COMP8811 : Data Analytics And Big Data Assignment2

S.P.S Priskila Pathirana
Student Id : 1583607
Unitec Institute of Technology

Contents

1 Abstract

There are several environmental factors and life style patterns that are impact on lung cancers in human body. Some of them are age, air pollution, smoking habits and presence of chronic lung diseases. This proposed system introduces Business Intelligence solution based on machine learning to predict the risk of lung cancer by considering different factors. The system mainly uses a neural network model within a Business Intelligence framework. Here the risk levels are categorized into 3 levels as low, medium and high. This helps the user to be aware of lung cancer risk and to be prepared for health decisions. User friendly interfaces helps the user to interact more with the system. And also this data driven approach is beneficial for health care providers to identify high risk patients and provide pre medical care. Moreover, the Business Intelligence capabilities in the proposed system helps health care organizations in analysing lung cancer risk in the population level, optimizing resource allocation, designing healthcare products, improving public health practices to reduce lung cancer among the society. Overall, the solution given by predictive analysis in Business Intelligence shows an innovative approach to health care and patients.

2 Introduction

Lung Cancer is one of the challenging health issue in the world. It is estimated that there had been around 10000 deaths due to lung cancer and 12000 new lung cancer cases in Argentina in the year of 2023. When focus on the economy of Argentina 1.4% of total healthcare expenditure has been allocated for the lung cancer patients [1]. One of the main reason for death of lung cancer patients is late stage detection. There is an urgent need of tools that assess cancer risk. Busy life style and various factors such as air pollution, smoking, excessive alcohol consumption have contribute to risk of lung cancers. Specially, 19.3% of cancer cases are represented by smoking people [2]. Traditional methods that help to perform risk assessment are time consuming and need clinical involvement. So its difficult to provide service for large amount of population. This proposed system is an solution to address this issue through Business Intelligence(BI) techniques. Cancer Prediction System enables individuals to know the risk of lung cancer at the early stage.

The Cancer Prediction System is implemented using machine learning algorithm and has combined BI techniques with user friendly interfaces. Neural network model is trained on several factors to deliver accurate risk prediction. This system classifies users into three categories as low, medium, high and provides quick and understandable way to know the risk of having a lung cancer in future. The predictive analysis of the system assists not only individuals but also health care providers to consider the number of high risk patients in resource allocation and taking various important decisions. This system aims to provide predictions in order to improve prevention of lung cancer and patient care.

3 Business Problem

Normally lung cancer develops slowly with the time. But factors such as age, less immunity, smoking history affect the cancer cells to quickly spread[3]. The main reasons for growing lung cancer are smoking, pollution, harmful gas exposure[4]. It very difficult to identify symptoms from outside at this stage. But if its detected by any medical checkups, the treatments doing at this stage are highly effective. Initial identification of lung cancer gives the chance of living for at least five years for 56% of the lung cancer patients. This disease progresses with the time and starts to show symptoms such as chronic cough, chest pain, shortness of breath or fatigue. Even at these stage, people don't care about the symptoms as they are not aware of the disease. Detection at this stage offers the 5 year survival rate for around 35% of the patients. In the advanced stage (Stage IV) cancer cells spread to other organs like brain, bones or liver. Its more painful at this stage as it causes bone pain, weight loss and neurological symptoms. At this stage the survival rate is very low. So timely treatment in early stages greatly improves survival rates and quality of life.

Generally there are regular screenings for people over 55. But there can be patients younger than 55 and having a risk of lung cancer. Its important to do the treatments for patients with high risk of lung cancer. So there is an urgent need to identify the lung cancer risk at the initial stage. So this proposed system helps to know the risk of having a lung cancer as low, medium or high. And also can know about the percentage of having a lung cancer in future. When people get to know the risk of having a lung cancer in future, they can can take necessary steps to prevent from lung cancer. And also they can even stop the bad health habits that make the disease worse.

4 ABI Technique Review

This implementation represents an advanced application of Business Intelligence (BI) techniques by performing a predictive analysis in healthcare field. This is a proacvtive and data driven approach to predict lung cancer risk. By using machine learning model to predict lung cancer risk, the proposed system goes beyond traditional BI practices such as descriptive and diagnostic analytics. Instead the system focuses on the predictive analysis and prescriptive analysis. By displaying the risk level, the system provides interactive dashboards that can used by individuals as well as healthcare organizations as a powerful BI implementation.

