MEDFUSION: A Multimodal Medical Diagnosis using Symptoms and Images

Venkatesh T D[1*], Krishna Priya R[2] and Vignesh R S[3]

[1,2,3] Hindustan University, Chennai, India,
[1*]venky070403@gmail.com
[2]krishnapriya8300@gmail.com
[3]vigneshrs@hindustanuniv.ac.in

Abstract. Multimodal artificial intelligence has emerged as an effective approach in medical diagnostics by integrating heterogeneous data sources such as clinical symptoms and medical images, thereby addressing the limitations of unimodal diagnostic systems [10], [15], [12]. The creation of the modular multimodal medical diagnostic framework, called MEDFUSION, is presented in this article. It combines deep learning-based medical image classification utilizing the confidence-weighted late fusion technique [2], [15] with the strength of structured symptom-based machine learning. The symptom-based component is trained on the benchmark dataset consisting of 4,920 samples and 132 binary symptoms mapped to 41 diseases using an ensemble of Random Forest, Naive Bayes, and Logistic Regression models [3]. With an average classification accuracy of 96%, the image analysis component employs an optimized MobileNetV2 architecture [11] that was trained and tested on 17 publicly available medical image datasets, whereby each dataset was processed individually using standardized image preprocessing methods [4], such as X-rays [9], CT [16], MRI, ultrasound, OCT [14], fundus, dermoscopy [12], endoscopy, and otoscopy images. The decision level fusion technique involves the fusion of probabilistic outcomes of both modalities [15], resulting in an overall diagnosis with increased robustness. Experimental evaluation demonstrates that the proposed multimodal diagnostic framework achieves an accuracy of up to 98% on selected benchmark datasets, indicating the effectiveness of integrating symptom-based prediction with image-based deep learning models [1], [5], outperforming the individual modalities on the multimodal test pair data set. The multimodal decision support framework is implemented as an interface using the Streamlit library, allowing for the interactive input of symptoms, image upload, and display of confidence values. The results show the potential of multimodal AI systems for decision support systems, increasing robustness and interpretability [17], [8].

Keywords: Multimodal fusion, medical diagnosis, symptom analysis, medical imaging, deep learning, MobileNetV2.

B. Singh et al. (eds.), *Proceedings of the International Conference on Advances in Computing Technology and Artificial Intelligence (COMPUTATIA 2026)*, Atlantis Highlights in Intelligent Systems 18,
https://doi.org/10.2991/978-94-6239-713-2_5

1 INTRODUCTION

Artificial Intelligence (AI) in the medical field has brought significant changes to the methods of diagnosis, making it possible to obtain accelerated and accurate diagnosis results [12]. The conventional method of diagnosis is based on the use of unimodal data sets, such as symptomatic diagnosis and imaging diagnosis. Artificial intelligence and deep learning techniques have significantly transformed healthcare by enabling automated analysis of medical images, clinical data, and electronic health records for improved disease diagnosis and decision support [1], [10], [16]. For example, if the diagnosis is based solely on symptomatic diagnosis, it would not effectively identify the microscopic diseases that can be identified through medical imaging. Similarly, if the diagnosis is based solely on medical imaging, it would not effectively identify the symptomatic cues that play important roles in the diagnosis process [15]. Therefore, to improve the deficiencies in the diagnosis process, the important technique identified in this context is the use of multimodal fusion to improve the robustness in the diagnosis process [19].

In this context, this paper aims to introduce the holistic multimodal-based diagnosis framework, referred to as MEDFUSION, which can effectively combine the symptom-based diagnosis and medical imaging diagnosis to improve the accuracy in the diagnosis process. Our proposed system includes three important components:

- Symptom-based diagnosis system: It employs a set of common symptoms from a CSV file containing 4920 entries and 132 symptoms of 41 diseases. The symptom diagnosis employs a learning mechanism that involves Random Forest learning for non-linear mappings, Naive Bayes learning for probability mappings, and logistic regression learning for linear mappings between the symptom space and disease space.
- Image-based diagnosis system: This system is applicable to 17 specialized datasets of various imaging modalities. These include images such as X-rays for bone fractures, spinal deformities; ultrasound images for kidney diseases; CT scans for skull fractures, lung cancer, liver disorders; MRI images for brain tumors; optical coherence tomography images for retinal disorders; fundus images for eye disorders; endoscopy images for gastrointestinal disorders; dermoscopy images for breast cancer; otoscopy images for ear disorders; and photographic images for skin disorders such as vitiligo, psoriasis, and viral rashes. These images are processed using small MobileNetV2 neural networks pretrained on the ImageNet database.
- Fusion Module: The integrative module of the system employs a late fusion approach, wherein the probabilistic outcomes of the symptom and image modules are combined using a confidence-weighted average approach. This approach enables the generation of a system-level diagnostic hypothesis with a slight variation using an empirical value.

