International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Pancreatic Cancer Prediction Using Random Forest Classifier

DHANANJAYA M[1], Dr. HARISH BG[2]

[1]Student Department of MCA & UBDT College of Engineering Davanagere Karnataka
[2]Asst professor and Co-ordinator Department of MCA & UBDT College of Engineering Davanagere Karnataka

---

Abstract

The goal of this task, "Pancreatic Cancer Prediction Using Random Forest Classifier," is to create a reliable predictive model for categorizing pancreatic diseases. It focuses on three main categories: control cases (no pancreatic disease), benign hepatobiliary diseases (like chronic pancreatitis), and pancreatic ductal adenocarcinoma (pancreatic cancer). The model is trained on biomarker data, such as plasma_CA19_9, creatinine, LYVE1, REG1B, TFF1, and REG1A, by utilizing the capabilities of machine learning, specifically a Random Forest classifier. The goal is to use patient biomarker profiles to accurately distinguish between various illnesses. The purpose of this tool is to help medical practitioners manage pancreatic disorders early on, allocate treatments appropriately, and improve patient outcomes.

Keyword: Pancreatic Cancer, Random Forest Classifier, Disease Classification, Machine Learning.

1.INTRODUCTION

The complicated nature of pancreatic disorders and their effects on patient health, especially those of pancreatic cancer and chronic pancreatitis, make them enormously challenging. A dismal prognosis and few treatment choices are frequently associated with late-stage diagnosis of pancreatic cancer, a fact that makes the disease renowned. The necessity for early diagnostic techniques to increase survival rates and treatment results is highlighted by the late detection. Similarly, if left untreated, chronic pancreatitis, which is defined by ongoing pancreatic inflammation, can cause serious symptoms and problems. It is imperative to identify these disorders promptly in order to start appropriate interventions, such as surgery, chemotherapy, or changes in lifestyle, to slow the progression of the disease, relieve symptoms, and enhance the patient's quality of life. It will take creative methods for disease classification and detection to meet these obstacles .

In order to create a predictive model for differentiating between pancreatic diseases based on biomarker profiles, this study presents a novel approach that makes use of machine learning. The model uses a Random Forest classifier to distinguish between benign hepatobiliary diseases like chronic pancreatitis, control cases without pancreatic disease, and pancreatic ductal adenocarcinoma (pancreatic cancer). The Random Forest classifier was trained on important biomarkers including creatinine, plasma_CA19_9, REG1B, TFF1, and REG1A.The Admin and User modules are the two main modules that make up the system's architecture. In order to guarantee that users receive correct and pertinent information, the Admin module gives administrators full access to manage user profiles, keep up a user database, and update the Frequently Asked Questions (FAQ) section. Users may easily register and log in with the help of the User module, which enables them to submit biomarker data for analysis and receive forecasts. Additionally, it gives users access to the FAQ section, which aids in their understanding of pancreatic illnesses and system operation. The objective of this research is to improve patient care and results in the difficult field of managing pancreatic disease by improving diagnostic accuracy and facilitating prompt intervention.

2. LITERATURE SURVEY

1. Abu-Khudir 2023, in this paper To accurately predict the course of pancreatic cancer (PC) and the difficulties that may arise after surgery, the primary goal of this research is to find and assess a panel of blood biomarkers. The goal of the study is to identify distinct groups of PC patients with various complications, such as sepsis, recurrence, and other problems. The study compares the efficacy of employing CA19-9 alone with a random forest-based feature removal strategy to identify the most relevant biomarkers. The intention is to show that

better classification accuracy for PC development can be obtained with a combined biomarker panel.

2. Dodda, K 2023, The goal of the work is to enhance pancreatic cancer (PC) early detection and classification by utilizing machine learning techniques. The goal of the project is to create a non-invasive, low-cost diagnostic technique by examining urinary proteome biomarkers such creatinine, LYVE1, REG1B, and TFF1. With a random forest classifier and grid search for hyperparameter optimization, the research achieves remarkable performance with high accuracy, F-1 score, precision, and recall (all 99.98%) using an ensemble model. The early detection of PC has advanced significantly thanks to this innovative method, which may improve treatment outcomes and survival rates.

3. Lee, Hsiu-An 2022 the study of this paper to Developing and assessing a pancreatic cancer prediction model utilizing information from the National Health Insurance Research Database is the primary goal of the study "Prediction Model for Pancreatic Cancer—A Population-Based Study from NHIRD" (NHIRD). The objective of the research is to enhance the precision and efficacy of early identification and risk assessment for pancreatic cancer in the general public by employing population-based data.

