e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

URINARY BIOMARKERS FOR PANCREATIC CANCER PREDICTION USING DATA SCIENCE TECHNIQUE

J. Jacob Daniel[*1], Viji Vinod[*2], V. Sarala Devi[*3]

[*1]MCA Student, Department Of Computer Applications Dr. M.G.R Educational And Research Institute, Chennai, India.

[*2]Professor, Department Of Computer Applications Dr. M.G.R Educational And Research Institute, Chennai, India.

[*3]Asst. Professor, Department Of Computer Applications Dr. M.G.R Educational And Research Institute, Chennai, India.

DOI : https://www.doi.org/10.56726/IRJMETS51943

ABSTRACT

Pancreatic cancer is the fourth most common cancer-related cause of death in the United States and is usually asymptomatic in early stages. There is a scarcity of tests that facilitate early diagnosis or accurately predict the disease progression. To this end, biomarkers have been identified as important tools in the diagnosis and management of pancreatic cancer. Despite the increasing number of biomarkers described in the literature, most of them have demonstrated moderate sensitivity and specificity and are far from being considered as screening tests. More efficient non-invasive biomarkers are needed to facilitate early-stage diagnosis and interventions. Multi-disciplinary collaboration might be required to facilitate the identification of such markers. Data mining is a commonly used technique for processing enormous data. Researchers apply several data mining and machine learning techniques to analyse huge complex data, to helping the prediction of pancreatic cancer. Different algorithms are compared and the best model is used for predicting the outcome.

Keywords: Pancreatic Ductal Adenocarcinoma (PDAC), Biomarkers, Machine Learning Techniques, Random Forest Feature Extraction.

I. INTRODUCTION

PANCREATIC cancer is one of the most lethal malignant tumours, characterized by delayed diagnosis, difficult treatment, and high mortality . The overall five-year survival rate among patients is less than 9%. In the process of diagnosis and treatment, an accurate segmentation of pancreatic cancer plays an important role. Atiyeh et al. performed 3D modelling based on accurately segmented pancreatic cancer, calculated the tumour size, extracted texture features, and built survival prediction models for patients.

Magnetic resonance-guided radiation therapy [4] has potential advantages in treating locally advanced pancreatic cancer. The critical step in this method is contouring, whereby a target volume is generated at the beginning of each treatment stage. However, manual segmentation of pancreatic cancer with small size and blurred boundary is challenging, as it is time consuming to obtain pixel-level annotated data. Given the high level of expertise required to read medical images and perform early diagnosis, senior radiologists, who are limited in number, are under great pressure to segment the ever-increasing number of cases manually. Thus, it is necessary to develop automated and objective segmentation algorithms for pancreatic cancer to overcome human errors and obtain timely segmentation results.

II. REVIEW OF LITERATURE SURVEY

Title: Predicted Prognosis of Pancreatic Cancer Patients by Machine Learning

Author: Julius M. Kernbach, and Victor E. Staartjes

Year: 2020

We recently read the article by Yokoyama and colleagues (1), in which the authors report a predictive model integrating DNA methylation status of three mucin genes to predict overall survival at a designated 5-year interval in pancreatic cancer. They collected samples from 191 patients and compared support vector machines (SVM) with various kernels of ranging complexity and a neural network. Models were trained using k-fold cross-validation (with various choices of k) or leave-one-out cross-validation or a 50/50 split to evaluate

generalizability. However, common discriminative performance measures, including AUC, F1-Score, sensitivity or specificity are not reported for neither the training nor testing set; resampling was not used consistently; model selection and evaluation was not performed separately; model calibration was not assessed; class imbalance and imputation were not considered; approaches to combat overfitting, such as dropout in neural networks, were not included–in total, the prognostic value of the model cannot safely be evaluated on the basis of the presented results.

Title: Management of Metastatic Pancreatic Adenocarcinoma

Author: Ahmad R. Cheema, Ahmad R. Cheema

Year: 2016

Progress in the treatment of pancreatic ductal adenocarcinoma (PDAC) has been incremental, mainly ensuing from cytotoxic systemic therapy, which continues to be a standard of care for metastatic disease. Folinic acid, 5-fluorouracil (5-FU), irinotecan, and oxaliplatin (FOLFIRINOX) and gemcitabine plus nab-paclitaxel have emerged as new standard therapies, improving survival in patients with good performance status. Novel therapeutics targeting the peritumoral stroma and tumour-driven immune suppression are currently a major focus of research in metastatic disease. Identification of reliable and validated predictive biomarkers to optimize therapeutics continues to be a challenge. Continued efforts toward better understanding of tumour biology and developing new drugs are warranted because a majority of patients succumb within a year of diagnosis, despite an increasing number of therapeutic options available today.

