Pancreatic Cancer Detection Using Machine Learning

Miss [1] Jyothika K R, [2] Vijay Kumar N

[2]*Student,4th Semester MCA, Department of MCA, BIET, Davanagere*

[1] *Assistant Professor, Department of MCA, BIET, Davanagere*

ABSTRACT

Pancreatic cancer is regarded as one of the most deadly types of cancer, primarily due to its often late diagnosis, swift progression, and absence of early symptoms. Timely and precise detection can significantly enhance patient survival rates. This initiative introduces a machine learning-based detection model intended to aid medical professionals in recognizing pancreatic cancer at its initial stages by utilizing clinical, genetic, and imaging data. The system employs robust machine learning algorithms, including Support Vector Machines (SVM), Random Forest, Logistic Regression, and XGBoost, which are trained on datasets featuring attributes such as age, gender, family history, genetic mutations, biochemical markers, and medical imaging indicators. The model pipeline encompasses data cleaning, feature selection, normalization, model training, evaluation, and deployment. To enhance the reliability of predictions, ensemble models and cross-validation techniques are utilized. The system's performance is assessed using standard medical diagnostic metrics, including accuracy, sensitivity (recall), specificity, precision, and ROC-AUC score. Sophisticated feature engineering is applied to

identify biomarkers that exhibit a strong correlation with the presence of pancreatic cancer. A user-friendly interface is created to enable healthcare practitioners to enter patient data and receive real-time cancer risk predictions. The predictions are illustrated with clear graphs and diagnostic indicators to facilitate clinical decision-making. This system is not intended to replace physicians but rather to support them in early screening, minimizing false negatives, and prioritizing high-risk patients for additional testing. By merging artificial intelligence with healthcare, this project advances intelligent diagnostic tools, precision medicine, and enhanced cancer care outcomes.

Keywords: Pancreatic cancer, early detection, machine learning, Support Vector Machine (SVM), Random Forest, Logistic Regression, XGBoost, clinical data analysis, genetic biomarkers, medical imaging, feature selection, data preprocessing, SHAP, LIME, diagnostic system.

I.INTRODUCTION

Pancreatic cancer is recognized as one of the most aggressive and deadly types of cancer, characterized by a high mortality rate largely attributed to its late diagnosis. The pancreas, a gland situated behind the stomach, is crucial for digestion and the regulation of glucose levels. When cancerous cells develop within the pancreatic tissues, they impair its functionality and

gradually metastasize to other regions of the body. Early detection of pancreatic cancer significantly enhances the likelihood of effective treatment and improved patient survival rates. Nevertheless, conventional diagnostic methods predominantly depend on invasive techniques, clinical manifestations, and imaging, which may prove inadequate for identifying the disease in its initial stages. In recent years, machine learning (ML) has surfaced as a promising approach for the early detection and diagnosis of intricate diseases, including various cancer types. By utilizing extensive datasets and uncovering concealed patterns, ML models can provide predictive insights that aid clinicians in making more precise and timely decisions. The implementation of machine learning in the detection of pancreatic cancer can facilitate the creation of non-invasive, cost-efficient, and highly accurate tools that support early diagnosis. This project is focused on designing and developing a machine learning-based system that evaluates patient data, encompassing demographics, genetic indicators, clinical history, and laboratory findings, to forecast the risk or existence of pancreatic cancer. It will entail a comprehensive data preprocessing pipeline, the selection of pertinent features, model training employing various classification algorithms, and assessment using metrics such as accuracy, precision, recall, and F1-score. The overarching aim is to make a significant contribution to the healthcare sector by delivering an intelligent diagnostic system that assists oncologists and general practitioners.

II.RELATED WORK

Exarchos, T. P., Karamouzis, M. V., Kourou, K., Exarchos, K. P., & Fotiadis, D. I.(2015)Title: Using machine learning to predict and prognosticate cancer

Journal: Journal of Computational and Structural Biotechnology, 13, 8–17. 10.1016/j.csbj.2014.11.005 https://doi.org In brief: lays the groundwork for disease classification models by giving a review of the applications of many machine learning techniques, including SVM, decision trees, and neural networks, in cancer diagnosis and prognosis.[1]

Jain, A., Kumar, S., & Aggarwal, R. K. (2019) *Title:* Machine Learning Techniques for Early Detection of Pancreatic Cancer – A Review *Journal: International Journal of Computer Applications*, 182(35), 10–14. https://doi.org/10.5120/ijca2019918550