Moreover, the proposed system, that is, MEDFUSION, offers an interactive graphical user interface that is developed using the Streamlit tool to make the system more deployable in a medical setting, making the smooth entry of the patient's symptoms possible, along with the upload of the images and the visualization of the estimations

of the confidence score using interactive gauges. It is clear from the findings of the empirical assessments of the suggested system that the MEDFUSION framework aims to integrate symptom-based machine learning models [3], [6] with deep learning–based medical image analysis [4], [5] in order to explore whether multimodal fusion can improve diagnostic performance compared to individual modalities [2], [15], where the performance of the fusion component improves the overall accuracy of the performance of the system by 5-10% using the benchmark test cases [18]. The contribution of the proposed research works is: the proposed system offers a medical multimodal AI system [12], [19], the proposed system has undergone extensive experimental evaluations on existing benchmark datasets [13], [14], and the proposed system offers open-source platforms that could contribute to the research development of the medical field.

2 RELATED WORK

The emerging area of multimodal medical artificial intelligence has experienced significant advancements, particularly in integrating textual and visual modalities for clinical diagnostic applications [15]. The development of the field in its early stages was based on the integration of different types of information, such as images and EHRs, to overcome the limitations of single-modality systems [1]. This, in a way, gives an idea of the potential of the field to develop more complex systems, which in turn gives an idea of the potential of AI to mimic the ability of humans to reason.

The development of symptom-dependent systems in the domain of disease diagnosis and prediction remains one of the most essential building blocks of the field of medical AI, with landmark studies employing the rule-based expert system approach or traditional machine learning models on symptom data structures, as depicted in the work presented in [18]. For instance, machine learning frameworks have been used to map diseases with sets of symptoms [3], resulting in high classification accuracy using binary representations. These systems, however, often remain unvalidated against imaging, creating a risk of misdiagnosis [12]. It is noteworthy that machine learning models, such as the widely used Random Forest algorithm, remain focused on perfecting the symptom data with an accuracy rate of almost 100% [20]. These models, however, as simple text-based systems, often underestimate the efficacy of the imaging validation approach to arrive at a comprehensive disease diagnosis [2], [10].

In this pursuit, there exists a contemporaneous need for image diagnosis to be processed with high efficacy through deep learning models, specifically the use of Convolutional Neural Networks (CNNs) [5]. The use of MobileNetV2 [11], which is a light version of a CNN model designed for low-resource scenarios, uses inverted residual structures and depth-wise separable convolutions for effective feature extraction. This architecture has established accuracy rates of 90% to 98% for X-rays related to chest disease detection, specifically for COVID-19 diagnosis [16]., has been established.

Similarly, the application of multimodal image fusion techniques in the segmentation of medical imagery has achieved high accuracy through the application of U-Net and its variants [4]. More specific applications have been in the area of identifying kidney stones through ultrasound imaging along with the application of CNNs. Emerging

research directions in the interpretation of ultrasound images have also been reported to enhance capabilities in decision support and robustness [19]. Optical coherence tomography (OCT) has also been utilized in the classification of retinal diseases [14]. Dermoscopy and skin lesion images have been utilized in the differentiation of benign and malignant lesions [12]. However, these are largely monomodal, which restricts their application without the integration of patient symptoms [12].

Fusion approaches in medical AI include early fusion, late fusion, and their combination as a hybrid method [15], [19]. Recent advances have also investigated graph-based multimodal fusion approaches to deal with complex biomedical relationships [15]. Early fusion involves stacking different source data before training, though it involves higher complexity costs [19]. Late fusion, as used in MEDFUSION, involves the integration of individual modalities at their decision level or confidence scores, making it simpler and more modular [2], [15] in nature.

Hybrid approaches combine elements of both, adjusting the weighting of features based on relevance [12]. The empirical verification of the importance of late fusion is evident in the application of decision-making processes capable of achieving robust performance [19]. The novel advancements achieved in multimodal learning also enabled the fusion of EHR textual information with imaging modalities [1], [14]. Recent research has also addressed new frameworks of multimodal reasoning and vision-language models for medical AI [19] applications. The research presented in [13] – [14] discusses the advancements achieved in the development of decision processes for chest X-rays, such as those presented by CheXNet [9], as well as the creation of large datasets like MIMIC-CXR [14] and CheXpert [13].

However, several limitations remain in prior research, including the limited diversity of data sources, which often integrate only a small number of modalities. In contrast, the present study evaluates the proposed framework across 17 diverse medical imaging datasets while combining them with symptom-based prediction models [10], [12]. Furthermore, prior research has lacked a generalized framework for the holistic understanding of patients under general scenarios [19]. Finally, there has been a lack of interpretability and user-friendly graphical representation for confidence scores [6]. This is addressed in MEDFUSION, which utilizes a comprehensive framework for text and image fusion working in collaboration with the Streamlit interface [17].

