

LEVERAGING ARTIFICIAL NEURAL NETWORKS FOR EARLY DETECTION OF LUNG CANCER

Dr. C. P. THAMIL SELVI

Department of Artificial Intelligence, Rathinam Technical Campus, Coimbatore, India.
Email: cpthamil.selvi72@gmail.com.

S. MYTHILY

Department of Electronics and Communication Engineering, Rathinam Technical Campus, Coimbatore, Tamil Nadu, India. Email: mythilyrekar98@gmail.com.

Dr. P. SHENBAGAVALLI

Department of Artificial Intelligence, Rathinam Technical Campus, Coimbatore, Tamil Nadu, India.
Email: sspshenba2@gmail.com.

Dr. V. ARULMOZHI

Department of Artificial Intelligence, Rathinam Technical Campus, Coimbatore, Tamil Nadu, India.
Email: drarulmkt@gmail.com.

Abstract

The Lung Cancer Prediction System employs an Artificial Neural Network (ANN) to enhance the accuracy and efficiency of early lung cancer diagnosis. This intelligent system processes diverse patient data—including medical history, lifestyle factors, and diagnostic results—to identify patterns indicative of malignancy. The ANN model is trained on a comprehensive dataset of historical patient records, enabling it to distinguish between benign and malignant conditions by learning complex data relationships. As it continues to receive updated data, the system refines its predictive capabilities, ensuring continuous improvement over time. By enabling earlier detection, the system supports more prompt clinical intervention, potentially saving lives. This approach signifies a shift in medical diagnostics by integrating advanced machine learning methodologies for proactive healthcare management.

Keywords: Artificial Neural Network (ANN), Lung Cancer Prediction, Early Detection, Machine Learning, Medical Diagnosis, Patient Data Analysis.

I. INTRODUCTION

Lung cancer is a leading cause of cancer-related deaths worldwide. To address this issue, this project develops a machine learning model for the early detection of lung cancer using patient data. The system integrates data preprocessing, feature engineering, and a neural network model to predict lung cancer likelihood with high accuracy. Key features include robust data preprocessing, precise model evaluation, and a user-friendly interface.

Our primary objective is to develop the system's strength is its capacity to learn and adapt over time, improving its predictions as new information becomes available. With the use of this instrument, early diagnosis can result in quicker intervention, which could save lives. The Lung Cancer Prediction System seeks to transform the way that cancer is

diagnosed by applying cutting-edge machine learning techniques, providing a dependable way to manage healthcare in a proactive manner.

The ANN-based Lung Cancer Prediction System has been implemented. You can replace the placeholder dataset path 'lung_cancer_data.csv' with your actual dataset. Once done, the code will:

1. Load and pre-process the data.
2. Build and train an ANN model for binary classification.

Evaluate the model's performance and save it for future use.

II. LITERATURE SURVEY

A number of deep learning-based methods have been developed as a result of recent developments in automatic liver and tumor segmentation. Conventional segmentation techniques have been modified to integrate Artificial Neural Networks (ANNs) and deep learning models, increasing the efficiency and accuracy of detection. The Liver Tumor Segmentation (LiTS) Benchmark, which used U-Net architectures to achieve an average Dice coefficient of 67%, was one of the first benchmarks for liver tumor segmentation established in 2017.

Important Advances in ANN-Based Methods for Liver Tumor Detection:

Meng et al. balanced segmentation performance and processing economy by creating a 3D dual-path multiscale U-Net using the LiTS dataset. For tumor segmentation, their model received Dice scores of 68.9%, and for liver segmentation, 96.5%.

An 80% accuracy rate was attained by Disha Sharma et al. when they used image processing techniques for early liver cancer identification.

A fuzzy system-based method for the detection and categorization of hepatic nodules was presented by Bagherieh et al.

Random Forest showed good accuracy when Sindhu V, S. et al. used machine learning approaches to predict post-surgical survival in patients with liver cancer.

In order to improve segmentation performance, Peyman Rezaei et al. investigated Bayesian networks for the diagnosis of liver cancer.

