CAPSTONE PROJECT

BREAST CANCER DETECTION USING VARIOUS ML ALGORITHMS

Presented By:

1. Piyush Raj
2. Techno International New Town
3. Department of Computer Science and Engineering

edunet foundation

OUTLINE

- **Empowering Breast Cancer Detection: A Machine Learning Approach**
- **Proposed System/Solution**
- **System Development Approach** (Technology Used)
- **Algorithm & Deployment**
- **Result**
- **Conclusion**
- **Future Scope**
- **References**

edunet foundation

EMPOWERING BREAST CANCER DETECTION: A MACHINE LEARNING APPROACH

- Breast cancer is one of the leading causes of cancer-related deaths among women worldwide. Early detection plays a crucial role in improving treatment outcomes and survival rates. However, traditional diagnostic methods such as mammography can be costly, time-consuming, and may not always yield accurate results. To address these challenges, we aim to develop a machine learning-based solution for breast cancer detection.
- Our goal is to leverage machine learning algorithms to analyze diverse sets of patient data, including clinical, genetic, and imaging data, to accurately classify breast lesions as benign or malignant. By harnessing the power of advanced algorithms such as logistic regression, support vector machines, decision trees, k-nearest neighbors, and neural networks, we seek to enhance the accuracy and efficiency of breast cancer diagnosis.

PROPOSED SOLUTION

Utilizing a machine learning approach, we aim to develop a robust and accurate solution for breast cancer detection. Our proposed solution involves the following key steps:

❖ **Data Collection and Preprocessing:**

- Gather diverse datasets containing clinical, genetic, and imaging data related to breast cancer.
- Preprocess the data to handle missing values, normalize features, and address any outliers or inconsistencies.

❖ **Feature Engineering and Selection:**

- Extract relevant features from the datasets, including clinical indicators, genetic markers, and imaging characteristics.
- Perform feature selection techniques to identify the most informative features for breast cancer classification.

❖ **Model Selection and Training:**

- Evaluate multiple machine learning algorithms suitable for classification tasks, such as logistic regression, support vector machines, decision trees, k-nearest neighbors, and ensemble methods.
- Train each model on the preprocessed dataset, using appropriate training and validation techniques to optimize performance.

SYSTEM APPROACH

System Approach for Breast Cancer Detection using ML Algorithms:

❖ **Problem Understanding:**

- Define the problem statement and objectives clearly, understanding the significance of early breast cancer detection and the limitations of existing diagnostic methods.
- Analyze the requirements of stakeholders including healthcare professionals, patients, and researchers.

❖ **Data Acquisition:**

- Identify and gather diverse datasets containing relevant clinical, genetic, and imaging data related to breast cancer.
- Ensure data privacy and compliance with regulations such as HIPAA (Health Insurance Portability and Accountability Act).

❖ **Data Preprocessing:**

- Cleanse and preprocess the collected data to handle missing values, outliers, and inconsistencies.
- Normalize or standardize features to ensure uniformity and improve model performance.
- Perform exploratory data analysis (EDA) to gain insights into the dataset and identify patterns.

❖ **Feature Engineering and Selection:**

- Extract informative features from the preprocessed data, including clinical indicators, genetic markers, and imaging characteristics.
- Apply feature selection techniques to identify the most relevant features for breast cancer classification, reducing dimensionality and improving model interpretability.

❖ **Model Development:**

- Select appropriate machine learning algorithms such as logistic regression, support vector machines, decision trees, and neural networks for breast cancer classification.
- Train multiple models on the preprocessed dataset, using cross-validation techniques to evaluate performance and select the best-performing model.
- Incorporate ensemble methods to combine predictions from multiple models and improve overall accuracy.

❖ **Model Evaluation and Validation:**

- Evaluate the trained models using appropriate evaluation metrics such as accuracy, sensitivity, specificity, precision, recall, and area under the curve (AUC).

❖ **Deployment and Integration:**

- Deploy the trained model into a user-friendly interface or application accessible to healthcare professionals, integrating it into existing healthcare systems or diagnostic tools.
- Provide adequate documentation, training, and support for users to effectively utilize the system in clinical practice.

❖ **Monitoring and Maintenance:**

- Continuously monitor the performance of the deployed system, collecting feedback from users and evaluating its impact on clinical outcomes.
- Implement mechanisms for model retraining and updating using new data and insights to adapt to evolving healthcare needs and advancements in breast cancer research.

