Promising Urinary Biomarkers to predict Pancreatic Ductal Adenocarcinoma using Machine Learning Techniques

H.S.Saraswathi[1*], Mohamed Rafi[2*]

[1]*Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, India*
[2]*Department of Studies in Computer Science and Engineering, UBDT College of Engineering, Davangere,India*
* Corresponding author's Email: saraswathi@jitd.in

Abstract: one of the most prevalent tumors that are considered incurable is a pancreatic tumor. It is one of the most common polyps that are likely to be lethal. It's been predicted to become the second deadliest disease by 2030. Currently, the Food and Drug Agency (FDA) approved only CA19-9 as a biomarker as a part of the screening program. But, the sensitivity and Specificity of CA19-9 are below 90%. The extensive growth of artificial intelligence techniques enables solutions in medical health systems including the automatic diagnosis of disease to monitor the progression of the disease. In this work, we proposed a novel panel of biomarkers LYVE1, REG1B, TFF1, HbA1C, CALCIUM, MAGNESIUM, ZINC, and COPPER through a modified random forest feature extraction technique. The missing values of the dataset are handled by the hybrid KNN and iterative imputation method, which gives a better standard derivation of about 0.0600. We also propose a modified Random Forest Machine learning Classifier, differentiating the pancreatic ductal Adenocarcinoma patients from healthy controls and Chronic Pancreatitis in the early stages such as Stage I and Stage II. The proposed techniques achieved a Sensitivity of 96.15% and a Specificity of 91.1% for the SS Institute of Medical Sciences and Research Centre dataset of 560 urine samples.

Keywords: CA-199, Biomarkers, Classification,HbA1C,LYVE1, Machine Learning, PDAC.

1. Introduction

The pancreas is an auxiliary organ, an endocrine gland that produces hormones, and an exocrine gland of the digestive system. The pancreas weighs about 100g, measures 14 to 25 cm in length, and has a volume of about 72.4 25.8 cm3 in a healthy adult person [1]. It is lobular and elongated in shape. The head, uncinate process, neck, body, and tail are the five anatomical divisions. The exocrine pancreas' main function is to release digestive enzymes that facilitate the breakdown of fatty meals. The endocrine gland aids in controlling blood sugar levels and nutrition absorption by cells [2]. The phrase "pancreatic cancer" refers to exocrine pancreas cancer. It is one of the more prevalent malignancies, especially in western nations and Japan. The next most frequent melanoma in the United States is pancreatic cancer. African Americans are more likely to suffer from it, which accounts for 5% of all cancer fatalities in that nation. Males experience it more frequently than females do. After the age of 50, the incidence rises, type 2 diabetes mellitus may tend to pancreatic ductal adenocarcinoma [3]. When the pancreas cells experience aberrant DNA changes, it results in pancreatic cancer, which causes the cells to expand and divide uncontrollably, resulting in a tumor. This tumor may occasionally spread to the liver, abdominal wall, lymph nodes, lungs, or bones. Smoking, being overweight, having long-term diabetes, having a significant family history of the disease, high dietary intake of processed food, and red meat, and chronic pancreatitis are the risk factors for developing pancreatic cancer [4].It ranks as the fourth-leading cause of cancer-related death in the western world. Within ten years, it is anticipated to be the second-deadliest disease. It accounts for 3% of all cancer cases in America with an average yearly incidence rate of 12.50 per 100,000 people [5]. Using data from the United States Surveillance, Epidemiology and

End Results Program (SEER), Saad et al. found that between 1973 and 2014, the age-standardized incidence rates of pancreatic cancer increased by 1.03% year. The survival rate is less than 9% for five years and it is desirable to identify novel biomarkers to diagnose the disease in the early stages [6]. For men and women respectively, pancreatic cancer is ranked 21st and 17th in India. Mizoram, Mumbai, Thiruvananthapuram, and Delhi have the highest rates for men, while Mumbai, Delhi, Bengaluru, and Thiruvananthapuram have the highest rates for women. With a 94% mortality and incidence ratio, pancreatic cancer is one of the more aggressive cancers. To improve the survival rate, early detection, and awareness are crucial [7]. Because of where it is located anatomically, pancreatic cancer is challenging to find and diagnose. Testing may be required if a patient exhibits any of the following signs and symptoms: jaundice (yellowing of the skin and whites of the eyes), abdominal discomfort, back pain, lack of appetite, pancreatitis, or unintended weight loss. There is no standard test to early diagnose the disease [8]. However, symptoms will appear only when the disease is at a later stage. In terms of prognosis/diagnosis, the patient may undergo a variety of tests, such as ultrasonography, CT (computerized tomography) scans, MRIs, and even PET (positron emission tomography) scans, to detect pancreatic cancer. The majority of pancreatic cancers are discovered to be metastatic at the time of initial diagnosis, making it challenging to make an early diagnosis. The advanced stage of diagnosis is mostly to blame for the bad prognosis. Research is going on in the identification of promising biomarkers for PDAC prediction [9]. Therefore before undergoing scans, the disease is to be predicted with the help of biomarkers. Although MDCT methods are used to diagnose the disease, smaller tumors will not be diagnosed with multiple sittings. Thus, biomarkers play a vital role in predicting the disease. Only 9.7% of cases of pancreatic cancer are diagnosed when they are still localized. For nearly 40 years, these low survival rates have remained mostly unchanged. According to data from the International Agency for Research on Cancer (IARC) and the American Cancer Society (ACS), pancreatic cancer is the leading cause of fatal disease, and one in five men and one in six women worldwide will get cancer at some point in their lifetime. Men die from cancer at a rate of 1:8, whereas women die at a rate of 1:11. In countries with high HDI, pancreatic cancer ranks eighth among all causes of death for women and seventh among all causes of death for men. GLOBOCAN figures show that more cancer cases have been found and more of those instances have resulted in fatalities. Lung, Bronchus, Breast, Colon, Cervix, Prostate, and Pancreatic cancer are said to be the cancers that kill most people worldwide. Estimated causes of death in affluent nations include diseases of the liver, pancreas, breast, and colon. Cancers of the lungs, liver, breast, cervix, colon, pancreas, ovary, and other organs are considered the leading causes of death in developing nations. Some researchers work on a high-risk group such as new onset diabetic patients [10].