The artificial intelligence model used here is a Multi-Class Classification Neural Network built using Keras library. This model architecture has three layers. First one is input layer which takes the selected feature parameters as input. This is a Dense layer with 32 neurons and Rectified Linear Unit(ReLU) activation. ReLU can be defined as below[5].

f(x)=max(0,x)

The second layer is called hidden layer which is a Dense layer with 16 neurons and ReLu activation. This is common in deep learning due to its ability to introduce non-linearity while avoiding issues like vanishing gradients. The Third layer is the output layer which is a Dense layer with 3 neurons and softmax activation. This layer is allocated for multi class classification such as low, medium and high. The softmax function in the output layer converts the raw output scores into probabilities for each class. When considering the loss function, Categorical Crossentropy is used there that is well suitable for multi class classification probelms. Here the model's goal is to minimize the distance between the predicted probabilities and the true labels. Adaptive Moment Estimation(Adam) Optimizer is used in this model which has a learning rate of 0.001 and well suitable for adaptive learning in neural networks. This is an efficient and adaptive optimizer and also popular in deep learning, which combines the benefits of both AdaGrad and RMSProp algorithms. Permutation feature importance is used to identify the top 10 most important features before retraining the model on these features. Initially the model is trained for 5 epochs to calculate feature importance through permutation. After identifying the top features the model is retrained for 10 epochs, and its performance is evaluated by accuracy on test data. Validation of the model is done for the test data to monitor its performance on unseen data and to reduce the risk of overfitting. The training and validation accuracy levels 1and loss 2 are displayed over the epochs in charts below.