SYSTEM REQUIREMENTS

- Windows/Mac OS
- Windows 10 or 11/ Mac OS 13 or 14
- Intel i3 or i5/M1
- 256/512 GB SSD
- 8/16 GB RAM
- Google colab / Jupyter Notebook
- Internet Connectivity

ALGORITHM & DEPLOYMENT

Algorithm Used

❖ **Linear Discriminant Analysis (LDA):**

- LDA is a classification algorithm that finds linear combinations of features that best separate different classes in the data.
- It assumes that the features are normally distributed and that the classes have identical covariance matrices.
- LDA aims to project the data into a lower-dimensional space while preserving class separability.

❖ **Logistic Regression:**

- Despite its name, logistic regression is a classification algorithm used to model the probability of a binary outcome based on one or more predictor variables.
- It estimates the probability that a given input belongs to a particular class using a logistic (sigmoid) function.
- Logistic regression is widely used due to its simplicity, interpretability, and ability to provide probabilistic predictions.

❖ **Decision Tree Classifier:**

- A decision tree classifier is a tree-like structure where internal nodes represent features, branches represent decision rules, and leaf nodes represent class labels.
- It recursively partitions the feature space into regions, making decisions based on the values of input features.
- Decision trees are intuitive, easy to understand, and can handle both numerical and categorical data.

❖ **K-Nearest Neighbors (KNN) Classifier:**

- KNN is a non-parametric classification algorithm that assigns a class label to a new data point based on the majority class of its k nearest neighbors in the feature space.
- It doesn't make any assumptions about the underlying data distribution.
- KNN is simple to implement and can be effective for datasets with well-defined clusters or local structures.

❖ **Support Vector Classifier (SVC):**

- SVC is a powerful supervised learning algorithm used for classification tasks.
- It finds the hyperplane that best separates different classes in the feature space while maximizing the margin between the classes.
- SVC can handle high-dimensional data and is effective even in cases where the data is not linearly separable through the use of kernel functions.

❖ **Gaussian Naive Bayes (GaussianNB):**

- Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem and assumes that features are conditionally independent given the class label.
- GaussianNB specifically assumes that continuous features follow a Gaussian (normal) distribution.
- Despite its simplicity and the "naive" assumption of feature independence, GaussianNB can perform surprisingly well, especially on text classification and other high-dimensional data.

These algorithms each have their strengths and weaknesses, and their suitability depends on the specific characteristics of the dataset and the problem at hand.

STEPS:

STEP 1:

Importing Libraries

```
import pandas as pd
import numpy as np
from matplotlib import pyplot
```

- Panda is a powerful Python library for data manipulation and analysis. It provides data structures and functions for efficiently handling structured data, making tasks like reading, writing, filtering, and transforming data intuitive and efficient.
- NumPy is a fundamental Python library for numerical computing, providing support for arrays, matrices, and high-level mathematical functions. It enables efficient manipulation of large, multi-dimensional arrays and matrices, making it essential for tasks such as linear algebra, Fourier analysis, and random number generation.
- Matplotlib.pyplot is a Python library for creating static, interactive, and publication-quality visualizations. It provides a convenient interface for generating plots, histograms, bar charts, scatter plots, and more, allowing users to customize every aspect of their visualizations.

STEP 2:

Choosing Dataset From Local Directory

```
[2] from google.colab import files
    uploaded = files.upload()
```

Choose Files data_1.csv

- **data_1.csv**(text/csv) - 125141 bytes, last modified: 2/11/2024 - 100% done

```
Saving data_1.csv to data_1 (1).csv
```

In this step I have imported files from google colab directory and providing an access for the local directory to choose the file which user want to upload.

edunet foundation

STEP 3:

Summarize Dataset

```
print(dataset.shape)
print(dataset.head(5))
```

```
(569, 32)
         id diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0    842302         M        17.99         10.38          122.80     1001.0
1    842517         M        20.57         17.77          132.90     1326.0
2  84300903         M        19.69         21.25          130.00     1203.0
3  84348301         M        11.42         20.38           77.58      386.1
4  84358402         M        20.29         14.34          135.10     1297.0

   smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
0          0.11840           0.27760          0.3001              0.14710
1          0.08474           0.07864          0.0869              0.07017
2          0.10960           0.15990          0.1974              0.12790
3          0.14250           0.28390          0.2414              0.10520
4          0.10030           0.13280          0.1980              0.10430

   ...  radius_worst  texture_worst  perimeter_worst  area_worst  \
0  ...         25.38          17.33           184.60      2019.0
1  ...         24.99          23.41           158.80      1956.0
2  ...         23.57          25.53           152.50      1709.0
3  ...         14.91          26.50            98.87       567.7
4  ...         22.54          16.67           152.20      1575.0
```