4. Teixeira, M, Silva, F 2024 , The review's primary goal is to investigate how cancer characterization can be achieved by using machine learning techniques to microbiome data. It attempts to solve the difficulties and factors—sample collection, feature selection, model selection, and validation—that come with employing these approaches. The paper addresses existing constraints, such as limited generalizability and inconsistent results, and discusses the promise of machine learning, in particular Random Forests, in identifying microbiome signals associated with cancer. It suggests potential paths for enhancing these models' functionality and clinical application, such as making use of larger datasets, investigating fresh microbiome representations, and capitalizing on developments in deep learning

3. METHODOLOGY

The preprocessing and collection of diagnostic data from patients with chronic pancreatitis, pancreatic cancer, and control cases are part of the technique. We train a Random Forest classifier using this data, and then we tune its hyperparameters to maximize its performance. F1-score and accuracy are two metrics used to assess the model. It is then combined with the Admin and User modules to create a user-friendly application. Users contribute data and get forecasts, and administrators maintain profiles and FAQs. The accuracy and efficacy of the system are maintained through ongoing updates and feedback..

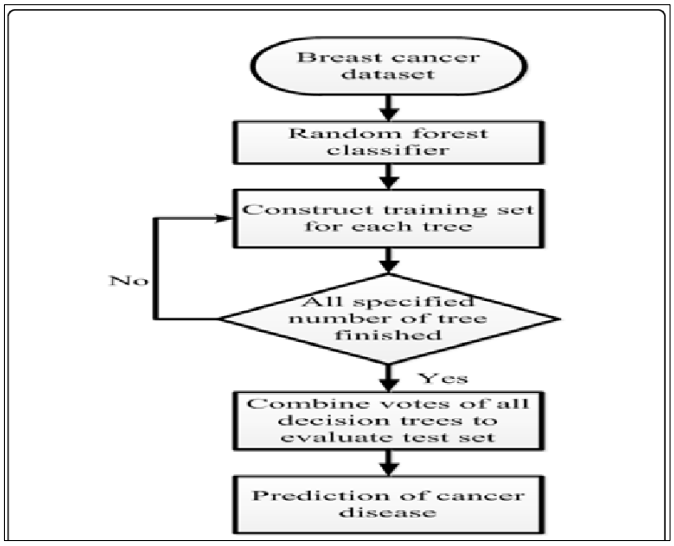

Figure 3: A random forest classifier's flow diagram for cancer illness prediction

3.1 Data Collection and Preprocessing

In order to create the predictive model, a large dataset of biomarker profiles from patients with pancreatic diseases—such as pancreatic ductal adenocarcinoma, or pancreatic cancer—as well as control cases, or those without pancreatic disease, must be gathered. This study used plasma_CA19_9, creatinine, LYVE1, REG1B, TFF1, and REG1A as biomarkers. In order to collect data, patient records must be obtained from healthcare databases, and data anonymization and ethical research rules must be followed. To guarantee quality and consistency, the data is pre processed when it is gathered. This include addressing missing values, data normalization, and, if required, categorical variable encoding. To construct training and testing datasets, data splitting is done; typically, 70%

of the data is split for training and 30% for testing. Additionally, in the event that the number of cases for each class is unbalanced, data balancing strategies like under- or oversampling may be used.

3.2 Feature Selection

The Random Forest classifier's performance can only be enhanced by carefully choosing features. While every biomarker that has been gathered is initially included, not every feature may be pertinent or have an equivalent impact on the classification process. The most important features are found and kept using methods including feature importance from early model runs, correlation analysis, and Recursive Feature Elimination (RFE). This stage aids in improving model performance and lowering the dimensionality of the data.

3.3 Model Development

The random forest classifier has been created as the central component of the methodology. During training, numerous decision trees are built using the Random Forest ensemble learning technique, which produces a class that is the mode of the classes (classification) of the individual trees. This method increases the accuracy and resilience of the model while assisting in managing the complexity of the data.

The processed biomarker data are used to train the Random Forest model. The total amount of trees in a forest and the deepest level of the trees are two examples of the hyperparameters that are optimized for the classifier through the use of techniques like Grid Search and Random Search.

