Lung cancer detection using Machine Learning

Dipanwita Adhikary

Machine Learning Internship

at

Feynn Lab

17.03.2025

Step 1.

Prototype Development

Objective:

Develop a small-scale prototype model to validate the product idea of AI-driven lung cancer detection.

Approach:

1. **Data Collection**:
    - Utilize publicly available datasets like LIDC-IDRI, NLST, and SPIE-AAPM.
2. **Data Preprocessing**:
    - Normalize, resize, and remove noise from CT and X-ray images.
    - Annotate the data with help from radiologists.
3. **Model Selection**:
    - Begin with a basic Convolutional Neural Network (CNN) for initial classification tasks.
    - Implement simple object detection (like YOLO) for nodule detection.
4. **Model Training & Validation**:
    - Train the model on a subset of the data to detect and classify lung nodules.
    - Validate the model's performance using standard metrics like accuracy, sensitivity, and specificity.
5. **Prototype Testing**:
    - Test the prototype on a small set of unseen data to confirm model viability.

Step 2.

Developing a robust business model is crucial for the success of AI-driven lung cancer detection services. Drawing from the resources provided and industry practices, here is a comprehensive business model tailored in this project:

1. Value Proposition

- **Early Detection**: Facilitate timely identification of lung cancer, improving patient survival rates.
- **Enhanced Diagnostic Accuracy**: Utilize AI to reduce human error and subjectivity in medical imaging analysis.
- **Operational Efficiency**: Streamline radiology workflows, allowing healthcare professionals to focus on patient care.

2. Target Customer Segments

- **Healthcare Providers**: Hospitals, clinics, and diagnostic centers seeking to enhance diagnostic capabilities.
- **Medical Professionals**: Radiologists and oncologists aiming for accurate and swift diagnostic support.
- **Patients**: Individuals desiring accessible and reliable diagnostic services for early lung cancer detection.

3. Revenue Streams

- **Subscription Model**: Tiered subscription plans to healthcare institutions based on usage volume and feature access.
- **Per-Scan Fee**: Charge a fixed fee for each scan analyzed, suitable for smaller clinics with variable workloads.
- **Licensing**: License the AI technology to medical imaging equipment manufacturers for integration.

4. Cost Structure

- **Research and Development**: Continuous improvement of AI algorithms and software updates.
- **Compliance and Certification**: Ensure adherence to medical standards and obtain necessary certifications.
- **Marketing and Sales**: Promote the service to potential clients and maintain customer relationships.

- **Operational Expenses**: Costs related to cloud computing, data storage, and customer support.

5. Key Activities

- **AI Model Training**: Utilize diverse and extensive datasets to enhance diagnostic accuracy.
- **Clinical Validation**: Conduct studies to validate the AI's performance against established diagnostic methods.
- **Regulatory Approvals**: Navigate the regulatory landscape to ensure the service meets all legal requirements.
- **Partnership Development**: Collaborate with healthcare institutions for pilot programs and feedback.

6. Key Resources

- **Technical Team**: Data scientists, AI specialists, and software developers.
- **Medical Advisors**: Radiologists and oncologists providing domain expertise.
- **Computing Infrastructure**: High-performance servers and cloud services for data processing.

7. Key Partnerships

- **Healthcare Institutions**: Hospitals and clinics for data sharing and pilot testing.
- **Medical Device Manufacturers**: Integration of AI tools into imaging equipment.
- **Regulatory Bodies**: Ensure compliance with healthcare regulations and standards.

8. Customer Relationships

- **Training and Support**: Provide onboarding and continuous support to medical staff.
- **Feedback Loops**: Establish channels for user feedback to drive improvements.
- **Community Building**: Create forums or user groups for knowledge sharing among clients.

9. Channels

- **Direct Sales**: Engage with healthcare providers through a dedicated sales force.

- **Online Platform**: Offer a user-friendly interface for service access and management.
- **Medical Conferences**: Present at industry events to showcase the technology and network.