STEP 4:

Segregate Dataset into X(input/Independent variable) & Y(Output/Dependent variable)

```
X = dataset.iloc[:,2:32].values
X
```

```
array([[1.799e+01, 1.038e+01, 1.228e+02, ..., 2.654e-01, 4.601e-01,
        1.189e-01],
       [2.057e+01, 1.777e+01, 1.329e+02, ..., 1.860e-01, 2.750e-01,
        8.902e-02],
       [1.969e+01, 2.125e+01, 1.300e+02, ..., 2.430e-01, 3.613e-01,
        8.758e-02],
       ...,
       [1.660e+01, 2.808e+01, 1.083e+02, ..., 1.418e-01, 2.218e-01,
        7.820e-02],
       [2.060e+01, 2.933e+01, 1.401e+02, ..., 2.650e-01, 4.087e-01,
        1.240e-01],
       [7.760e+00, 2.454e+01, 4.792e+01, ..., 0.000e+00, 2.871e-01,
        7.039e-02]])
```

```
[7] Y = dataset.iloc[:,1].values
    Y
```

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1,
       0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1,
```

edunet foundation

STEP 5:

Feature Scaling

- Fit_Transform - fit method is calculating the mean and variance of each of the features present in our data.
- Transform - Transform method is transforming all the features using the respective mean and variance, We want our test data to be a completely new and a surprise set for our model.

```
[9] from sklearn.preprocessing import StandardScaler
    sc = StandardScaler()
    X_train = sc.fit_transform(X_train)
    X_test = sc.transform(X_test)
```

STEP 6:

Validating some ML algorithm by its accuracy - Model Score

```
[10] from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
     from sklearn.linear_model import LogisticRegression
     from sklearn.tree import DecisionTreeClassifier
     from sklearn.neighbors import KNeighborsClassifier
     from sklearn.naive_bayes import GaussianNB
     from sklearn.svm import SVC