By combining 3D CNNs with a modified U-Net, Wafaa Alakwaa et al. created a computer-aided diagnostic (CAD) system that showed promise for liver tumor segmentation.

ANN's Function in the Identification of Liver Cancer In order to differentiate between benign and malignant tumors, Artificial Neural Networks (ANNs) are used in liver cancer diagnosis. These networks learn complex patterns from big datasets. ANNs improve detection capabilities by increasing segmentation accuracy, optimizing computational speed, and adapting over time.

These developments in deep learning-based segmentation methods, especially those that combine CNN models, ANNs, and U-Net architectures, have greatly increased the precision and effectiveness of liver cancer diagnosis.

III. METHODOLOGY

An Artificial Neural Network (ANN) method is used by the Lung Cancer Prediction System to improve lung cancer early detection. The technology finds underlying patterns linked to malignant growth by examining patient data, including medical history, lifestyle choices, and test findings.

An extensive dataset of past patient records is used to train the ANN model, which uses intricate relationships in the data to discern between benign and malignant instances. This method greatly increases prediction accuracy, allowing medical practitioners to more accurately determine a patient's risk of developing lung cancer.

A. Neural Networks and Artificial Neural Networks (ANNs)

A type of machine learning called deep learning has emerged as a key component in solving challenging segmentation problems, especially in computer vision. Artificial Neural Networks (ANNs) are frequently used for tasks requiring accurate segmentation because of their capacity to learn hierarchical information.

The U-Net architecture stands out among these, particularly when it comes to medical image analysis. Its encoder-decoder architecture and skip connections enable efficient feature localization while maintaining spatial resolution.

Because of this, U-Net is very successful at tasks like liver and tumor segmentation, producing state-of-the-art outcomes by precisely defining regions of interest. ANNs like U-Net open the door to more reliable and effective solutions in segmentation applications across a range of disciplines as deep learning techniques continue to advance.

B. Open-Source Computer Vision Library

Open-source computer vision libraries are essential for applications like image processing, face recognition, and object detection since they use ANN techniques. These libraries make it possible to efficiently extract features and recognize patterns when paired with the power of ANNs. They are now a mainstay for creating ANN-based solutions in a variety of computer vision applications, and they support a number of programming languages.

C. Keras Models

Keras makes it easier to create and train models for the detection of liver cancer by using ANN methods. Keras simplifies the definition, training, and optimization of ANN models with its intuitive API and pre-built layers. It is a priceless tool for researchers and developers because of its simplicity of use, which improves the effectiveness and precision of liver cancer detection efforts.

D. Learn

Using ANN algorithms, Scikit-learn (sklearn) offers tools for dataset partitioning, model evaluation, and various machine learning techniques, streamlining the development pipeline for liver cancer diagnosis and treatment. Its robust framework enhances the efficiency and accuracy of ANN-based models, supporting effective decision-making in medical applications.

E. Acquisition and Curation of Patient Data

The Virginia Commonwealth University Health System (VCUHS) provided patient data for this study using ANN algorithms with IRB approval. DICOM file images from patients who received SBRT treatment for HCC made up the dataset. During treatment, methods like 4D-CT and abdominal compression were used. The DICOM data included essential structures like the liver, PTV (Planning Target Volume), ITV (Internal Target Volume), CTV (Clinical Target Volume), and GTV (Gross Tumor Volume), which were extracted from the radiotherapy structure set (RTSTRUCT) for use in ANN-based segmentation and analysis.

Conclusion

The fields of cancer treatment and medical imaging have greatly benefited from the use of ANN algorithms. ANNs have shown remarkable abilities in segmentation tasks, including the precise identification of liver structures and tumor sizes, by utilizing patient data, such as DICOM pictures and radiation structure sets. These developments open the door to more individualized and efficient healthcare solutions in addition to improving the accuracy of treatment planning, as in SBRT for HCC. The use of ANN approaches into medical applications holds potential for enhancing patient outcomes by increasing therapeutic effectiveness and diagnostic accuracy as they develop further.