The prevalence of cancer and its mortality rate are both rising quickly globally. Table 1 displays the global percentages of cancer incidence and fatalities among men and women in various geographic areas. The likelihood that a population will survive cancer depends on several circumstances, including the type of cancer, the stage at which it is discovered, the prevalence of early detection or screening, the availability of therapy, etc. The survival rate is the proportion of cancer patients who avoid dying and remain alive for a predetermined amount of time. GLOBACAN 2018 estimates 458,918 new cases and causing 432,242 deaths [11,12]. Artificial Intelligence (AI) has been increasingly prevalent over the past ten years, including in the medical industry. To analyze a sizable dataset and produce a prediction model, machine learning and deep learning are two key AI methodologies. The development of AI in the field of gastroenterology has significantly impacted the identification and prognosis of pancreatic cancer. The development of biomarkers and imaging techniques with adequate sensitivity and specificity to correctly identify early-stage PDAC is a primary goal of pancreatic cancer research. This will raise the proportion of patients whose cancer is discovered at an early stage and enhance five-year survival. ML techniques are emerging, especially in the healthcare environment. Several ML-based techniques have

been used to extract prediction patterns from this virtually limitless amount of data. In this article, we will go into more detail on how AI is being used to diagnose pancreatic cancer by non-invasive biomarkers with non-organic compounds such as CALCIUM, COPPER, MAGNESIUM, and ZINC in the AI field after a detailed review of pancreatic cancer with promising urine biomarkers such as LYVE1, TFF1, REG1A, REG1B, and creatinine. Currently used CA 19-9 biomarker serum levels <35 U/mL indicates normal levels, >35 U/mL indicates median survival, <100 U/mL resectable disease, whereas >100 U/mL metastatic disease. But still CA 19-9 gives more false positive rate [13]. While some researchers worked on MicroRNAs and proteomics to improve the performance, but Sensitivity is up to 83% [14]. It has been reported that insulin-like growth factor-binding protein IGFBP2 and IGFBP3 are better than CA 19-9 alone [15]. In the survey the existing models Sensitivity, Specificity, Recall are comparatively low and most of the techniques follows invasive method of collecting the data samples. The currently approved biomarker by Food Drug Agency is CA 19-9 and existing systems have used it for the early diagnosis. Model provides low performance for CA19-9 alone. So, in our proposed work we are looking for multiple biomarkers instead of the currently approved CA 19-9 alone and efficient machine learning classifier to improve the performance of a model in predicting a PDAC in the early stages for non-invasive data samples. In our study, the dataset contains missing values which may affect the performance of a model. So we handled it with a Hybrid KNN and Iterative Imputation method. Some of the instances are behaved too differently to the other sets of instances of a dataset. The presence of outliers may increase the depth of a tree and the model may get over fitted. In our system we handle outliers by a modified interpolated percentile method. A novel modified random forest classifier is applied for the prediction. Then we compare the result with the existing work.

2. Proposed Work

2.1 Dataset and its augmentation

The original dataset was obtained from the Kaggle repository; intern these are collected from BPTB, university college research center, London. Under the supervision of a domain expert, these were combined. Around 300 instances are PDAC stage I-II, 158 instances are Healthy controls, and 102 samples with chronic pancreatitis are used for the biomarker panel analysis. The information covers 560x14 features and contains intriguing indicators like plasma such as CA199 (FDA Approved), LYVE1, REG1B, TFF1, and REG1A. A urine indicator of renal function is creatinine. Lymphatic vascular endothelial hyaluronan receptor 1, or LYVE1, is a protein that is found in the urine and may contribute to the spread of tumors. Urinary levels of a protein associated with pancreatic regeneration (REG1A), pancreatic regeneration-associated protein (REG1B), and pancreatic regeneration-associated protein (REG1B), were measured [16]. TFF1 urinary Trefoil Factor 1 level were measured, and they may be related to urinary tract healing and regeneration of Pancreatic cancer, the stages of pancreatic cancer are stage, stage IB, stage IIA, stage IIIB, stage III, and stage IV. Among all of them, a few input variables need to be designated as useless. Not every study sample contains the variable corresponding to the biomarker REG1A. That is why we decided to classify it as unused as well. The model will not suffer as a result of this choice because the biomarker REG1B enhances the outcomes. We can determine the data distribution of the variables once the data set has been set up. The number of patients with cancer and those without is shown in the following graph. We only take into consideration the data of IA, IB, and IIA stages because our focus is on constructing an artificial intelligence system for an early prediction of pancreatic cancer.

To uplift the enactment of an artificial intelligence system and to reduce the over fitting problem, data augmentation is carried out to increase the training data and the dataset is expanded with non-organic chemicals including CALCIUM, COPPER, MAGNESIUM, and ZINC according to the Domain Expert Advice. Data augmentation also leads to some

issues which need to be handled carefully. The biases of data augmentation are handled clearly by domain knowledge experts. The output quality is observed by the domain experts. Some medical research journals like PUBMED published papers about the usage of major and trace elements for the diagnosis of PDAC using urine samples. The elevated levels of Copper, Zinc and decreased levels of Calcium, Magnesium indicate PDAC.

Data for machine learning models are produced using the data augmentation method. It is a method for producing new data in a range of data orientations. Data augmentation increases the amount of data from a small amount of data and it reduces over fitting. The dataset contains 300 PDAC Stage I to II and 158 Healthy Controls. After being cleaned, the data was one-hot encoded. The patient's age, gender, and urine biomarkers were among the features. In contrast to the undesirable production class, which epitomized certainly not identification, the positive output class represented a pancreatic cancer diagnosis. 20% of the data was utilized as the test set, while 80% of the data was used as the training set.