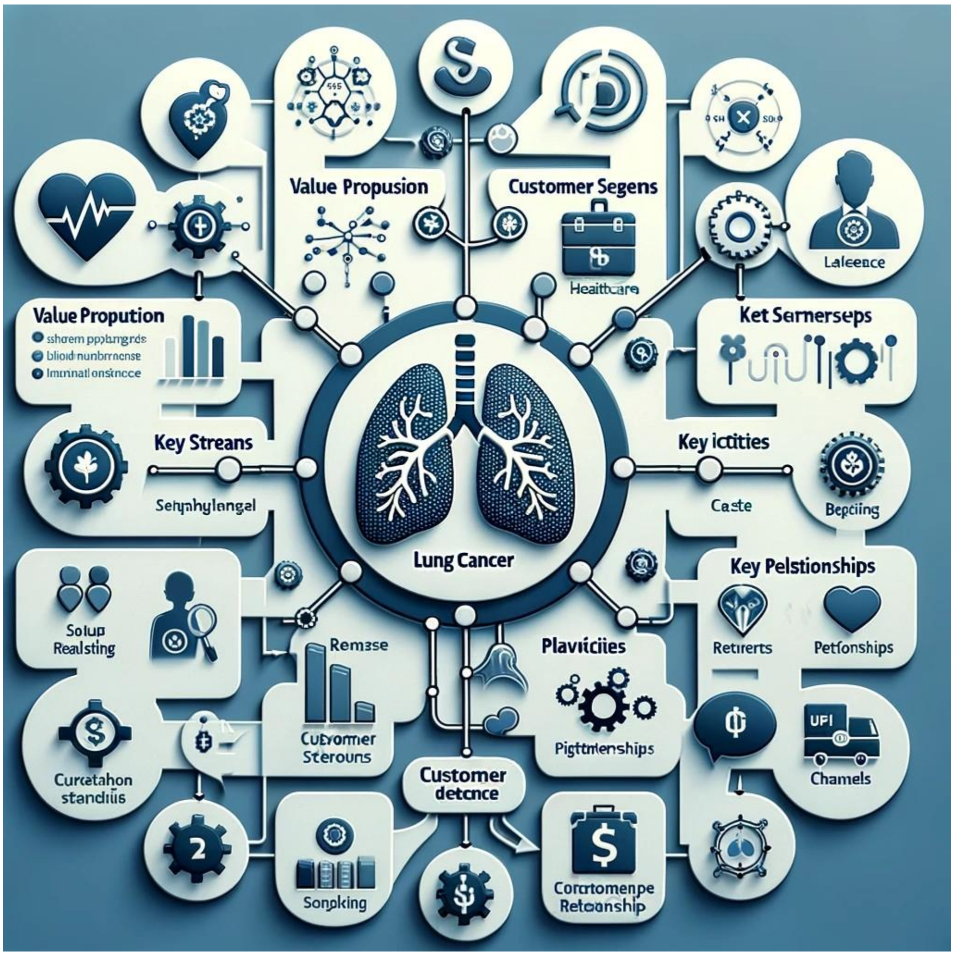

Fig. 1. Here is a visual representation of the business model for the AI-driven lung cancer detection project. The diagram illustrates key components such as Value Proposition, Customer Segments, Revenue Streams, and more

Step 3.

Financial Modelling with Machine Learning & Data Analysis

a. Market Identification

- The primary market is the Indian healthcare diagnostics sector, particularly radiology centers and hospitals.

b. Data Collection

- Collect online statistics on the number of lung cancer cases, the growth rate of diagnostic imaging centers, and the adoption of AI in healthcare.

c. Forecasting with Machine Learning

- Use historical data to perform time-series forecasting for market trends and potential sales growth.

d. Financial Equation Design

- **Assumptions**:
  - Product price per unit (service): ₹500.
  - Monthly operational cost: ₹2000.
  - Expected monthly sales: 300 units.
- **Revenue Equation**:

  y=500x−2000y = 500x - 2000y=500x−2000

  Where:

  - yyy = Total monthly revenue
  - xxx = Number of units (or services) sold

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA

# Load the dataset
data = pd.read_json('dataset.json')

# Convert 'LUNG_CANCER' to binary
data['LUNG_CANCER'] = data['LUNG_CANCER'].map({'YES': 1, 'NO': 0})

# Simulate monthly sales data over 12 months
diagnosed_cases = data[data['LUNG_CANCER'] == 1]
monthly_sales = diagnosed_cases.groupby(diagnosed_cases.index %
12).size()
monthly_sales.index = pd.date_range(start='2023-01-01', periods=12,
freq='M')

# Fit ARIMA model
model = ARIMA(monthly_sales, order=(1, 1, 1))
model_fit = model.fit()

# Forecast the next 12 months
forecast = model_fit.get_forecast(steps=12)
forecast_index = pd.date_range(start=monthly_sales.index[-1] +
pd.offsets.MonthBegin(), periods=12, freq='M')
forecast_values = forecast.predicted_mean

# Convert dates to numeric indices for plotting
historical_numeric_index = np.arange(len(monthly_sales))
forecast_numeric_index = np.arange(len(forecast_index))

# Plotting the historical and forecasted sales
plt.figure(figsize=(12, 6))
plt.plot(historical_numeric_index, monthly_sales.values,
label='Historical Sales', marker='o')
plt.plot(forecast_numeric_index + len(monthly_sales),
forecast_values.values, label='Forecasted Sales', color='green',
marker='o')
plt.title('Monthly Sales Forecast for Lung Cancer Diagnosis Service')
plt.xlabel('Month Index')
plt.ylabel('Number of Diagnosed Cases (Sales)')
plt.legend()
plt.grid(True)
plt.show()
```

```
C:\Users\Dipanwita\AppData\Local\Temp\ipykernel_9936\3328503582.py:15:
FutureWarning: 'M' is deprecated and will be removed in a future
version, please use 'ME' instead.
  monthly_sales.index = pd.date_range(start='2023-01-01', periods=12,
```

```
freq='M')
C:\Users\Dipanwita\AppData\Local\Temp\ipykernel_9936\3328503582.py:23:
FutureWarning: 'M' is deprecated and will be removed in a future
version, please use 'ME' instead.
  forecast_index = pd.date_range(start=monthly_sales.index[-1] +
pd.offsets.MonthBegin(), periods=12, freq='M')
```