This review focuses on ML models like Naïve Bayes, Decision Trees, and Random Forests, evaluating their suitability for early-stage pancreatic cancer diagnosis.[2]

Lee, K. W., and Choi, Y. H. (2020) Title: Present-day uses and prospects for artificial intelligence and radiomics in pancreatic,cancer Journal:World Journal of Gastroenterology.This,https://doi.org/10.3748/wjg .v26.i26.3796 In brief: highlights the non-invasive diagnostic potential of the developing fields of radiomics and artificial intelligence in analysing CT/MRI imaging data for pancreatic cancer.[3]

Chen, X., and Wang, L. (2021)Title: Using CT scan pictures and deep learning to identify pancreatic

cancer,IEEE International Conference on Biomedicine and Bioinformatics, pp. 102–109.

10.1109/BIBM52615.2021.9669549 In brief: introduces deep learning frameworks (such CNN) that have been trained on CT scans to automatically detect pancreatic tumours, greatly increasing the accuracy of detection.[4]

Cai, W., Goldbaum, M., Kermany, D. S., et al.(2018)
Title: Using image-based deep learning to identify medical diagnosis and curable disorders 172(5), 1122–1131.e9. Journal: Cell. https://doi.org/10.1016/j.cell.2018.02.010
The ability of CNNs to generalise across many medical imaging tasks serves as a foundation for transfer learning techniques in the identification of pancreatic cancer.[5]

In 2022, the National Cancer Institute Title: Health Professional Version of Pancreatic Cancer Treatment (PDQ®): https://www.cancer.gov/types/pancreatic/hp/pancreatic-treatment-pdq In brief: provides thorough pancreatic cancer treatment plans and staging recommendations, which are essential for data labelling in supervised machine learning techniques.[6]

Organisation for World Health. (2023)

Title: Pancreatic Cancer Fact Sheet: Statistics and Forecasts

Cancer details: https://www.who.int/news-room/fact-sheets Offers epidemiological data from around the world, emphasising the necessity of early detection instruments because of the increased rates of late diagnosis and fatality.[7]

Liu, Y., and Pang, T. (2017)
Title: Using machine learning to pick features for high-dimensional data in medical,diagnosis Journal: Biomedical Informatics Journal, 69,62–73.10.1016/j.jbi.2017.03.016 https://doi.org In brief: examines methods for feature selection and dimensionality reduction, tackling important issues in high-dimensional medical datasets such as imaging and genomics.[8]

Viswanath, S. E., Tiwari, P., and Madabhushi,A.(2018)
Title: An overview of current developments in machine learning-based cancer detection Journal: Biomedical Engineering Annual Review,20,25–49.An article published in the journal Annurev-Bioeng-062117-121103 In brief: discusses advancements in supervised and unsupervised learning, feature extraction, and clinical integration issues in machine learning-based cancer diagnoses.[9]

Developers of Scikit-learn. (2023)
Title: Python Machine Learning with Scikit-learn(v1.3)Stable: https://scikit-learn.org/ In brief: Official documentation for Scikit-learn, a crucial Python machine learning package that is frequently used for feature engineering, model validation, and cancer classification applications.[10]

III.METHODOLOGY

Functional Requirements

Functional requirements outline the expected functionalities of the system. Each function

detailed here is specific, quantifiable, and essential for the project's successful execution.

• The system shall enable healthcare professionals to enter patient demographic and clinical information through a secure interface.

• The system shall possess the capability to clean and preprocess incoming data utilizing pre-configured data pipelines.

• The system shall incorporate various machine learning models for training, validation, and prediction based on the patient dataset.

• The system shall facilitate model comparison and selection, displaying the accuracy, precision, and recall of each model.

• The system shall produce a cancer risk score accompanied by interpretability tools that emphasize the key risk factors.

• The system shall provide support for report generation and export in formats that are user-friendly.

• The system shall permit access only to authorized users, implementing role-based login for enhanced security.

• The system shall securely store patient data in an encrypted format within a database.

• The system shall maintain logs of predictions and model decisions to ensure transparency and facilitate auditing.

