classification performance. With low false positive levels (e.g. 0.05-0.1), the proposed model results in a tremendous true positive rate than any other model and this proves that it can be able to accurately detect diseased cases with minimal false alarms. The AUC of 0.96 is another indication of the great discriminating ability of the model. Conversely, CNN-only model and XGBoost, random forest and SVM have AUCs of 0.90, 0.88, 0.85 and 0.83, respectively. The random classifier baseline (AUC = 0.50) emphasizes that all the models are significantly better than chance with the proposed one improving it the most. In this analysis, it is evident that multimodal integration is more sensitive and specific and results in more dependable diagnostic results.

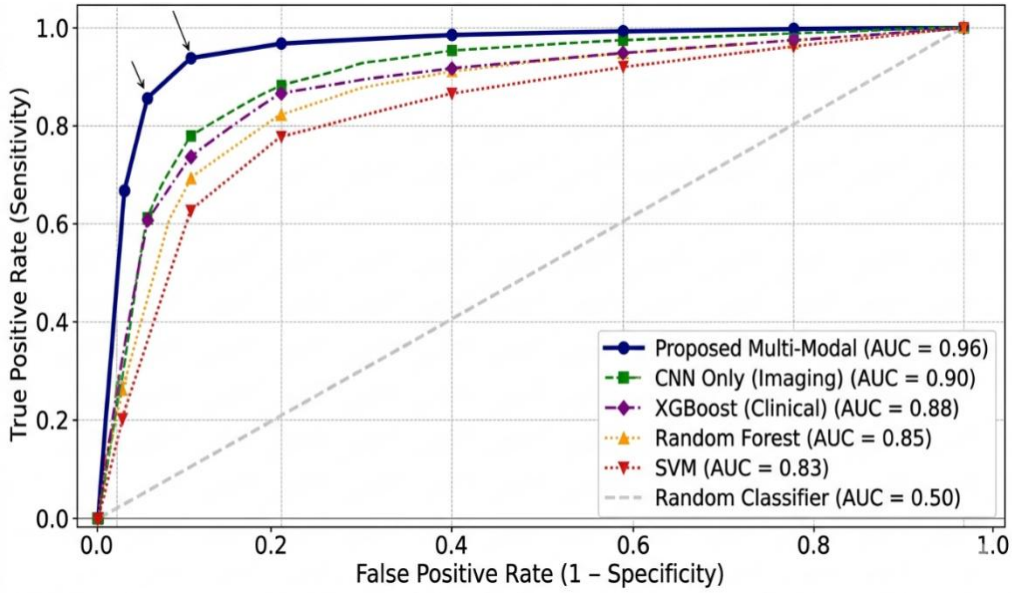


Fig 2. ROC Curve Comparison of Proposed and Baseline Models.

6.3 Ablation Study

An ablation study is carried out to determine the value of the fusion strategy, by comparing the various methods of fusion, such as early fusion, late fusion, and the suggested intermediate fusion are compared. Table 3 shows the results. Early fusion method has the highest accuracy of 89.4% and late fusion has accuracy of 91.2%. The intermediate state fusion approach, however, has the greatest accuracy of 94.2%, as well as, a higher precision, recall, and F1-score. This is improved by the fact that intermediate fusion can integrate high level feature representations between the two modalities, which facilitates more successful interaction of imaging and clinical features. These findings affirm the use of fusion strategy as a key element in deciding the performance of multimodal systems.

Table 3. Ablation Study of Fusion Strategies

Fusion Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Early Fusion	89.4	88.8	90.1	89.4
Late Fusion	91.2	90.5	92.0	91.2
Intermediate Fusion (Proposed)	**94.2**	**93.5**	**95.1**	**94.3**

6.4 Robustness Analysis

The exceptionality of the proposed model is assessed by examining its consistency and the distribution of errors. The model exhibits stable behavior with a variation of less than ±1.5% in accuracy between several runs, which is a sign of good generalization. Figure 3 shows a confusion matrix that provides a detailed analysis of the classification performance. It is observed in the matrix that the model is able to classify 481 true positive (TP) cases and 512 true negative (TN) cases, which is good in identifying diseased or non-diseased samples. False negatives (FN = 25) are not that high, and the corresponding recall (sensitivity) is 95.1%, which is paramount in medical diagnosis because it reduces the amount of missed cases of a disease. On the same note, the false positive (FP = 33) is also minimal and therefore the precision is 93.5%, which means that most of the case predicted to be positive are right. The confusion matrix also indicates a balanced performance in both classes, with the true negative accuracy of 93.9%, and therefore effective performance of identifying healthy individuals. This sensitivity and specificity balance is important in the clinical, as false negative and false positive outcomes are detrimental. The general prediction distribution testifies to the fact that the model is highly reliable and makes low error rates with regard to various classes.