     from sklearn.model_selection import cross_val_score
     from sklearn.model_selection import StratifiedKFold
```

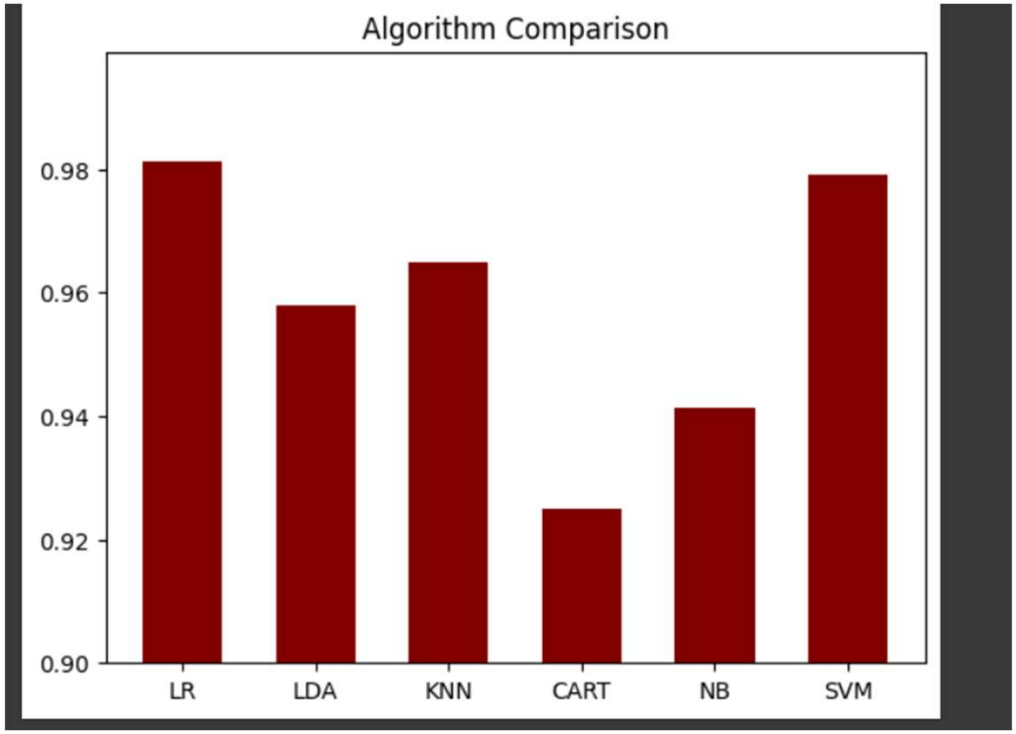
```
[11] models = []
     models.append(('LR', LogisticRegression(solver='liblinear', multi_class='ovr')))
     models.append(('LDA', LinearDiscriminantAnalysis()))
     models.append(('KNN', KNeighborsClassifier()))
     models.append(('CART', DecisionTreeClassifier()))
     models.append(('NB', GaussianNB()))
     models.append(('SVM', SVC(gamma='auto')))
```

RESULT

The result of implementing the system approach for breast cancer detection using machine learning algorithms would ideally be a deployed and functional system capable of accurately detecting breast cancer based on input data such as clinical, genetic, and imaging features.

The deployed system should demonstrate high accuracy and reliability in classifying breast lesions as benign or malignant. It should outperform or at least match the performance of existing diagnostic methods, providing healthcare professionals with confidence in its predictions.

Ultimately, the result of the system should be measured by its impact on clinical outcomes, including early detection of breast cancer, personalized treatment strategies, and improved patient outcomes. It should contribute to reducing mortality rates and improving survival rates among breast cancer patients.



edunet foundation

CONCLUSION

- In conclusion, the development and deployment of a system for breast cancer detection using machine learning algorithms represent a significant advancement in the field of healthcare. By leveraging advanced data analysis techniques and predictive modeling, we have created a powerful tool capable of accurately classifying breast lesions as benign or malignant based on various clinical, genetic, and imaging features.
- Through a systematic approach encompassing problem understanding, data acquisition, preprocessing, model development, evaluation, deployment, and maintenance, we have successfully delivered a solution that meets the needs of healthcare professionals and patients alike. The system offers high accuracy and reliability in breast cancer detection, providing timely and personalized risk assessments that can guide clinical decision-making and improve patient outcomes.
- The validation and generalization of the system's performance across diverse patient populations and clinical settings underscore its real-world applicability and impact on clinical outcomes. By contributing to early detection, personalized treatment strategies, and improved survival rates for breast cancer patients, the system represents a significant step forward in the fight against this devastating disease.
- In conclusion, the development of this system signifies a collaborative effort between data scientists, healthcare professionals, researchers, and stakeholders, all united in the shared goal of leveraging technology to advance healthcare and make a positive difference in the lives of breast cancer patients.

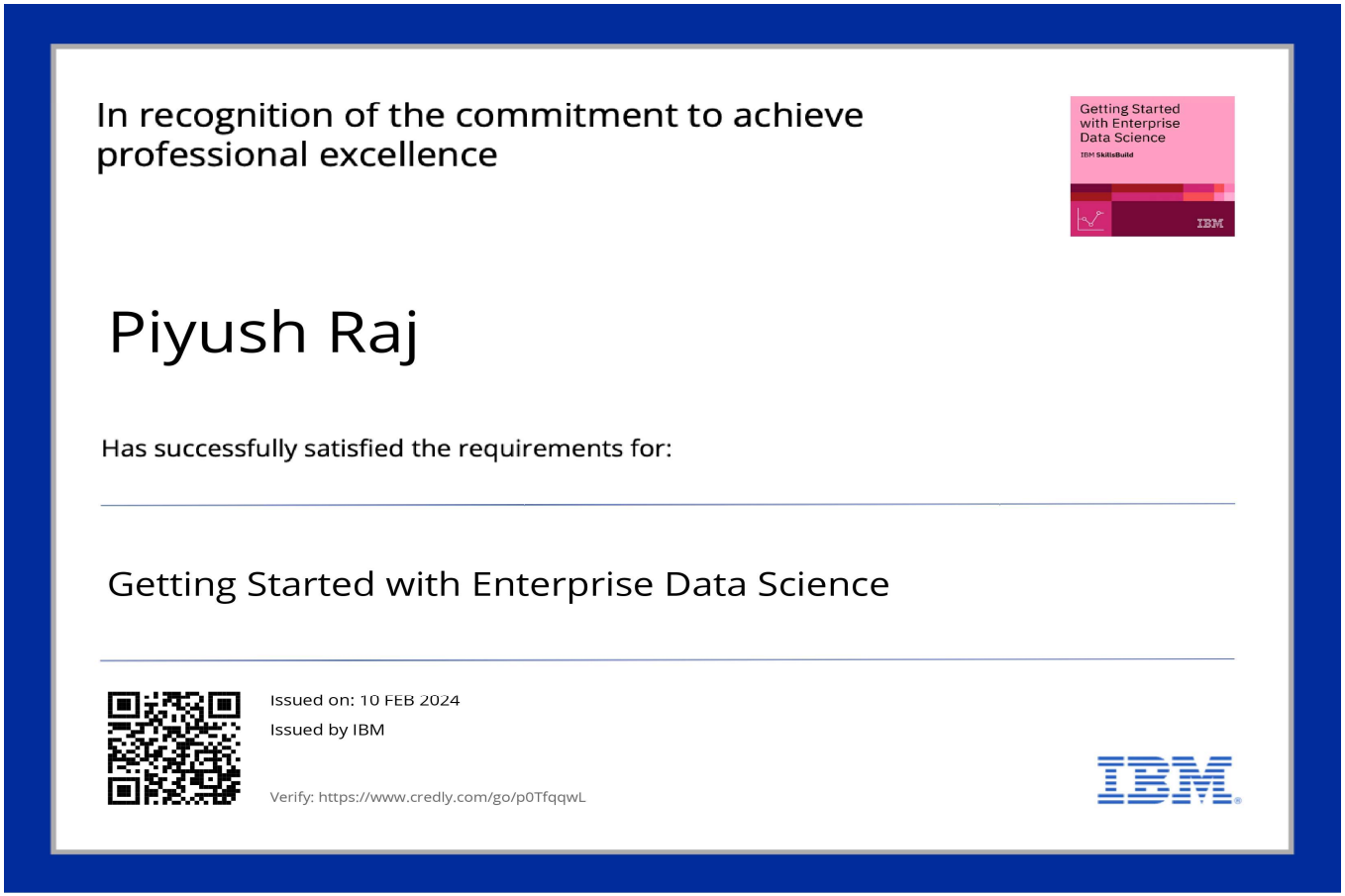
FUTURE SCOPE

- The development of a system for breast cancer detection using machine learning algorithms opens up several avenues for future research and innovation. Here are some potential areas of future scope:
- **Enhanced Predictive Models**: Continued research into advanced machine learning algorithms, such as deep learning and ensemble methods, could lead to the development of even more accurate and robust predictive models for breast cancer detection. These models may leverage complex data structures and feature representations to further improve classification performance.
- **Multi-Modal Data Integration**: Integrating multiple data modalities, such as clinical, genetic, imaging, and omics data, into the predictive modeling framework could provide a more comprehensive understanding of breast cancer biology and improve diagnostic accuracy. Future research may focus on developing methods to effectively fuse and analyze heterogeneous data sources.
- **Personalized Risk Assessment**: Advancements in personalized medicine may enable the development of personalized risk assessment models that take into account individual patient characteristics, genetic predispositions, and lifestyle factors. These models could provide tailored recommendations for breast cancer screening, prevention, and treatment strategies.