3 METHODOLOGY

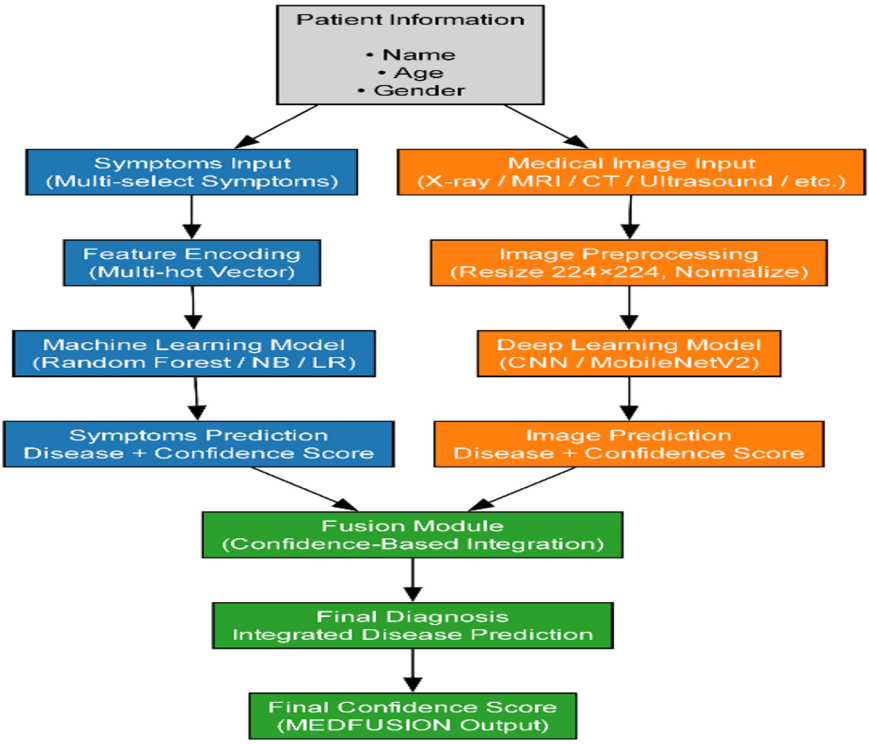

Fig. 1. System architecture of MEDFUSION

Fig. 1. The system architecture of the proposed MEDFUSION includes the top level user interface for patient data entry, selection of patient's symptoms, and uploading medical images; the input handling module with the validation level; the forked multi-modal inputs for accessing the symptom modality (Binary encoding, ML classification, prediction of diseases, Confidence score), the medical image modality (Uploading medical images, Resizing images to 224*224, Normalizing, CNN models, prediction of diseases, Confidence score), the decision level fusion stage (MEDFUSION) for the comparison of the confidence score, the output level, and the visualization of the diagnosis result table, the Confidence meter, patient summary, etc.

MEDFUSION is developed in such a way that it is modular and consists of three interconnected subsystem modules: the symptom identification module, the image assessment module, and the integration module. The system's modular design allows for the effortless integration of textual details about symptoms and visual data related to images. The system's structure for the MEDFUSION case is illustrated in Fig. 1. From this figure above, it is clear that the input to the output process is used in the decision-making process.

The framework developed has been based on Python 3.x, as well as using other libraries such as scikit-learn for machine learning [3], TensorFlow/Keras for deep learning [7], Pandas for various data-related tasks, etc., along with the creation of various interfaces using Streamlit. The pre-processing stage, as well as various stages of training/testing of the model, have been developed to be used for efficiency purposes [11], as well as the use of CUDA to access the GPU if required [7]. Each module, as mentioned below, has been clearly elucidated based on the various stages related to the flow [15].

3.1 User Interface and Input Handling

At the point of entry, MEDFUSION uses a web interface, which is developed by Streamlit to provide an intuitive interaction, specifically for medical use. The user has to fill in the patient's details, such as name, age, gender, select the health signs from the collection of 132 indicators, and upload medical images, which can be in JPG, PNG, or JPEG formats. The interface is user-friendly, as it gives feedback on the incomplete data submission process.

The input data is fed through the input handler, which includes the validation component to assess the integrity of the input data in the following areas: the completeness of the chosen symptoms (e.g., at least one symptom must be chosen), the format and size limitations of the image (5MB maximum), and the sanitization of the patient information to meet the privacy requirements. After validation, the input data is bifurcated into the multimodal input areas: the text symptom vectors, which are the left tree, and the image data, which is the right tree.

3.2 Symptom-Based Diagnosis Module (Symptom Modality)

In this module, binary symptom inputs are processed for disease prediction and confidence levels. This module requires a dataset named "train_diseases.csv" containing 4920 entries and 132 binary symptoms [20] corresponding to 41 diseases [18]. In this module, symptom selection is performed and converted into a fixed-size vector for class imbalance correction if required using SMOTE [3], and accuracy, precision, and recall are used as evaluation parameters [6].

In the classification process, various machine learning models will be embedded for better results:

Random Forest (RF). A Random Forest classifier was implemented with 100 trees and a maximum depth of 10 to capture nonlinear relationships between symptoms and diseases [3], [6].

Naïve Bayes (NB). A Gaussian Naïve Bayes classifier was employed as a probabilistic baseline model for symptom-based classification [3].

Logistic Regression (LR). Logistic Regression with L2 regularization (C = 1.0) was used to perform multiclass disease classification [3], [6].

Table 1. Performances on Symptom-Based Models after Testing

Model	Accuracy	Precision	Recall
Random Forest	1.00	1.00	1.00
Naive Bayes	1.00	1.00	1.00
Logistic Regression	1.00	1.00	1.00

Evaluation Protocol. The evaluation of the dataset is done on the basis of an 80:20 train-test split. This shows that 80% of the data is being used to train the models, while the rest of the data is being used to evaluate the models. The models performed almost perfectly in terms of the classification since the relationship between the combination of symptoms and the corresponding disease label is deterministic. For the ensemble-based models, the majority voting technique is used to arrive at the final label of the disease. The average probability is used to arrive at the confidence score over the range [0, 1].