2.2 Hybrid KNN-IM Method for Missing Value Handling

Missing observations are frequently encountered during the analysis and processing of real-world data using machine learning algorithms. The presence of missing values presents the biggest obstacle to information extraction from databases. The imputation approach should be used to impute the missing values in a dataset to increase the performance and accuracy of machine learning models. Since our original data is collected from various hospitals in London and was stored in the Kaggle repository, it includes some entries with value NaN indicating missing values. We aim to apply an efficient hybrid missing value handling technique to improve the performance of a machine learning model.

There are methods for imputing missing data that use the k-nearest neighbor algorithm, however, choosing the right k value can be difficult. Iterative imputation and k-nearest neighbors are both included in the proposed hybrid missing data imputation approach. Using the k-nearest neighbor technique, the optimal collection of nearest neighbors for each missing record is found based on how similar the records are (kNN). A suitable k value for the kNN is automatically estimated to increase similarity. The global correlation structure among the chosen records is then used to estimate the missing values of the incomplete records using the iterative imputation approach. The proposed method is evaluated using root mean square error.

The proposed Artificial Intelligence System for predicting the PDAC using urinary biomarkers consists of four phases as shown in figure 1. The data collection and its preparation with feature extraction. Handling the missing values and outliers plays a crucial phase in achieving good performance. Then applying multiple classifiers such as Logistic Regression, Naïve Bayes, SVM, KNN, and Decision Tree along with Novel Modified Random Forest followed by the evaluation of the model through validation set with the ground truth.

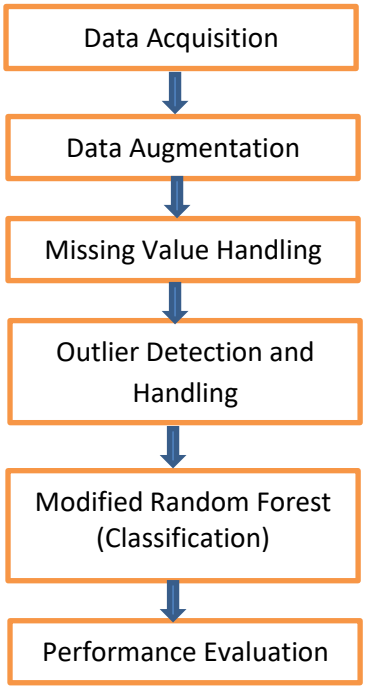

Figure 1: Proposed Methodology

The kNN is based on the Euclidean distance

$$d(P,Q) = \sqrt{\sum_{i=1}^{n}(P_i - Q_i)^2} \qquad \text{Eq.(1)}$$

To assign the missing values, the suggested method combines the concept of kNN and iterative imputation techniques. The kNN method selects the best set of nearest neighbors for each missing record based on how similar the data are. An optimal k value is automatically estimated for the kNN to increase the similarity without any user input. The overall correlation structure between the chosen records is then used to impute the missing values of incomplete records using the iterative imputation approach. This method employs only a degree of similarity. The accuracy of the imputation of missing data can be improved.

Algorithm IKNN-IM(DI, DF)
Input: DI incomplete dataset i.e. contains records with missing values in some attributes.
Output: DF final dataset in which missing values are filled by hybrid approach of KNN and Iterative imputation methods.
1. Determine a collection of P instances which has NAN values in dataset x 2. Select ith record where i is an incomplete record in set P. 3. Determine the attributes A of the corresponding missing values in i^{th}record. 4. Determine a collection Z of instances in dataset X which have values in the corresponding attributes 5. ADD ith record to the set Z. 6. Create a missing value at an index Z in ith record 7.Apply KNNI for imputing values at Z for values K from to Np-1. 8. Calculate RMSE values using imputed value and original values. 9. Best K is determined from minimum RMSE value. 10. Derive Subset with records nearest to ith record (Ri) for values of K to P. 11. Compute S= mi U Ri. 12. Impute missing values in Ri Iterative imputation to get new Ri. 13. Remove old Ri from dataset and add Ri new to dataset 14. Remove old Ri from the set Y. 15.If set Y contains no records then return final dataset otherwise repeat the above steps

Figure 2: Hybrid KNN and Iterative Imputation method for missing value handling.

The K-nearest neighbours (KNN) imputation and iterative imputation are combined in the IKNN-IM algorithm. Where X_ij(t+1) denotes the imputed value at position (i, j) in iteration (t+1), X_i1, X_i2,..., X_ik denotes the K nearest neighbours of X_ij, and X_ij(t) denotes the imputed value at position (i, j) in iteration (t).

The Root Mean Square Error (RMSE) is calculated by an Equation

$$\text{RMSE} = \text{sqrt}(\left(\frac{1}{\text{N}}\right) * \sum(Oi - Pi)\text{^}2) \qquad (2)$$

The anticipated value (P) and the observed value (O) are two terms, respectively. The N represents the total number of missing values for which imputation is made. O represents the true or observed values of the missing data. P represents the imputed values obtained from the imputation method.

The squared difference between each true value and its associated imputed value is measured, added, divided by the total number of missing values, and the square root of the result is used to determine the RMSE.

The original PDAC dataset is divided into four parts. The experimentation of 6 different missing value handling methods such as Mean, KNN, Iterative Imputation, Missing Indicator, Random Imputation, and Proposed Method is conducted individually on all four PDAC parts.

2.2 Outlier Detection and Handling- interpolation Approach

A key idea in the study of medical data is outlier detection. Data objects that depart from typical data behaviour can be found in a database. Outlier scanning is the process of analysing outlier data, which includes those data objects. Dealing with outliers in a dataset is one of the most important phases in data pre-processing for two reasons.