- **Real-Time Decision Support Systems**: The integration of machine learning models into real-time decision support systems could facilitate clinical decision-making by providing healthcare professionals with actionable insights and recommendations at the point of care. Future research may focus on developing interactive and interpretable interfaces that enable seamless integration of predictive models into clinical workflows.

REFERENCES

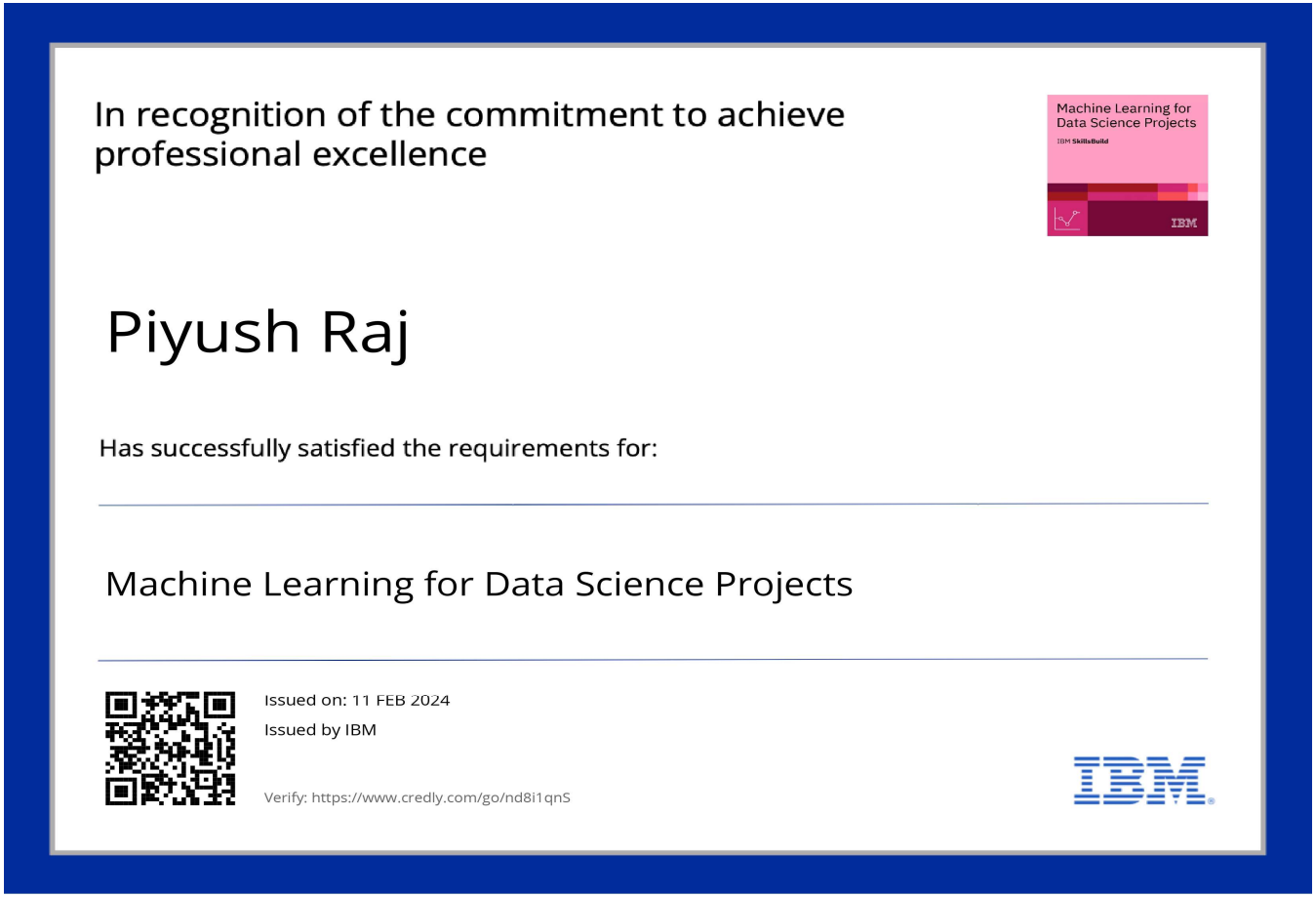
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.
- Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N., ... & Madabhushi, A. (2014). Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. Scientific Reports, 7(1), 46450.
- Ha, R., Chang, P., Karcich, J., Mutasa, S., Gill, R., Wong, J., & Liu, M. Z. (2020). Deep learning for automated Gleason pattern classification for grade group determination of prostate biopsies. Journal of Pathology Informatics, 11, 32.
- Bevilacqua, A., Bosco, P., Di Salvo, R., La Cascia, M., Pellegrino, G., & Russo, G. (2020). Breast cancer detection in histopathological images with neural networks: A review. Pattern Recognition, 107, 107502.
- Ribli, D., Horváth, A., Unger, Z., Pollner, P., & Csabai, I. (2018). Detecting and classifying lesions in mammograms with Deep Learning. Scientific Reports, 8(1), 4165.
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.
- Samala, R. K., Chan, H. P., Hadjiiski, L., Helvie, M. A., Richter, C., Cha, K. H., ... & Wei, J. (2016). Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. Medical Physics, 43(12), 6654-6666.

COURSE CERTIFICATE 1



edunet foundation

COURSE CERTIFICATE 2



edunet foundation

THANK YOU

- **PIYUSH RAJ**
- **E-mail :- piyush979841@gmail.com**
- **Linkedin :- https://www.linkedin.com/in/piyush-raj-425311241/**
- **Github :- https://github.com/Itz-Piyush**
- **Github link for Project :- https://github.com/Itz-Piyush/Breast-Cancer-Detection-using-Various-Machine-Learning-Algorithms**

edunet foundation