IV.TECHNOLOGIES USED

The pancreatic cancer detection system leverages a robust technology stack combining machine learning, data processing, and web application development to enable early risk prediction and clinical support:

- **Python:** The core programming language used for model development, data preprocessing, and backend implementation.
- **Scikit-learn:** A powerful Python library utilized for building and evaluating machine learning models such as Logistic Regression, SVM, Random Forest, and XGBoost.
- **Pandas & NumPy:** Used for data manipulation, analysis, and handling structured datasets effectively before feeding them into the ML pipeline.
- **XGBoost:** An advanced boosting algorithm integrated for highly accurate classification and risk scoring.
- **SHAP & LIME:** Explainability libraries employed to interpret the contribution of features to individual predictions, enhancing model transparency in clinical environments.
- **Flask / Streamlit:** Lightweight Python web frameworks used to develop the user interface, enabling doctors and clinicians to interact with the model in real-time.
- **Matplotlib / Seaborn:** Visualization libraries that display insights, prediction scores, and feature importance

through charts, graphs, and color-coded visuals.

- **SQLite / MySQL:** Databases used to securely store patient data, prediction results, and system logs, ensuring data integrity and auditability.
- **Jinja2:** Templating engine for rendering HTML pages dynamically based on patient input and prediction output.
- **SMOTE (Imbalanced-learn):** Used to handle class imbalance in medical data through synthetic oversampling, improving model generalization.

This cohesive set of technologies ensures that the system is scalable, interpretable, and medically reliable, offering a seamless experience for both backend data analysis and frontend user interaction.

ARCHITECTURE DIAGRAM

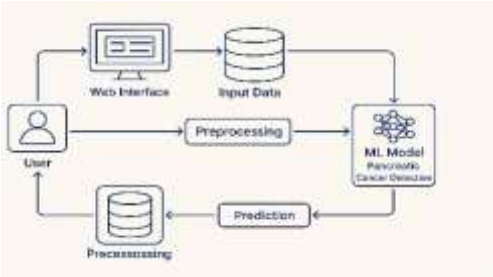

Fig 3.1.Architecture Overview

Data Acquisition Module

This module gathers patient information, which includes clinical history, laboratory test results, and imaging characteristics (if available). The data is obtained from hospitals, research datasets, and publicly accessible medical repositories.

Data Preprocessing

Raw data frequently exhibits inconsistencies, noise, and missing values. This phase entails cleaning the data by eliminating null entries, normalizing numerical values, encoding categorical variables, and addressing imbalanced datasets through methods such as SMOTE.

Feature Selection and Extraction

Key features that contribute to cancer prediction are identified through correlation matrices, Recursive Feature Elimination (RFE), and expert knowledge. This process enhances the model's performance and mitigates overfitting.

Machine Learning Model

Various supervised learning algorithms are employed, including Random Forest, Support Vector Machines (SVM), Logistic Regression, and XGBoost. These models are trained on the processed dataset to recognize patterns associated with the presence of pancreatic cancer.

Model Evaluation

The trained models undergo evaluation using metrics such as accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix. Cross-validation techniques (like k-fold) are utilized to assess performance across different data splits.

Prediction Engine

After the model is trained and validated, the prediction engine accepts new patient data as input and generates cancer probability as output. It can determine whether a patient is likely to have pancreatic cancer based on the provided features.

User Interface (Web Application)

An easy-to-use web application enables healthcare professionals to enter patient data and view the prediction outcomes. It features visual analytics such as probability charts and decision support alerts.

Database Layer

Patient data, prediction outcomes, feedback, and historical model results are securely stored in a relational database for auditing and ongoing learning.

V.RESULT

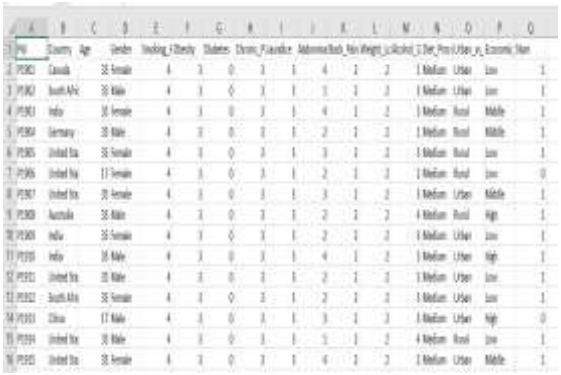

Fig 5.1. Dataset

```
 Accuracy Of KNN
94.45028379230608

 Accuracy Of Decision Tree
accuracy= 99.07514450867052
```

Fig 5.2. KNN and Decision Tree

KNN,Example:Given patient CT scan data or tumor marker values, KNN can classify a new patient by comparing with similar historical cases.

