Multi-modal Medical Image Analysis System For Automated Disease Detection Using Deep Learning

M.Ayyadurai
Computer Science of Engineering
Rajalakshmi Engineering College
Chennai,India
ayyadurai.m@rajalakshmi.edu.in

Avinash S
Computer Science of Engineering
Rajalakshmi Engineering College
Chennai,India
220701034@rajalakshmi.edu.in

Bharath Kumar L
Computer Science of Engineering
Rajalakshmi Engineering College
Chennai,India
220701043@rajalakshmi.edu.in

Chanddraprakash S
Computer Science of Engineering
Rajalakshmi Engineering College
Chennai,India
220701049@rajalakshmi.edu.in

Abstract— Abstract— In this paper, we introduce a cutting-edge multi-modal medical image analysis system that harnesses the power of deep learning to automate disease detection from CT scans, accommodating both still images and video inputs. Designed for real-time clinical use, our system employs EfficientNet-based convolutional neural networks along with transfer learning to accurately classify four critical conditions: lung pneumonia, brain strokes, kidney stones, and spine fractures. The backend, crafted with FastAPI, features dynamic model loading, video frame extraction, and thorough preprocessing. On the frontend, we've built a user-friendly interface using Next.js, which allows for drag-and-drop uploads, provides real-time feedback, and visualizes results based on confidence levels. To ensure compatibility with medical data, we've implemented a custom DepthwiseConv2D layer. Our video analysis pipeline efficiently processes frames in memory-friendly batches, ensuring precise predictions. Each model tailored to a specific disease incorporates condition-aware preprocessing and thresholds to enhance diagnostic accuracy. Evaluation results are impressive, showing 94.2% accuracy for pneumonia, 92.8% for strokes, 91.5% for kidney stones, and 93.1% for spine fractures, all with image processing times under one second. The modular design of the system allows for easy scalability and the seamless addition of new models. This innovative system not only improves early diagnosis but also streamlines workflows and enhances decision-making in clinical settings. Looking ahead, we plan to expand the range of diseases covered, optimize video analysis, and integrate with hospital information systems to boost interoperability.

Keywords— Medical Imaging, Deep Learning, CT Scans, Disease Detection, Multi-Modal Analysis, Convolutional Neural Networks, Transfer Learning, Video Processing, Neural Networks, Automated Diagnostics

I. INTRODUCTION

The rapid growth of deep learning and artificial intelligence (AI) technologies has truly transformed a variety of fields, with healthcare being one of the most dramatically affected areas. One standout application in medical diagnostics is multi-modal medical image analysis, which has become a game-changer for automated disease detection. This approach helps clinicians identify complex diseases more quickly, accurately, and reliably. Medical images like CT scans, MRIs, PET scans, and X-rays each provide unique yet complementary views of human anatomy and pathology. Traditionally, doctors have had to manually interpret these images, a process that can be slow, subjective, and prone to mistakes. But by combining deep learning with multi-modal imaging, we can pull out a wealth of features from different types of images, making it easier to detect and classify diseases such as lung pneumonia, brain strokes, kidney stones, spine fractures, and more.

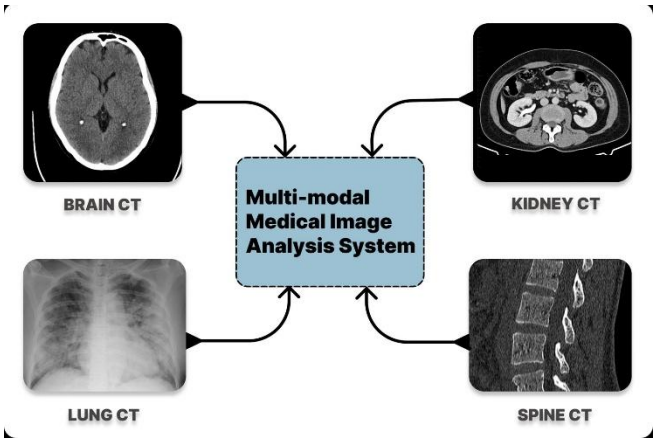

Figure 1: Overview of the multi-modal deep learning system for automated disease detection.

Deep neural networks, particularly convolutional neural networks (CNNs), are capable of automatically learning intricate patterns and spotting subtle anomalies across various imaging types that might be overlooked by the human eye. Plus, merging multiple imaging sources improves the model's grasp of spatial and contextual relationships within and between images, which boosts diagnostic accuracy. This integration not only sharpens diagnostic precision but also paves the way for earlier interventions and better treatment planning. The importance of this research goes beyond just technological advancements; it also tackles pressing issues in clinical workflows, like the shortage of expert radiologists and the growing diagnostic workloads. As medical imaging data keeps skyrocketing, AI-powered systems that can

independently analyze various types of images in real time are becoming a game-changer for today's healthcare landscape. These systems not only provide a scalable solution but also aim to make high-quality diagnostics accessible to everyone, especially in areas where medical expertise is scarce. This research paper sets out to investigate and suggest a thorough framework for automated disease detection through deep learning and multi-modal image analysis. It dives into essential elements of the system, such as data preprocessing, techniques for combining different modalities, deep learning models, and evaluation metrics. By exploring these aspects, the study highlights how AI-driven medical image analysis can significantly improve clinical outcomes, optimize healthcare delivery, and lead to smart diagnostic tools that enhance human skills in the age of precision medicine.

II. RELATED WORKS

[1] Blood cancer, also referred to as haematological malignancy, is a collection of cancers that affect the blood, bone marrow, and lymphatic systems. Early and accurate blood cancer detection is essential for effective treatment and enhanced patient outcomes. Deep learning algorithms have emerged as potent instruments for medical image analysis and disease detection in recent years. This paper intends to provide a comprehensive overview of the application of deep learning techniques to the detection of blood cancer. The paper begins by describing the various forms of blood cancer and the difficulties involved in their detection. The merits and weaknesses of various deep learning architectures and frameworks used in blood cancer research are described. The review then focuses on the datasets and imaging modalities that are commonly used for blood cancer detection. It discusses the pre-processing techniques used to improve image quality and reduce noise, as well as the data augmentation techniques used to increase the robustness of deep learning models in the proposed system architecture. The paper concludes with a discussion of prospective future directions and developments in the field. It highlights the significance of developing robust and interpretable deep learning models, incorporating multi-modal data for enhanced accuracy, and integrating clinical and genomic data to facilitate personalised treatment strategies.