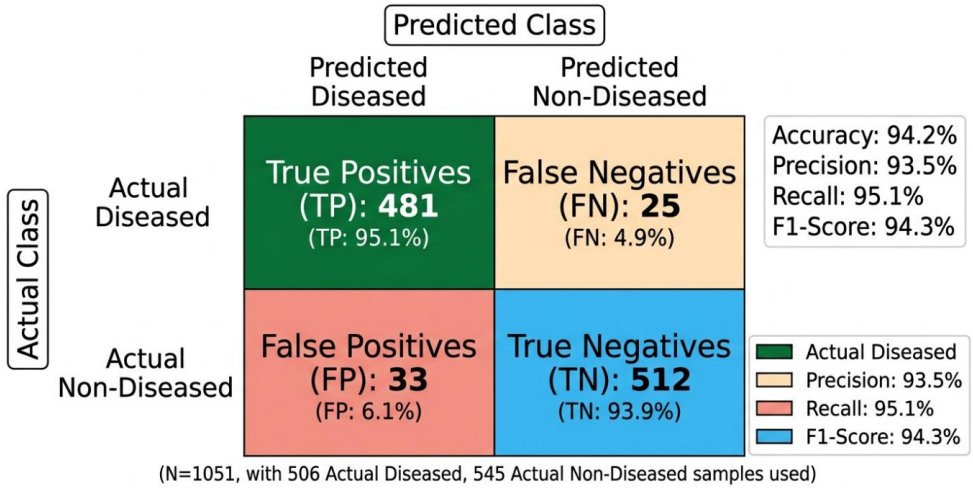


Fig 3. Confusion Matrix of the Proposed Multi-Modal Model.

7. Applications

The suggested multi-modal deep learning system has a high potential in a range of healthcare applications because of the possibility of unifying heterogeneous data streams and predicting diagnoses with high accuracy. A major use is in clinical decision support systems (CDSS), where the model may be used to assist the healthcare professional in making an informed diagnostic decision. The system can be used to offer a more detailed evaluation of a patient by integrating the imaging data with patient clinical records, thus enabling accurate diagnosis and better treatment plans. Its accuracy and reliability are very high and hence the model is applicable in implementation in hospital settings to aid the physicians in real-time decision-making.

The framework is also very suitable in making early diagnosis of diseases, which is vital in enhancing patient outcomes. The model can identify any slight patterns of diseases at an early stage which a single-modality analysis can be unable to identify because of the two imaging features and clinical indicators utilized. This will allow timely intervention, decrease the progression of the disease, and increase survival especially in diseases like cancer, cardiovascular diseases, and neurological disorders. The second application that is crucial is to use the framework in multi-hospital AI, as the framework can be installed in distributed healthcare facilities. Its capacity to process and deal with varied data types facilitates its use in large scale healthcare networks where the diagnostic work across and across multiple hospitals can be performed uniformly and to the same standard. This enables cooperative health care settings and allows the study of vast amounts of medical data to enhance the management of population health.

Moreover, the suggested system can be incorporated into intelligent healthcare monitoring systems, such as remote and wearable-based systems. The model will allow real-time monitoring of the health condition of patients and giving them proactive warning of probable medical ailments by constantly examining the patient information obtained through several sources. This can be especially useful in the field of chronic disease management and remote patient care, where constant monitoring is also needed to control the disease and prevent it. In general, the multi-modal framework proposed is a versatile and robust framework that can be utilized to support the development of the next-generation intelligent healthcare systems and provide real-time, accurate, and scalable medical diagnosis and monitoring.

8. Limitations

Although the proposed multi-modal deep learning framework has been working incredibly well, there are some limitations that must be taken into consideration. Computational complexity of the model is one of the main issues. Multi-data modalities: Medical images and structured clinical data of high dimensions are data modalities that demand substantial computational resources to train and infer. Deep convolutional networks and extra fusion and classification layers also add to memory and processing time, potentially limiting their use to real-time applications in resource-limited settings like small clinics or edge devices. The other limitation is due to the problems of data heterogeneity. Medical images and clinical data vary vastly in form, magnitude, and quality. The differences in imaging protocols, equipment specifications, and patient demographics across datasets may cause inconsistencies that may impact model performance. Moreover, the clinical data is usually full of missing values, noise, and disproportionate distributions, which may affect the integrity of the learned representations. Even with preprocessing and normalization techniques, it is challenging to completely deal with these heterogeneity challenges.

More so, the framework has not been extensively tested in the real world. Although the model has shown a good performance on benchmark data sets, its performance on actual clinical set-ups might be different since the data distribution, workflow integration, and operational constraints might be diverse. Other factors that affect the practical implementation include interoperability with hospital information systems, data privacy regulations, and user acceptance. Scalability, reliability, and applicability to clinical settings of the model require extensive validation in real-world healthcare settings. On the whole, these drawbacks indicate that further optimization, standardization, and practical testing are necessary in order to achieve maximum potential of the suggested multi-modal diagnostic framework.