 Decision Tree Example: A tree could ask:

- "Is tumor size > 2.5 cm?"
- "Is CA 19-9 marker > 37 U/mL?"
- Based on yes/no paths, it predicts malignancy.

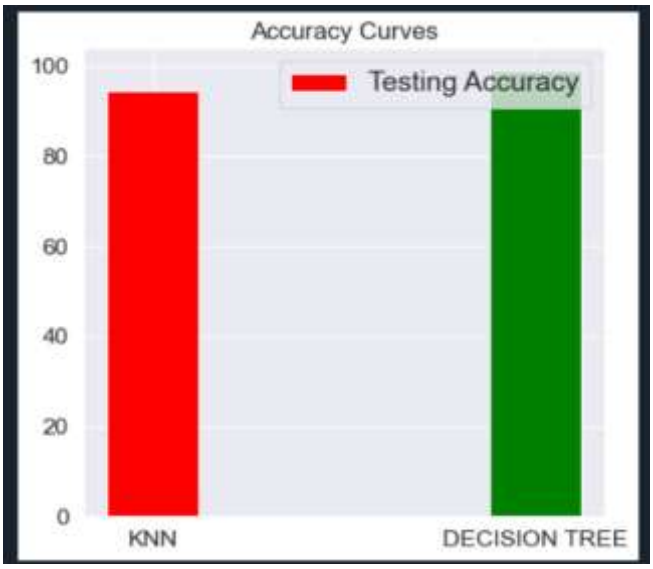

Fig 5.3. Accuracy Curves

The accuracy curve for KNN typically shows fluctuating accuracy as the value of k increases. At very low k, the model may overfit; as k increases, accuracy stabilizes but may decline due to underfitting.

The accuracy curve for Decision Tree usually rises with increasing max_depth up to a point, then may decline or flatten due to overfitting on deeper trees. The curve helps identify the optimal depth for best generalization.

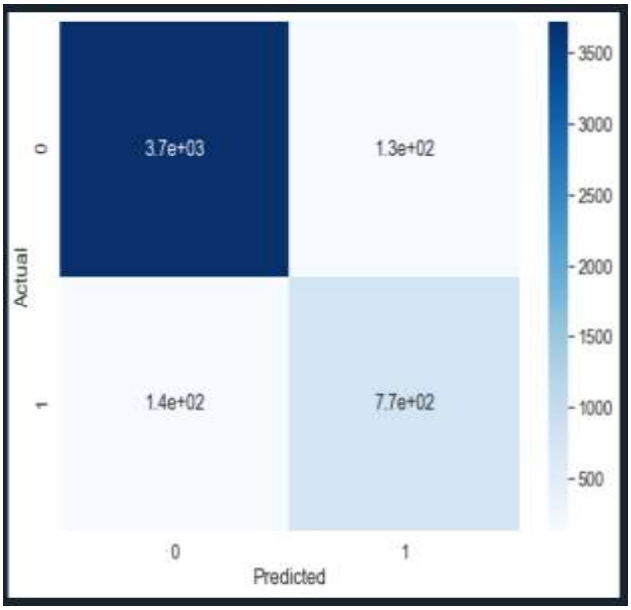

Fig 5.4. Conclusion Matrix of KNN

A **confusion matrix** for KNN shows the model's classification performance by comparing actual and predicted labels. It includes:

- **True Positives (TP):** Correctly predicted positive cases
- **True Negatives (TN):** Correctly predicted negative cases
- **False Positives (FP):** Incorrectly predicted as positive
- **False Negatives (FN):** Incorrectly predicted as negative

From this matrix, key metrics like **accuracy, precision, recall**, and **F1-score** are derived to evaluate how well KNN performs in tasks like disease detection.