Firstly, outliers may have a significant impact on the results of an analysis. Second, outliers are frequently the result of measurement or recording errors; some of them may even reflect interesting events or other important things from the perspective of the application domain. The outliers may provide useful information for domain specialists, who may learn more about the data by scrutinizing them. Since we combined the Kaggle repository data which comprises features such as CA199 (FDA Approved), LYVE1, REG1B, TFF1, and REG1A with non-organic chemicals including CALCIUM, COPPER, MAGNESIUM, and ZINC under the supervision of

pathologist of SS Institute of Medical Sciences Research Centre, Davangere, Karnataka, India. It is very much required to identify and handle the outliers.

A modified interpolated percentile statistical method is used for outlier detection. The lower and upper thresholds based on percentiles and inter quartile range (IQR) are calculated to handle the outliers. Q1 is representing the first quartile (25^{th} percentile) and Q3 is representing the third quartile (75^{th} percentile). N representing the width of the outlier detecting range and value of 1.5 to 1.6 is providing the good sensitivity in the training phase.

An interpolation approach acts as an efficient method to calculate the percentile, which produces an optimal Inter Quartile Range.

Algorithm to calculate an modified interpolated percentile

- Rank R is determined for the usage of percentile.
- Determine the data point which matches to the position in first step if it is an integer, then use the same value the percentile.
- Interpolate between the two nearest samples whereas in ranking does not contain an integer.
- Calculate the difference of these two observations by the rank's fractional component.
- To obtain the imputed value for the percentile, add the value of the previous step to the lower-ranked value

Figure 3: Modified Interpolated Percentile

Rank R is determined with the help of sample size (n) and percentile (p), i.e.

$$R = P(n + 1) \quad (3)$$

If R points to an integer then use that value pointed by R for percentile, otherwise calculate

$$X = (ob1 - ob2) * fractional\ component \quad (4)$$

Then the imputed value (I) for percentile is calculated based on

$$I = X + LRV \quad (5)$$

Where LRV is a lower ranked value.

The mean-variance test and the box-plot test are the two outlier tests for normal distributions that are most frequently used. When the population has a mean and variance and the distribution is Gaussian, outliers are points that deviate from the mean by three standard deviations or more (i.e., by three or more). The box-plot test is used to visually represent the scattering of data.

2.3 Modified Random Forest for identification of a new panel of Urine Biomarkers and its usage for PDAC prediction

Choosing a subset of relevant features for a machine learning model is a process known as feature selection. The performance on the train and score is improved using pertinent characteristics in the categorization. Considerably lessen the variation and stay away from over-fitting. Based on the weights assigned to the characteristics, supervised techniques are used. Contrarily, unsupervised techniques will attempt to choose the features that translate the majority of the dataset's information into as few features as possible. Designing feature selection procedures with care is important since feature extraction might impair the performance or even cause the produced model to fail.

Algorithm: Modified Random Forest

Algorithm MRFFE(df)

Input: Dataset with twenty biomarkers of PDAC

Output: Relevant features for the diagnosis of PDAC

- Initialize random forest, random tree and feature vector.
- Compute global weight according to weight of a tree.
- Entropy based ranking is done.
- Sort features and place top l in related set R.
- Place (initial feature vector} – {top l } in unrelated set UR.
- Initialize y = 0.
- UntilURn> efix from calculate to increment n.
- Compute μ and Ƞ from UR.
- Calculate μ-2 Ƞ from UR and take away such attributes whose universal weights are below μ-2 Ƞ and place it in new set x.
- Then update new feature = initial feature vector – x.
- After updating feature vector update random forest.
- Updating of feature vector and random forest is done with n iterations.
- Final forest grows with P tress and Q feature vectors.
- Find weight of a tree and determine the rank all features in Q.
- Increment y.

- End of While.
- Print all selected features with its Rank.

Figure 4: Modified Random Forest Feature Extraction method

Random forest is one of the most widely used classifiers for classification and regression problems. Its ability to handle high-dimensional data and perform well on unbalanced datasets gives it a significant advantage over other machine learning methods [17]. In order to avoid feature strength saturation, the proposed method seeks to determine the ideal number of trees and reduces features in a random forest. Instead of accuracy, improving correlation is the main goal of adding more trees. It simultaneously reduces the number of features and adds trees. To reduce the chance of missing critical characteristics, features for the forest are chosen from both important and minor feature bags. The suggested method builds each decision tree using a bootstrap sample. If the bootstrap sample has F features, f features that are smaller than F are chosen for tree construction. Only one feature is selected for node splitting from the subset of features f that were previously selected from the whole set of features F. Our method is unique in that we add the trees recursively. The approach splits features into two groups, distinguishing between important and unimportant features. The random forest method begins with a basic number of trees and goes through iterations. New significant and insignificant traits are added to the random forest with each cycle. Assigning weights to each feature, separating significant from irrelevant features based on a threshold, ranking the features, locating the most trees possible, and converging the forest are the main components of our method. The procedure is carried out repeatedly until the forest has converged and the best categorization has been attained. In order to enhance classification performance and avoid feature strength saturation, our study suggests a modified Random Forest technique that combines feature reduction and the inclusion of trees.

3. Result and Discussion

3.1 Data Exploration

The distribution of sex versus diagnosis, is shown in the below figure 5 and figure 6. The data distribution related to diagnosis and age are shown in figure 7 and figure 8.