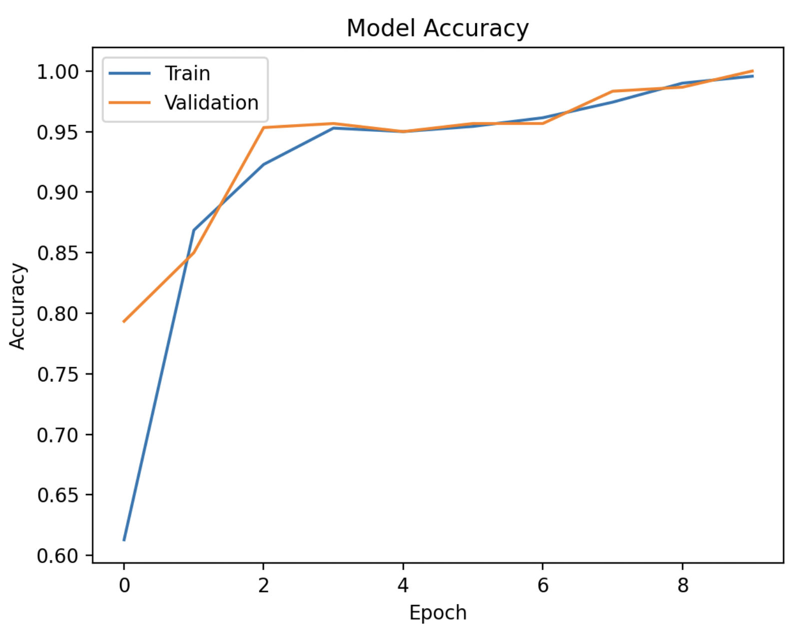


Figure 1: Model Accuracy

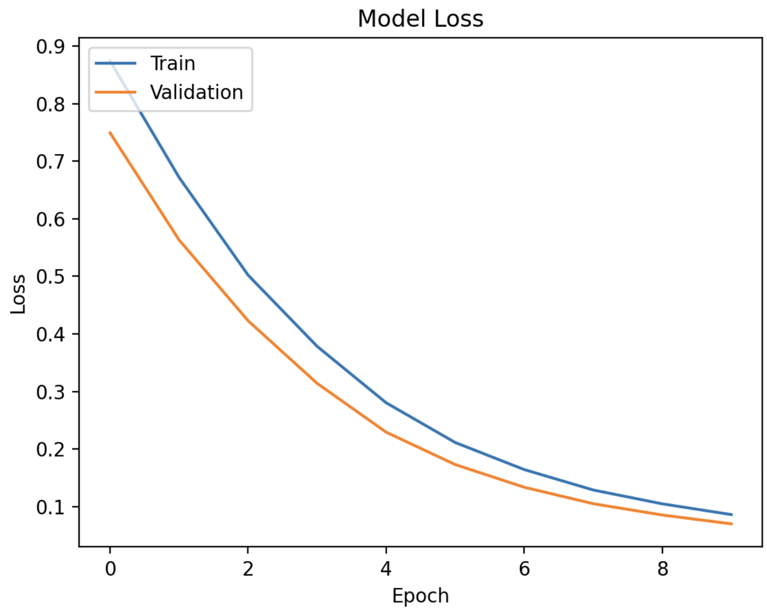


Figure 2: Model Loss

5 Proposed Solution and Implementation

The proposed solution is a machine learning based lung cancer prediction system that uses lifestyle of a person and various other parameters to check the risk level of lung cancer in the body. The system is implemented using nueral network model and provided with csv file to that model. Through the user friendly interfaces this system gives the output of risk levels as low, mdeium and high with the percentages.

From an operational standpoint, the solution integrates key machine learning and deployment practices to ensure accurate predictions and a smooth user experience. By preprocessing the data, performing feature selection, and scaling inputs consistently, the model is optimized to generalize well across different user inputs[6]. Furthermore, the system's ability to save and load the trained model and scaler allows for efficient deployment and reuse without requiring real-time model training. The Streamlit interface is designed to make this solution accessible to both individuals and healthcare providers, simplifying the process of gathering relevant health data and delivering actionable insights. This approach not only aids individuals in understanding their cancer risk but also provides healthcare professionals with a valuable tool to enhance preventive care strategies. By using this system people can make data driven decisions that aim to reduce lung cancer incidence in the future and improve patient outcomes.

5.1 Dataset

The dataset used here is a csv file that contains 25 variables namely Index, Patient Id, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, chronic Lung Disease, Balanced Diet, Obesity, 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 and Level. These values are considered as reasons for having a risk of a lung cancer in human body. It consists of numerical as well as categorical values. Here the target variable is the Level that show the value low or medium or high. These feature allow the model to predict accurate risk level of having a lung cancer. The quality and the balance of data is very important to get the correct output.

5.2 Data Preprocessing

Data preprocessing is done before training the model. This includes mapping categorical target to numerical values using a dictionary. Here the low level is mapped to 0.0, medium level is mapped to 1.0 and high level is mapped to 2.0. This step is done to make the target variable compatible with the model. Another task is feature selection done by excluding first two columns index, Patient Id and last target column Level. Rest of the feature column are considered as X and target variable as Y. Feature scaling is the next step done and StandardScaler is used for this step. This scaling ensures that these feature column's mean equals to 0 and standard deviation equal to 1. This helps to improve the performance of the model and speed. This step is very important for neural networks. The dataset is split into two part as training and testing for ratio of 70 and 30 respectively. Other critical step is the calculation done to identify the important features through permutation. The most influential columns are recognized by shuffling each feature column one at a time and observing the impact on model accuracy. Then its identified top 10 most important features and continued the training. This helps to reduce dimensionality and improve model efficiency. Then the process of scaling is repeated for these features. These are the main data pre processing steps done in order to achieve better performance, reduce computation complexity, and ensure compatibility with the neural network structure.

5.3 Integration with User Interface

The technology used in this system are mainly python libraries. The trained ML model is saved as .h5 file and loaded using load_model, and the scaler is saved as .pkl file and loaded using pickle. The scaler involved in matching user inputs with the data distribution on the model trained. Then the front end part is done by streamlit libraries which includes button, containers, navigation in pages and result display. Session state variables are used to manage page navigation, showing the result, and storing input values temporarily. Buttons such as Home, Services, Blog, About and FAQ allow users to navigate between different pages. In the "Enter Patient Details" section, user inputs are collected through sliders and input boxes for attributes like age, air pollution level, weight loss and stored in st.session_state for consistency and ease of access in the prediction function 3. Then the predict button allows the predict function to be called and display the result in a new page. This show the risk level and the percentages in pie chart 4, "Back" button allows users to go back to the input page and the "Consult a Doctor in Your Area" button opens Google Maps in a new tab to search for nearby medical centers. Other pages such as Services5, Blog7, About6 and FAQ8 provide information for the users with images and links to relevant articles. For these interfaces, CSS is used to style various elements as buttons. This system provides a complete integration between the user interface and the machine learning model.