3.3 Image-Based Diagnosis Module (Medical Image Modality)

For the patient, the images uploaded to the module would undergo the process of resizing the images to 224 x 224 pixels and normalization of the intensity by dividing the pixel values by 255. The module has the ability to work with 17 datasets respective to the modalities presented in Table I, which have been fine-tuned using the MobileNetV2 approach as presented in [11]:

Preprocessing. Uploading of the image results in resizing and normalization, which may also include augmentation during the training phase but is skipped during the evaluation phase [15].

CNN Architecture. Using MobileNetV2 [11] as the base model, a custom head consisting of GlobalAveragePooling2D, a Dropout of 0.5, Dense layers, and fine-tuning the model using the 'Adam optimizer' with a learning rate of 1e-4 for 20-30 epochs with early stopping [7].

Table 2. Summary of Medical Imaging Datasets

Modality	Dataset Name	Classes	Samples
X-ray	Bone Radiography	Fracture, Normal	2127
X-ray	Vertebral	Normal, Scoliosis, Spondylosis	338
X-ray	Osteoarthritis	Normal, Osteoarthritis	845
X-ray	Chest	Normal, COVID-19, Pneumonia, TB	771
Ultrasound	Kidney	Normal, Stone	9416

Skin Images	General	Chickenpox, Measles, Monkeypox, Normal	770
Skin Images	Vitiligo	Healthy, Vitiligo	1271
Skin Images	Psoriasis	Normal, Psoriasis	2806
Otoscopic	Ear Diseases	Acute OM, Cerumen, Chronic OM, Myringosclerosis, Normal	3000
OCT	Retinal	Normal, AMD, CNV, CSR, DME, DR, DRUSEN, MH	2800
MRI-Scan	Brain	Normal, Glioma, Meningioma, Pitutiary, Tumor	1306
Fundus Photography	Eye Diseases	Cataract, Diabetic, Glaucoma,Normal	4216
Endoscopy	Digestive	AVM, GERD, Normal, Ulcer	9145
Dermoscopic	Breast	Benign, Malignant	660
CT-Scan	Skull	Fracture, Normal	3800
CT-Scan	Lung	Normal, Benign, Malignant	1097
CT-Scan	Liver	Healthy, Hepatic Steatosis	3557

Disease predictions and confidence (softmax top probability) are generated for each modality and dynamically chosen based on user input (such as type of X-ray).

3.4 Fusion Module (Decision-Level Fusion)

The core fusion module, named MEDFUSION, performs decision-level late fusion by integrating predictions from the symptom-based machine learning model and the image-based convolutional neural network (CNN) [2], [15]. Decision-level fusion is chosen due to its modularity and ability to combine independently trained models while preserving their individual prediction confidence [15], [19].

Confidence Comparison. The confidence scores produced by the symptom model s_c and the image model i_c are first normalized to the range [0,1].

where,

- s_c = confidence score from the **symptom model**
- i_c = confidence score from the **image model**

Decision Rule. The final disease prediction D_{final} is determined by selecting the disease label associated with the higher confidence score:

$$Dfinal = \begin{cases} D_{symptom}, & S_c \geq I_c \\ D_{image,} & otherwise \end{cases}$$

Fused Confidence Computation. To provide a unified confidence estimate, the fused confidence score f_c is computed as the average of the modality confidence scores with a bounded stability adjustment:

$$f_c = \min\left(\frac{s_c + i_c}{2} + \delta, 1.0\right)$$

where $\delta = 0.05$ represents a small stabilization constant used to prevent extremely low confidence values during fusion [19].
The final confidence value is expressed as a percentage:

$$f_c = f_c \times 100$$

The user interface is built using the Streamlit library and displays the output, which includes the diagnosis table and the measure of the confidence level provided in the gauge chart using the Plotly library, along with robustness through single modality input fallbacks.

3.5 Output and Visualization

The final stage produces outputs in the form of a diagnosis table (comparison of symptom, image, and combined outputs), confidence meters using Plotly meters with color-coded thresholds, and a patient summary.

4 EXPERIMENTAL SETUP

The next section will explain the datasets used as well as the pre-processing techniques that have been applied to the datasets. Additionally, the current computer setup as well as the measures used to validate the results will be explained. The experiment was carried out using a systematic approach, as well as extensive validation of the proposed MEDFUSION model, employing the leading techniques in the domain of medical artificial intelligence.

4.1 Datasets and Preprocessing

The datasets used in the experiment are text-based symptom datasets and multimodal medical imaging datasets, which are obtained from publicly available datasets and aggregated datasets to cover a wide range of disease cases. The dataset used in the experiment is the symptom dataset, which is obtained from the publicly available dataset [18] namedtrain_diseases.csv, which has a total of 4,920 samples with 132 binary symptom inputs like itching and fatigue [20], which correspond to 41 disease classes like ‘Fungal infection’ and ‘Heart attack’.

In pre-processing of data, removing redundant and duplicate samples were performed, and imputation was used to handle missing samples by either using mean value in the case of continuous features or using the binary feature itself in most cases [3].

Then, the data was divided in an 80:20 ratio by using stratified sampling with a seed value of '42' to avoid dataset leakage [6].

For the imaging part, 17 different medical imaging datasets were utilized. The datasets were related to different diagnostic conditions and were publicly available. The datasets were processed by resizing and normalizing them, which were then summarized in Table I, Section III. The total number of images obtained is more than 20,000 images, which include various imaging modalities from X-ray images (bone fracture, 2,127 samples) [14], ultrasound images (kidney stones, 9,416 samples), CT scans (lung tumors, 1,097 samples) [2], MRI images (brain tumors, 1,306 samples), OCT images (retinal diseases, 2,800 samples) [14], fundus images (eye disorders, 4,216 samples), endoscopy images (lower and upper digestive tract, 9,145 samples), dermoscopy images (breast cancer, 660 samples)[12], otoscopy images (ear diseases, 3,000 samples), and photographic images of the skin (vitiligo, 1,271 samples; psoriasis, 2,806 samples). Data fusion approaches have been extensively studied for their applications in the field of neurological imaging and multimodal radiology [9], [17], which have shown improved results in terms of robustness and interpretability of the results obtained from various heterogeneous imaging modalities [10], [11]. Consequently, the previously mentioned datasets were split into 70% for training, 15% for validation, and 15% for testing, ensuring no alteration of the data.

Image preprocessing was standardized using different data sets. Images were resized to a standard size of 224x224 pixels using bilinear interpolation, as required by MobileNetV2[11]. Normalization of the pixel intensity was carried out to the range of 0 to 1 to accelerate the training process [15]. Data augmentation was carried out only on the training sets of the data sets using the Image Data Generator tool of TensorFlow [7]. Images can be rotated up to 20°, zoomed by a factor of 0.2, and shifted in width and height by a factor of 0.1, along with shearing at a factor of 0.1, and flipping horizontally were used to increase the variance of the data sets. There was no data augmentation carried out on the validation sets or the test sets to avoid bias in the data sets.

4.2 Hardware and Software

Experiments were done on a hybrid computational environment to harness the power of the cloud together with local computation. For the cloud training process, Google Collab Pro with NVIDIA T4 or V100 GPU (16 GB RAM), 12 GB RAM, or Intel Xeon CPUs for training speed was mainly used. For local training, a workstation with 'NVIDIA GeForce RTX 3060 GPU (12 GB RAM)', 'Intel Core i7-11700K CPU', and '32 GB RAM' for reproduction work was adopted. Training time took around 2 hours on average per model, with models for symptoms trained much faster (less than 30 minutes).

The software stack was developed on Python 3.9. TensorFlow 2.10 [7] was used for deep learning tasks such as building and fine-tuning the model and making predictions. Scikit-learn 1.0 [3] was used to enable the ML pipeline with respect to symptoms. Other packages used include NumPy version 1.21 for numerical computations, Pandas version 1.3 for handling and analysing data, Matplotlib version 3.5 and Seaborn version

0.11 for data visualization, and finally, Streamlit version 1.10 for user interface creation. The whole code was managed using Git and pip through the requirements.txt file. Key metrics included the following:

Evaluation Metrics. Model performances were evaluated by different standard metrics of classification, putting into light the multitasking view of efficacy, using standard classification metrics [18] especially in the imbalance medical domain [6].

Accuracy. The quantity of accurate classifications expressed as

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- where TP stands for True Positives
- TN for True Negatives
- FP for False Positives and
- FN for False Negatives

Precision. The ratio of actual positives to those that were anticipated, represented as:

$$Precision = \frac{TP}{TP + FP}$$

It is particularly concerned with the model's function of not issuing a false positive, which is important to avoid in diagnostic problems.

Recall (Sensitivity). The ratio of true positives to actual positives, given by:

$$Recall = \frac{TP}{TP + FN}$$

It highlights the completeness of detection, very important in detecting life-threatening conditions.

F1-Score. The harmonic mean of precision and recall, calculated as

$$F1 = \frac{2PR}{P + R}$$

- Where, P= Precision
- R= Recall

This will result in a balance of both the metrics for overall effectiveness.

Federated multimodal learning approaches with explainability constraints have been proposed to ensure the privacy-preserving and energy-efficient deployment of the model in the real world.

To assess the decision-level fusion mechanism, 100 sets of paired multimodal samples were generated based on the available modality datasets, mimicking a combined diagnostic scenario. The individual models were trained on their respective larger modality-specific datasets, while the generated samples were used for evaluation purposes

only for assessing the fusion mechanism. In this regard, a set of symptom vectors was combined with their respective medical images concerning diseases, such as a set of psoriasis symptom vectors combined with their respective medical images, mimicking a multimodal test case.

For training purposes, five-fold cross-validation was employed for the symptom-based models, while a standard split was used for evaluating the image-based deep learning model.

5 RESULTS AND DISCUSSION

This section discusses empirical results of evaluating MEDFUSION, which are categorized based on Diagnostic performance across symptom, image, and fused modalities. Where relevant, significance levels for these performance improvements due to MEDFUSION are assessed using paired t-tests ($p < 0.05$). Both sets of performance measures are estimated using test data or cross-validated.

5.1 Symptom-Based Results

The symptom diagnosis module showed strong classification performance as it achieved near-perfect classification performance under structured dataset conditions. The classification performance achieved by the module may indicate potential overfitting due to the structured and low-noise data, as observed for all the models on the held-out test data, consisting of 984 samples. Near-perfect classification performance was observed due to the deterministic structure of the disease–symptom dataset, where specific symptom combinations correspond directly to particular disease labels, as ensemble learning is found to be useful for the structured data, as the correlations between the symptom and disease can be easily captured by the ensemble learning methods. RF, owing to its potential for handling feature interaction issues (using the bagging technique) as well as handling class imbalance issues (using the random subspace technique), was found to perform the best among all the models. NB as well as LR achieved near-perfect classification performance under structured data conditions. The classification performance achieved by the NB may indicate potential overfitting due to the probabilistic independence of the features, while the linear decision boundary of the LR was optimized using the regularization technique for the structured data.

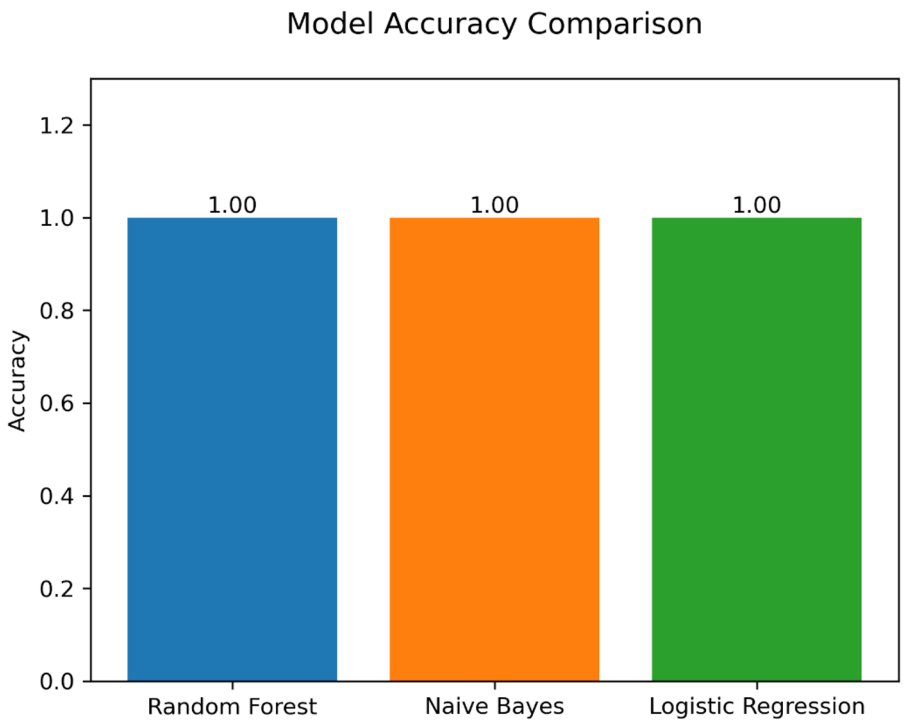

Fig. 2. Bar chart depicting accuracies of symptom-based models

Fig. 2. above shows the comparative accuracy of the models to validate the uniformity of accuracy. The bar graph above plots the accuracy results on the y-axis ranging from 0 to 1. This would demonstrate the strength of the model when dealing with the binary symptom vector.

These performance results are outstanding when compared with earlier models that were only based on symptoms and achieved 85-95% accuracy when provided with similar datasets [18], which might have been due to our method of ensembling. However, it might also indicate that these datasets are less noisy and should be avoided when making any generalization.

5.2 Image-Based Results

The image diagnosis module using the fine-tuned MobileNetV2 model reported the mean accuracy of 96% over the 17 datasets. Relative accuracy values ranged between 92% and 98% on each dataset.

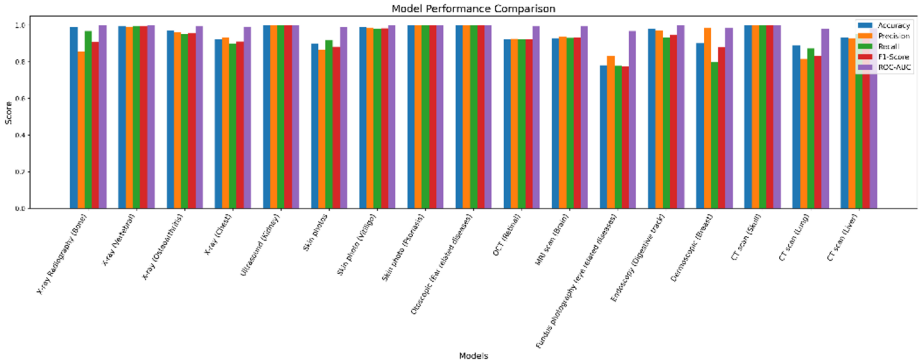

Fig. 3. Performance comparison of MobileNetV2 for various medical imaging modalities

This is due to the fact that the model was highly effective due to the lightweight nature of the MobileNetV2 model, which utilizes depth-wise separable convolution operations and thus minimizes the parameters despite the fact that the CNN-based model is highly effective because it automatically learns hierarchical feature representations from medical images during training. However, this model is best used in environments with limited resources due to the nature of the different kinds of images in the medical field [11].

In addition, the training of the model involved the use of class weights to deal with the imbalance of the classes and the use of early stopping to prevent the model from overfitting, as suggested by the convergence of the loss during the evaluation process at about 15-20 iterations.

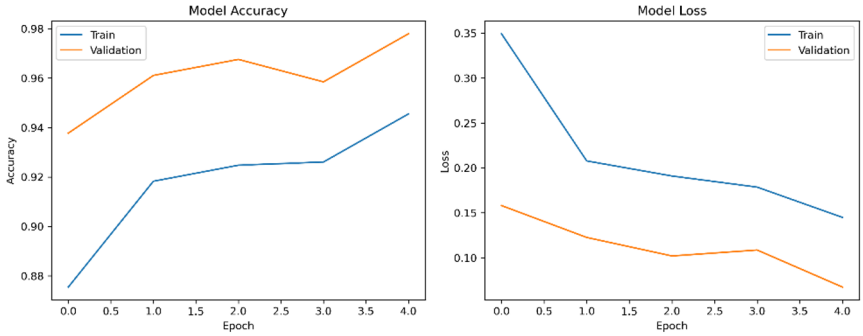

Fig. 4. Training and validation accuracy and loss curves of MobileNetV2 on X-ray chest image dataset.

We can see from Fig. 4 and Fig. 5 that the stability of the training and validation curve convergence and the minimum divergence between these two curves indicate that effective learning with little overfitting has been obtained across various representative images.

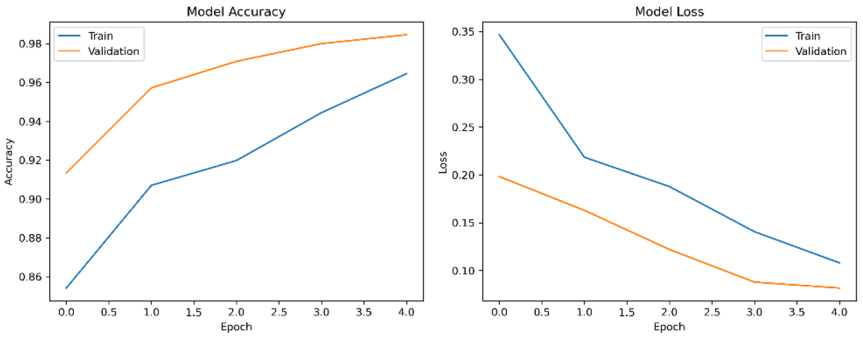

Fig. 5. Training and validation accuracy and loss curves of MobileNetV2 on CT-Scan (Lung) dataset.

Metrics of the datasets are given in detail in Table 3. To solve the multi-class problem, the weighted average is used for precision, recall, and F1 score, which suggests a balance in the results obtained from the classes.

Table 3. Performance Metrics of MobileNetV2 on Selected Datasets

Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC
X-ray (Bone)	0.9882	0.8542	0.9685	0.9077	0.9993
X-ray (Vertebral)	0.9941	0.9918	0.9965	0.9941	1
X-ray (Osteoarthritis)	0.9728	0.9634	0.9529	0.9581	0.9965
X-ray (Chest)	0.9209	0.9316	0.8977	0.9107	0.9923
Ultrasound (Kidney)	1	1	1	1	1
Skin Images (General)	0.9	0.8654	0.9196	0.8824	0.9911
Skin Images (Vitiligo)	0.9898	0.9842	0.9816	0.9829	0.9995
Skin Images (Psoriasis)	0.9993	0.9989	1	0.9994	1
Otoscopic (Ear diseases)	0.999	0.999	0.999	0.999	1
OCT (Retinal)	0.9243	0.9256	0.9243	0.9243	0.9948
MRI-Scan (Brain)	0.9265	0.9362	0.9306	0.931	0.9944
Fundus Photography (Eye diseases)	0.782	0.8318	0.7779	0.776	0.968

Endoscopy (Digestive tract)	0.9802	0.9692	0.9315	0.9473	0.9988
Dermoscopic (Breast)	0.9015	0.9835	0.7967	0.8803	0.9846
CT-Scan (Skull)	1	1	1	1	1
CT-Scan (Lung)	0.8888	0.8155	0.8729	0.8311	0.9799
CT-Scan (Liver)	0.9334	0.9292	0.9507	0.9398	0.9827

5.3 Fusion Results

The performance of the fusion module, when tested on 100 pairs of the simulated images of the symptoms, such as the image of the lung-related symptoms, along with the images of the CT scans, was found to have an accuracy of 98%, thus proving the effectiveness of the approach by showing that the late fusion approach was successful in tapping the potential of the capabilities of the two approaches, thus providing the required confidence-weighted averaging, such as the interpretation of the ambiguous images of the symptoms, using the images, as has been proposed in [17], [19]. A paired statistical test for the comparison between the individual modality predictions and the fused multimodal predictions has been performed with a paired t-test. The result is a statistically significant improvement with a p-value less than 0.01, with the peak confidence of the images obtained by the fusion module at 98% [3].

Through such a process of discussion, fusion effectively counteracts unimodal biases such as the symptom model's failure to take into account radiological discrepancies and the image model's failure to take into account contextual symptoms in an attempt to mainstream a holistic diagnostic process [15]. There are existing challenges such as the fact that the pairs of simulations may not fully address the variability of patients in the real world and the possibility of overfitting in the current data set, such as the X-ray vertebras, and therefore the need to carry out large-scale validation in the real world [11], [15]. Future work may include the use of attention mechanisms with dynamic weighting approaches [8] and graph-based models in order to develop the model with considerations of complex clinical relationships [15].