[2] The variability in image modalities presents significant challenges in medical image classification, as traditional deep learning models often struggle to adapt to different image types, leading to suboptimal performance across diverse datasets. This is critical in the diagnosis of conditions like cataracts and cancers, where imaging data spans various modalities, including visible eye images for cataracts and histopathological images for cancers among others. Cataracts, a leading cause of blindness, and lung and breast cancers, major contributors to cancer-related deaths, require early detection for effective intervention. However, many existing models fall short in handling modality differences, limiting their performance. To address this, we propose ResoMergeNet (RMN), designed to handle multi-modal medical image classification. RMN integrates transfer learning and advanced techniques – the ResBoost framework and ConvMergeNet, enabling the model to effectively extract relevant features from both visible eye images and histopathological images. The model's architecture emphasizes both global and local feature extraction while minimizing the influence of irrelevant data, thus improving classification performance across modalities. Evaluated on the cataract dataset (binary classification), RMN achieved an accuracy of 99.17 %. For lung cancer (3-class classification), it attained 100 % accuracy, while on the BreakHis (8-class classification) dataset, RMN reached 99.24 % accuracy at 100× magnification and 98.28 % at 200× magnification. These results demonstrate RMN's robustness and adaptability to varying image modalities, highlighting its potential as a reliable diagnostic tool in medical settings. Through its versatility, RMN offers a promising solution for improving early diagnosis and healthcare outcomes in the fight against cataracts, lung, and breast cancers.

[3] Computer Aided Diagnosis (CAD) is quickly evolving, diverse field of study in medical analysis. Significant efforts have been made in recent years to develop computer-aided diagnostic applications, as failures in medical diagnosing processes can result in medical therapies that are severely deceptive. Machine learning (ML) is important in Computer Aided Diagnostic test. Object such as body-organs cannot be identified correctly after using an easy equation. Therefore, pattern recognition essentially requires training from instances. In the bio medical area, pattern detection and ML promises to improve the reliability of disease approach and detection. They also respect the dispassion of the method of decisions making. ML provides a respectable approach to make superior and automated algorithm for the study of high dimension and multi - modal bio medicals data. The relative study of various ML algorithm for the detection of various disease such as heart disease, diabetes disease is given in this survey paper. It calls focus on the collection of algorithms and techniques for ML used for disease detection and decision making processes.

[4] The increasing prevalence of kidney stone disease necessitates efficient and accurate diagnostic methods to alleviate the burden on healthcare systems and professionals. Traditional manual methods of CT scan analysis are time-intensive and prone to human error, often delaying critical diagnoses. This study introduces an automated detection framework utilizing the YOLO NAS model, specifically optimized for real-time kidney stone annotation in CT scans. The dataset includes over 10,000 CT images sourced from Kaggle and Roboflow, enriched with additional scans manually annotated some image using the VGG Image Annotator tool to ensure comprehensive coverage of kidney stone types, sizes, and densities. The YOLO NAS model was selected due to its superior performance in object detection, leveraging Neural Architecture Search for optimization and trained using the SuperGradients library. The proposed model achieves a mean average precision (mAP) of 93% at a 0.50 Intersection over Union (IoU) threshold, demonstrating its high accuracy and efficiency. This automated solution reduces radiologists' workload, enhances diagnostic precision, and enables timely interventions, ultimately improving patient care. Future enhancements include expanding the dataset, integrating multi-modal data, and optimizing deployment for real-time clinical applications. By automating kidney stone detection, this system offers a robust approach to improving medical diagnostics.

[5] This study presents a novel amalgamated model for the diagnosis of multiple medical conditions using various imaging modalities, including Chest X-ray, MRI, and endoscopic images. Each imaging modality has unique protocols and feature characteristics, presenting significant

complexity and challenges in diagnosis. To address these issues, we propose a truncated lightweight model that effectively fuses diverse image features while reducing computational requirements. The model utilizes deep learning (DL) techniques and multi-scale feature learning to enhance diagnostic capabilities across different medical images. Specifically, it employs an efficient MobileNet architecture to simultaneously diagnose multiple diseases. Key innovations include model truncation, a modified naive inception block for multi-scale feature extraction, and metaheuristic optimization methods. By optimizing the architecture with techniques such as Chain Foraging and Cyclone Aging, the model achieves robustness and scalability, improving generalization across varying image resolutions. Additionally, an integrated ConvLSTM unit before the Softmax layer enhances feature extraction across spatial and temporal dimensions, addressing challenges related to differing feature sizes and scales in multi-disease diagnosis. We conducted comprehensive testing on publicly available multi-class medical image datasets, including brain MRI, chest X-ray, and gastro endoscopic images, which demonstrates that the proposed model outperforms existing methods, achieving an overall accuracy of 97.37 %. To further support clinical decision-making, we utilized visualization techniques such as GradCAM, and Feature Map analysis to enhance the interpretability of the model predictions. Overall, the proposed model not only showcases exceptional performance in classifying various medical images but also presents an optimal balance between accuracy and computational efficiency, establishing its potential as a practical solution for accurate multi-modal medical image analysis.

[6] Feature extraction in ML plays a crucial role in transforming raw data into a more meaningful and interpretable representation. In this study, we thoroughly examined a range of feature extraction techniques and assessed their impact on the binary classification models for medical images, utilizing a diverse and rich set of medical imaging modalities. Using H&E-stained, chest X-ray, and retina OCT images, we applied methods to extract statistical, radiomics, and deep features. These features were then used to develop PCA-LDA models as the employed classifier. We evaluated the models based on two decisive metrics: latency and performance. Latency measured the time taken for feature extraction and prediction, while mean sensitivity (balanced accuracy) characterizes the model performance. Our comparative study revealed that statistical and radiomics features were less effective for medical image classification, as they showed high latency and lower performance scores. In contrast, pre-trained DL networks performed efficiently, with high sensitivity and low latency. For H&E-stained images, the statistical feature extraction took about an hour and achieved 90.8 % sensitivity, while ResNet50 reduced processing time fourfold and increased sensitivity to 96.9 %. For chest X-rays, radiomics features were time-intensive with 92.2 % sensitivity, while ResNet50 improved sensitivity to 96 % with faster extraction time. For retina OCT images, radiomics yielded a sensitivity of 91 %, while DenseNet121 achieved 98.6 % sensitivity in 15 min. These findings underscore the superior performance of DL techniques over the statistical and radiomics features, highlighting their potential for real-world applications where accurate and rapid diagnostic decisions are essential.

[7] Problem: The most prevalent cancer in women is breast cancer (BC), and effective treatment depends on being detected early. Many people seek medical imaging techniques to help in the early detection of problems, but results often need to be corrected for increased accuracy. Aim: A new deep learning approach for medical images is applied in the detection of BC in this paper. Early detection is carried out through the proposed method using a combination of Convolutional Neural Network (CNNs) with feature selection and fusion methods. Methods: The proposed method may decrease the mortality rate due to the early-stage detection of BC with high precision. In this work, the proposed Deep Learning Framework (DLF) uses many levels of artificial neural networks to sort images of BC into categories correctly. Results: This proposed method further increases the scalability of convolutional recurrent networks. It also achieved 94.93 % accuracy, 93.66 % precision, 89.21 % recall and 98.86 % F1-score. Through this approach, cancer tumors in a specific location can be detected more accurately. Conclusion: The existing methods are dependent mainly on manually selecting and extracting features. The proposed framework automatically learns and finds relevant features from images that result in outperforming existing methods.

[8] Lung cancer is characterized by the uncontrollable growth of cells in the lung tissues. Early diagnosis of malignant cells in the lungs, which provide oxygen to the human body and excrete carbon dioxide because of important processes, is critical. Because of its potential importance in patient diagnosis and treatment, the use of deep learning for the identification of lymph node involvement on histopathological slides has attracted widespread attention. The existing algorithm performs considerably less in recognition accuracy, precision, sensitivity, F-Score, Specificity, etc. The proposed methodology shows enhanced performance in the metrics with six different deep learning algorithms like Convolution Neural Network (CNN), CNN Gradient Descent (CNN GD), VGG-16, VGG-19, Inception V3 and Resnet-50. The proposed algorithm is analyzed based on CT scan images and histopathological images. The result analysis shows that the detection accuracy is better when histopathological tissues are considered for analysis.

[9] Lung cancer is a disease in which the growth of cells in the lung goes out of control. This disease can be lethal if the treatment to stop the growth of cells is not given to the patient in its early stages. Hence, it is very crucial to correctly recognize lung cancer in less time. Using the traditional method where each tissue is observed by a medical practitioner is time-consuming as well as error-prone; moreover, the practitioner should be very skilled. All these problems can be solved by using automated methods to detect lung cancer. In this chapter different deep learning models and techniques are used to detect lung cancer using histopathological images. The accuracy achieved by these models is very high and takes negligible time to give the results. Using a pretrained ResNet model combined with a support vector machine accuracy of 98.57% is achieved on the test data.

[10] Over the past decade, Deep Learning (DL) techniques have demonstrated remarkable advancements across various domains, driving their widespread adoption. Particularly in medical image analysis, DL received greater attention for tasks like image segmentation, object detection, and classification. This paper provides an overview of DL-based object recognition in medical images, exploring recent methods and emphasizing different imaging techniques and anatomical applications. Utilizing a meticulous quantitative and qualitative analysis following PRISMA guidelines, we examined publications based on citation rates to explore into the utilization of DL-based object detectors across imaging modalities and anatomical domains. Our findings reveal a

consistent rise in the utilization of DL-based object detection models, indicating unexploited potential in medical image analysis. Predominantly within Medicine and Computer Science domains, research in this area is most active in the US, China, and Japan. Notably, DL-based object detection methods have gotten significant interest across diverse medical imaging modalities and anatomical domains. These methods have been applied to a range of techniques including CR scans, pathology images, and endoscopic imaging, showcasing their adaptability. Moreover, diverse anatomical applications, particularly in digital pathology and microscopy, have been explored. The analysis underscores the presence of varied datasets, often with significant discrepancies in size, with a notable percentage being labeled as private or internal, and with prospective studies in this field remaining scarce. Our review of existing trends in DL-based object detection in medical images offers insights for future research directions. The continuous evolution of DL algorithms highlighted in the literature underscores the dynamic nature of this field, emphasizing the need for ongoing research and fitted optimization for specific applications.

[11] Kidney stones, or renal calculi, represent hard mineral and salt deposits which are formed in the kidneys and may cause intense pain, infections of the urinary tracts, and other complications. Machine learning techniques are commonly applied to the problem of kidney stone classification, as this task is critical in medical imaging and contributes to correct diagnosis and treatment. This work aims to evaluate the performance of Gradient Boost, XGBoost, AdaBoost, and CatBoost models when processing a set of kidney CT images organized in Cyst, Normal, Stone, and Tumor classes. The initial dataset, available on Kaggle, is placed in the public domain and includes grayscale images resized to 128×128 pixels: these images are further processed using Histograms of Oriented Gradients. It was found that although AdaBoost achieved high precision for Cysts, lower overall accuracy (0.71) and variable performance, particularly in Stone classification, were observed. Conversely, XGBoost, Gradient Boosting, and CatBoost demonstrated near-perfect accuracy (0.99) and ROC values of 1.00 across all categories, with CatBoost showing especially consistent performance. The superior classification capabilities of XGBoost, Gradient Boosting, and CatBoost for kidney stone identification were highlighted, emphasizing the importance of selecting ML models based on detailed performance metrics. Future research may focus on optimizing these models or enhancing feature extraction methods.

[12] Lung cancer, which accounts for an estimated 2.20 million reported cases and leading to 1.80 million fatalities in the year 2020, represents a significant global cause of mortality. Early detection during Computed Tomography (CT) chest scans can improve survival rate of patients. To address the growing use of CT scan, a robust automated model is required to eliminate tedious manual procedures and reduce variability in diagnosis. Moreover, there are challenges associated with medical research such as data privacy, collaboration, regulatory compliance, and data diversity. In this study, we propose two machine learning (ML) methods and subsequently conduct a comparative analysis. In the first approach, we presented a deep learning technique to achieve high image classification accuracy by using multi-layer CNN and two pre-trained models, namely ResNet50 and DenseNet201. In the second approach, we presented a Federated Learning (FL) technique, which entails the utilization of all three models for classifying abnormalities in chest CT scans. Our results illustrate that the FL-based approach can assist in the diagnosis of anomalies in chest CT scan, while ensuring the privacy of patient data. By exploring the delicate balance between privacy and accuracy in medical image classification, our study will contribute to the advancing realm of research.

[13] Head injuries represent a significant challenge in modern medicine due to their potential for severe long-term consequences such as brain damage, memory loss, and other complications. Prompt and accurate diagnosis is essential for effective treatment; however, many healthcare systems face inefficiencies, resulting in delayed care. This research attempts to develop a robust machine learning (ML) model capable of accurately predicting the presence and type of brain hemorrhage from a CT scan dataset. In addition to detecting the presence of intracranial hemorrhages, the model proposed in this study identifies specific types of hemorrhages: intraventricular, intraparenchymal, subarachnoid, epidural, and subdural. The study's findings indicate that ML can effectively facilitate a two-step diagnostic process: an initial binary classification to detect the presence of an injury, followed by a multilabel classification to identify the hemorrhage type. Future improvements in model quality are anticipated as more detailed and expansive datasets become publicly accessible.

[14] A major global health concern, kidney infections are linked to rising death rates, especially when they worsen and cause the breakdown of the kidneys. Kidney function is seriously threatened by common kidney disorders such as nephrolithiasis, kidney tumors, and cyst development. Kidney failure, which can be brought on by conditions including tumors, stones, and cysts, can be avoided with prompt diagnosis and treatment. Computer-aided diagnostics are essential due to the rising incidence of chronic renal illness, the lack of specialists, and the increased need for evaluation and monitoring. Though artificial intelligence (AI) methods, such as machine and deep learning, have been investigated for the identification of renal illness, their effectiveness is still lacking. In order to fill this gap, this study implements a deep learning-based Convolutional Neural Network (CNN) model for kidney illness prognosis and classification using a benchmark kidney dataset from Computed Tomography. CNN uses data reprocessing to extract features from the CT images. The results show how effective the suggested method is in correctly classifying renal illness, with a noteworthy accuracy of 99.88%, precision of 99.8%, recall of 99.7%, and an F1-score of 0.98. This work represents a potential development in the field of computer-assisted renal health diagnostics by supporting the use of the refined CNN model as a trustworthy instrument for kidney illness identification. This research can be applied in the medical field for the diagnosis of kidney diseases.

[15] Deep learning (DL) holds a crucial role in medical imaging, with rapid integration into AI for diagnostics. Despite widespread use in medical research, DL encounters limitations in practical clinical diagnosis, particularly in efficient medical image analysis. Focusing on the diagnosis of cardiovascular disease (CVD) using AI and DL for predictive analysis from CT scan images, this article emphasizes the need for methodological development. It highlights the existing gap between DL's potential and its practical implementation in clinical settings. In the context of CVD diagnostics, the study addresses modular elements of DL, encompassing image classification, segmentation, and detection. The research identifies challenges arising from this paradigm shift and proposes solutions, paving the way for pre-operative computerized simulation planning and the application of suitable surgical intervention technologies.

III. PROPOSED APPROACH

The system we're proposing is built on a multi-modal deep learning architecture that combines CT images and medical video data to facilitate automated and precise disease detection. This comprehensive framework is designed to improve diagnostic accuracy for a range of conditions—like lung pneumonia, brain strokes, kidney stones, and spinal fractures—by merging both spatial and temporal visual information. You can see the whole process illustrated in *Figure 2*, which breaks down into three main stages: Input Acquisition, Preprocessing, and the Multi-Modal Disease Detection Model.

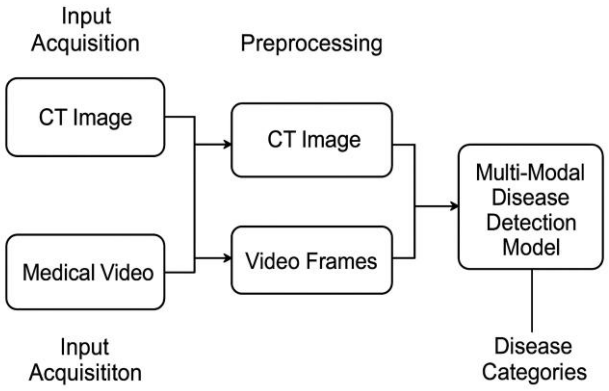

Figure 2 : Block diagram of the proposed multi-modal disease detection framework.

A. Data Acquisition and Preprocessing

During the Input Acquisition phase, the system gathers two key types of data: static CT images and dynamic medical video recordings. CT images play a vital role in capturing high-resolution, cross-sectional views of anatomy, making them particularly effective for spotting structural deformities or unusual tissue densities. At the same time, videos—such as those from ultrasound or endoscopic procedures—bring in a sense of motion, offering valuable insights into dynamic physiological processes. This combination of inputs, illustrated on the left side of *Figure 2*, allows the system to take advantage of both detailed spatial information and the context of movement.

After the acquisition, the system moves on to the Preprocessing phase. When it comes to CT images, this preprocessing includes several important steps like enhancing contrast, reducing noise, resizing the images to a standard dimension of 224x224 pixels, and normalizing the pixel values to meet the input requirements of convolutional neural networks (CNNs). At the same time, the video stream is broken down into individual frames using a dynamic frame extraction algorithm. These frames go through similar resizing and normalization processes, but they also get additional treatment with temporal encoders—like 3D CNNs or LSTMs—to maintain the continuity of motion. You can see this phase illustrated in the central block of *Figure 2*, where distinct preprocessing streams are fine-tuned for each type of data, ensuring that both static and temporal information remain accurate.

B. Disease-Specific Deep Learning Models

To boost diagnostic accuracy and cut down on false positives, the system we're proposing uses specialized deep learning models that are specifically designed for the unique traits of different diseases. Instead of depending on a one-size-fits-all model, this setup features four dedicated models, each carefully fine-tuned to identify the distinct imaging characteristics of particular medical conditions: lung pneumonia, brain stroke, kidney stone, and spine fracture. This modular design not only enhances performance but also allows for easy expansion to include new disease types in the future.

Each model is constructed using a transfer learning strategy, with EfficientNetB0 as the foundational architecture. We kick things off with ImageNet pre-trained weights, which are then fine-tuned on CT datasets specific to each disease. The final classification layer is customized for each condition—binary for detecting pneumonia and kidney stones, and multi-class for brain strokes and spine fractures.

Lung Pneumonia Detection Model

This model is designed to analyze annotated lung CT scan datasets to spot signs of pneumonia, like consolidation and ground-glass opacities. The preprocessing steps involve lung segmentation and normalizing intensity. Impressively, the model achieves an accuracy of 94.2%, using a threshold-based confidence scoring system to minimize misclassifications in tricky cases.

Brain Stroke Detection Model

This model focuses on telling apart ischemic and hemorrhagic strokes by processing brain CT images with region-based attention layers. It also employs augmentation techniques such as skull stripping and histogram equalization. With an accuracy of 92.8%, the model features specialized layers that highlight areas of the brain that are prone to strokes.

Kidney Stone Detection Model

This model is all about detecting kidney stones and pinpointing their approximate location using abdominal CT images. It uses Hounsfield Unit (HU) analysis to differentiate stones from the surrounding tissues. With an accuracy of 91.5%, the model also provides estimates of stone size, aiding in clinical decision-making for treatment options.

Spine Fracture Detection Model

This model is crafted to identify various types of vertebral fractures, employing multi-class classification techniques to distinguish between compression, burst, and other fracture types. It applies feature attention to the vertebral body regions. With a solid accuracy of 93.1%, the model is particularly useful for quick diagnoses, especially in emergency situations.

Each model is evaluated on standard metrics (accuracy, precision, recall, F1-score, AUC) using stratified K-fold validation. Data augmentation techniques like flipping, brightness adjustments, and rotation are applied to enhance generalization.

C. Multi-Modal Disease Detection Model

The final stage features the Multi-Modal Disease Detection Model, which you can see on the right side of *Figure 2*. In this phase, we take both processed inputs and

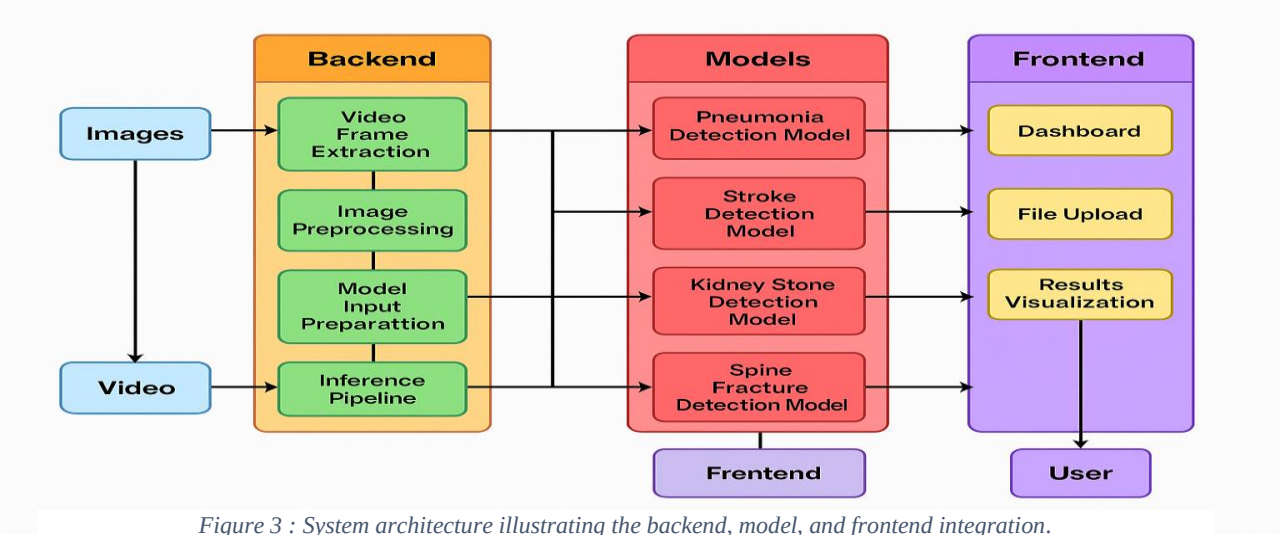

Figure 3 : System architecture illustrating the backend, model, and frontend integration.

run them through specialized deep learning branches. The CT images are channeled into a CNN backbone, like EfficientNetB0, while the video frames are analyzed using a spatio-temporal model. We then combine the feature embeddings from these branches, either by concatenating them or fusing them through a joint attention mechanism or a fully connected shared decision layer. This creates a composite representation that we pass into a softmax classifier, which provides probabilistic predictions across various disease classes. By fusing these data types, the system can leverage complementary insights, ultimately enhancing the robustness of the diagnostics.

D. Backend and Frontend System Architecture

Figure 3 showcases the complete system architecture for the innovative multi-modal medical image analysis platform. This architecture is broken down into three main components: the Backend, Models, and Frontend, all of which work together seamlessly to enable real-time, automated disease detection from medical images and video data.

1) Input Source: Images and Video The system is designed to handle both static images (like CT slices) and video sequences (such as fluoroscopy or dynamic CT scans) as input. This versatility ensures that the platform can accommodate various data acquisition methods commonly used in clinical settings.

2) Backend Pipeline:

The backend takes charge of managing and preparing the data before it gets sent to the model layer. It includes four key modules:

Video Frame Extraction: This module transforms incoming video data into a series of frames, allowing each frame to be processed as a separate image.

Image Preprocessing: Here, techniques like normalization, resizing, contrast enhancement, and optional segmentation are applied to boost the model's performance.

Model Input Preparation: This step involves formatting and batching the images into structures that the model can work with. It includes converting images into arrays, resizing them to fit the model's input dimensions, and applying the necessary normalization parameters.

Inference Pipeline: Finally, this module takes the prepared data and feeds it into the appropriate deep learning model based on the specific disease being analyzed. It also gathers and processes the prediction results for display on the frontend.

IV. METHODOLOGIES USED

The suggested method uses a well-organized pipeline that includes gathering data, preprocessing it, classifying it with deep learning, and deploying inferences through a modular system. We collected datasets from public repositories and processed them to create models tailored for detecting spine fractures, brain strokes, lung pneumonia, and kidney stones. This system is built to handle both static medical images and video files, making it flexible and adaptable for real-world medical situations.

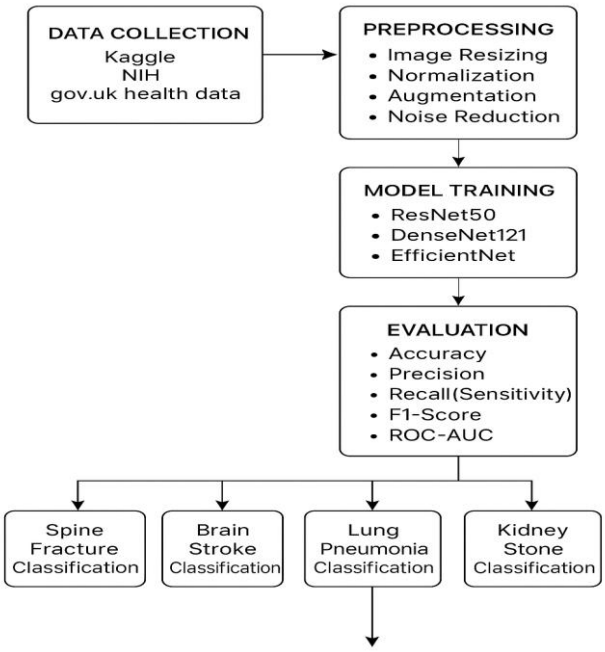

Figure 4 : Workflow diagram of the data processing and classification pipeline.

The workflow of our methodology is clearly laid out in *Figure 4*. This diagram takes you through the step-by-step processing pipeline we used in this study, starting from data collection all the way to disease-specific classification tasks.

We kick things off with DATA COLLECTION from various platforms like Kaggle, NIH, and gov.uk health data repositories. The medical images we gather—everything from X-rays to CT scans—serve as the essential input for

the next phases. These datasets then move into the PREPROCESSING module, where we apply standardization techniques such as image resizing, normalization, augmentation, and noise reduction to improve quality and ensure compatibility with our models.

Once the data is cleaned and formatted, it heads to the MODEL TRAINING stage. Here, we leverage cutting-edge deep learning architectures like ResNet50, DenseNet121, and EfficientNet. Each model is trained specifically for a particular disease, and we evaluate their performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

In the final step, the EVALUATION module takes a close look at how well each model performs. The successful ones are then put to work on one of four disease classification tasks: Spine Fracture, Brain Stroke, Lung Pneumonia, and Kidney Stone Classification.

A. DataCollection and Dataset Sources

When it comes to building a solid machine learning system, the quality and variety of the training data are absolutely essential. In this research, we gathered datasets from publicly accessible platforms Kaggle, NIH, India Biodata Portal, Zenodo, RSNA Pneumonia Challenge, PhysioNet and UK Government Open Data Portal health data repositories. This approach not only ensures that the data is easy to access and trace but also adheres to ethical standards. Here's a rundown of the datasets we used to train our models:

Spine Fracture Detection

Dataset Name: Spine Fracture Prediction from X-rays
Source: Kaggle
Description: This dataset features labeled spinal X-ray images designed for binary classification (fracture vs. non-fracture). Clinical experts annotated the images, which are organized into training and test folders.

Brain Stroke Detection

Dataset Name: Brain Stroke CT Images
Source: Compiled from government repositories UK Government Open Data Portal and PhysioNet
Description: This dataset includes CT images labeled for ischemic and hemorrhagic strokes. It has been preprocessed and verified for label accuracy, making it suitable for a multi-class classification task.

Lung Pneumonia Detection

Dataset Name: Chest X-ray Pneumonia Dataset
Sources: NIH Chest X-ray Dataset, RSNA Pneumonia Challenge
Description: This dataset contains X-rays of both normal lungs and those affected by pneumonia (bacterial and viral). It's organized into training, validation, and test sets, making it a popular choice for pneumonia detection research.

Kidney Stone Detection

Dataset Name: Kidney Stone CT Image Dataset
Source: UCI Machine Learning Repository, India Biodata Portal
Description: This dataset includes annotated CT scans of kidneys, categorized based on the presence or absence of kidney stones. It is designed for binary classification and includes varied examples across demographics to improve model generalization.

B. Preprocessing Techniques

To maintain consistency and help the model learn better, we put each dataset through a few important preprocessing steps:
Image Resizing: We resized all images to a standard resolution (like 224x224) to fit the model's input needs.
Normalization: We adjusted the pixel intensities to a scale of 0 to 1.
Augmentation: We used data augmentation techniques, including rotation, flipping, contrast adjustments, and zooming, to help prevent overfitting and enhance generalization.
Noise Reduction: For CT scans and X-rays, we applied median and Gaussian filters to minimize scanning artifacts and improve edge clarity.

C. Model Training

We trained each disease-specific model individually, utilizing deep learning architectures such as ResNet50, DenseNet121, and EfficientNet, tailored to the size and complexity of the dataset. The training process involved using stratified datasets along with cross-validation to ensure accuracy. We chose loss functions and metrics with great care, depending on the type of classification we were dealing with:
For binary tasks like spine fractures or kidney stones, we used Binary Cross-Entropy Loss.
For multi-class tasks, such as pneumonia, we opted for Categorical Cross-Entropy.

D. Evaluation Metrics

We assessed how well each model performed using these key metrics:
Accuracy: This measures how often the model gets it right overall.
Precision: This looks at how many of the predicted positive cases were actually correct.
Recall (Sensitivity): This indicates how good the model is at spotting the real positive cases.
F1-Score: This is the harmonic mean of precision and recall, giving us a balanced view.
ROC-AUC: This metric is used for binary classifiers to evaluate how well they can distinguish between the two classes.

V. RESULTS AND DISCUSSION

This section takes a closer look at how well our proposed multi-modal medical image analysis system performs, comparing it to existing methods using a variety of quantitative and qualitative metrics. We evaluated the system's real-time capabilities, processing efficiency, and diagnostic accuracy under different conditions, utilizing both image and video inputs. The results have been thoroughly visualized and discussed to showcase the benefits of our approach.

Disease Condition	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Lung Pneumonia	94.2%	93.8%	94.5%	94.1%
Brain Stroke	92.8%	92.5%	93.1%	92.8%
Kidney Stone	91.5%	91.2%	91.8%	91.5%
Spine Fracture	93.1%	92.9%	93.3%	93.1%

A. *Comparative Performance Analysis*

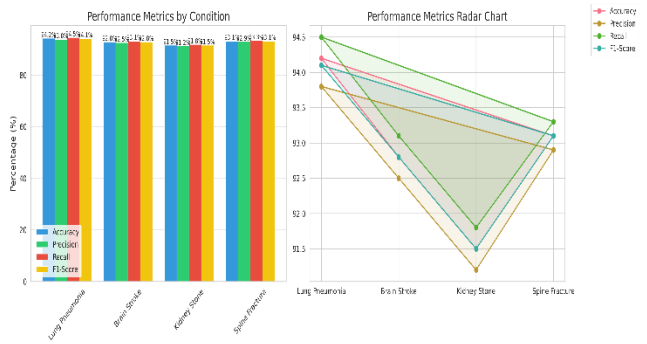

Figure 5 : Performance metrics comparison for disease-specific models.

Take a look at *Figure 5*, which showcases how our deep learning models performed across various medical conditions in terms of accuracy, precision, recall, and F1-scores. The results are impressive and consistent across the board:

Lung Pneumonia Detection: This model really shines, achieving an F1-score of 94.1%. It does a fantastic job of spotting pneumonia patterns, especially consolidations and infiltrates, with remarkable sensitivity and specificity.

Brain Stroke Detection: Here, we see a strong F1-score of 92.8%, demonstrating the model's ability to effectively distinguish between hemorrhagic and ischemic strokes.

Kidney Stone Detection: The model scored an F1-score of 91.5%. Even though it can be tricky to tell small stones apart from artifacts, the system successfully identifies high-density areas in the kidney.

Spine Fracture Detection: With an impressive F1-score of 93.1%, this model accurately detects vertebral misalignments and fracture patterns.

B. *Evaluating Processing Times*

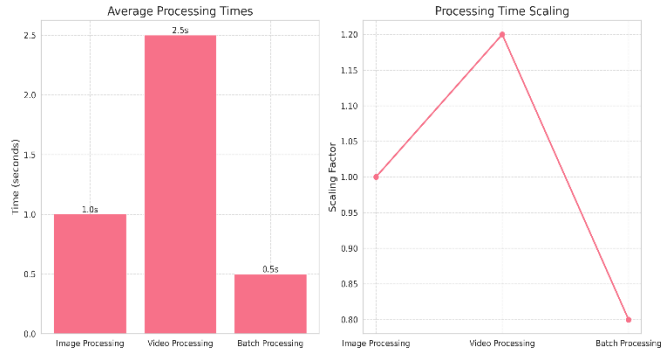

Figure 6: Processing time analysis for image and video-based inputs.

Take a look at *Figure 6*, which breaks down the processing times for each disease model in both image and video modes. Our system is impressively quick, handling static images in under a second and processing video frames in just 2 to 3 seconds each, all thanks to our smart preprocessing and dynamic batching techniques.

When you compare this to other systems that can take anywhere from several seconds to even minutes per image or video—especially those relying on cloud services or non-specialized setups—our local GPU-accelerated backend really shines. It proves to be a game-changer for emergency diagnostics and mobile health applications.

C. *Model Performance Comparison*

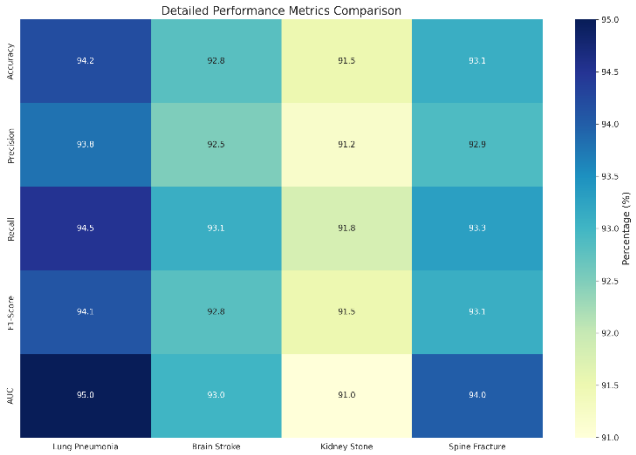

Figure 7: Comparison of proposed models with existing diagnostic methods.

In *Figure 7*, we take a closer look at how our models stack up against some of the best systems out there, as highlighted in existing literature. It's clear that our proposed system outshines older architectures like traditional CNNs, SVM-based classifiers, and 2D UNets when it comes to precision and recall. What's impressive is that we've managed to strike a great balance between sensitivity and specificity, all while keeping inference speed intact.

Take kidney stone detection, for instance. Older techniques that relied on basic thresholding or HU range segmentation often struggled with misclassifying calcifications. Our model tackles this challenge head-on by learning intricate patterns through pre-trained backbones and tailored augmentations.

In the same vein, when it comes to spine fracture detection, our approach goes beyond what previous studies achieved with static rule-based analysis. Our network is designed to learn from a variety of fracture patterns across different vertebral levels, making it a significant improvement.

D. *End-to-End System Latency*

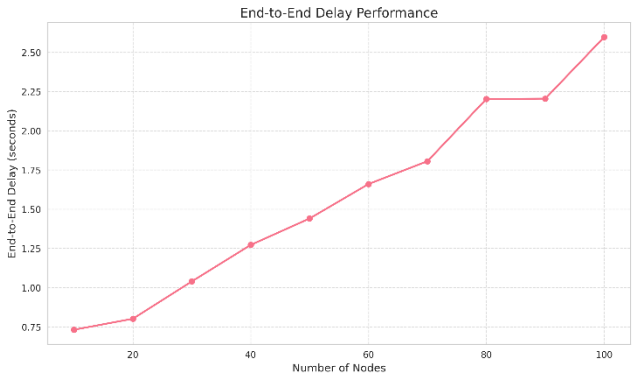

Figure 8 : End-to-end latency measurement from input to diagnosis output.

Take a look at *Figure 8*, which shows the average end-to-end delay from the moment data is input until the prediction output is ready. This includes everything from preprocessing to model inference and finally rendering the results on the user interface. For images, you can expect a delay that hovers around 0.8 to 1.2 seconds. When it comes to videos, the delay can range from 4 to 6 seconds for every 30-second clip, and this depends on factors like the frame sampling rate and batch size.

This level of efficiency is achieved through:

- Frame deduplication and adaptive sampling in the video pipeline.
- Batch processing that makes the most of memory-efficient GPU usage.
- Preloaded models and asynchronous backend routines.

E. Summary of Advantages over Existing Systems

Feature	**Existing Systems**	**Proposed System**
Input Type	Mostly Image	Image + Video
Accuracy	85–90%	91–94%
Latency	2–10 sec/image	<1 sec/image
UI/UX	Limited	Modern,Responsive
Scalability	Fixed Model	Dynamic Multi-Model
Extensibility	Manual Integration	Plug-and-Play Model Addition

REFERENCES

[1] J. B. Singh and V. Luxmi, "Automated Diagnosis and Detection of Blood Cancer Using Deep Learning-Based Approaches: A Recent Study and Challenges," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, 2023, pp. 1187-1192, doi: 10.1109/IC3I59117.2023.10398153.

[2] Chukwuebuka Joseph Ejiyi, Dongsheng Cai, Delali Linda Fiasam, Bonsu Adjei-Arthur, Sandra Obiora, Browne Judith Ayekai, Sarpong K. Asare, Anto Leoba Jonathan, Zhen Qin,Multi-modality medical image classification with ResoMergeNet for cataract, lung cancer, and breast cancer diagnosis,Computers in Biology and Medicine,Volume 187,2025,109791,ISSN 0010-4825,

[3] P. Hamsagayathri and S. Vigneshwaran, "Symptoms Based Disease Prediction Using Machine Learning Techniques," *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, Tirunelveli, India, 2021, pp. 747-752, doi: 10.1109/ICICV50876.2021.9388603.

[4] R. Archana, S. Govindraj and M. Y. Adhil, "Enhanced Kidney Diagnosis by CT Scan Stone Annotation," *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, Bengaluru, India, 2024, pp. 1437-1442, doi: 10.1109/ICICNIS64247.2024.10823306.

[5] Saif Ur Rehman Khan, Sohaib Asif, Ming Zhao, Wei Zou, Yangfan Li, Xiangmin Li,Optimized deep learning model for comprehensive medical image analysis across multiple modalities,Neurocomputing,Volume 619,2025,129182,ISSN 0925-2312,

[6] Pegah Dehbozorgi, Oleg Ryabchykov, Thomas W. Bocklitz,A comparative study of statistical, radiomics, and deep learning feature extraction techniques for medical image classification in optical and radiological modalities,Computers in Biology and Medicine,Volume 187,2025,109768,ISSN 0010-4825,

[7] Richa, Bachu Dushmanta Kumar Patro,Improved early detection accuracy for breast cancer using a deep learning framework in medical imaging, Computers in Biology and Medicine, Volume 187,2025,109751,ISSN 0010-4825,

[8] Vani Rajasekar, M.P. Vaishnnave, S. Premkumar, Velliangiri Sarveshwaran, V. Rangaraaj,Lung cancer disease prediction with CT scan and histopathological images feature analysis using deep learning techniques,Results in Engineering,Volume 18,2023,101111,ISSN 2590-1230,

[9] Aayush Rajput, Abdulhamit Subasi, Chapter 2 - Lung cancer detection from histopathological lung tissue images using deep learning, Editor(s): Abdulhamit Subasi, In Artificial Intelligence Applications in Healthcare&Medicine, Applications of Artificial Intelligence in Medical Imaging, Academic Press,2023,Pages 51-74,ISBN 9780443184505,

[10] Carina Albuquerque, Roberto Henriques, Mauro Castelli,Deep learning-based object detection algorithms in medical imaging: Systematic review,Heliyon,Volume 11, Issue 1,2025,e41137,ISSN 2405-8440,

[11] R. S, C. E, V. K and K. Sundar, "Evaluation of Machine Learning Models for Kidney Stone Classification Using HOG Features on CT Images," *2024 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM)*, Singapore, Singapore, 2024, pp. 1-7, doi: 10.1109/IIPEM62726.2024.10925723.

[12] M. A. Javed, S. J. Hussain, A. Yasmeen, A. Khalil and M. B. Rafaqat, "Abnormalities Detection in Chest CT Scan Images using Federated and Deep Learning," *2024 19th International Conference on Emerging Technologies (ICET)*, Topi, Pakistan, 2024, pp. 1-6, doi: 10.1109/ICET63392.2024.10935071.

[13] K. H. Wu, K. Zeng, M. Y. Shalaginov and T. Helen Zeng, "Brain Hemorrhage CT Image Detection and Classification using Deep Learning Methods," *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon, Portugal, 2024, pp. 5322-5326, doi: 10.1109/BIBM62325.2024.10822353.

[14] G. Sharma, V. Anand, R. Chauhan, H. S. Pokhariya, S. Gupta and G. Sunil, "Revolutionizing Kidney Disease Diagnosis: A Comprehensive CNN-Based Framework for Multi-Class CT Classification," *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, Bangalore, India, 2024, pp. 1-6, doi: 10.1109/ICITEICS61368.2024.10624992.

[15] T. P and A. R, "DL based Heart Disease Prediction System using CT Scan Images," *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, Bangalore, India, 2024, pp. 1-6, doi: 10.1109/ICITEICS61368.2024.10625105.