Figure 3: Home Page

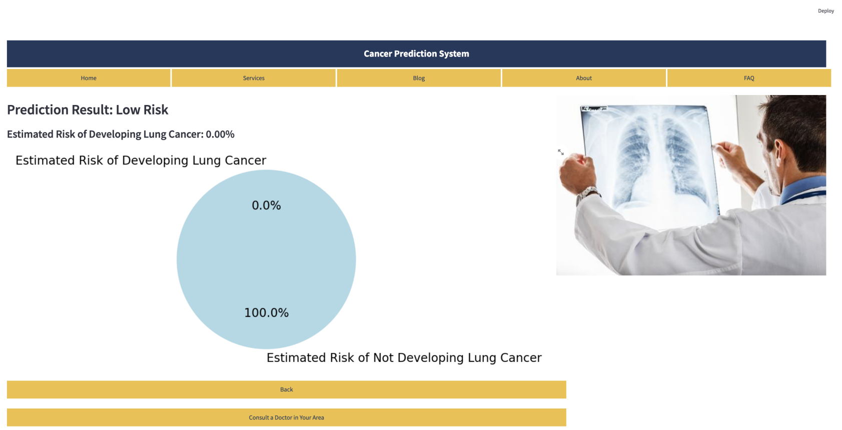


Figure 4: Result Page

Figure 5: Services Page

Figure 6: About Page

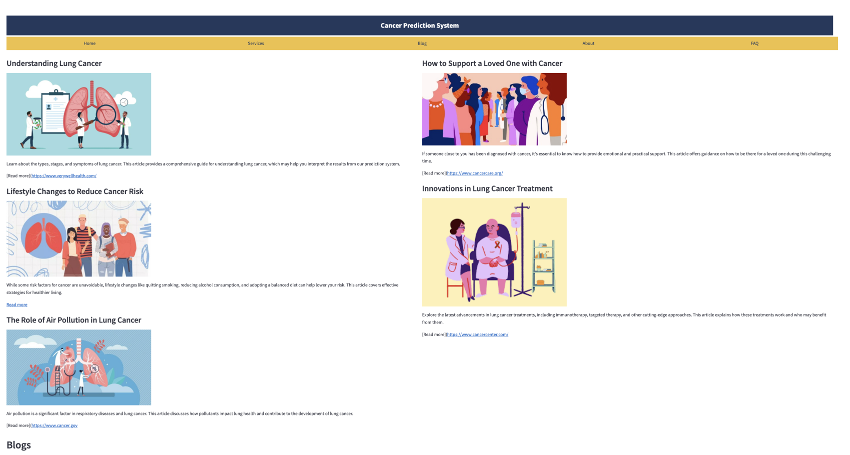


Figure 7: Blog Page

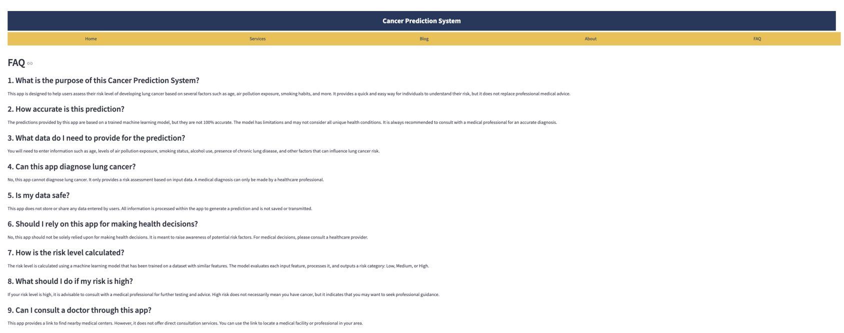


Figure 8: FAQ Page

5.4 Process Flow Chart

The flow chart 9 shows each stage starting from data collection and result display.

1. Data Collection: Load the data from the CSV file
2. Data Preprocessing: Map target values, select features, scale data, and split into training and testing sets
3. Feature Selection: Train initial model to identify the most important features
4. Model Training: Train the final model using the selected features
5. Save Model and Scaler: Save the trained model and scaler to use in the Streamlit app
6. Load Model and Scaler in UI: Load the saved model and scaler in the Streamlit app
7. Collect User Input: Get the inputs from the user
8. Process and Scale Input Data: Convert user input into the required format and scale it
9. Generate Prediction: Use the model to predict the risk level
10. Display Results: Show the risk level and percentages through pie chart

This flow shows each stage in building, training, deploying, and interacting with the model in the user interface.

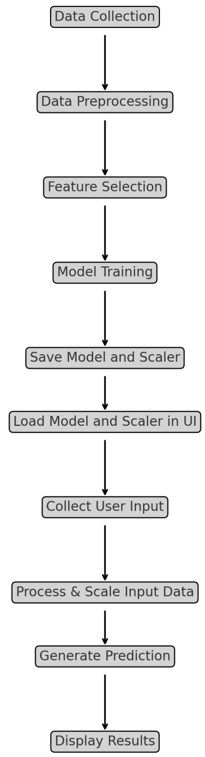


Figure 9: Model Loss

6 Evaluation

The proposed system shows the effectiveness in multiple stages such as feature selection, model training, performance and validation.Initially the model is trained with some key features that selected and got accuracy of 96.67%. The figure 10 below shows the model training before.

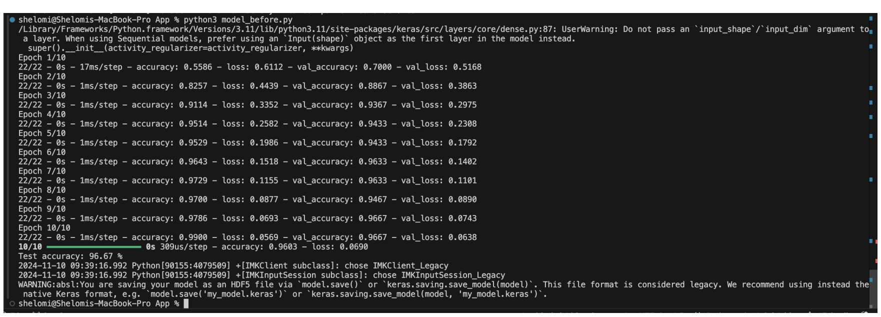
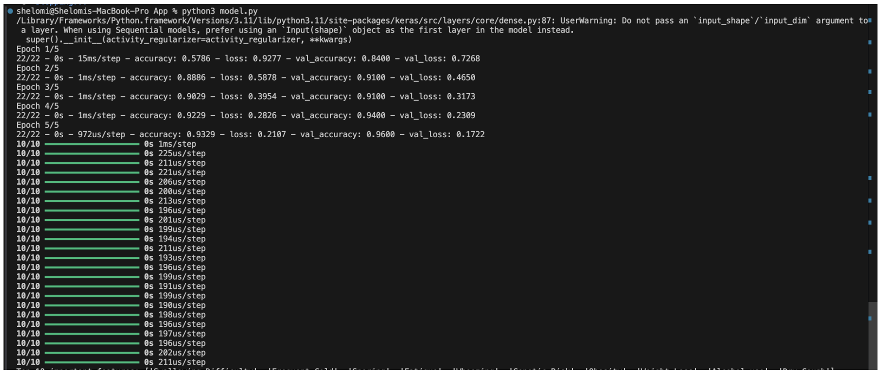