Figure 5: Sex versus Diagnosis

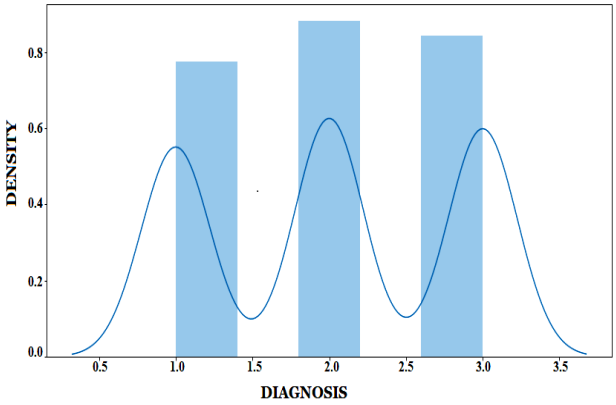

Figure 6: Diagnosis Distribution

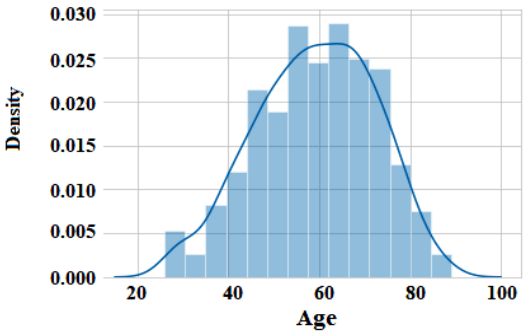

Figure 7: AgeDistribution

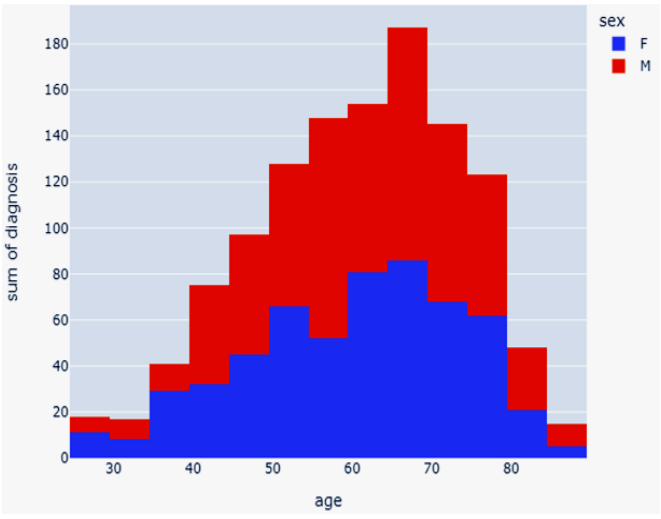

Figure 8: Histogram showing diagnosis for male and female group

The figures 10 to figure 12 show that the target class doesn't have much uneven distribution of observations. The little imbalanced data is skewed to uplift the performance.

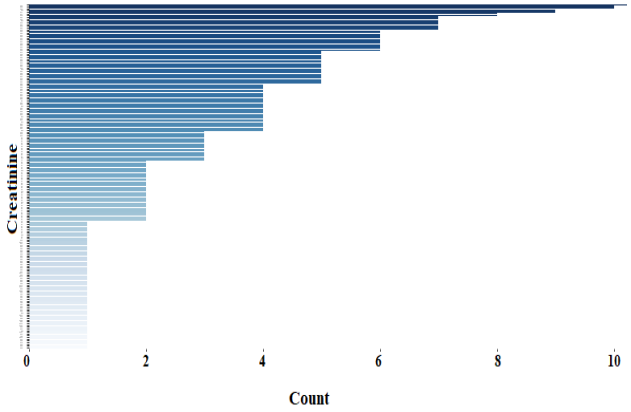

Figure 9: count plots for categorical features of creatinine

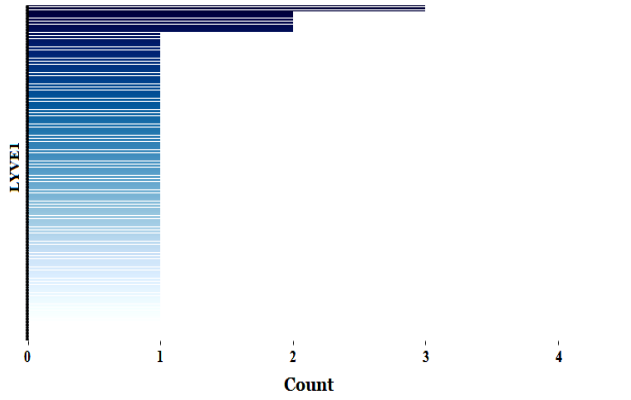

Figure 10: count plots for categorical features of LYVE1

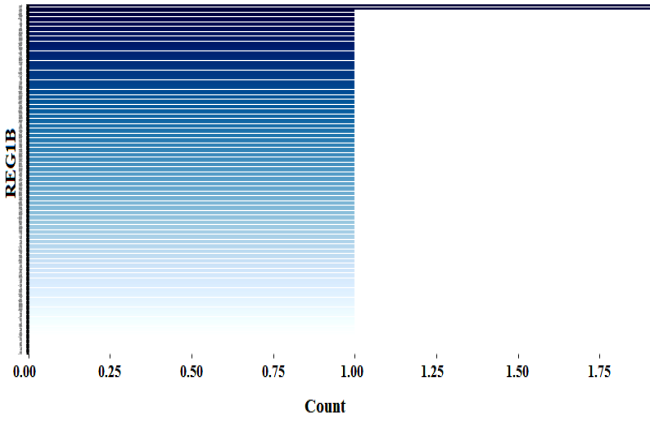

Figure 11: count plots for categorical features of REG1B

In essence, the PDAC dataset with Stage I and Stage II records is divided into four sets. Each set has experimented with different missing value handling techniques such as Mean, KNN, Iterative Imputation, Missing Indicator, Random Imputation, and the proposed hybrid algorithm IKNN-IM. The whole dataset is divided into four parts in our experimental setup to apply and analyze various methods to handle the missing values of the dataset.