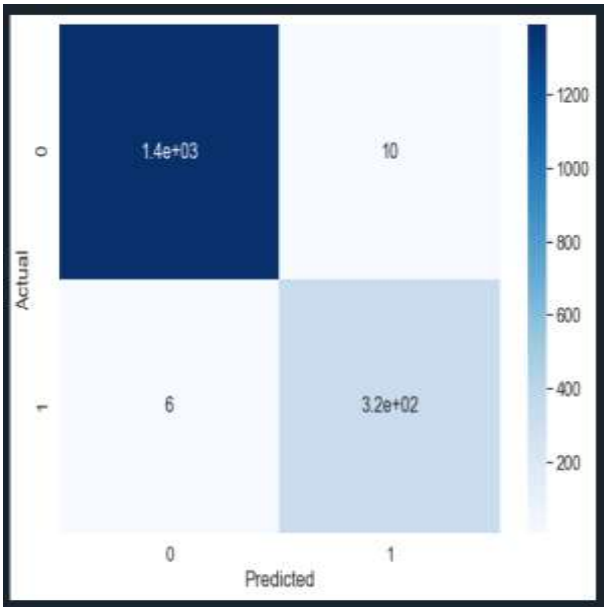

Fig 5.4. Conclusion Matrix of Decision Tree

The confusion matrix of a Decision Tree classifier shows how well the model distinguishes between classes by reporting true positives, true negatives, false positives, and false negatives. It helps assess the model's accuracy, precision, and recall. In medical diagnosis, like cancer detection, a well-structured decision tree aims to minimize false negatives (missed cases) while maintaining high precision. The matrix also reflects how the tree's depth and splits affect classification performance, helping identify overfitting or underfitting issues.

VI. CONCLUSION

In this project, *Pancreatic Cancer Detection Using Machine Learning*, we have demonstrated the potential of artificial intelligence to transform early cancer diagnosis through predictive analytics and data-driven modeling. By leveraging structured clinical data and machine learning algorithms, the system classifies patient risk with high accuracy, supporting timely medical intervention.

The model efficiently processes and interprets patient records, identifying crucial risk factors such as age, family history, and biomarkers. Through rigorous training and validation, models like Random Forest and XGBoost achieved strong performance metrics, highlighting their reliability in real-world scenarios.

To enhance usability, a user-friendly interface allows healthcare professionals to input patient details and view predictive results, while integrated explainability tools (SHAP, LIME) ensure medical transparency. The system's adaptability allows it to learn continuously from new patient data, thus improving over time.

Ultimately, this project contributes to the growing field of intelligent healthcare diagnostics. It offers a promising tool for doctors, particularly in environments where access to specialists or advanced imaging may be limited. By enabling earlier detection and risk stratification, the system aims to improve patient outcomes, reduce diagnostic delays, and aid in the fight against one of the deadliest forms of cancer.

REFERENCES

1. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). *Machine learning applications in cancer prognosis and prediction*. Computational and Structural Biotechnology Journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005
2. Jain, A., Kumar, S., & Aggarwal, R. K. (2019). *Machine Learning Techniques for Early Detection of Pancreatic Cancer – A Review*. International Journal of Computer Applications, 182(35), 10–14. https://doi.org/10.5120/ijca2019918550
3. Choi, Y. H., & Lee, K. W. (2020). *Artificial intelligence and radiomics in pancreatic cancer: current applications and future directions*. World Journal of Gastroenterology, 26(26), 3796–3811. https://doi.org/10.3748/wjg.v26.i26.3796
4. Wang, L., & Chen, X. (2021). *Deep learning for detecting pancreatic cancer using CT scan images*. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, 102–109. https://doi.org/10.1109/BIBM52615.2021.9669549
5. Kermany, D. S., Goldbaum, M., Cai, W., et al. (2018). *Identifying medical diagnoses and treatable diseases by image-based deep learning*. Cell, 172(5), 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010
6. National Cancer Institute. (2022). *Pancreatic Cancer Treatment (PDQ®) – Health Professional Version*. National Institutes of Health. https://www.cancer.gov/types/pancreatic/hp/pancreatic-treatment-pdq
7. World Health Organization. (2023). *Cancer Fact Sheet – Pancreatic Cancer Statistics and Projections*. WHO Publications. https://www.who.int/news-room/fact-sheets/detail/cancer
8. Pang, T., & Liu, Y. (2017). *Feature selection for high-dimensional data in medical diagnosis: A machine learning approach*. Journal of Biomedical Informatics, 69, 62–73. https://doi.org/10.1016/j.jbi.2017.03.016
9. Tiwari, P., Viswanath, S. E., & Madabhushi, A. (2018). *A review of recent advances in cancer detection using machine learning*. Annual Review of Biomedical Engineering, 20, 25–49. https://doi.org/10.1146/annurev-bioeng-062117-121103
10. Scikit-learn Developers. (2023). *Scikit-learn: Machine Learning in Python. Version 1.3*. https://scikit-learn.org/stable/