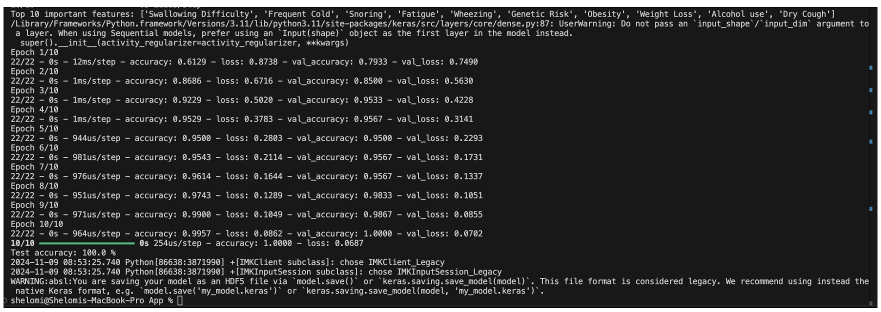
Figure 10: Model Loss

Then I have applied permutation feature importance to identify the top 10 ten features that influence more on the target. Below images shows the model training after and the accuracy reach up to 100%.



```
shelomi@Shelomis-MacBook-Pro App % python3 model.py
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/keras/src/layers/core/dense.py:87: UserWarning: Do not pass an `input_shape`/`input_dim` argument to
 a layer. When using Sequential models, prefer using an `Input(shape)` object as the first layer in the model instead.
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
Epoch 1/5
22/22 - 0s - 15ms/step - accuracy: 0.5786 - loss: 0.9277 - val_accuracy: 0.8400 - val_loss: 0.7268
Epoch 2/5
22/22 - 0s - 1ms/step - accuracy: 0.8886 - loss: 0.5878 - val_accuracy: 0.9100 - val_loss: 0.4650
Epoch 3/5
22/22 - 0s - 1ms/step - accuracy: 0.9029 - loss: 0.3954 - val_accuracy: 0.9100 - val_loss: 0.3173
Epoch 4/5
22/22 - 0s - 1ms/step - accuracy: 0.9229 - loss: 0.2826 - val_accuracy: 0.9400 - val_loss: 0.2309
Epoch 5/5
22/22 - 0s - 972us/step - accuracy: 0.9329 - loss: 0.2107 - val_accuracy: 0.9600 - val_loss: 0.1722
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 225us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 211us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 221us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 206us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 200us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 213us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 196us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 201us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 199us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 194us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 211us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 193us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 196us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 199us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 191us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 199us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 190us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 198us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 196us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 197us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 196us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 202us/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 211us/step
```

Figure 11: Model Loss



```
Top 10 important features: ['Swallowing Difficulty', 'Frequent Cold', 'Snoring', 'Fatigue', 'Wheezing', 'Genetic Risk', 'Obesity', 'Weight Loss', 'Alcohol use', 'Dry Cough']
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/keras/src/layers/core/dense.py:87: UserWarning: Do not pass an `input_shape`/`input_dim` argument to
 a layer. When using Sequential models, prefer using an `Input(shape)` object as the first layer in the model instead.
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
Epoch 1/10
22/22 - 0s - 12ms/step - accuracy: 0.6129 - loss: 0.8738 - val_accuracy: 0.7933 - val_loss: 0.7490
Epoch 2/10
22/22 - 0s - 1ms/step - accuracy: 0.8686 - loss: 0.6716 - val_accuracy: 0.8500 - val_loss: 0.5630
Epoch 3/10
22/22 - 0s - 1ms/step - accuracy: 0.9229 - loss: 0.5020 - val_accuracy: 0.9533 - val_loss: 0.4228
Epoch 4/10
22/22 - 0s - 1ms/step - accuracy: 0.9529 - loss: 0.3783 - val_accuracy: 0.9567 - val_loss: 0.3141
Epoch 5/10
22/22 - 0s - 944us/step - accuracy: 0.9500 - loss: 0.2803 - val_accuracy: 0.9500 - val_loss: 0.2293
Epoch 6/10
22/22 - 0s - 981us/step - accuracy: 0.9543 - loss: 0.2114 - val_accuracy: 0.9567 - val_loss: 0.1731
Epoch 7/10
22/22 - 0s - 976us/step - accuracy: 0.9614 - loss: 0.1644 - val_accuracy: 0.9567 - val_loss: 0.1337
Epoch 8/10
22/22 - 0s - 951us/step - accuracy: 0.9743 - loss: 0.1289 - val_accuracy: 0.9833 - val_loss: 0.1051
Epoch 9/10
22/22 - 0s - 971us/step - accuracy: 0.9900 - loss: 0.1049 - val_accuracy: 0.9867 - val_loss: 0.0855
Epoch 10/10
22/22 - 0s - 964us/step - accuracy: 0.9957 - loss: 0.0862 - val_accuracy: 1.0000 - val_loss: 0.0702
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 254us/step - accuracy: 1.0000 - loss: 0.0687
Test accuracy: 100.0 %
2024-11-09 08:53:25.740 Python[86638:3871990] +[IMKClient subclass]: chose IMKClient_Legacy
2024-11-09 08:53:25.740 Python[86638:3871990] +[IMKInputSession subclass]: chose IMKInputSession_Legacy
WARNING:absl:You are saving your model as an HDF5 file via `model.save()` or `keras.saving.save_model(model)`. This file format is considered legacy. We recommend using instead the
 native Keras format, e.g. `model.save('my_model.keras')` or `keras.saving.save_model(model, 'my_model.keras')`.
shelomi@Shelomis-MacBook-Pro App % 
```

Figure 12: Model Loss

7 Challenges

One of the challenge is getting a quality data set without any uncleaned data. Because the accuracy of the model is highly depends on the dataset and it affects the performance of the model. Also its difficult to find a personal health care information related dataset. Identifying most relevant features can be complex. One other challenge is the transparency of the model. Normally neural network models are "Black Box" and it makes difficult to understand the basis of its prediction.