Table 2: The Average RMSE Values over 6 methods on PDAC Dataset (with 4 parts)

Dataset	M1	M2	M3	M4	M5	M6
PDAC1	0.211	0.594	0.181	0.300	0.333	**0.075**
PDAC2	0.349	0.390	0.231	0.200	0.077	**0.045**
PDAC3	0.032	0.487	0.433	0.022	0.542	**0.059**
PDAC4	0.419	0.224	0.360	0.542	0.233	**0.198**
SD	0.148	0.135	0.100	0.434	0.168	**0.060**

M1:Mean, M2:KNN,M3:IterativeImputation, M4:Missing Indicator,M5:Random Imputation, M5: Proposed Method IKNN-IM.

The average difference between the imputed values and the truth values is measured by the RMSE. Better imputation accuracy is shown by a lower RMSE since it shows a smaller overall difference between the imputed and true values for the missing data.

Compared to the existing methods Mean, KNN, Iterative Imputation, Missing Indicator, Random Imputation our proposed method IKNN-IM gives betters performance of standard deviation of 0.0600 than for existing methods.

The histogram and box plot for sample biomarkers REG1B, is shown in the figure below for handling of outliers.

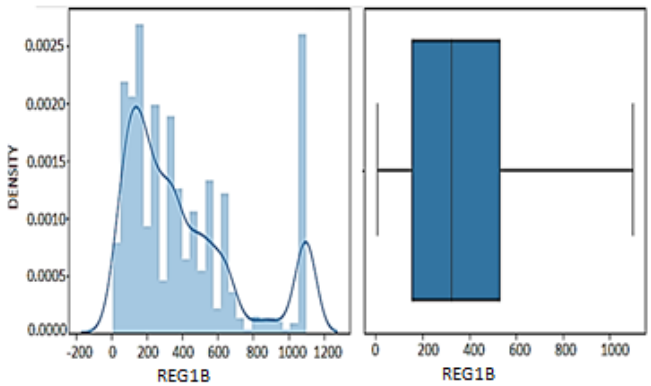

Figure 5: A Histogram and Box Plot for REG1B

The proposed model selects the relevant features and drops the irrelevant features from the dataset with the help of producing creating the optimal number of trees during the model prediction of Pancreatic

Ductal Adenocarcinoma. Urine biomarkers LYVE1, REG1B, TFF1, Creatinine, Age, CALCIUM, MAGNESIUM, ZINC, and COPPER are high-ranked features derived from the modified Random Forest method. Based on the selected features, the classification is done by the proposed model.

A test's sensitivity and specificity are indicators of how well it can identify whether a person has a Pancreatic Ductal Adenocarcinoma or not. The ability of a test to identify a positive result in a sick patient is referred to as sensitivity. Because fewer false negative results arise from a highly sensitive test, fewer cases of illness are missed. The ability of a test to classify as negative someone who does not have a condition is how specific it is. A highly specific test will yield minimal false positive outcomes.

Table 3: The Proposed Modified Random Forest Feature Extraction method results with a novel panel of urine biomarkers for PDAC dataset with Stage I, Stage II, and HC

Dataset	Model	Acc.(%)	Sen.(%)	Spec.(%)	Precision	F Measure	AUC
PDAC (StageI and StageII)	LR	88.42	86.66	91.77	95.23	90.75	0.90
	NB	90.17	91.66	87.34	93.22	92.43	0.92
	KNN	90.61	90.00	91.77	95.40	92.62	0.92
	DT	92.35	93.33	90.50	94.91	94.11	0.93
	SVM	88.96	90.66	87.34	92.20	91.42	0.89
	Proposed Model	**96.15**	**96.00**	**91.03**	**96.32**	**96.16**	**0.95**

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$FMeasure = \frac{2TP}{2TP + FP + FN}$$

Where *TP= True Positive, FN= False Negative, TN= True Negative, FP= False Positive*

To show the performance of the proposed modified random forest model for classification, results regarding the accuracy, sensitivity, specificity, precision, Recall, F measure, and AUC values are used for each model concerning the dataset used for experimentation. The experimentation dataset is verified and certified by the domain expert, which contains 300 PDAC Stage I and Stage II records with 158 healthy controls. First, the experimentation was done only on recommended biomarkers by the FDA, i.e., CA19-9, which tends to result in 81% and 92% for sensitivity and specificity respectively. Then LYVE1+REG1B+TFF1+Creatinine+age panel was used, which produces the performance of 82% and 90% Sensitivity and Specificity, respectively. The experimentation is also done on metals, i.e., major and trace elements CALCIUM, MAGNESIUM, COPPER, and ZINC, which results in 90% and 67% sensitivity and specificity respectively. To improve the true positive rate, our work includes the panel ofbiomarkersLYVE1+REG1B+TFF1+creatinine+age+metalsand the model results in 93% and 92% Sensitivity and Specificity respectively.

However, even though researchers are identifying many promising biomarkers over the period Diagnosing- Pancreatic Ductal Adenocarcinoma, the current clinical setting uses CA19-9. Finally, we present our work using the LYVE1+REG1B+TFF1+creatinine+age+metals+CA19-9 model, which has improved sensitivity and specificity compared to other machine learning models used in this study. Additionally, other performance evaluation criteria like recall, accuracy, and F measure outperform other models already in use.

4. Comparison of Proposed work with an existing work

Our proposed strategy performs well when compared to previous studies by other researchers using various biomarkers and machine learning algorithms. The in-depth comparison is depicted in the picture below. Most of the previous work includes CA 19-9 in their panel of biomarkers [18, 19, and 20]. The pancrisk framework was developed by Debernardi, S [14] which includes urinary biomarkers with 94% of sensitivity, whereas the work of Arasaradnam, R.P [21] gives 82% with the help of non-organic volatile compounds. The proposed work gives the best performance of 96.15% of sensitivity in association with promising CA 19-9 with urinary and non-organic volatile biomarkers. The current work uses a modified random forest and which produces an optimal number of trees by selecting the best features. The Proposed method of classification is not data-dependent because, unlike other approaches, the number of trees is not preset. The proposed method is

quick. Regarding feature reduction, the suggested method employs an iterative procedure. The least important elements are automatically eliminated after each cycle. The proposed method uses the optimal number of trees, unlike standard Random Forest, which makes the unpractical assumption that adding more trees will inevitably increase accuracy. It increases the number of trees in the forest, removes pointless features, and evaluates the model's classification performance in each step.

Table 4: Comparison of Proposed Work with an existing work.

Sl No	**Paper Reference**	**Specimen Type**	**Biomarkers**	**Sensitivity (%) Specificity (%) AUC**
1	Yu et al., 2019	Plasma	d-signature: EV long RNA	93.68,91.57,0.936
2	Nolen et al., 2014	Serum	CA19-9	25.7, 95.0,0 .68
			CA19-9 + CEA	26.7, 95.0,0.67
			CA19-9 + CEA + Cyfra 21-1	32.4,95.0 0.69
3	Debernardi et al., 2015	Urine	miR-143	83.3,88.5,0.86
			miR-223	83.3,76.9,0.80
			miR-30e	83.3,80.8,0.85
			miR-143+miR-30e	83.3,96.2,0.92
4	Radon et al., 2015	Urine	LYVE1(StageI-IV)	67.0,91.8,0.84
			LYVE1(StageI-IV)	67.0,91.8,0.84
			REG1A(StageI-IV)	75.0,68.9,0.75
			TFF1(StageI-IV)	78.6,52.5,0.70
			LYVE1 + REG1A + TFF1+Creatinine +Age (StageI-II)	82.1,88.5,0.90
			Plasma CA19-9	83.1,92.9,0.88
			Panel (LYVE1 + REG1A + TFF1) (Stage I-II)	93.0,92.9,0.97
			Panel + Plasma CA19-9(StageI-II)	94.4,100,0.99
5	Arasaradnam et al.2018	Urine	Volatile organic compounds	091,0.78,0.89
				0.82,0.89,0.92
6	Proposed work for stage I-II patients	Urine	Plasma CA19-9	81.12,92.8,0.86
			LYVE1 + REG1B + TFF1 + (Creatinine +Age)+HbA1c	82.45,90.2,0.90
			Calcium+magnesium+zinc+copper+ HbA1c	90,67,0.84
			Proposed panel I (LYVE1+REG1B + TFF1 +HbA1c Creatinine +Age+ Calcium+magnesium+zinc+copper	93.34,92.8,0.97
			Proposed panel I + Plasma CA19-9 with modified Random Forest Classifier	96.15,91.1 (AUC = 0.9579)

We evaluated the feasibility of identifying urinary biomarkers and metals with Food and Drug Agency-approved biomarkers for the early-stage detection of PDAC, through the analysis of urine samples. Data was prepared with 560 urine samples including 300 PDAC, 102 Chronic Pancreatitis, and 158 Healthy controls. The modified random forest analyses were applied to determine the discriminatory potential of the candidate biomarkers. The modified random forest with Plasma CA19-9 was able to differentiate PDAC patients from non-PDAC with sensitivity (SN) of 81.12% and specificity (SP) of 92.8 %(AUC=0.86). The same classifier is used for LYVE1, REG1B, TFF1, creatinine, and HbA1c and can differentiate PDAC patients from non-PDAC with sensitivity (SN) of 82.45% and specificity (SP) of 90.2 %(AUC=0.90). The metals such as Calcium, Magnesium, Zinc, and Copper with HbA1c for the proposed classifier resulted in sensitivity (SN) of 90.00% and specificity (SP) of 67.0 %(AUC=0.84).

Whereas LYVE1, REG1B, TFF1, creatinine, and HbA1c with metals can differentiate from PDAC and NON-PDAC with sensitivity (SN) of 93.34% and specificity (SP) of 92.8 %(AUC=0.97). But in association with the Food and Drug Agency-approved biomarker and our proposed panel of biomarkers with the proposed classifier gives out an extremely good performance with sensitivity (SN) of 96.15% and specificity (SP) of 91.1 %(AUC=0.95). The model is evaluated with the help of a Domain Expert, Pathology Department, SSIMS, Research Center, Davangere for the efficacy of the approach. The results of the validation set are verified. Through this study, we were able to establish for the first time the value of urine biomarkers for the early, non-invasive identification of PDAC in urine samples from newly diagnosed diabetic population.

5. CONCLUSION AND FUTURE ENHANCEMENT

The primary factor in pancreatic cancer fatality in PDAC. The diagnosis is not straightforward because the symptoms do not manifest while the tumor is in its primary phases. Even the absence of techniques for early disease diagnosis makes the situation worse. Because of their poor sensitivity/specificity, biomarkers utilised for diagnosis and treatment response monitoring, such as carbohydrate antigen 19-9 (CA19-9) and carcinoembryonic antigen (CEA), are insufficient as early detection markers of PDAC. New PDAC biomarkers are therefore desperately needed. To forecast the PDAC at stages I–II, we thus concentrated on a novel panel of biomarkers with a modified random forest classifier. Our systematic work on missing treatment, outlier handling, feature extraction, and classification processes indicates numerous interesting non-invasive urine PDAC indicators for predicting pancreatic ductal adenocarcinoma using a machine learning model with a high performance of sensitivity and specificity. Promising results are obtained using a panel of new biomarkers, including LYVE1, REG1B, TFF1, Creatinine, CA19-9, and Age, together with metals including calcium, magnesium, copper, and zinc. When compared to healthy controls, we found that PDAC patients had considerably lower levels of urine calcium and magnesium and higher amounts of copper and zinc. Further we extend the work by implementing Deep learning models on CT images for the best diagnosis to further increase performance shortly.