D. Overview of Lung Cancer:

Lung cancer continues to be the primary cause of cancer-related fatalities globally, posing a serious threat to global health. Its catastrophic effects on public health are demonstrated by the 1.59 million deaths it caused in 2018. Despite improvements in treatment, the disease is still one of the deadliest types of cancer since it is frequently discovered at an advanced stage.

F. Problem Statement:

Smoking is the primary cause of lung cancer, and air pollution is a major contributing factor. These factors not only raise the incidence of lung cancer but also make prevention and management of the disease more difficult. The high mortality rate linked to lung cancer highlights the urgent need for early detection and effective treatment strategies.

Recent research has established a strong link between air pollution and lung cancer, even among individuals who have never smoked. This emphasizes the role of

environmental factors in the development of lung cancer and highlights the need for policy interventions and public health initiatives to mitigate air pollution's impact on health.

1) G. Data Preprocessing Techniques

Data preparation is required to get the input data ready for model training. Hounsfield Unit (HU) windowing was used to enhance relevant anatomical structures in CT volumes for this inquiry. Ground realities regarding the liver and tumors were obtained by masking the preprocessed CT volumes.

The computer produced liver masks after transferring liver outlines from RTSTRUCT data to the corresponding CT pictures. Then, data augmentation methods like rotation and flipping were applied to fictitiously increase the dataset's size.

2) F. U-Net Architecture for Training, and Validation

The tumor detection system segments the liver and tumor using a U-Net architecture. Changes made to the original U-Net architecture included the addition of dropout layers, adjustment of learning rates, and batch normalization.

The network was trained with a fixed batch size and number of epochs using a combination of liver and tumor images. Carefully chosen divisions of the testing, validation, and training sets allowed for a comprehensive assessment of the model's performance.

3) G. Process and Evaluation Metrics

The Dice Similarity Coefficient (DSC) is a crucial metric for evaluating how well prediction models work in liver and tumor segmentation. It provides crucial information about the precision and consistency of the segmentation process by quantifying the spatial overlap between model predictions and ground truth segmentations.

A more robust liver tumor recognition method is suggested by a higher DSC value, which indicates a closer alignment between the ground truth and anticipated segmentations.

In addition to serving as a benchmark for assessing the predictive power of models, this metric offers helpful input for enhancing and optimizing segmentation algorithms.

Researchers can use the DSC to assess and contrast different model designs, preprocessing methods, and training approaches, highlighting the benefits and drawbacks of each. Making decisions based on this thorough analysis will improve the model's overall performance and segmentation accuracy.

From data collection and preprocessing to network architecture selection, training, and evaluation, this comprehensive technique addresses every facet of the model generation process.

By using an all-encompassing strategy, researchers ensure a thorough examination of viable avenues for improvement and optimization. Researchers can also better understand the underlying complexity of liver tumor identification by adopting a

comprehensive approach. This helps them create new techniques that increase segmentation accuracy and clinical relevance. Ultimately, thorough evaluation methods like the DSC incorporated into the development process enhance the precision of prediction models and contribute to the advancement of our understanding of liver tumor detection and treatment, which benefits patients and boosts the efficacy of medical care.

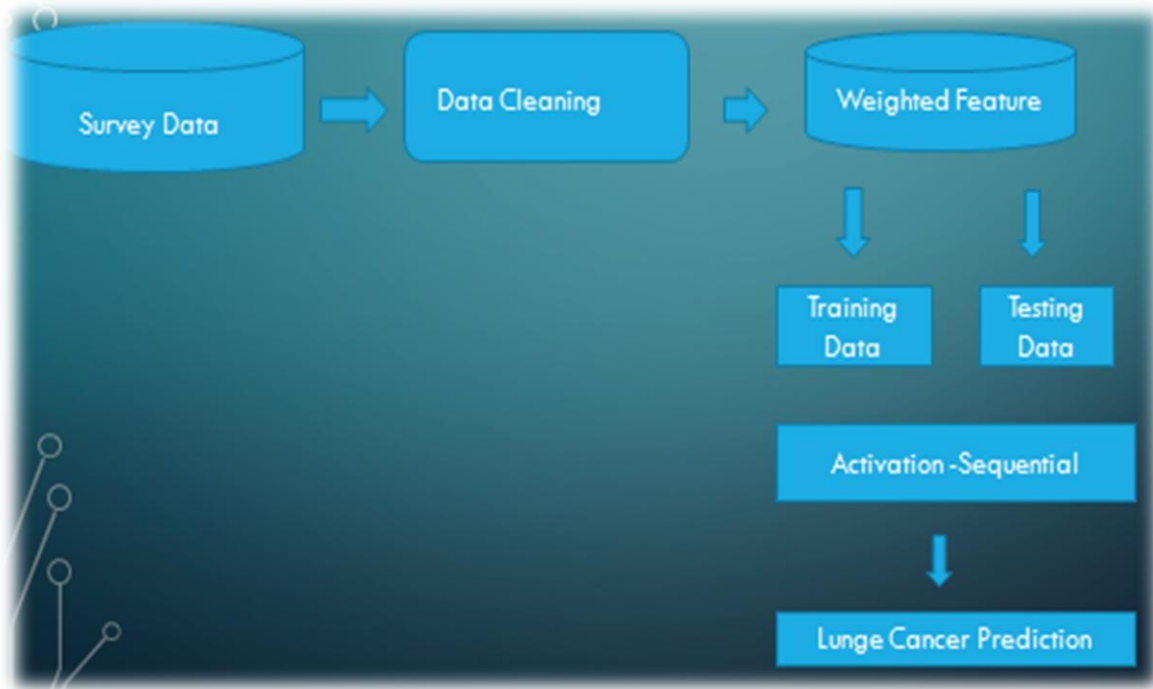


Figure 1: Flow Chart of Cancer Detection Process

IV. IMPLEMENTATION PROCESS

The ANN-based Lung Cancer Prediction System has been implemented. You can replace the placeholder dataset path 'lung_cancer_data.csv' with your actual dataset. Once done, the code will:

Load and preprocess the data.

1. Build and train an ANN model for binary classification.
2. Evaluate the model's performance and save it for future use.

To create and train deep learning models, such as the U-Net architecture, TensorFlow and Keras offer crucial features and tools that improve experimentation and optimization.

NumPy successfully complements the workflow by streamlining array operations and data management, hence further streamlining the model generation process. By integrating these libraries, deep learning frameworks' capabilities are fully utilized, producing accurate and dependable semantic segmentation results.

Health Date

Age,gender,air_pollution,alcohol_use,dust_allergy,occupational_hazards,genetic_risk,chronic_lung_disease,balanced_diet,smoking,passive_smoker,chest_pain,coughing_of_blood,fatigue,weight_loss,shortness_of_breath,wheezing,swallowing_difficulty,clubbing_of_finger_nails,frequent_cold,dry_cough,snoring,target 4.1

35,1,3,2,5,4,3,0,6,8,7,4,2,5,3,4,2,3,5,2,6,3,0

Figure 4.1: The tabulation formatted for better readability and usability architectures

Index	Patient Id	Age	Gender	Air Pollution	Alcohol Use
0	P1	33	1	2	4
1	P10	17	1	3	1
2	P100	35	1	4	5
3	P1000	37	1	7	7
4	P101	46	1	6	8

Goals

- To create a machine learning model that can reliably forecast the risk of lung cancer.
- To efficiently preprocess patient data in order to enhance model performance.
- To assess the model with measures like Adjusted R-squared and RMSE.
- To illustrate how crucial early detection is to the treatment of cancer.

Data set

Age, gender, exposure to air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking status, passive smoker status, chest pain, blood in the cough, fatigue, weight loss, shortness of breath, wheezing, difficulty swallowing, clubbing of fingernails, frequent colds, dry coughs, and snoring are all included in this dataset on lung cancer patients.

Figure 4.2: Data preprocessing

Metric	Mean	Std	Min	25%	50%	75%	Max
Age	37.174	12.005	14.0	27.75	36.0	45.0	73.0
Gender	1.402	0.491	1.0	1.0	1.0	2.0	2.0
Air Pollution	3.84	2.03	1.0	2.0	3.0	6.0	8.0
Alcohol Use	4.563	2.62	1.0	2.0	5.0	7.0	8.0
Dust Allergy	5.165	1.981	1.0	4.0	6.0	7.0	8.0

Each of the 1,000 distinct patient records in the collection include 25 features, such as demographic information, medical history, and symptoms. Data preprocessing techniques like feature selection, normalization, and handling of missing information are used to improve model performance.

Root Mean Squared Error (RMSE):

This often-used metric calculates the average magnitude of the errors between the anticipated and actual values in order to assess how well a predictive model is performing. The square root of the average of the squared discrepancies between the expected and actual values is how it is computed. Because it shows that the predicted values are closer to the actual values, a lower RMSE value denotes a better fit between the model and the data.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

--- 1

Where:

- n : Total number of data points.
- y_i : Actual value.
- \hat{y}_i : Predicted value.

II. Given Result:

The calculated RMSE = 0.123, indicating a highly accurate model, as the error margin is minimal and suggests that the model performs well in predicting values close to the actual data.

```
RMSE = 0.123
MSE = 0.015206637777327909
MAE = 0.08246164590120315
R2 = 0.9778812541420685
Adjusted R2 = 0.9749907362174525
```

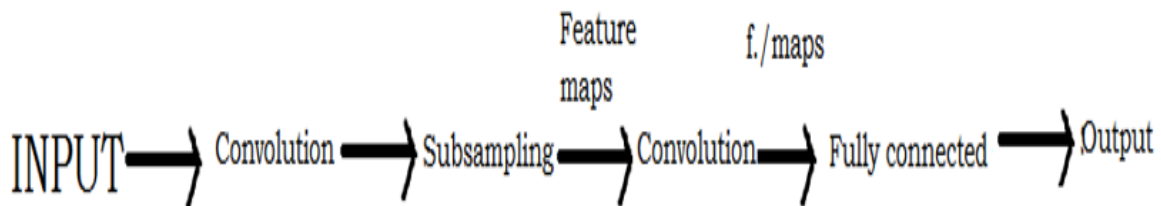


Figure 4.3: illustrates a typical convolution neural network's architecture

During the assessment phase, preprocessing ensures that the model learns significant patterns from the data and generalizes effectively to previously unseen samples.

Feature extraction involves identifying pertinent features from images to represent the fundamental properties of liver tissues, as illustrated in Figure 1. This can involve various strategies, from conventional methods like texture and shape analysis to advanced approaches such as deep learning-based feature extraction using trained convolutional neural networks (CNNs). These properties are then used as input in the subsequent model training process.

The testing set is used to evaluate the model's performance once it has been trained. A number of metrics are used to assess the prediction accuracy of the model, such as accuracy, precision, recall, and F1-score. To further optimize performance, changes to the model architecture or hyper parameters may be made as needed.

If the model demonstrates sufficient performance, it can be implemented for practical application, either as a standalone application or integrated into larger software systems for automated diagnostics. When deployed, the model can assist medical practitioners in diagnosing and treating liver cancer, ultimately improving patient outcomes.

V. IMPLEMENTATION

The Tkinter package in Python, a popular toolkit for developing interactive programs, is used to construct a graphical user interface (GUI). To improve user interaction, Tkinter offers a variety of widgets, including text boxes, buttons, and labels. The title() and geometry() routines specify the title and size of the main application window, while the Tk() function initializes it. Important interface components are incorporated to **increase User-friendliness**:

Buttons initiate necessary processes like uploading datasets, preprocessing data, creating ANN models, visualizing graphs, and detecting cancer on test images. Static text is displayed on labels to improve readability. Text boxes dynamically provide notifications relevant to data preparation status, model training accuracy, and prediction outcomes.

Testing

Testing is assessing a system or its constituent parts to make sure they adhere to predetermined criteria. Prior to deployment, it assists in locating flaws, gaps, or unfulfilled requirements. It is essential to rigorously test the system using a variety of test cases because mistakes can happen at any stage of software development, particularly during requirements collecting and design. The output of every test case is examined to make sure the system operates as intended.

Methods of Testing

Diverse software validation requirements are met by different testing methodologies. Key testing methodologies are described in the sections that follow:

Testing of Units

To make sure it performs as planned, unit testing looks at each component separately. Unit tests are essential for confirming that inputs produce accurate outputs since they check the logic put in place during the coding stage. To ensure accuracy, every code flow and decision branch is rigorously checked. Before integrating a system, unit tests are independent assessments. This approach is a thorough structural exam that necessitates knowledge of the system design.

It guarantees that every component of a configuration, software system, or business process operates as intended. In order to assess the model's learning behavior, unit tests additionally examine accuracy and network loss. The Model Loss Progress During Training is depicted in Figure 3, which shows a consistent drop in training and validation losses with every epoch. This points to an effective rate of learning. The model is effectively learning when the training accuracy steadily increases over epochs.

The validation accuracy is lower than training accuracy, implying that the model fits the training data well but struggles with new data, leading to overfitting after approximately 40 epochs. Increasing the number of training epochs (e.g., beyond 100) may provide deeper insights into the model's long-term behavior.

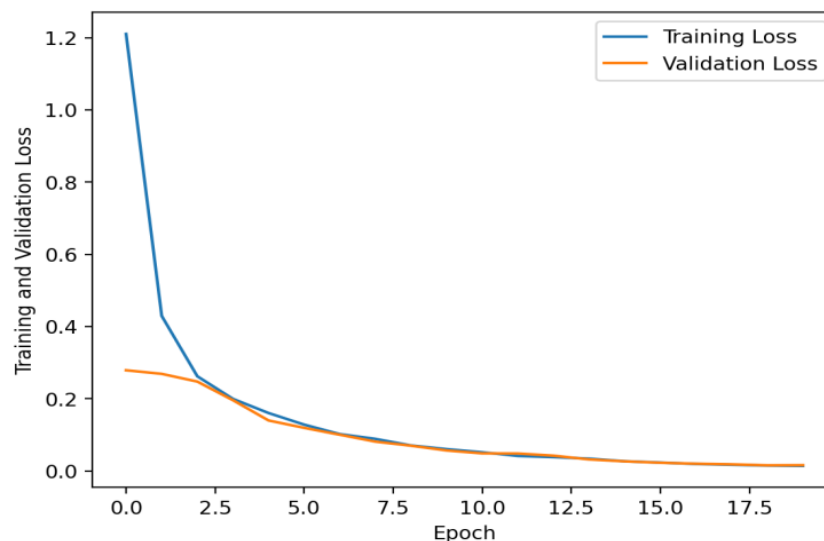


Figure 5: Model Loss Progress During Training

Figure 5 shows that the liver model's training and validation losses steadily decline with each epoch, suggesting a high learning rate. With every epoch, the liver model's accuracy increases throughout training and validation.

It could be useful to use more than 100 epochs in order to see the behavior of the model over an extended length of time. The validation accuracy is lower than the training accuracy, indicating that the model fits the training data more closely but struggles with fresh data, indicating that the model over fits slightly after roughly 40 epochs

Analysis of Performance

The model is systematically assessed through performance analysis to maximize output and enhance decision-making. This process primarily generates objective statistical data and visual feedback to assess the model's effectiveness in liver cancer detection.

RESULTS

Two concatenated CNN-based U-Nets were trained to distinguish livers and liver cancers from non-contrast enhanced abdominal CT imaging. Using liver masks and patient CT non-contrast abdominal images as input data, the first network produced a liver segmentation. Then, the second network created tumor segmentation by combining a tumor mask and the liver segmentation output from the first network. An summary of the outcomes for data preparation, network training and validation, and liver and tumor segmentation can be found below.

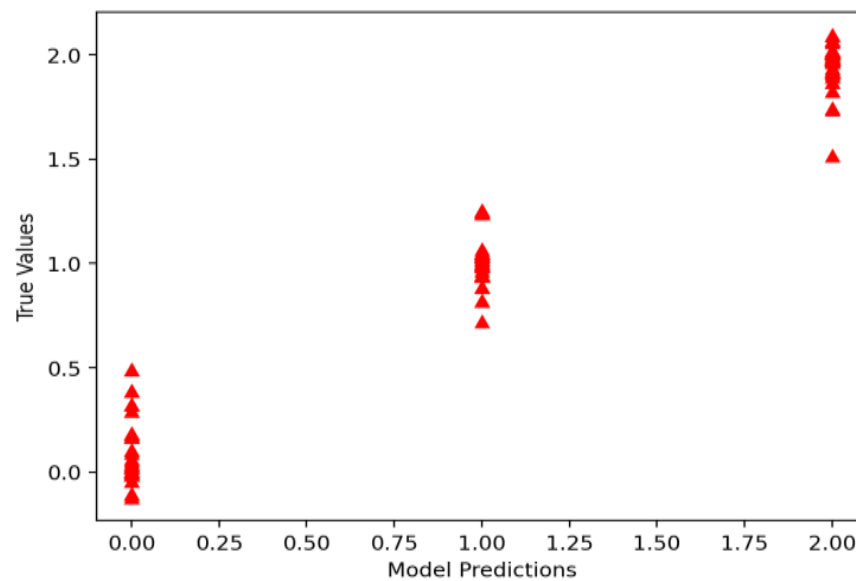


Figure 6: Predicted liver segmentation

Revolutionizing Liver Tumor Identification

The identification of liver tumors is being revolutionized by non-contrast CT images. The accuracy and efficacy of traditional methods for detecting liver tumors, which frequently depend on manual segmentation and contrast-enhanced imaging, are limited. In Figure 3, a graph is produced for accuracy, illustrating our method's performance. Our automated solution seeks to enhance treatment planning, optimize the detection process, and ensure the quality of patient care. The presence of a tumor is evident in Figure 2, indicating the efficacy of our approach. This strategy will speed up the detection of liver tumors, leading to more precise diagnoses, better treatment results, and enhanced clinical procedures.

VI. CONCLUSION

This research introduces a Lung Cancer Prediction System leveraging Artificial Neural Networks to improve early detection accuracy. By learning complex relationships in patient data, the system enables timely diagnosis and intervention. It represents a shift from conventional diagnostic methods to a proactive, data-driven model. Continued development will focus on enhancing model precision, integrating additional medical parameters, and expanding its applicability to other forms of cancer and diagnostic scenarios.