8 limitations

Limitation of this system is the model's generalization capability, which is highly dependent on the range and variety of the training dataset. Overfitting can occur if the model becomes too regular and common to the training data. It may impacts on the real world application due to reducing of performance on new data. Additionally, if the system's preprocessing steps such as scaling can be inconsistent and due to that prediction accuracy[7] can be affected. Also when user input through interfaces are incorrect and then it directly affects the accuracy of the prediction. The other important aspect is Privacy and regulatory constraints such as like HIPAA and GDPR [8],[9],[10]. These can policies can add complexity to the implementation of the system as they limit the type and amount of data that can be used.

9 Future Development

This system can be improved by using real world related and more quality dataset that impacts on the accuracy level of the result. This can be tested by other different machine learning model in the future in order to address the black box concern. Furthermore, the interface can be more arranged by using front end technologies like React. Moreover, Advanced BI features can be included to the system such as automated reporting in future. The system can develop more to store data into a database like SQL, and intergrate database to a visualization tool like PowerBI to visualize the result in a more attractive way.

10 Conclusion

In conclusion, this cancer prediction system uses machine learning and Business Intelligence techniques to provide data driven method to predict the risk of lung cancer. By using personal health and lifestyle data and machine learning model with a good accuracy, it gives ability to know the risk of having a lung cancer, percentages of having and not having a lung cancer and helps to identify high risk patients. Though there are some challenges and limitation, this cancer prediction system contributes to health care field in a much better way.

References

[1] A. Alcaraz, F. Rodriguez-Cairoli, C. Colaci, C. Silvestrini, C. Gabay, and N. Espinola, “Lung cancer in argentina: a modelling study of disease and economic burden,” *Public Health*, vol. 232, pp. 86–92, 2024.

[2] A. Fuller, “Prevention over cure almost half of all cancer cases are caused by 6 lifestyle factors – are you at risk?.”

[3] A. Shankar, A. Dubey, D. Saini, M. Singh, C. P. Prasad, S. Roy, S. J. Bharati, M. Rinki, N. Singh, T. Seth, *et al.*, “Environmental and occupational determinants of lung cancer,” *Translational lung cancer research*, vol. 8, no. Suppl 1, p. S31, 2019.

[4] Y. Xue, L. Wang, Y. Zhang, Y. Zhao, and Y. Liu, “Air pollution: A culprit of lung cancer,” *Journal of hazardous materials*, vol. 434, p. 128937, 2022.

[5] J. He, L. Li, J. Xu, and C. Zheng, “Relu deep neural networks and linear finite elements,” *arXiv preprint arXiv:1807.03973*, 2018.

[6] A. Oulasvirta, N. R. Dayama, M. Shiripour, M. John, and A. Karrenbauer, “Combinatorial optimization of graphical user interface designs,” *Proceedings of the IEEE*, vol. 108, no. 3, pp. 434–464, 2020.

[7] H. Guan, “Cabonate scaling prediction: The importance of valid data input,” in *NACE CORROSION*, pp. NACE–10132, NACE, 2010.

[8] N. Terry, “Existential challenges for healthcare data protection in the united states,” *Ethics, Medicine and Public Health*, vol. 3, no. 1, pp. 19–27, 2017.

[9] S. S. Bakare, A. O. Adeniyi, C. U. Akpuokwe, and N. E. Eneh, “Data privacy laws and compliance: a comparative review of the eu gdpr and usa regulations,” *Computer Science & IT Research Journal*, vol. 5, no. 3, pp. 528–543, 2024.

[10] C. J. McKinstry, “The hipaa privacy rule: flawed privacy exposed when compared with the european union’s general data protection regulation,” *Journal of Health Care Finance*, 2018.