9. Future Work

Even though the suggested multi-modal deep learning model reveals remarkable results in terms of combining clinical and imaging to diagnose a disease, a number of promising solutions can be envisaged to boost its capacity and practical implementation. The addition of transformer based multimodal learning architectures is one of these directions. Recent developments in models of transformers have demonstrated greater capability in modelling long-range dependencies and complex cross-modal interactions. The combination of vision transformers processing imaging data and multimodal transformers learning to represent joints can be a great way to enhance the feature interaction and the overall accuracy in making diagnoses. This method can provide a better management of heterogeneous data and offer more contextual knowledge than traditional CNN-based techniques.

The other important area that can be developed in the future is real time deployment in hospitals. Even though the existing model is very accurate, it is necessary to optimize the architecture towards low-latency inference and efficient use of resources in order to use it in practice in a clinical setting. Some of the techniques that can be considered include model compression, pruning and hardware acceleration which can facilitate smooth integration with the hospital information

systems and allow healthcare professionals to make real-time decisions.

Combining both wearable and remote monitoring devices with the suggested framework is also an important opportunity. The system can facilitate constant health tracking and prevent the occurrence of disease development by adding real-time physiological data provided by wearable sensors and imaging and clinical records. This extension would be devoutly useful in the treatment of chronic illnesses and facilitating individualized medical care solutions.

Also, new methods of federated learning could be implemented to overcome the privacy and security issues. In the medical sector, patient sensitive information is usually shared among various institutions and thus, it is difficult to centralize training. Federated learning enables training models to be collaboratively trained on decentralized datasets without exchanging raw data, and models are privacy-preserving and enhance the generalization of the model in different populations. All in all, these future directions will transform the scalability, efficiency and privacy of the proposed multi-modal framework, and the next generation intelligent healthcare frameworks that would provide accurate, real time, and secure medical diagnosis.

10. CONCLUSION

This paper presented a new multi-mode deep learning architecture to perform a comprehensive diagnosis of the disease through an integration of medical images and formal clinical records. The framework successfully generates an effective use of the complementary information of both modalities by a solid feature extraction and intermediate fusion plan, which creates a single representation that is used to perform correct classification. Experimental findings reveal that the given approach is much better at diagnostic performance as it is more accurate, sensitive, and has a higher AUC than the traditional single-modality models. Combining heterogeneous data sources does not only lead to a higher reliability in prediction, but also increases the robustness of models and their generalization in different conditions. Clinically, the framework offers a potent instrument to assist health professionals to make beneficial decisions at an early stage and accurately diagnose the condition. Altogether, the suggested system emphasizes the possibilities of multi-modal deep learning to evolve intelligent, data-driven, and scalable health solutions.

REFERENCES

1. Atek, S., Mehidi, I., Jabri, D., &Belkhiat, D. E. C. (2025). Deep learning for multi-modal medical image segmentation: a survey and comparative study. Brain Imaging and Behavior, 1-26.
2. Bhattacharya, S., Prusty, S., Pande, S. P., Gulhane, M., Lavate, S. H., Rakesh, N., &Veerasamy, S. (2025). Integration of multimodal imaging data with machine learning for improved diagnosis and prognosis in neuroimaging. Frontiers in Human Neuroscience, 19, 1552178.
3. Chen, W., Zhao, Z., Yao, J., Zhang, Y., Bu, J., & Wang, H. (2025). Multi-modal medical diagnosis via large-small model collaboration. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 30763-30773).
4. Chen, Y. W., Lin, L., & Jain, R. K. (2025). Multimodal Deep Learning for Medical Image Analysis. In Recent Advances in Deep Learning for Medical Image Analysis: Paradigms and Applications (pp. 129-150). Cham: Springer Nature Switzerland.
5. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K.,& Dean, J. (2019). A guide to deep learning in healthcare. Nature medicine, 25(1), 24-29.
6. Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., &Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine, 3(1), 136.
7. Khader, F., Mueller-Franzes, G., Wang, T., Han, T., Arasteh, S. T., Haarburger, C., ...&Truhn, D. (2022). Medical diagnosis with large scale multimodal transformers: Leveraging diverse data for more accurate diagnosis. arXiv preprint arXiv:2212.09162.
8. Lahat, D., Adali, T., &Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE, 103(9), 1449-1477.
9. Rajkomar, A., Dean, J., &Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.
10. Shurrab, S., Nepal, A., John, T. J. L. S., Ghazi, N. G., Piechowski-Jozwiak, B., &Shamout, F. E. (2025, July). Multimodal Deep Learning for Stroke Prediction and Detection using Retinal Imaging and Clinical Data. In 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1-7). IEEE.
11. Yao, Z., Lin, F., Chai, S., He, W., Dai, L., &Fei, X. (2024, August). Integrating medical imaging and clinical reports using multimodal deep learning for advanced disease analysis. In 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE) (pp. 1217-1223). IEEE.

12. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018, September). Unet++: A nested u-net architecture for medical image segmentation. In International workshop on deep learning in medical image analysis (pp. 3-11). Cham: Springer International Publishing.