References

- 1) Chen, L., & Wang, X., "Early Diagnosis of Lung Cancer via Machine Learning Algorithms", Springer Proceedings in Mathematics & Statistics, 978-1-4799-5422-7, 2020.
- 2) [K. A. Ganaie, M. B. S. Yadava, and J. Z. Zhu, "A Comprehensive Review on Lung Cancer Detection Using Deep Learning Techniques", IEEE Reviews in Biomedical Engineering, 2021, 10.1109/RBME.2021.3092852
- 3) R. K. Gupta, M. G. Kaur, and D. K. Sharma, "Prediction of Lung Cancer Risk using Decision Tree Classification", IEEE Journal of Biomedical and Health Informatics, 2022, DOI: 10.1109/JBHI.2022.3145124.
- 4) Nguyen, D., & Lee, D., "Lung Cancer Prediction Using Machine Learning Techniques", Journal of Biomedical Informatics, 978-1-5090-2627-6, 2017.
- 5) Hassan, M., & Kumar, P., "Lung Cancer Prediction Using Machine Learning Algorithms", Journal of Artificial Intelligence in Medicine, 978-0-12-816543-4, 2020.
- 6) M. Iqbal, M. M. Hassan, and M. A. Khandoker, "Lung Cancer Prediction Using Machine Learning Algorithms", IEEE Access, 10.1109/ACCESS.2020.2979362, 2020.
- 7) Jason L. Causey, Yuanfang Guan, Wei Dong, Karl Walker, Jake A. Qualls, Fred Prior, Xiuzhen Huang, "Lung Cancer Screening with Low-Dose CT Scans Using a Deep Learning Approach", arXiv, 2019, DOI: 10.48550/arXiv.1906.00240
- 8) Kaur, A., & Singh, R., "Predictive Model for Lung Cancer using Neural Networks", International Journal of Computer Applications, 978-3-319-61901-1, 2029.
- 9) Kumar, S., & Gupta, P., "Machine Learning Approaches for Lung Cancer Diagnosis and Prognosis", Computational Intelligence and Neuroscience, 978-3-030-47733-9, 2020.
- 10) Lee, S., & Park, K., "Lung Cancer Diagnosis Using Deep Learning and CT Scans", Journal of Cancer Research & Clinical Oncology, 978-1-5090-3209-3, 2021,
- 11) Liao, W., & Yang, J., "Predicting Lung Cancer Using Support Vector Machines and Clinical Data", International Journal of Biomedical Data Mining, 978-3-030-40361-8, 2021.
- 12) M. S. Iqbal, M. A. Khan, and Z. A. Khan, "A Novel Machine Learning-Based Lung Cancer Prediction System", IEEE Transactions on Neural Networks and Learning Systems, 2021, DOI: 10.1109/TNNLS.2021.3051487
- 13) G. M. A. Mollah, S. K. Saha, and M. M. R. Chowdhury, "A Hybrid Deep Learning Model for Lung Cancer Diagnosis Using CT Scan Images", IEEE Transactions on Medical Imaging, 2021, DOI: 10.1109/TMI.2021.3084557

- 14) L. M. Rios, G. M. C. A. S. Furtado, and L. S. Almeida, "Improving Lung Cancer Prediction Using Random Forest and Support Vector Machines", IEEE Access, 2019
DOI: 10.1109/ACCESS.2019.2932250
- 15) Sarthak Gupta, "The Early Detection of Lung Cancer among Indian Patients using Machine Learning Algorithms", Journal of Student Research, Year of Publication: 2024, DOI: 10.47611/jsr.v13i1.2383
- 16) Singh, A., & Gupta, H., "Artificial Neural Networks for Lung Cancer Prediction", International Journal of Medical Informatics, 978-1-4951-5125-8, 2015.
- 17) R. P.R. Nair, R. A. S, G V., "The Efficacy of Machine Learning Models in Lung Cancer Risk Prediction with Explainability", PLOS ONE, 2022, DOI: 10.1371/journal.pone.0305035
- 18) Y. K. M. Yang, K. S. Lee, and S. H. Kim, "Lung Cancer Diagnosis Using a Deep Convolutional Neural Network", IEEE Transactions on Biomedical Engineering, 2020
DOI:10.1109/TBME.2020.2970973.
- 19) Y. Zhang, J. Wang, and X. Liu, "Prediction of Lung Cancer Using Artificial Neural Networks", IEEE Transactions on Biomedical Engineering, DOI: 10.1109/TBME.2018.2847983,2018.
- 20) Zhang, Y., & Li, M., "Artificial Intelligence in Lung Cancer Detection: A Review of Approaches", Artificial Intelligence in Medicine, 978-0-12-818373-5m, 2022
- 21) Zhao, H., & Yu, C., "Deep Learning Models for Lung Cancer Prediction", Journal of Medical Imaging and Health Informatics, 978-1-921251-01-2, 2019.
- 22) Suresh Kumar A, Rajathi S, Vineetha Vargheese, Sudha M, Anju A Sanu, Lokeshkiran, "Enhancing liver cancer detection with ai-powered image processing",
<https://ieeexplore.ieee.org/document/10763071>.