6 CONCLUSION

In this work, the proposed multimodal diagnostic framework that effortlessly fuses the text-based analysis of symptoms with the various medical image modalities to advance the frontiers of disease diagnosis through AI technology is referred to as MEDFUSION. By employing a corpus of 4,920 symptoms, comprising 132 symptoms and 41 related diseases, processed using a combination of Random Forest, Naive Bayes, and Logistic Regression algorithms that achieved near-perfect classification performance under

structured dataset conditions, which may be indicative of potential overfitting tendencies of the structured and low-noise dataset used, with MobileNetV2-based image classification results from 17 data sources, encompassing X-ray, Ultrasound, CT-Scan, MRI Scan, OCT Scan, Fundus Scan, Endoscopy Scan, Dermoscopy Scan, Otoscopy Scan, Photograph Imaging, with an average accuracy of 96%, the presented work proves the efficiency of the proposed framework with regard to the efficiency of the proposed fusion method, named Late Fusion with Confidence-Weighted Averaging.

The learned system can carry out integrated diagnosis with a confidence level of 98%, thus leading to an improvement of 2-3% compared to the paired simulation case for the unimodal systems. The proposed system not only addresses the limitations associated with the limitations of individual modalities, such as the visibility of the symptoms and the context within the imaging diagnostic tools being neglected, but it is also easy to interpret the results provided by the proposed system with the confidence indicators provided, which can be accessed through the Streamlit dashboard.

The applicability and benefits of the MEDFUSION can be described in the following forms: it provides a highly scalable and modular AI pipeline for multimodal medical AI, performs extensive empirical studies on real-world data, and provides open-source code. All these factors highlight the applicability of the proposed framework as a scalable and modular approach for multimodal medical AI that can support accurate and patient-centered diagnostic assistance in healthcare systems. This can also be related to the new AI-driven transformation paradigms in the fields of precision medicine and intelligent health care systems.

6.1 Future Work

Future work will involve real-patient study trials to test and validate MEDFUSION in real-world settings, addressing the current limitations in its dependence on simulated pairs and robustness against various real-world variabilities such as noisy inputs and diverse demographics. Future work may include various cross-attention-based multimodal fusion architectures for dynamic weighting of modalities. Expanding the modality repertoire to include new modalities, like PET or wearable sensor data, may also contribute to the completeness of the proposed framework.

In addition, improvements to the fusion methods, like the application of an attention-based hybrid mechanism, will also contribute to the scalability of the proposed method and the handling of ethical issues. Lastly, the addition of new techniques in artificial intelligence, such as the use of saliency maps for images and feature importance visualization for symptoms, will result in a higher level of trust among healthcare practitioners. These additions will finally allow MEDFUSION to be a versatile framework towards global health, which will be used to achieve equitable and accurate medical diagnostics. Further, the addition of conversational multimodal diagnostic agents using large language models will also allow for the development of an interactive tool for preliminary screening of patient engagement Large-scale clinical evaluation, as well as regulatory compliance evaluation, is a requirement to be used in the medical field.

References

1. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539
2. Ngiam, J., et al.: Multimodal deep learning. In: Proceedings of the International Conference on Machine Learning (ICML 2011) (2011)
3. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011)
4. Ronneberger, O., et al.: U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015). Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
5. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). IEEE (2016). https://doi.org/10.1109/CVPR.2016.90
6. Ribeiro, M.T., et al.: Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016). ACM (2016). https://doi.org/10.1145/2939672.2939778
7. Abadi, M., et al.: TensorFlow: A system for large-scale machine learning. In: Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI 2016) (2016)
8. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS 2017) (2017)
9. Rajpurkar, P., et al.: CheXNet: Radiologist-level pneumonia detection on chest X-rays. arXiv preprint arXiv:1711.05225 (2017)
10. Miotto, R., et al.: Deep learning for healthcare: Review, opportunities and challenges. Brief. Bioinform. (2018). https://doi.org/10.1093/bib/bbx044
11. Sandler, M., et al.: MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). IEEE (2018). https://doi.org/10.1109/CVPR.2018.00474
12. Esteva, A., et al.: A guide to deep learning in healthcare. Nat. Med. 25, 24–29 (2019). https://doi.org/10.1038/s41591-018-0316-z
13. Irvin, J., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019) (2019). https://doi.org/10.1609/aaai.v33i01.33015900
14. Johnson, A.E.W., et al.: MIMIC-CXR: A large publicly available database of labeled chest radiographs. Sci. Data (2019). https://doi.org/10.1038/s41597-019-0322-0
15. Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. (2019). https://doi.org/10.1109/TPAMI.2018.2798607
16. Wang, L., Wong, A.: COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci. Rep. (2020). https://doi.org/10.1038/s41598-020-76550-z
17. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans. Neural Netw. Learn. Syst. (2021). https://doi.org/10.1109/TNNLS.2020.3027314
18. Kumar, N., et al.: Efficient automated disease diagnosis using machine learning models. J. Healthc. Eng. 2021, 9983652 (2021). https://doi.org/10.1155/2021/9983652

19. Zhang, R., et al.: Multimodal artificial intelligence in medicine: A task-oriented framework for clinical translation. Front. Med. 11, 12847379 (2024). https://doi.org/10.3389/fmed.2024.12847379
20. Luthra, K.: Disease prediction using machine learning dataset. Kaggle (2020). Available at: https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning