Aswan University Journal of Science and Technology Volume 5, issue 1, March 2025

Lung Cancer Detection System with XAI

Tamer Abdel Latif Ali[1,2], Abdelrahman Mohamed Ibrahim[2], Mostafa M. Elsherbini[3]

[1] Department of Software Engineering, College of Computing and Information Technology, Arab Academy for Science, Technology &Maritime Transport, Aswan, Egypt

[2] Department of Computer Science, College of Computing and Information Technology, Arab Academy for Science, Technology &Maritime Transport, Aswan, Egypt

[3] Department of Electronics and Communication, College of Engineering and Technology, Arab Academy for Science, Technology &Maritime Transport, Aswan, Egypt

Abstract

Lung cancer detection using deep learning significantly improves patient outcomes by enabling timely treatment and diagnosis that crucial for survival. Deep learning techniques introduce high efficiency by allowing the analysis of complex medical data, making them valuable in enhancing diagnostic precision. This study presents a proposed system leveraging artificial intelligence (AI) algorithms for the early detection of lung cancer, addressing the significant global mortality associated with the disease. In addition, the system utilizes nail image analysis for curvature assessment and user-provided data such as smoking history and genetic predisposition. A complementary dataset was created to enable rapid and accurate detection. Advanced AI models, Explainable AI (XAI) using the SHAP model, were employed to extract important details from the data, enhancing detection accuracy. The system also identifies additional diagnostic markers, such as clubbing nails, a key indicator of lung cancer. The results show the superiority of the proposed system with a detection accuracy of 96.0% and a loss rate of 0.07.

Keywords: Lung Cancer; Machine Learning; Deep Learning; XAI.

Introduction

Lung cancer is considered as one of the most severe health challenges facing our society, with its incidence rates continuing to rise. However, with advancements in treatment techniques and methodologies, managing the disease remains a challenge, often due to late-stage diagnoses resulting from inadequate early detection and monitoring [1]. Consequently, the disease may progress significantly, making treatment a prolonged and uncertain process. The widespread prevalence of smoking has heightened the risk of lung cancer, not only among active smokers but also among non-smokers exposed to secondhand smoke [2], [3].

Technical intervention helps assess an individual's likelihood of developing lung cancer using diagnostic data. The system uses an initial diagnostic technique to evaluate the probability, advising medical consultation for moderate or high-risk cases while deeming immediate attention unnecessary for low-risk cases. This approach ensures timely treatment decisions based on the assessed risk, [1], [4]. Machine learning combined with deep learning improves the accuracy of lung cancer detection by enabling precise tumor identification through AI models trained on extensive datasets. This collaboration across radiology, oncology, and computer science enhances early diagnosis and patient care. The integration also streamlines healthcare processes, ensuring advanced technology supports efficient disease detection and management, [5, 6].

*Corresponding author E-mail: mostafa.elsherbini@aast.edu
Received November 20, 2024, received in revised form, December 15, 2024, accepted December 15, 2024.

The proposed system uses advanced AI techniques, including convolutional neural networks and deep learning, to analyze medical images for precise and rapid lung cancer detection. By identifying subtle features in early stages, it aims to enhance diagnostic accuracy and timeliness, improving patient outcomes, [7, 8].

The rest of this article is organized as follows: Section II presents a literature review of existing techniques and their limitations. Section III outlines the proposed technique along with the paper's contributions. In section IV, the datasets utilized in the proposed technique are described. Section V focuses on the results and evaluation of the proposed approach, including comparisons with alternative methods. Finally, the conclusion and potential directions for future work are discussed.

Literature review

The author of this study proposed a hybrid deep learning model combining ResNet101 and SqueezeNet to extract feature mappings from CT images, using mRMR to optimize feature selection. The refined features were classified with SVM and KNN, achieving 99.09% accuracy with SVM. Results demonstrated the model's exceptional performance on the IQ-OTH/NCCD dataset [5].

This study introduces a deep learning model that uses CNNs, separable CNNs, and residual blocks to achieve 98% accuracy in lung cancer detection, surpassing pathologists' performance (100% vs. 79%). The model not only improves diagnostic precision but also offers potential as a training tool for less experienced pathologists, [9].

This study presents "DeepXplainer," a hybrid deep learning model combining CNN and XGBoost for lung cancer detection and explainable predictions using SHAP. The model achieved 97.43% accuracy, 98.71% sensitivity, and a 98.08 F1-score on the "Survey Lung Cancer" dataset, outperforming existing methods. By offering interpretable predictions at local and global levels, "DeepXplainer" supports clinicians in diagnosis and treatment decisions [10].

This study demonstrates a hybrid feature extraction method for lung cancer detection, combining GLCM, Haralick, and autoencoder-derived features. Supervised machine learning models, particularly SVM polynomial with 99.89% accuracy, showcased near-perfect performance, highlighting the method's potential to improve diagnostic precision and support treatment planning, [11].

This study introduces an ensemble of deep learning models, combining ResNet-152, DenseNet-169, and EfficientNet-B7 with a novel weight optimization scheme to classify lung nodule severity. Achieving 97.23% accuracy and 98.6% sensitivity on the LIDC-IDRI dataset, the method outperforms existing approaches and reduces false negatives, enhancing diagnostic precision, [12].

Proposed System

SVM is a powerful learning method used in binary classification. Its main task is to find the best hyperplane that can perfectly separate the data into two classes. We first stored all the data except the "level" attribute in "X" and stored the "level" in "Y". Then we split the data at a rate of 20% for testing and at a rate of 80% for training. Then we fed this data into the "SVM" for training and the value was "kernel = rbf". Then we did a "fit" for "x_train, y_train" and then evaluated the training and calculated the "accuracy" to show whether the training was effective or not. Recently, multi-class classification has been achieved by combining multiple binary SVMs [11]. The structure of the SVM is shown in Figure 1.

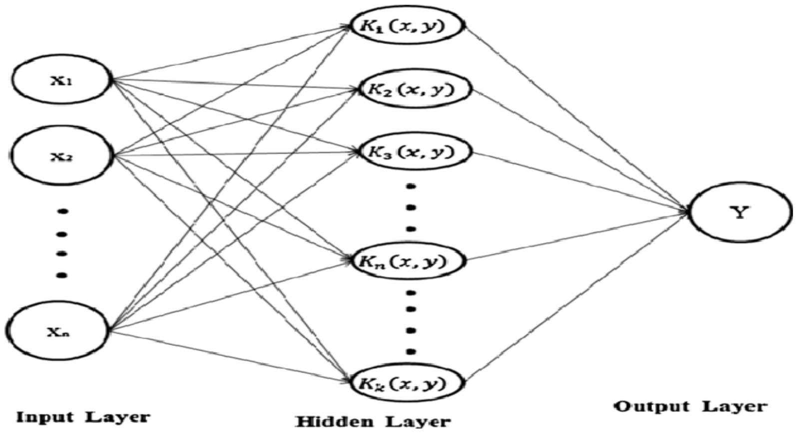

Figure 1: SVM architecture [8]

XAI (Explainable Artificial Intelligence): It is important to make decisions about a machine learning model that humans can understand, and it is important to build trust in real AI systems such as in the financial and medical domains. We used XAI to extract the most important features and learn them and ignore the features that have low importance through the SHAP model.

SHAP (Shapley Additive exPlanations): is a type of XAI technique used to visualize predictions made by individuals that are made by complex models such as Deep Neural Network [13].

• SHAP does not involve building a separate explanation model.

• It works by interacting with the existing model you want to explain, which is the SVM model that the dataset was trained on.

• SHAP performs calculations to estimate the contributions of features in a single prediction or across multiple predictions Transfer learning, hyper parameter tuning, and exceptions are techniques used to enhance and improve the performance of machine learning models [13].

• Transfer learning reuses a previously trained model on a new project or task, leveraging the knowledge gained from a related problem to improve performance on a new but similar task [13].

Hyper parameter tuning optimizes the settings that control the model training process. These settings do not directly encode knowledge but can significantly impact the model's performance. Transfer learning, hyper parameter tuning, and exceptions involve three operations [13].:

• Transforming the data distribution.

• Task-specific learning rates.

• Regularization strength.

There is no specific architecture, but it contains three methods:

• Fine-tuning using initial hyper parameters.

• Transfer learning libraries.

• Transfer learning research.

We used the Xception architecture to build the Clubbin nail model and before training the model we did preprocessing of the images by resizing the data to "(128, 128)" and by increasing the number of images by we made “augmentation” of the images. This was done by adding the following values: "rotation_range=30, width_shift_range=0.1, height_shift_range=0.1, shear_range=0.2, zoom_range=0.2, horizontal_flip=True "So, he produced a number of “3552” images, between infected and healthy, and in the next step we added the original images to the images produced from “augmentation”, so the number of images that we will use in building the model became “473”. 6” images, and this was the last step in the “preprocessing” process.

Then we trained the model through This step is the most important step, and it is the step of building the model. We have used the “transfer learning” method, and we have tried “four architectures.” The first is “VGG16,” the second is the developed version of it, which is “VGG19,” the third is “RESNET50,” and the fourth and final is “Xception.” We have used the "hyper parameters" method to obtain the best results. We used three hidden layers, and each layer had its own parameters as follows: "min_value=64, max_value=1024, step=32", and the activation function was "relu" and also consisted of three dropouts, and the special values were They are as follows: "min_value=0.2, max_value=0.9, step=0.1" and the last hidden layer, which is used in classification, contains the activation function, which is "sigmoid", and the best model among them was "Xception". It was the model used in the end. The detailed steps of the proposed system are represented in figure 2.

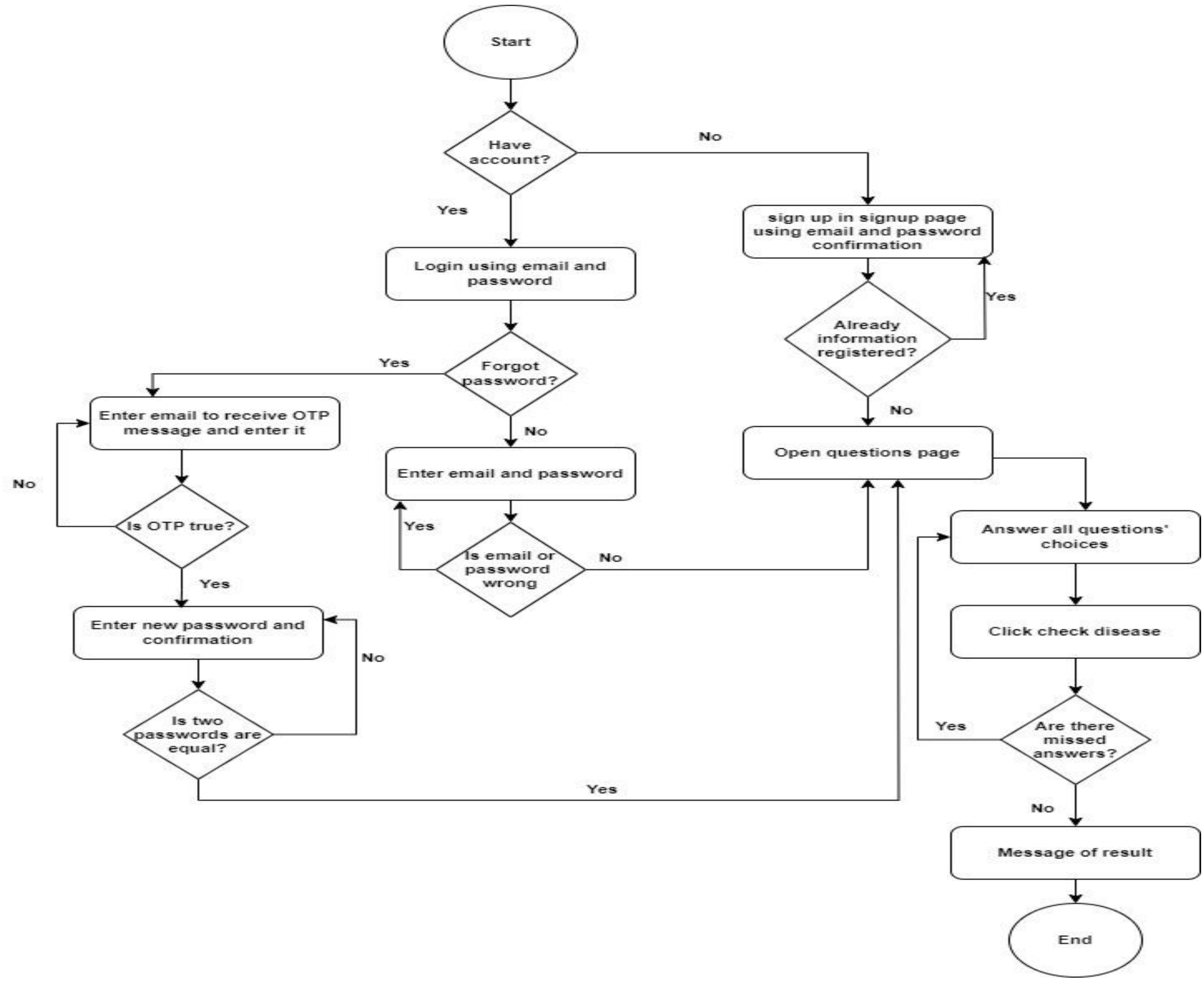

Figure 2: Flowchart of Lung Cancer Detection System

Dataset

The dataset used in this system is divided into two types. The first type contains our dataset. We collected data containing 24 features: "age, gender, exposure to air pollution, alcohol consumption, dust allergy, occupational hazards, genetic risks, chronic lung diseases, balanced diet, obesity, smoking status, passive smoking status", chest pain, bloody cough, fatigue levels, weight loss, shortness of breath, wheezing, difficulty swallowing, nail curvature, recurrent colds, dry cough, and snoring. The data was collected from the International Cancer Database, and the available results for this data are medical questionnaires and international medical opinions. The second type is used in the clubbing nail feature. It contains images and images divided into two types: images of healthy nails and images containing disease. It contains 1480 images between this and that [14].

Results and Discussion

Figure 3 shows the results of using XAI, from which we used the SHAP model by applying it to the SVM model, and it gave us the importance of all the features in the dataset and their percentage in each level, in order to extract the most important features and remove the unimportant features or those of weak importance.

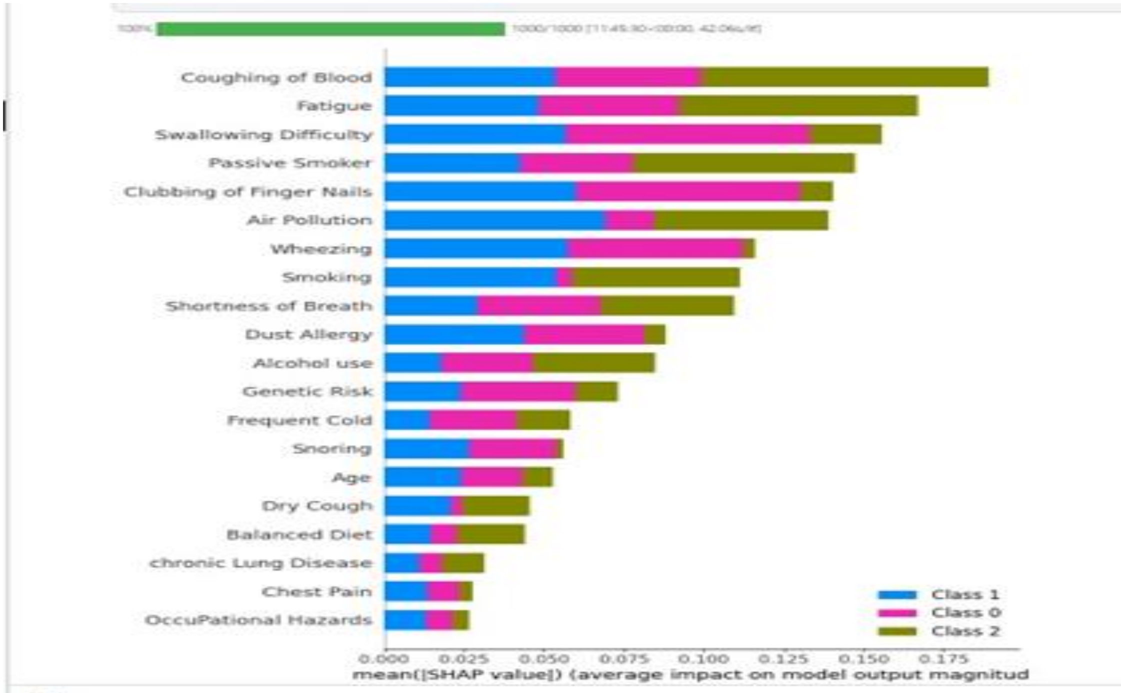

Figure 3: Data Selection

These weak features were represented in ["index", "Patient Id","Gender","Obesity","Weight Loss","OccuPational Hazards","Chronic Lung Disease","Chest Pain","Wheezing"], and thus the quality of the model increased and their number changed from 25 features to 17 features with the label.

Figure 4 shows the results that the system achieved. The evaluation was done by calculating the accuracy for each of these models, which are KNeighbors, and its accuracy was 100%. Here, there is overfitting, as was also found in the Logistic Regression model, and its accuracy was also 100%. As for the Gaussian NB model, it had the lowest results, and its accuracy was 87.5%. As for the best result, it was for the SVM model, and its accuracy was 96.0%. This was the model used.

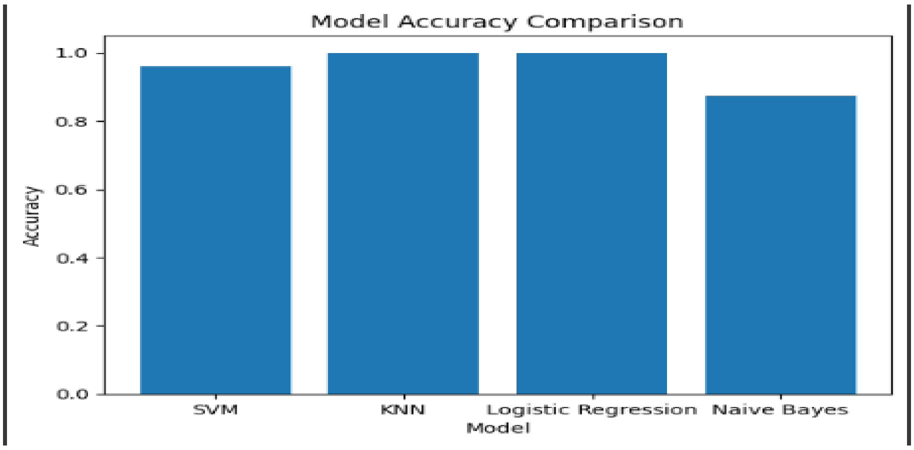

Figure 4: Model Accuracy

Figure 5 shows the results that the system reached, and the evaluation was by calculating the Loss for each of those models, which are KNeighbors, and its Loss was 0.0, and here there is overfitting, as was also found in the Logistic Regression model, and its Loss was 0.057 as well.

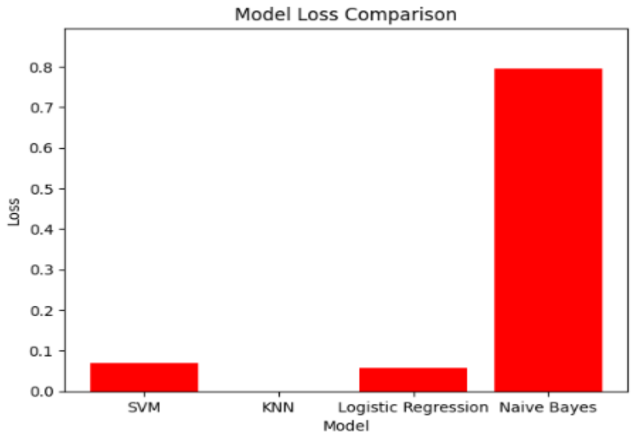

Figure 5: Model Loss

As for the Gaussian NB model, it had the highest results, and its Loss was 0.79. As for the best result, it was for the SVM model, and its Loss was 0.07.

Figures 6 and 7 show a comparison between Training and Validation accuracy, which were 0.9926 and 0.9561, respectively. And a comparison between Training and Validation loss, which were 0.0189 and 0.2341, respectively.

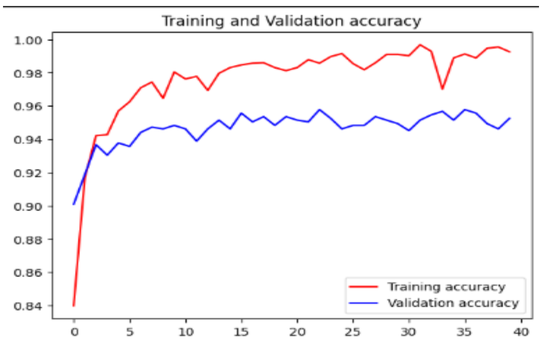

Figure 6: Training and Validation accuracy

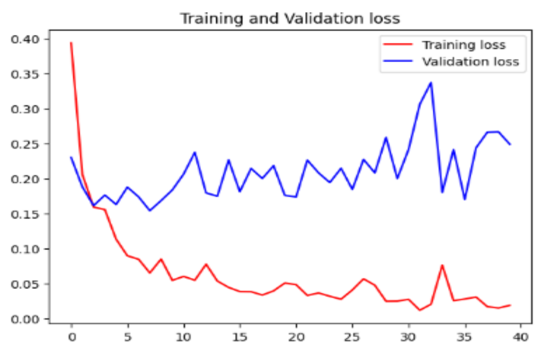

Figure 7: Training and Validation loss

- Login in so that user can access to app through email and password. Signup: that user can create a new account through email and
- Password and confirm password if don't have previous account.
- Question Page: the camera takes a photo for the user or patient to check clubbing nail and in another pages put questions that user must answer it.
- Another Question Pages: make choices for the patient in check box or radio buttons to check his pains
- Forget Password page: If a user
- forgets his password can recover an account by entering the email.
- Check Validation page: After entering the email application send via email the OTP number (verification code) to ensure the OTP number delivered to the account personally.

Figures 8, 9 and 10 are samples from the mobile application pages.

Figure 8: Login page

Figure 9: Sign up page

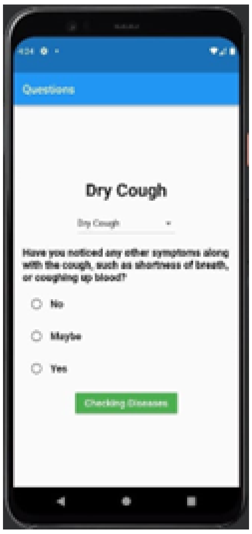

Figure 10: Questions

The Attributes in order of importance features that divided into 3 blocks high, medium and low original data that was used to create the model is in the form of numbers before it is converted into understandable data from which information can be extracted as shown in figure 11 and figure 12 respectively.

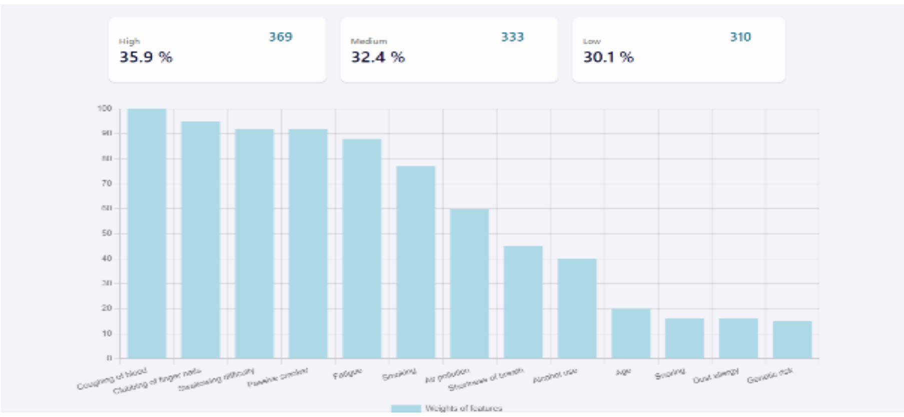

Figure 11: column chart (Weights of features)

INDEX	PATIENT ID	AGE	GENDER	AIR POLLUTION	ALCOHOL USE	DUST ALLERGY	OCCUPATIONAL HAZARDS	GENETIC RISK	CHRONIC LUNG D
0	P1	33	1	2	4	5	4	3	2
1	P10	17	1	3	1	5	3	4	2
2	P100	35	1	4	5	6	5	5	4
3	P1000	37	1	7	7	7	7	6	7
4	P101	46	1	6	8	7	7	7	6
5	P102	35	1	4	5	6	5	5	4
6	P103	52	2	2	4	5	4	3	2
7	P104	28	2	3	1	4	3	2	3
8	P105	35	2	4	5	6	5	6	5
9	P106	46	1	2	3	4	2	4	3
10	P107	44	1	6	7	7	7	7	6

Figure 12: table (the real data)

Figure 13: Dashboard sample

The pie chart that describes 3 levels of result may be high. medium or low. column chart that describes the relation between age and level of cancer by detecting each type of level with average of age that they have. Figure 13 shows a sample from the dashboard.

Doughnut chart that views clubbing of finger nails describing it to yes (clubbed nails) or no (not clubbed) scatter chart that displays coughing of blood and in relation with level of cancer Displaying coughing of blood table with patients ID, age of patient and result of cancer level to view the relation between all coughing of blood results with cancer level.

Pie chart the view the percentage of each type of result level and in relation with coughing of blood, meaning how many patients have coughing and have one of each result level as shown in figure 4.

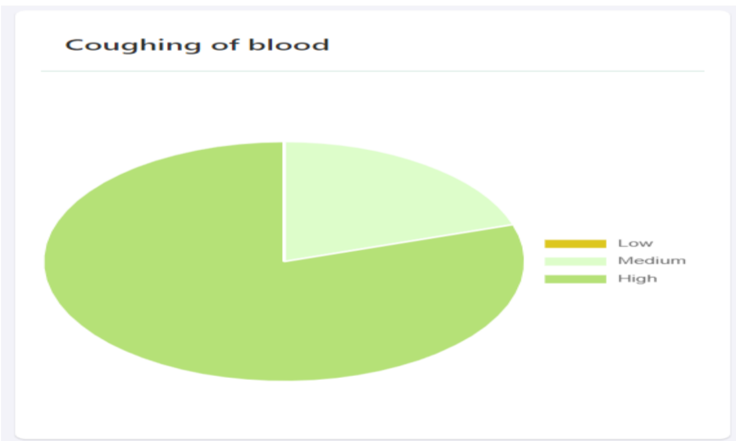

Figure 14: pie chart of coughing of blood pie chart

Conclusion & Future work

This study underscores the transformative potential of integrating radiology, oncology, and computational systems to improve healthcare delivery. The developed system not only facilitates early disease detection but also offers a pathway to integrating additional diagnostic capabilities. By addressing the identified limitations, the system can be further refined to meet evolving healthcare needs.

Future efforts will focus on increasing the efficiency of the clubbing nail detection model, developing a deep learning model to assess fatigue levels for user convenience, and enhancing the application's user interface. Additional goals include enabling platform-based sign-up functionality, integrating advanced security features for improved data protection, and ensuring offline accessibility of the application. These enhancements aim to expand the system's usability, security, and adaptability for diverse healthcare scenarios.

References

1. Ahmad, S.A.; Mayya, A.M. A new tool to predict lung cancer based on risk factors. Heliyon 2020, 8, e03402. https://doi.org/10.1016/j.heliyon.2020.e03402.
2. World Health Organization. Lung Cancer. World Health Organization, 2024. Available online: https://www.who.int/news-room/fact-sheets/detail/lung-cancer (accessed on 27 January 2024).
3. Perra, N. Modeling and Predicting Human Infectious Diseases. Natl. Libr. Med. 2014, 854, 37–48.
4. Amin, M.E.K.; Nørgaard, L.S.; Cavaco, A.M.; Witry, M.J.; Hillman, L.; Cernasev, A.; Desselle, S.P. Establishing trustworthiness and authenticity in qualitative pharmacy research. Res. Soc. Adm. Pharm. 2020, 16, 1472–1482. https://doi.org/10.1016/j.sapharm.2020.02.005.
5. Zafar, S.; Ahmad, J.; Mubeen, Z.; Mumtaz, G. Enhanced Lung Cancer Detection and Classification with mRMR-Based Hybrid Deep Learning Model. J. Comput. Biomed. Inform. 2024, 7, 02. https://doi.org/10.56979/702/2024.
6. Hussein, A.; Abd-Elhafiez, W.; Zanaty, E.; Hussein, M. Medical Image Segmentation Using Deep Learning. Aswan Univ. J. Sci. Technol. 2023, 3, 87–108. https://journals.aswu.edu.eg/stjournal.
7. Dahab, Y.A. Improving Glaucoma Detection: Harnessing the Power of Ensemble Semantic Segmentation for Optic Disc and Optic Cup with Deep Learning. Aswan Univ. J. Sci. Technol. 2024, 4, 1–14. https://journals.aswu.edu.eg/stjournal.
8. Wang, H.; Shao, Y. Sparse and robust SVM classifier for large scale classification. Appl. Intell. 2023, 53, 19647–19671. https://link.springer.com/article/10.1007/s10489-023-04511-w.
9. Ahmed, A.A.; Fawi, M.; Brychcy, A.; Abouzid, M.; Witt, M.; Kaczmarek, E. Development and Validation of a Deep Learning Model for Histopathological Slide Analysis in Lung Cancer Diagnosis. Cancers 2024, 16, 1506. https://doi.org/10.3390/cancers16081506.
10. Wani, N.A.; Kumar, R.; Bedi, J. DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. Comput. Methods Programs Biomed. 2024, 243, 107879. https://doi.org/10.1016/j.cmpb.2023.107879.
11. Li, L.; Yang, J.; Por, L.Y.; Khan, M.S.; Hamdaoui, R.; Hussain, L.; Iqbal, Z. et al. Enhancing lung cancer detection through hybrid features and machine learning hyperparameters optimization techniques. Heliyon 2024, 10, e26192. https://doi.org/10.1016/j.heliyon.2024.e26192.

12. Gautam, N.; Basu, A.; Sarkar, R. Lung cancer detection from thoracic CT scans using an ensemble of deep learning models. Neural Comput. Appl. 2024, 36, 2459–2477. https://link.springer.com/article/10.1007/s00521-023-09130-7.
13. Belle, V.; Papantonis, I. Principles and practice of explainable machine learning. Front. Big Data 2021, 4, 688969. https://doi.org/10.3389/fdata.2021.688969.
14. Clubingnail Dataset. Available online: https://universe.roboflow.com/mini-project-xlocq/nailcheck (accessed on 27 January 2024).