Title: Diagnostic, Predictive and Prognostic Molecular Biomarkers in Pancreatic Cancer: An Overview for Clinicians

Author: Dimitrios Giannis , Dimitrios Moris

Pancreatic ductal adenocarcinoma (PDAC) is the most common pancreatic malignancy and is associated with aggressive tumor behavior and poor prognosis. Most patients with PDAC present with an advanced disease stage and treatment-resistant tumors. The lack of noninvasive tests for PDAC diagnosis and survival prediction mandates the identification of novel biomarkers. The early identification of high-risk patients and patients with PDAC is of utmost importance. In addition, the identification of molecules that are associated with tumor biology, aggressiveness, and metastatic potential is crucial to predict survival and to provide patients with personalized treatment regimens. In this review, we summarize the current literature and focus on newer biomarkers, which are continuously added to the armamentarium of PDAC screening, predictive tools, and prognostic tools.

Title: A Survey On Prediction Of Survival Time On Pancreatic Cancer Using Machine Learning Paradigms Towards Big Data

Author: Santosh Reddy P

Year: 2020

Resistance to impending disease can be created using paradigms built through previous procedures of different types, included measurable multi-variate relapses and AI. In any case, such a strategy does not offer the most predictive displays for every cases. To represent mechanized meta-training approaches that define how to predict the best performed system for all patients. The extremely selected procedure is used to maintain patient resistance. We evaluated the proposed approaches in a database of review records of careful resections of pancreatic disease.

Title: Pancreatic cancer: Clinical presentation, pitfalls and early clues

Author: E. P. DiMagno

Year: 1999

The diagnosis of pancreatic cancer usually depends upon symptoms; consequently it is late when there is no chance for cure. At this point, pain, anorexia, early satiety, sleep problems and weight loss are present. Back pain also may be prominent, which predicts unresectability and shortened survival after resection. However, earlier recognition of symptoms of pancreatic cancer might improve early detection of the cancer. For example, 25% of patients have symptoms compatible with upper abdominal disease up to 6 months prior to diagnosis and 15% of patients may seek medical attention more than 6 months prior to diagnosis. These symptoms

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

erroneously may be attributed to problems such as irritable syndrome. Symptoms, however, may be less common. For example a quarter of patients with pancreatic cancer may have no pain at diagnosis, and half, particularly those with pancreatic head tumors, may have little pain compared with patients with body-tail tumors. However, if the tumor is suspected because of predisposing conditions, earlier diagnosis may be possible. These conditions include diseases such as chronic pancreatitis, intraductal papillary mutinous tumor (BPMT), and recent onset of diabetes mellitus, particularly if the diabetes occurs during or beyond the sixth decade.

III. PROPOSED SYSTEM

The proposed method is to build a machine learning model for classification of pancreatic cancer. The process carries from data collection where the past data related to Pancreatic Cancer are collected. Data mining is a commonly used technique for processing enormous data in the healthcare domain. The Pancreatic Cancer if found before proper treatment can save lives. Machine learning is now applied and mostly used in health care where it reduces the manual effort and better model makes error less which leads to save the life. The data analysis is done on the dataset proper variable identification done that is both the dependent variables and independent variables are found. Then proper machine learning algorithm are applied on the dataset where the pattern of data is learnt. After applying different algorithms a better algorithm is used for the prediction of outcome.

System Architecture:

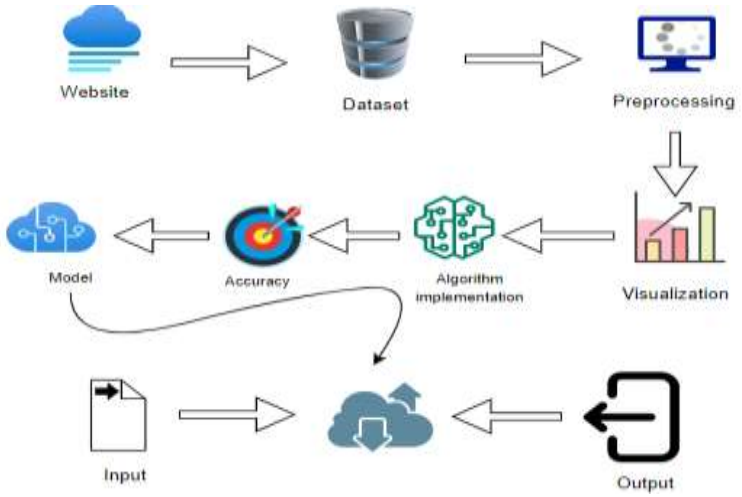

Figure 1:

IV. MODULE DESCRIPTION

➢ Data Pre-processing

➢ Data Analysis of Visualization

➢ Implementing Algorithm

➢ Deployment

Data Pre-processing:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data 32 analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

Data Analysis of Visualization:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning models and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Implementing Algorithm:

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labelled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabelled new data

Used Python Packages:

sklearn: In python, sklearn is a machine learning package which include a lot of ML algorithms. Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score. **NumPy:** It is a numeric python module which provides fast maths functions for calculations. It is used to read data in numpy arrays and for manipulation purpose**. Pandas**: Used to read and write different files. Data manipulation can be done easily with data frames. **Matplotlib:** Data visualization is a useful way to help with identify the patterns from given dataset. Data manipulation can be done easily with data frames.

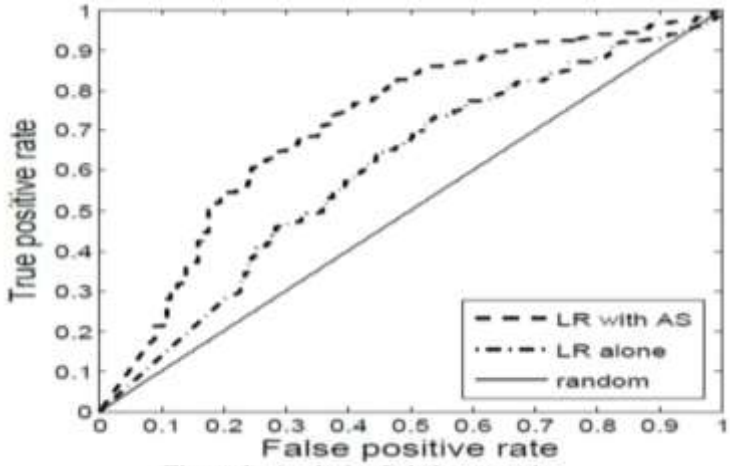

Figure 2. Analysis of gistic regression

K-Nearest Neighbour:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-

parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it ieither cat or dog category.

V. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set of higher accuracy score algorithm will be find out. The founded one is used in the application which can help to find the Pancreatic cancer of the patient.

VI. FUTURE WORK

Deploying the project in the cloud. To optimize the work to implement in the IOT system.

VII. REFERENCES

[1] H. Manuel, "Pancreatic cancer," New England J. Med., vol. 362, no. 17, pp. 1605–1617, 2010.

[2] P. Rawla, T. Sunkara, and V. Gaduputi, "Epidemiology of pancreatic cancer: Global trends, etiology and risk factors," World J. Oncol., vol. 10, no. 1, p. 10, 2019.

[3] M. A. Attiyeh et al., "Survival prediction in pancreatic ductal adenocarcinoma by quantitative computed tomography image analysis," Ann. Surgical Oncol., vol. 25, no. 4, pp. 1034–1042, Apr. 2018.

[4] O. Bohoudi et al., "Identification of patients with locally advanced pancreatic cancer benefitting from plan adaptation in MR-guided radiation therapy," Radiotherapy Oncol., vol. 132, pp. 16–22, Mar. 2019.

[5] Z. Zhu, C. Liu, D. Yang, A. Yuille, and D. Xu, "V-NAS: Neural architecture search for volumetric medical image segmentation," in Proc. Int. Conf. 3D Vis. (3DV), Sep. 2019, pp. 240–248.

[6] Z. Zhu, Y. Xia, L. Xie, E. K. Fishman, and A. L. Yuille, "Multiscale coarse to-fine segmentation for screening pancreatic ductal adenocarcinoma," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2019, pp. 3–12.

[7] Y. Zhou et al., "Hyper-pairing network for multi-phase pancreatic ductal adenocarcinoma segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2019, pp. 155–163.

[8] J. Wang, R. Engelmann, and Q. Li, "Segmentation of pulmonary nodules in three-dimensional CT images by use of a spiral-scanning technique," Med. Phys., vol. 34, no. 12, pp. 4678–4689, Nov. 2007.

[9] D. Smeets, D. Loeckx, B. Stijnen, B. De Dobbelaer, D. Vandermeulen, and P. Suetens, "Semi-automatic level set segmentation of liver tumors combining a spiral-scanning technique with supervised fuzzy pixel classification," Med. Image Anal., vol. 14, no. 1, pp. 13–20, Feb. 2010.

[10] T. Tan, B. Mus, and Platel, H. Huisman, C. I. Sánchez, R. N. Karssemeijer, "Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation," IEEE Trans. Med. Imag., vol. 31, no. 5, pp. 1034–1042, May 2012

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional net works for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.

[12] Z. Gu et al., "Ce-Net: Context encoder network for 2D medical image segmentation," IEEE Trans. Med. Imag., vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[13] M. Havaei et al., "Brain tumor segmentation with deep neural networks," Med. Image Anal., vol. 35, pp. 18–31, Jan. 2017. [14] H. Kim et al., "Abdominal multi-organ auto-segmentation using 3Dpatch based deep convolutional neural network," Sci. Rep., vol. 10, no. 1, pp. 1–9, Dec. 2020. 77