References

[1] Atkinson MA, Campbell-Thompson M, Kusmartseva I and Kaestner KH. "Organisation of the human pancreas in health and in diabetes". *Diabetologia*, 2020.

[2] Leung, P.S, "The Renin-Angiotensin System:Current Research Progress in The Pancreas", *Springer*,vol 690, 2010.

[3] Mohan H. *Textbook of pathology*, Jaypee Brothers Medical Publishers,2018.

[4] Moore, A., & Donahue, T. *Pancreatic Cancer*. JAMA, 322(14), pp.1426-1426,2019

[5] Zhang L, Sanagapalli S and Stoita A. *Challenges in diagnosis of pancreatic cancer*. World J Gastroenterol. Vol. 24, No. 19, pp.2047-2060,2018.

[6] McGuigan A, Kelly P, Turkington RC, Jones C, Coleman HG and McCain RS. "Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes", *World J Gastroenterol.* Vol. 24 N0 43, pp. 4846-4861,2018

[7] Gaidhani RH and Balasubramaniam G, "An epidemiological review of pancreatic cancer with special reference to India*", Indian Journal of Medical Sciences*, Vol 73,No 1,pp.99-109,2021.

[8] Moore, A & Donahue T. "Pancreatic Cancer". *JAMA*, Vol. 322,No. 14, pp.1426-1426,2019.

[9] R.Young, P.D.Wagner and S.Ghosh, " Validation of Biomarkers for Early Detection of Pancreatic Cancer: Summary of The Alliance of Pancreatic Cancer Consortia for Biomarkers for Early Detection Workshop", Pubmed, Vol. 47,No.2,pp.135-141,2019.

[10] S.P.Pereira ,L.Oldfield, A.Ney, P.A.Hart and M.G.Keane ,"Early-detection of pancreatic cancer", *Lancet Gastroenterol. Hepatol*, Vol. 5,No. 7, pp.698- 710,2020

[11] W.L. Da Costa, A.O. Oluyomi and A.P.Thrift,"Trends in the Incidence of Pancreatic Adenocarcinoma in All 50 United States Examined Through an Age-Period-Cohort Analysis",JNCI, Cancer Spectr. Vol. 4,No.4,pp.pkaa033,2020.

[12] P.Rawla,T.Sunkara and V. Gaduputi,"Epidemiology of Pancreatic Cancer: Global Trends, Etiology and

Risk Factors", *World J. Oncol*, Vol.10, No. 1,pp.10–27,2019.

[13] U.K.Ballehaninna and R.S.Chamberlain,"The clinical utility of serum CA 19-9 in the diagnosis,prognosis and management of pancreatic adenocarcinoma: An evidence based appraisal", J.Gastrointest. Oncol. Vol.3, No. 2,pp.105–119,2012.

[14] S.Debernardi, N.J.Massat, T.P. Radon, A. Sangaralingam,A.Banissi, D.P.Ennis, T. Dowe,C. Chelala, S.P. Pereira and H.MKocher,"Noninvasive urinary miRNA biomarkers for early detection of pancreatic adenocarcinoma",Am. J. Cancer Res. Vol. 5, No. 11,pp.3455–3466,2015.

[15] T.Yoneyama, S.Ohtsuki, K. Honda, M. Kobayashi,M.Iwasaki,"Identification of IGFBP2 and IGFBP3 As Compensatory Biomarkers for CA19-9 in Early-Stage Pancreatic Cancer Using a Combination of Antibody-Based and LC-MS/MS-Based Proteomics", *PLoS ONE*, Vol.11, No.8,pp.e0161009,2016.

[16] S.Debernardi, H. O'Brien, A.S.Algahmdi, N.Malats, G.D.Stewart and M.Plješa-Ercegovac,"A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study" *PLoS Med*, Vol.17, No.12,pp.e1003489,2020.

[17] C.LUO, Z Wang, S. Wang, J. Zhang and J. Yu, "Locating facial landmarks using probabilistic random forest. *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2324-2328, 2015.

[18] S. Kaur, L. Smith, A. Patel, M. Menning, and D.Watley, "A Combination of MUC5AC and CA19-9 Improves the Diagnosis of Pancreatic Cancer: A Multicenter Study". *Am. J. Gastroenterol*, 2017, Vol.112,No.1,pp.172–183,2017.

[19] J. Kim, W.R. Bamlet, A.L. Oberg, K.G. Chaffee, G. Donahue, X,J. Cao,"Detection of early pancreatic ductal adenocarcinoma with thrombospondin-2 and CA19-9 blood markers" *Sci. Transl. Med*, Vol. 9,No.398,pp.eaah5583, 2017.

[20] D.Dong, L.Jia, L. Zhang, N. Ma, A. Zhang and Y.Zhou, "Periostin and CA242 as potential diagnostic serum biomarkers complementing CA19.9 in detecting pancreatic cancer" *Cancer Sci*. Vol.109, No.9, pp.2841–2851,2018.

[21] R.P.Arasaradnam, A.Wicaksono, H. O'Brien,H.M. Kocher, J.A. Covington,"Noninvasive Diagnosis of Pancreatic Cancer Through Detection of Volatile Organic Compounds in Urine" Gastroenterology, Vol. 154, No. 3, pp.485–487, 2018.

Conflicts of Interest

"The authors declare no conflict of interest."

Author Contributions

"Conceptualization, H.S.Saraswathi and Mohamed Rafi; methodology, H.S.Saraswathi; software, H.S.Saraswathi; validation, H.S.Saraswathi, and Mohamed Rafi; formal analysis, H.S.Saraswathi; investigation, H.S.Saraswathi; resources, H.S.Saraswathi; data curation, H.S.Saraswathi; writing—original draft preparation, H.S.Saraswathi; writing—review and editing, H.S.Saraswathi and Mohamed Rafi ; visualization, H.S.Saraswathi; supervision, Mohamed Rafi