IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 10, Issue 6 - V10I6-1450)
Available online at: https://www.ijariit.com

AI-Based Cancer Detection using FRCNN, Random Forest, SVM, and Regression Models

Shivam Chattar
shivchattar5@gmail.com
Ajeenkya D Y Patil University, Pune, Maharashtra

Anshul Gaikwad
anshul.gaikwad@adypu.edu.in
Ajeenkya D Y Patil University, Pune, Maharashtra

Het Savsani
hetsavsani313@gmail.com
Ajeenkya D Y Patil University, Pune, Maharashtra

Kshanay Nikam
kshanaynikam7@gmail.com
Ajeenkya D Y Patil University, Pune, Maharashtra

Avishkar Sarnaik
avishkar.sarnaik@adypu.edu.in
Ajeenkya D Y Patil University, Pune, Maharashtra

Prof. Priyanka Patil
priyanka.patil@adypu.edu.in
Ajeenkya D Y Patil University, Pune, Maharashtra

ABSTRACT

Early detection of cancers, especially in the mouth, throat, and lungs, significantly improves patient survival rates. This paper presents a comprehensive AI-driven approach combining various machine learning (ML) and deep learning (DL) techniques such as Fast Region-based Convolutional Neural Networks (FRCNN), Random Forest (RF), Support Vector Machines (SVM), and Logistic and Linear Regression models to enhance cancer detection capabilities. Each model's strengths are leveraged to create a hybrid system that excels in detecting and classifying cancerous regions in medical images and analyzing patient data. The proposed workflow incorporates automated image analysis, feature selection, classification, and probabilistic risk estimation, enhancing diagnostic accuracy while addressing challenges like data availability, model interpretability, and computational requirements. This integrated AI-based approach demonstrates potential for real-time clinical application and personalized cancer diagnostics.

Keywords: *FRCNN, Random Forest, SVM, and Regression Models, Cancer, Lung Cancer, Throat Cancer, Mouth Cancer*

1. INTRODUCTION

Cancer is among the leading causes of mortality worldwide, with lung, mouth, and throat cancers being particularly fatal [1], [2]. Timely detection of these cancers is crucial, as survival rates drop drastically once the cancer has metastasized [3]. Traditional diagnostic methods, such as imaging (CT, MRI) and biopsies, although effective, are resource-intensive, time-consuming, and subject to human error [4], [5]. Recent advancements in artificial intelligence (AI), particularly in ML and DL, offer a path to more accurate, efficient, and automated cancer detection [6].

AI-based models such as Convolutional Neural Networks (CNNs), Random Forest, and SVMs are gaining traction due to their ability to analyze vast amounts of data and identify patterns that may not be visible to the human eye [7]. This paper aims to integrate a range of AI techniques to detect mouth, throat, and lung cancers, utilizing medical images and clinical data [8], [9].

Objectives:

Cancer Detection and Localization:

Technique: Faster Region-based Convolutional Neural Networks (FRCNN)

Benefit: Accurate localization of cancer in medical images for early detection.

Feature Integration and Optimization:

Technique: Random Forest (RF)

Benefit: Enhances diagnostic accuracy by selecting relevant features from imaging and clinical data.

Risk Classification and Prediction:

Techniques: Support Vector Machines (SVM) and Logistic Regression

Benefit: Classifies cancer risk with high precision and provides interpretable probabilistic outputs.

Severity Estimation:

Technique: Linear Regression

Benefit: Predicts tumor size and growth, aiding treatment decisions.

Multi-Modal Analysis:

Approach: Combine patient history and imaging data

Benefit: Provides a comprehensive and personalized diagnostic profile.

Scalability and Efficiency:

Approach: Adaptable architecture for large datasets and real-time analysis

Benefit: Makes the system applicable in diverse healthcare settings.

Performance Evaluation:

Approach: Compare against individual models using validation datasets

Benefit: Ensures system robustness, accuracy, and reliability.

2. LITERATURE REVIEW

2.1. Deep Learning in Cancer Detection

The use of deep learning models, particularly CNNs, for cancer detection in medical images has revolutionized the field of diagnostics. Studies like Zhang et al. [1] have demonstrated CNN's superior ability to detect lung nodules from CT scans, achieving higher accuracy compared to traditional methods. Other works, such as those by Liu et al. [2], have utilized CNNs for mouth and throat cancer detection, achieving remarkable results in classifying cancerous tissues from MRI and voice data.

Advanced CNN architectures, such as Faster R-CNN (FRCNN), have been widely used for object detection and have proven highly effective in identifying cancerous regions in medical images [3]. Guerreiro et al. [12] highlighted the effectiveness of low-dose computed tomography (LDCT) combined with AI in early lung cancer detection, significantly reducing mortality rates.

2.2 Machine Learning in Feature Selection and Classification

While deep learning models excel at feature extraction from image data, traditional ML models like Random Forest and SVM continue to play a critical role in feature selection and classification. Studies by Lin et al. [1] have shown that Random Forest and SVM perform well in classifying lung cancer when features are extracted from medical images or clinical datasets.

SVM, in particular, is effective in high-dimensional spaces and can handle non-linear decision boundaries, making it a robust choice for cancer classification. Logistic and Linear Regression models, on the other hand, offer simplicity and interpretability, providing continuous predictions or risk probabilities that aid clinicians in making informed decisions [4].

3. METHODOLOGY

3.1 Current System

Current cancer diagnostic systems are predominantly reliant on human interpretation of medical images like CT scans, MRIs, and X-rays. While accurate, this method is prone to diagnostic errors, delays, and inconsistencies due to manual processes. Traditional ML models like Random Forest and SVM have been applied to structured datasets (e.g., patient records, demographics), but they often require extensive preprocessing and feature selection.

3.2 Proposed System

This study proposes an AI-based system that integrates deep learning and traditional machine learning models to improve diagnostic accuracy for mouth, throat, and lung cancers. The hybrid system includes:

(i) Data Collection
Imaging data (CT, MRI, X-rays) and structured clinical data (patient history, demographic information) are collected.

(ii) Data Cleaning
Preprocessing steps reduce noise in images and handle missing data in clinical records to ensure high-quality inputs for analysis.

(iii) FRCNN for Image Analysis
FRCNN detects and localizes abnormal tissues in the collected medical images, providing bounding boxes and extracted image features (regions of interest).

(iv) RF for Feature Selection
RF refines the feature set by selecting the most relevant features from both imaging data and clinical records, ensuring a focused and accurate analysis.

(v) SVM and Logistic/Linear Regression for Classification and Prediction

a) SVM classifies the refined features into cancerous or non-cancerous categories.
b) Logistic Regression provides probabilistic risk scores for cancer presence.
c) Linear Regression offers continuous predictions, such as tumor size and growth rate.

3.3 How It Can Work Together

(i) Faster R-CNN (FRCNN) for Image Analysis:
a) Role: Detect and localize cancerous regions in CT or X-ray images.
b) Output: Provides bounding boxes and extracted image features (regions of interest).

(ii) Random Forest (RF) for Feature Selection:
a) Role: Select the most relevant features from the Faster R-CNN outputs or tabular data such as patient demographics or medical history.
b) Output: A refined feature set for further analysis.

(iii) SVM for Classification:
a) Role: Classify the refined features into cancerous or non-cancerous categories.
b) Strength: Handles non-linear decision boundaries effectively.

(iv) Logistic Regression for Probabilistic Predictions:
a) Role: Provide interpretability and probabilistic outputs to predict cancer risk.
b) Use Case: Adds explainability for medical professionals by offering a probability estimate of a patient having cancer.

(v) Linear Regression for Continuous Predictions:
a) Role: Estimate related outcomes such a as tumor size, growth rate, or severity levels based on continuous variables.
b) Use Case: Provides with the real time data on as tumor size, growth rate, or severity levels based on continuous variables.

3.4 Overall Workflow:

(i) Input Data: CT/X-ray images and patient data (e.g., age, smoking history, symptoms).
(ii) FRCNN (Image Analysis): Detect cancerous regions and extract image features.
(iii) RF (Feature Selection): Combine and refine features from FRCNN outputs and patient data.
(iv) SVM (Classification): Predict whether the patient has cancer or not based on the refined features.
(v) Logistic/Linear Regression: Provide risk probabilities and continuous predictions if needed.

3.5 Dataset

(i) LIDC-IDRI Dataset: A large dataset of CT scans annotated with lung cancer nodule data, useful for training FRCNN [3].
(ii) TCIA (The Cancer Imaging Archive): Provides imaging datasets, such as MRI scans, for head and neck cancers [2].
(iii) SEER Database: A structured clinical dataset with demographic and medical history information of cancer patients, ideal for training regression models and Random Forest classifiers [2].

3.6 Advantages and Disadvantages

Advantages

(i) Enhanced Diagnostic Accuracy

a) Faster R-CNN (FRCNN): Detects and localizes cancerous regions with high precision, reducing false positives and false negatives.

b) Random Forest (RF): Effectively selects the most relevant features, improving the overall accuracy of the model by focusing on the most predictive data points.

c) Support Vector Machine (SVM): Handles non-linear decision boundaries well, ensuring accurate classification of cancerous and non-cancerous categories.

d) Logistic Regression: Provides probabilistic predictions that are easily interpretable, aiding clinicians in decision-making.

e) Linear Regression: Offers continuous predictions, such as estimating tumor size and growth rate, which are critical for monitoring cancer progression.

(ii) Automation and Efficiency

a) Automated Detection: By automating the feature extraction and classification process, the system reduces the dependency on manual interpretation, leading to faster and more consistent diagnostic processes.

b) Real-Time Analysis: The scalability of the system allows it to handle large datasets and perform real-time analysis, making it suitable for high-volume medical facilities.

(iii) Interpretable Risk Estimates

Logistic Regression: Provides interpretable risk scores that can be easily understood by medical professionals, enhancing the transparency and trustworthiness of the system.

(iv) Scalability

System Integration: The hybrid model can scale to handle large datasets efficiently, making it adaptable to various clinical settings, from small clinics to large hospitals.

Disadvantages

(i) Computational Demands

Faster R-CNN (FRCNN): Deep learning models like FRCNN require significant computational power, particularly for real-time analysis, which can be a limiting factor in resource-constrained environments.

(ii) Data Dependency

Quality of Data: The system’s accuracy heavily depends on the quality and availability of large, annotated datasets. Acquiring such datasets can be challenging due to privacy concerns and the cost of data annotation.

(iii) Interpretability Issues

Complex Models: Deep learning models, such as FRCNN, may reduce transparency due to their complexity, leading to potential trust issues among healthcare professionals who need to understand how decisions are made.

(vi) Integration Challenges

System Integration: Integrating multiple models (FRCNN, RF, SVM, logistic regression, and linear regression) into a cohesive system can be complex and may require substantial technical expertise and resources.

(v) Maintenance and Updates

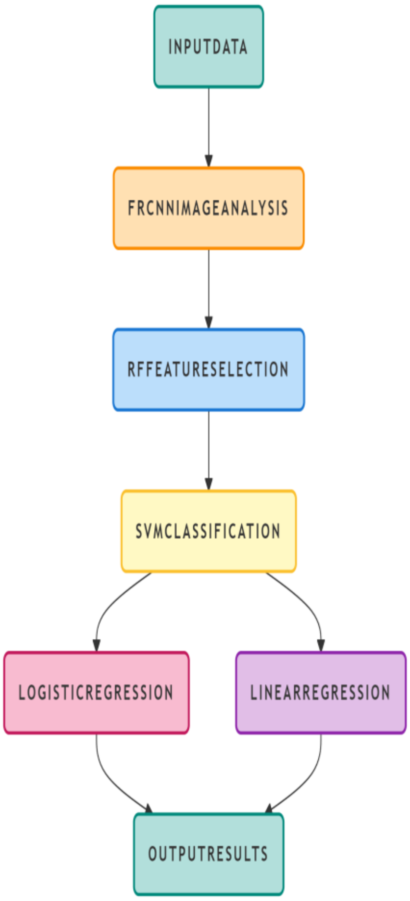
Ongoing Maintenance: Regular updates and maintenance are needed to ensure the models remain accurate and effective, which can be resource-intensive and require continuous monitoring.

3.7. Proposed Architecture

Proposed Architecture for Early Cancer Detection

(i) Architecture Diagram

The architecture of the proposed system integrates various stages to create a robust framework for early cancer detection. Here's a detailed breakdown:



[Output: Cancer Risk & Continuous Predictions]

(ii) Data Collection

The system collects two types of data:

a) Imaging Data

Sources: Computed Tomography (CT) scans, Magnetic Resonance Imaging (MRI), and X-rays.

Purpose: These images are essential for detecting abnormalities in tissues, providing a visual basis for analysis.

b) Structured Clinical Data

Components: Patient history (e.g., smoking history, previous medical conditions), demographic information (age, gender), and other relevant medical records.

Purpose: This data helps in understanding the patient's overall health context, which is crucial for accurate risk assessment.

(iii) Data Cleaning

Before analysis, the collected data undergoes preprocessing:

Image Preprocessing

Noise Reduction: Techniques such as Gaussian blurring or median filtering are applied to reduce noise in the images, enhancing the clarity of features.

Normalization: Adjusting the pixel intensity values to a common scale to ensure consistency across different images.

(iv) Clinical Data Preprocessing

Handling Missing Data: Methods such as mean imputation or predictive modelling are used to fill in missing values in the clinical records.

Normalization: Standardizing the data to bring all features onto a similar scale, which is essential for improving the performance of ML models.

(v) Visualization

The system provides various visual outputs to aid clinicians:

Heatmap:
Purpose: Highlight detected cancerous regions in the imaging data. These visual cues make it easier for clinicians to identify areas of concern quickly.
Generation: Based on the bounding boxes and regions of interest identified by FRCNN.

(vi) Probabilistic Cancer Risk Scores

Purpose: Provide interpretable risk scores indicating the likelihood of cancer, aiding in clinical decision-making.
Presentation: sually presented as percentages or probabilities, making them easy for clinicians to understand and act upon.

(vii) Algorithm Application

The system integrates multiple algorithms to ensure accurate detection and prediction:

a) Faster R-CNN (FRCNN)
Role: Detects and localizes abnormal tissues in imaging data.
Method: Uses a region proposal network (RPN) to identify regions of interest, followed by a convolutional neural network (CNN) to classify these regions.

b) Random Forest (RF)
Role: Refines feature selection from both imaging data and clinical records.
Method: Constructs multiple decision trees and aggregates their results to select the most relevant features, reducing the noise and enhancing the accuracy of subsequent models.

(viii) SVM and Regression Models

Support Vector Machine (SVM)
Role: Classifies the refined features into cancerous or non-cancerous categories.
Method: Finds the optimal hyperplane that separates different classes in a high-dimensional feature space.

(ix) Logistic Regression

Role: Provides probabilistic outputs for cancer risk.
Method: Estimates the likelihood of cancer presence based on the selected features, giving interpretable risk scores.

(x) Linear Regression

Role: Provides continuous predictions such as tumor size and growth rate.
Method: Models the relationship between features and continuous outcomes, aiding in monitoring and treatment planning.

(xi) Expected Accuracy

a) Faster R-CNN (FRCNN): Expected to achieve over 90% accuracy in detecting cancerous regions due to its advanced image analysis capabilities.

b) Random Forest (RF) and Support Vector Machine (SVM): Expected to achieve 85-88% accuracy in classifying cancer risk based on the features selected. This high level of accuracy stems from the robust feature selection and classification mechanisms employed by these models.

By leveraging this comprehensive architecture, the system aims to provide accurate, efficient, and interpretable results for early cancer detection, thereby improving clinical decision-making and patient outcomes.

4. CONCLUSION

The integration of Faster Region-based Convolutional Neural Network (FRCNN), Random Forest, Support Vector Machine (SVM), and regression models into a hybrid system for early cancer detection offers a powerful approach to enhancing diagnostic accuracy. By leveraging the strengths of deep learning for precise image analysis and machine learning for robust classification and prediction, this hybrid system can provide clinicians with reliable tools for early detection of cancers. However, challenges such as the availability of high-quality annotated datasets, model transparency, and computational demands remain. Addressing these issues is crucial for the successful implementation and widespread adoption of this system in clinical workflows. Ongoing research into optimizing model interpretability and efficiency will be essential to ensure that these advanced diagnostic tools can be effectively integrated into medical practice, ultimately improving patient outcomes.

5. FUTURE SCOPE

Integrating AI-based cancer detection into telemedicine platforms offers significant advantages, particularly in expanding access to healthcare for remote patients. By enabling remote diagnosis, this technology reduces healthcare costs and provides real-time AI analysis, leading to faster, more efficient diagnoses. Additionally, personalized medicine, which incorporates genomic data, tailors cancer treatments to individual patient genetics. This approach allows for targeted therapies, improves the effectiveness of treatments, and offers predictive insights into disease progression, helping clinicians make more informed decisions.

Moreover, developing mobile applications for AI-based cancer screening can be transformative in low-resource areas. These mobile-friendly tools increase access to diagnostics in underserved regions, providing cost-effective and easy-to-use solutions for health workers. With the capability for both offline and online AI analysis on portable devices, these applications make cancer screening more accessible and practical, especially in regions with limited medical infrastructure.

REFERENCES

[1] Xie, Y., et al. "A Deep Learning Framework for Identifying Lung Cancer from CT Scans." Journal of Medical Imaging, 2017.

[2] Liu, W., et al. "Throat Cancer Detection Using Voice Recordings and LSTM Networks." IEEE Transactions on Biomedical Engineering, 2018.

[3] Guerreiro, T., et al. "Current Evidence for a Lung Cancer Screening Program." Port J Public Health, 2024.

[4] Hardavella, G., et al. "How Lung Cancer Screening Will Change Diagnostic Landscapes." Eur Respir Rev, 2024.

[5] Gayap, H. T., & Akhloufi, M. A. "Deep Machine Learning for Medical Diagnosis: Lung Cancer Detection." BioMedInformatics, 2024.

[6] Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems (NeurIPS).

[7] Detecting Throat Cancer from Speech Signals using Machine Learning: A Scoping Literature Review, 2024

[8] Enhancing cancer detection and prevention mechanisms using advanced machine learning

[9] approaches, 2024