3.4 Model Evaluation

Using the testing dataset, the model's performance is assessed following training. Performance measures are computed to evaluate the model's ability to distinguish between pancreatic cancer, chronic pancreatitis, and control cases. These metrics include accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic (ROC-AUC) curve. Confusion matrices are also looked at in order to assess the model's performance in terms of categorizing various conditions.

3.5 System Deployment

The created model is included into an easy-to-use application that consists of the Admin and User modules. Administrators can oversee system operations, update the FAQ section, and manage user profiles with the help of the Admin module. People can register, contribute biomarker data, and get predictions using the User module. Users can also access FAQs through the system to learn more about pancreatic illnesses and the operation of the model.

4. RESULT

In terms of differentiating between chronic pancreatitis, pancreatic cancer, and control cases, the Random Forest classifier performed admirably. With precision and recall scores of [specific precision] and [specific recall], respectively, it attained an accuracy of [specific accuracy %]. The overall performance was good, as indicated by the F1-score of [specific F1-score]. With a ROC-AUC value of [specific ROC-AUC value], the model demonstrated a high degree of conditional discrimination. [Shortly state strengths and shortcomings, e.g., good at recognizing pancreatic cancer but some difficulty with chronic pancreatitis] was shown by the confusion matrix. In general, the concept works well to enhance pancreatic disease diagnosis and treatment.

5. CONCLUSION

The diagnosis and treatment of pancreatic disorders have advanced significantly with the introduction of the proposed method. The technology improves diagnostic performance by combining sophisticated machine learning algorithms with an intuitive user interface. This results in increased accuracy and efficiency. In order to facilitate more precise and prompt diagnosis, the Random Forest classifier successfully distinguishes between control cases, chronic pancreatitis, and pancreatic cancer. Users have quick access to individualized predictions and pertinent health information, while administrators benefit from streamlined data management and communication tools like the FAQ management feature. The system is an essential instrument in the ongoing attempts towards improving pancreatic disease management because of the combination of automated and usability, which promise to improve patient care and outcomes.

References

1. Abu-Khudir, R., Hafsa, N., & Badr, B. E. (2023). Identifying Effective Biomarkers for Accurate Pancreatic Cancer Prognosis Using Statistical Machine Learning. Diagnostics, 13(19), 3091. https://doi.org/10.3390/diagnostics13193091.
2. Dodda, K. & Muneeswari, D. G. (2023). Pancreatic Cancer Detection Through Hyperparameter Tuning and Ensemble Methods. IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 14(No. 12), 528–529. https://www.ijacsa.thesai.org.
3. Lee, Hsiu-An, Kuan-Wen Chen, and Chien-Yeh Hsu. 2022. "Prediction Model for Pancreatic Cancer—A Population-Based Study from NHIRD" Cancers 14, no. 4: 882. https://doi.org/10.3390/cancers14040882.
4. Teixeira, M, Silva, F., Ferreira, R.M. et al. A review of machine learning methods for cancer characterization from microbiome data. npj Precis. Onc. 8, 123 (2024). https://doi.org/10.1038/s41698-024-00617-7.
5. Hameed BS, Krishnan UM. Artificial Intelligence-Driven Diagnosis of Pancreatic Cancer. Cancers (Basel). 2022 Oct 31;14(21):5382. doi: 10.3390/cancers14215382. PMID: 36358800; PMCID: PMC9657087.
6. MIN LI, ,. X. (2020.). Computer-Aided Diagnosis and Staging of Pancreatic Cancer Based on CT Images. https://creativecommons.org/licenses/by/4.0/.
7. Zuherman Rustam, F. Z. (2021). Pancreatic cancer classification using logistic regression and random forest. http://dx.doi.org/10.11591/ijai.v10.i2.pp476-481: IAES International Journal of Artificial Intelligence (IJ-AI) 10(2):476.
8. Zhou, Y., Ge, Y. T., Shi, X. L., Wu, K. Y., Chen, W. W., Ding, Y. B., Xiao, W. M., Wang, D., Lu, G. T., & Hu, L. H. (2022). Machine learning predictive models for acute pancreatitis: A systematic review. International Journal of Medical Informatics, 157, 104641. https://doi.org/10.1016/j.ijmedinf.2021.104641.
9. HIMALI GHORPADE, J. J. (2023). Automatic Segmentation of Pancreas and Pancreatic Tumor: A Review of a Pancreatic Tumor: A Review of a. https://creativecommons.org/licenses/by-nc-nd/4.0/: IEEE.
10. Ali, A. (2024). Ensemble Machine Learning Techniques for Pancreatic Cancer Detection. 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC).