Pancreatic Cancer Detection Through Hyperparameter Tuning and Ensemble Methods

Koteswaramma Dodda, Dr. G. Muneeswari
School of Computer Science and Engineering, VIT-AP University, Amaravati 522237, India

***Abstract*—Computing techniques have brought about a significant transformation in the field of medical research. Machine learning techniques have facilitated the analysis of vast amounts of data, modeling of complex scenarios, and the ability to make well-informed decisions. This presents an opportunity to develop reliable and effective medical system implementations, which may include the automatic recognition of uncertain issues related to health. Currently, significant research efforts to be directed towards the prediction of cancer, particularly focusing on addressing the various health complications caused by this disease, which can adversely impact multiple organs within the body. Pancreatic Cancer (PC) stands out as a highly lethal form of tumor, with a rather discouraging global five-year survival rate of approximately 5%. The truth behind the early detection increases the survival rate and it also helps the radiologists to give better treatment to those who are affected at early stages. Creatinine, LYVE1, REG1B, and TFF1 are urine proteomic biomarkers that offer a promising non-invasive and affordable diagnostic technique for detecting pancreatic cancer. In this study, a novel model that combines gridsearchCV technique to search and find the optimal combination of hyperparameters for a random forest classifier. In this research a new ensemble method to enhance the performance for classification of pancreatic cancer and non-cancer by using urinary biomarkers which is collected from Kaggle. The implemented model achieved better results of Accuracy 99.98%, F-1 score 99.98, Precision 99.98, and Recall 99.98.**

Keywords*—*Pancreatic cancer; Machine learning (ML); urinary biomarkers; grid search hyper parameter tuning; Random Forest (RF)

I. INTRODUCTION

According to cancer statistics, Pancreatic Cancer (PC) is predicted to overtake all other causes of death globally in 2022 [1]. The pancreas is positioned posterior to the stomach and anterior to the spine, with the liver, spleen, and gallbladder encircling it. Pancreas releases hormones into the digestive tract to regulate blood sugar levels. The pancreas, positioned behind the lower stomach, is responsible for both regulating blood sugar levels through hormone production and releasing digestive enzymes. When cancer forms in the tissue of the pancreas, it is known as pancreatic cancer. This is due to its location in stomach which is becoming a difficult task for clinicians (researchers) to find PC. Most instances of Pancreatic Cancer occur in those over 45 like additional risk condition, specific genetic mutations, chronic pancreatitis, diabetes, and advancing age.

Now-a-days, PC has become more prevalent [2]. At initial stages Pancreatic cancer often remains asymptomatic. Symptoms including icterus back or bellyache pain, decreased appetite, unexplained weight loss, and digestive issues may appear as the condition worsens [3]. Hence, there exist significant variations in the structure and presentation of the pancreas among different individuals. Among the various types of PC, Pancreatic Ductal Adenocarcinoma (PDAC) is leads to dead. Pancreatic Cancer, which currently lacks a cure, ranks is among the most severe diseases with one of the poorest survival rates. PC stands out for having the lowest survival rate nearly five years of any cancer, primarily due to its late-stage diagnosis, which stands at 11% [1]. Because early indicators are rare, pancreatic cancer (PC) is infrequently detected in its early stages [4], [5]. As a result, both surgical interventions and treatments involving chemotherapy and radiation are generally ineffective in combating PC [6]. To enhance knowledge about the risk of pancreatic cancer, medical professionals advise patients to undergo additional diagnostic measures, such as biomarker testing and medical imaging scans [7]. There are number of difficulties are reported for imaging early pancreatic cancer. The expense associated with radiological image screening is considerable, making it an improbable choice for widespread pancreatic cancer (PC) screening.

Consequently, researchers are turning their attention towards the utilization of biomarkers as an initial approach to early PC detection. The rapid advancements in genomic sequencing and its diverse techniques, including proteomics, epigenomics, and transcriptomics, generate vast amounts of multi-omics data. Biomarkers in PC serve critical roles in early identification, prognosis, treatment decision-making, and research, resulting in more effective and customized care for patients suffering from this difficult disease. The essential biomarkers that play a pivotal role in the early identification of PC can be found in bodily fluids, which encompass cyst fluid, pancreatic juice, and bile. However, gathering these fluids necessitates invasive methods like surgery or endoscopy. In contrast, blood stands out as a low-risk, cost-effective, and easily repeatable source of tumor biomarkers [8]. Blood contains a wealth of proteomic biomarkers, including Carbohydrate Antigen 19-9 (CA19-9), and transcriptomic biomarkers like Circulating Micro RNAs (miRNAs), which can be detected through RNA sequencing. [9].

Historically, blood has served as the principal reservoir for these biomarkers, although urine presents itself as a viable alternative biological fluid. It makes it simple to collect a non-invasive sample in large quantities and do repeated measurements. As of right now, there is no reliable biomarker for detecting PDAC at early stage. The single biomarker used

in clinical practice, serum CA19-9, is used largely as a prognostic indicator and to track therapy effectiveness because it is neither specific nor sensitive enough for screening [10]. Urine stands as a hopeful substitute body fluid for the exploration of biomarkers. It proves to be an excellent option for broad diagnostic screening due to its non-invasive and cost-effective nature, as patients can readily provide ample samples [11]. Urine includes proteome indicators much like blood does.

A three-protein biomarker panel that was presented by Radon et al. in 2015 [12] suggests that urine samples can be utilized to diagnose people with early-stage PC. LYVE-1, TFF1, and REG1A were investigated putative proteomic biological markers. In a smaller study, scientists investigated that the miRNA in urine is also used for diagnosis of pancreatic ductal adenocarcinoma at early stage. By substituting REG1B for REG1A in 2020, Debernardi et al. [13] discover an additional protein biomarker panel. Furthermore, because of the comparable symptoms in early-stage PDAC cases, they can differentiate between cases of PDAC and benign hepatobiliary disease, which can be challenging. This analysis takes into account four important urine biomarkers: TFF1, LYVE1, REG1B, and creatinine. Creatinine is likely employed to assess function of kidney. Moreover, Lymphatic Vessel Endothelial Hyaluronan Receptor 1 has associations with tumor metastasis, REG1B with pancreatic regeneration, and Trefoil Factor 1 with the regeneration and urinary tract repair. The analysis of computer-aided diagnosis (CAD) systems has seen the emergence of ML and DL techniques, which serve as the foundation for handling a wide range of information about patients, including radiology, pharmacogenetics, and biological markers.

In the healthcare sector, machine learning (ML) has become a game-changing technology that is revolutionizing the way medical data is analyzed, diagnoses are made, and treatments are given. Random forest serves as a widely employed ML algorithm for addressing both Regression and Classification tasks. GridsearchCV, on the other hand, is the method used for fine-tuning hyperparameters to identify the most effective settings for a given model. AdaBoost, as a potent ensemble learning algorithm, harnesses the collective strength of multiple weak learners to construct a resilient predictive model.

The following is the paper organizational structure. In Section II, a number of related works are addressed. The proposed system methodology is examined in Section III. In Section IV the experimental results are discussed. Finally, Section V concludes the proposed research paper.

II. Literature Review

In the study, Lee et al. [14] detected miRNA indicators released into the bloodstream. The developed diagnostic system for PC by choosing 39 miRNA markers using a penalized support vector machine (SVM) is uses a smoothly trimmed absolute deviation. The model's accuracy was 93%, and AUC of 0.98.

Long et al. [15] identified and validated oncogenic indicators of pancreatic cancer at the genome level using data mining and multi-omics data. Random forest (RF) technique was used to build their prediction system because it is a simple methodology. After effectively uncovering the unseen biological information from multi-omics data, Scientists have pinpointed dependable biomarkers for the early observation, prediction, and diagnosis of PC. The suggested Random Forest (RF) model achieved an impressive 96% of efficiency.

Debernardi and colleagues [16] identified diagnostic miRNAs for detecting pancreatic ductal adenocarcinoma (PDAC) based on urine samples at early stage. The discriminatory potential miRNA biomarkers were identified using LR algorithms. The suggested models yielded the most favorable outcomes, demonstrating 83.3% sensitivity, 96.2% specificity, with AUC 0.92.

Exosomes were used by Ko et al. [17] to diagnose PC utilizing machine learning and nanofluidic technologies. In order to assess unrefined clinical samples, they created a multichannel nanofluidic device. Tenth, to aid in the definitive diagnosis of cancer patients, these exosomes are subjected to the linear discriminant analysis (LDA) technique. The AUC for this prediction model's classification of pancreatic tumours against healthy samples was 0.81.

Lee et al. [18] reported the construction of a prediction model for the people who are in advanced stage from population-based research to find at early stage. The diagnostic methodology acknowledged that will aid in educating the medical community about the risk of pancreatic cancer. NHIRD, Taiwan's health insurance database, served as the foundation for this study. Combination of four models is used to create predictive model: voting ensemble, ensemble learning, deep neural networks, and logistic regression (LR). The model's AUC ranged from 0.71 to 0.76, and its accuracy was between 73 and 75%.

To identify PC at the localized stage, D Agarwal et al. [19] proposed highly sensitive nano biosensors for protease/arginine detection. The hard hierarchical decision structure (HDS) and soft hierarchical decision structure (SDS) beat traditional multi-class classification methods, achieving an accuracy score of 92%.

Thanya et al. [20] employs color conversion and anis lateral filters on pancreatic cancer CT images from a PCCD database. FK-NNE segmentation and features based on histograms are extracted. A DCNN combined with a DBN classifier classifies pancreatic cell tumors as benign or malignant. The accuracy climbed to 99.6%, with a sensitivity of 100% and a specificity of 99.47%.

1D CNN-LSTM model accurately categorizes patients with pancreatic cancer, outperforming competing models in assessment measures by 97%, Karar et al. [21] based on urine biological markers. The study evaluates several risk prediction algorithms in order to create PancRISK, a biomarker-based risk score. Out of 379 patients were categorized into training and testing sets using urine biomarkers. When a number of machine learning algorithms were compared, none of them performed noticeably better than the others. The PancRISK

score was derived using the logistic regression model; however, it could be enhanced by include the PDAC biomarker CA19-9. Currently, Blyuss et al. [22] are investigating the utility of PancRISK for non-invasive patient risk assessment in pancreatic cancer.

In a separate investigation, Jiao et al. [23] delved into the connection between ITGA3 expression in the pancreas and the observations and unhealthy features of patients with pancreatic cancer. Their methodology encompassed data mining and the application of Chi-squared tests to evaluate associations. Their findings unveiled a notable elevation in ITGA3 expression in pancreatic cancer patients, showcasing correlations with microanatomy, stage, and recurrence. Most notably, heightened ITGA3 expressions analogous with diminished persistence rate are especially to find cancer at earlier stage.

In their study, Liu and colleagues [24] investigated 11 long non-coding RNAs (lncRNAs) associated with pancreatic cancer and identified plasma ABHD11-AS1 as a probable biomarker for the detection of pancreatic cancer. They observed that combining ABHD11-AS1 with CA199 improved the efficiency of pancreatic cancer diagnosis compared to using ABHD11-AS1 by itself, especially for early tumor detection. These findings play a potentially important role in lncRNAs for the diagnosis of cancer.

Iwano et al. [25] proposed to develop a PDAC-symptomatic model using serum anabolism from Japanese patients. Utilized liquid chromatography/electrospray ionization mass spectrometry, Researchers developed a diagnostic algorithm based on machine learning. By combining primary metabolites and phospholipids, they enhanced the accuracy of cancer diagnosis and the area under the receiver operating characteristic curve. The incorporation of 36 statistically significant metabolites as a combined biomarker led to a significant 97.4% improvement in results. This system offers an efficient screening approach for pancreatic ductal adenocarcinoma (PDAC).

Discovered by Takahashi et al. [26] Long noncoding RNAs (lncRNAs) play a role in the epithelial-mesenchymal transition (EMT) in pancreatic ductal adenocarcinoma (PDAC). A lncRNA, known as HULC, has been identified as a future biomarker for PDAC. HULC was observed to have elevated expression levels, induced by transforming growth factor-β, in both PDAC cells and their extracellular vesicles (EVs). Suppression of HULC led to reduced invasion and migration of PDAC cells by inhibiting the process of epithelial-mesenchymal transition (EMT). The encapsulation of HULC in EVs may hold promise for aiding in the diagnosis of human PDAC. PDAC poses a significant threat as it is corelated with mortality. Current diagnostic tests and prognostic tests are limited due to the lack of reliable biomarkers. The analysis of methylation in cell-free DNA (cfDNA) holds great promise as a non-invasive method for detecting specific patterns related to disease in pre-neoplastic lesions and chronic pancreatitis.

Brancaccio et al. [27] present a review that delves into the benefits and challenges of cfDNA methylation studies, as well as the latest developments in identifying early diagnostic or prognostic biomarkers for Pancreatic Cancer. In the provision of patient care with invasive malignancies, biomarkers are crucial. Pancreatic Cancer with a dismal prognosis is because of its advanced stage and lack of available treatments. The lack of predictive biomarkers and proven screening tools for early diagnosis and targeted therapy exacerbates this. In the last two decades, there has been a growing body of information on biomarkers related to pancreatic cancer. The limited sensitivity and specificity of CA 19-9 have hindered its exclusive approval as the sole biomarker for diagnosis and evaluating treatment response up to this point.

Hasan et al. [28] cover emerging and existing biomarkers in this article that may be vital for early diagnosis, prognostic, and predictive indications of pancreatic cancer. CD73, a protein that attenuates tumor immunity, has been studied for its role in PC.

Chen et al. [29] found that patients with higher CD73 expression had poorer overall survival and disease-free survival. CD73 was also associated with reduced methylation and higher PD-L1 expression. This suggests that CD73 may be a biomarker for response to anti-PD-1/PD-L1 treatment in PC. Blockade of CD73 could be a promising therapeutic strategy for PC.

The results that we have seen in the literature indicate that different biomarkers are employed to identify PC. Numerous publications present systems for PC categorizations and early detections is done by using noncoding RNSs and the biomarker CA 19-9 as input. Th3dese systems are based on various machine learning and deep learning techniques. Urine proteome indicators are the main focus of this proposed system.

III. METHODOLOGY

PC is a devasting disease with a high risk rete, predicting at early is very crucial for patient.Accurate diagnosis is improve the chances to give perfect treament,recently ML techniquies have gained prominence in the health sector. By consider the publicly available dataset urinary biomarkers related to PC from Kaggle repository. To create any model frist step is the collection and preprocessing of data. After preprocessing step, data is splitted as training data and testing data(80% & 20%) .The parameters which influences the performance of the model are tuned by the hyperparameter tuning method gridsearchCV, these parameters are pass to the model as input data.GridsearchCV helps to fine-tune the parameters and optimizing its performance.A hybrid model is developed for classification of PC, with random forest and Ada Boost giving a promising approach to enhance the accuracy of diagnosis.

The created hybrid model demonstrates the major influence on early PC detection and, in the end, enhances patient outcomes in the PC battle. Fig. 1 shows the symbolic representation for the suggested pancreatic cancer classification system.

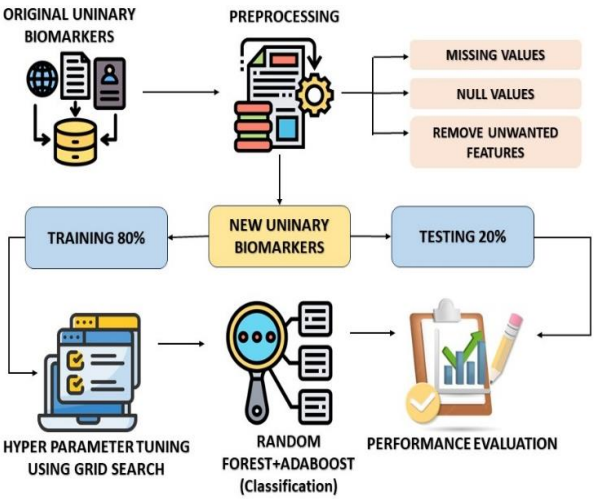

Fig. 1. Proposed pancreatic cancer classification system.

A. Dataset Description

The urinary biomarkers related to PC dataset are collected from the publicly available data repository Kaggle. Table I shows the urinary biomarkers dataset attributes and information. The data collection, supported by the biomarker panel, involved 590 urine samples gathered from various sources, including the University College London, the Barts Pancreas Tissue Bank, the University of Liverpool, Cambridge University Hospital, and the University of Belgrade, the Spanish National Cancer Research Centre. These samples comprised control samples 183, from individuals with 108 benign hepatobiliary conditions (including 119 from chronic pancreatitis patients), and 199 from individuals managed with pancreatic ductal adenocarcinoma (PDAC). The dataset has total 14 columns and 509 rows. The columns reflect numerous cancer risk factors, whereas the rows represent the study samples. The sixteen variables used in this data collection are as follows: the sample's ID. A cohort, which is a distinctive string used to identify each participant. Samples from cohort 1 have been utilized before. Samples from Cohort 2 have been added, and they are from the same above-mentioned sources. Sample_ Origin from where the sources are collected. Age, Sex, Diagnosis, Stage, Benign_Sample_Diagnosis, Blood plasma levels of the monoclonal antibody CA 19-9 are typically high in people with pancreatic cancer. There were just 350 subjects examined. Creatinine serves as an indicator of renal function in urine. Moreover, the presence of Lymphatic Vascular Endothelial Hyaluronan Receptor 1 (LYVE1) protein was identified in urine, indicating a potential significance in tumor metastasis. Additionally, the study quantified the urinary levels of proteins appended to pancreatic regeneration, specifically REG1A and REG1B. Furthermore, the urinary levels of Trefoil Factor 1 (TFF1), which might be related to the urinary tract, were assessed in a cohort of 306 patients.

Pancreatic cancer is categorized into stages such as IA, IB, IIA, IIB, III, and IV. However, in the dataset some input variables can be designated as irrelevant. Specifically, "sample id," automatically generated by Neural Designer, "Sample origin" signifies the source of patient samples, having no bearing on the ultimate diagnosis. "Stage" is a parameter exclusive to cancer patients. "Patient cohort" doesn't show any influence on the Sample Diagnosis, and "benign sample diagnosis" is an unrelated feature to the final sample diagnosis.

TABLE I. URINARY BIOMARKERS DATASET ATTRIBUTES AND INFORMATION

S. No	Feature	Specifies
1	Sample_id	Patient's id
2	Patient_cohort	To identify each participant
3	Sample_Origin	Where the patient samples came from
4	age	Patient's age, in years
5	sex	F-female M-male
6	diagnosis	Patient is having cancer
7	stage	Cancer stages IA, IB, IIA, IIIB, III, and IV
8	Benign_Sample_Diagnosis	Patient's not having cancer
9	plasma_CA19_9	Blood plasma levels of antibody CA_19-9
10	creatinine	Urine indicator of renal function
11	LYVE1	A protein in urine
12	REG1B	Pancreatic regeneration associated protein
13	TFF1	Urinary Trefoil Factor 1

The dataset consists of urine samples (590), classified into three variant groups of patients. These groups include healthy stage with 183 samples, and individuals with Stage of Benign and Stage of PDAC cases, 208 and 199 samples, respectively. Table II is a visual representation of this breakdown.

TABLE II. CHARACTERISTICS OF THE CLINICAL DATASET COMPRISING SAMPLES OF URINE FROM INDIVIDUALS WITH PANCREATIC CONDITIONS IN PROPOSED STUDY

Health Condition	Number of Sample	Gender	Range of age (median value)
Healthy stage	183	F (115) M (68)	26 – 89 (58) 30 – 87 (55)
Stage Benign	208	F (101) M (107)	26 – 82 (53) 29 – 82 (55)
Stage PDAC	199	F (83) M (116)	42 – 88 (68) 29 – 87 (67)

Fig. 2 shows the gender distribution of patients in different stages whereas Fig. 3 shows PC stage gender distribution on dataset.

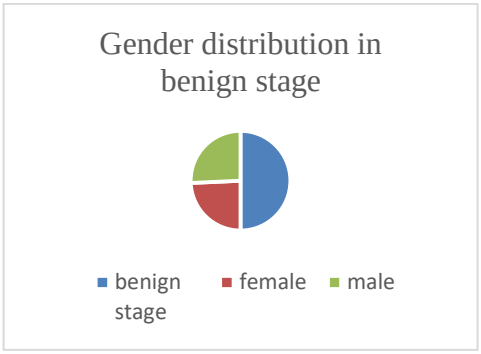

Fig. 2. Benign stage gender distribution on dataset.

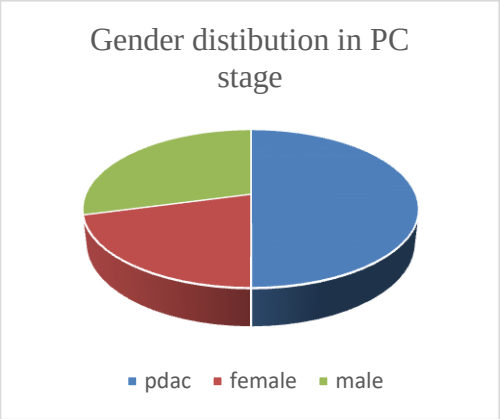

Fig. 3. PC stage gender distribution on dataset.

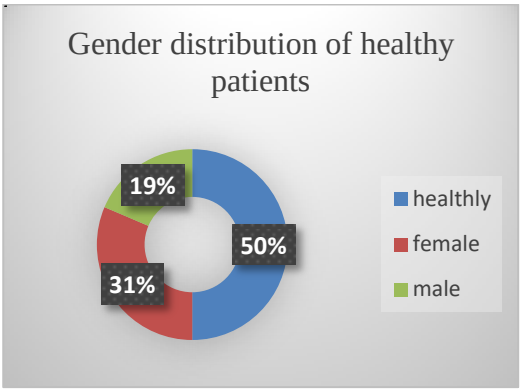

Fig. 4. Healthy patients gender distribution on dataset.

B. Preprocessing

Preprocessing is a crucial preliminary stage in data analysis. The situations where the dataset contains irrelevant, incomplete (missing), or inconsistent data, preprocessing becomes necessary. The preprocessing stage typically includes the following steps:

1) Data cleaning: This involves addressing missing values, which can be done by either omitting the entire data entry, replacing the missing value with a specific value, or employing strategies to manage null values. Inconsistencies in the dataset may also be resolved manually.

2) Data transformation: Converting data from one format to another format is called data transformation. When required, numerical data may undergo transformations for categorical variables, one-hot encoding is often employed to convert them into a suitable format for analysis.

In this study at data cleaning step the number of unwanted data, null data and missed data were cleaned. At transformation step one-hot encoding is used to covert char data into integer data.

Fig. 5 displays the graph that illustrates the correlation between each feature in the dataset and shows the degree of dependence between a variable to other variable used in the study. According to the color scale, the correlation relationship between the columns close to white is high, while the correlation relationship decreases gradually in the colors towards red.

C. Random Forest (RF)

RF is one of the best classification algorithms in Machine-learning. Instead of relying solely on output of a one decision tree, the random forest algorithm aggregates predictions from multiple decision trees then ultimately makes predictions based on the majority vote among these individual tree predictions. Fig. 6 shows random Forest classification. RF is a classification method that consists of a collection of decision trees, each built on different subsets of the dataset. By aggregating the results from these trees, often by taking their average, it enhances the predictive accuracy for the given dataset. But, the classification of random forest shows the low accuracy for the given dataset.

In RF classification Gini index is decide how nodes are distributed on a decision tree branch. It determines the branches is more likely to occurs and find the class and probability of a node in each branch.

$$\text{Gini} = 1 - \sum_{i=1}^{j} f(i)^2 \quad (1)$$

f is the class of frequency in the dataset.

c is the number of classes.

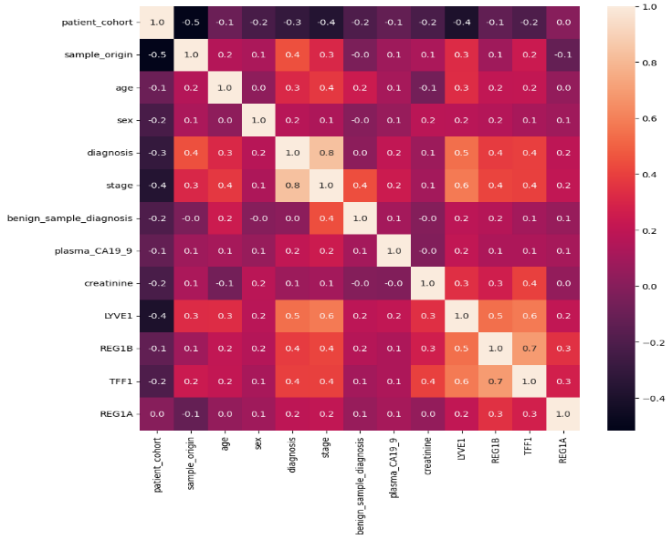

Fig. 5. Matrix of correlation.

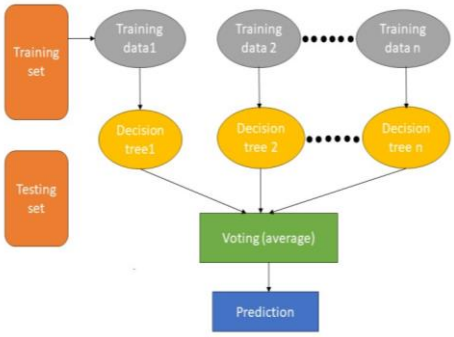

Fig. 6. Random forest classification.

There is another way to determine how nodes branch in a decision tree are determined is entropy. Entropy determines which branch the node should follow by analyzing the probability of a specific outcome. Its calculation involves the use of a logarithmic function, making it more mathematically sophisticated than the Gini index.

$$\text{Entropy } = \sum_{i=1}^{c} -p_i \log(p_i) \quad (2)$$

where, the p_i is the probability of randomly selecting a node in class i.

Information gain is a metric used to identify the most informative features in a dataset, and it is depended on the entropy value. It quantifies the difference in entropy before and after a data split and, in doing so, measures the impurity or disorder within class elements. Essentially, information gain helps determine which feature, when used to split the data, leads to the greatest reduction in uncertainty or entropy, making it a valuable criterion for feature selection in decision tree-based algorithms.

Information Gain= Entropy before splitting- Entropy after splitting

Consider the following pseudo code for RF classification.

Pseudo code of random forest classification.

Step 1: F features are select from the feature set randomly.

Step 2: In F each of x.

- Find the Gain.

$$\text{Gain } (s, x) = E(s) - E(s, x)$$

$$E(s) = \sum_{i=1}^{c} -f_i \log(f_i)$$

$$E(s, x) = \sum_{c \in X} P(c) E(c)$$

where E(s) two classes entropy, E(s,x) is the entropy of feature x..

- Choose the node with the highest information gain, denoted as y.
- Divide the node into sub-nodes.
- To build the tree, iterate through steps a, b, and c until the minimum required number of samples for splitting is reached.

Step 3: To create a forest consisting of n trees, replicate steps 1 and 2 n times.

The performance of RF model with all selected features is 83%. To improve the classification accuracy hyper parameter tuning is applied to evaluate the optimal parameters of the model. To find the best hyper parameters to build a single model which is suitable to improve the RF classification Grid search CV is one of the suitable hyper parameter techniques for the RF classification.

D. Hyperparameter Tuning

Hyperparameter tuning may be automated with the use of GridSearchCV, which also improves model performance and does away with manual trial-and-error. Fig. 7 shows grid search across two parameters. Grid Search combines each of the supplied hyperparameters and their values in a different way, calculates the performance of each combination, and then chooses the hyperparameters with the best value.

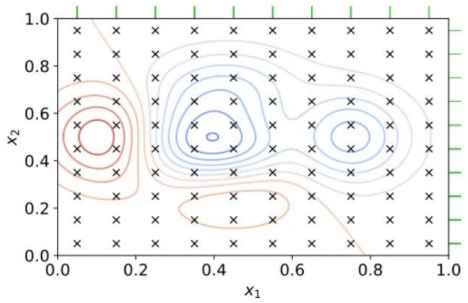

Fig. 7. Grid search across the two parameters.

Within the scikit-learn library's model_selection module, you can find a function known as GridSearchCV ().

Initiating the execution can be done by creating a GridSearchCV () object. In the scikit-learn model_selection class, there exists a function known as GridSearchCV (). The process begin can be begin by creating a GridSerachCV () object.

Clf=GridSearchCV (estimator, param_grid, cv, scoring).

The necessary four inputs are estimator, param_grid, cv, scoring.

These arguments are explained as follows:

1) estimator: A scikit-learn model

2) param_grid: A dictionary associating parameter names with lists of parameter values.

3) cv: A numeric value that specifies the number of folds in K-fold cross-validation.

4) scoring: Performance measure.

By using the GridSearchCV () prepare a best random model for predicting the pancreatic cancer efficiently. To obtain the better performance develop a hybrid model with the grid search cv random model with the help of Boosting algorithms. Ada Boost is one of the best boosting algorithms to improve the classification performance.

E. Boosting Algorithm

Ada Boost is one of the boosting algorithm in this one common estimator is used to learn for the decision trees in every split. By using with the splits, it builds a model with same weights in all levels if there are any high weights are assigned then those are called wrongly classified. In this context give the importance for high weights training with another model until reaches the minimum error. The argument is that while AdaBoost can be used independently of decision trees, it is frequently combined with them in reality because of the advantages that stumps have over other weak learners.

The mathematical summary of Ada Boost algorithm as follows:

1. Initializing Weights

   Dataset with samples of N, give each data point weight with $w_i = 1/N_s$.

   For each m=1 to M:

   Dataset sampled by the weight $w_i^{(m)}$ to attain training samples T_s

For each datapoint, AdaBoost initializes a weight of $1/N_s$. Once more, the weight of a datapoint indicates the likelihood that it will be chosen during sampling.

2. Training Weak Classifiers

   For all the training samples T_s, fit a classifier X_m

   X_m is a weak classifier then trained using the training dataset.

3. Here the AdaBoost algorithm is begins. The weights are updated as follows

$$\epsilon = \frac{\sum_{yi \neq K_m(x_i)} w_i^{(m)}}{\sum_{yi} w_i^{(m)}} \quad (3)$$

y_i is target variable ground truth value

w_i^m is the sample weight i at m^{th} iteration

the weight is updated at certain iteration m

$$\alpha_m = \frac{1}{2} \ln \frac{1-\epsilon}{\epsilon} \quad (4)$$

$$w_i^{(m+1)} = w_i^{(m)} e^{-\alpha_m y X_m(T_s)} \quad (5)$$

y represents the true values of the targent feature

$X_m(T_s)$ prediction made by the stump at iteration m

α_m predictive power of stump m

4. Making new overall Predictions

$$K(x) = \text{sign}[\sum_{m=1}^{M} \alpha_m X_m(x)] \quad (6)$$

The proposed hybrid classification model using a combination of a Random Forest classifier and an AdaBoost classifier developed. The Random Forest classifier is first tuned using gridsearchCV, and then the AdaBoost classifier is used to enhance the accuracy of RF. Finally, the working manner of the combined model is evaluated using metrics like accuracy, a classification report, and a confusion matrix.

IV. Results and Discussions

There are several researches are going on urinary biomarkers to determine whether they belonged to the pancreatic cancer or healthy patients (see Fig. 4). The publicly available dataset is used for the implementation of the proposed approach. The proposed system uses the hyperparameter tuning technique gridsearchCV to select the hyperparameters from the dataset, then perform the classification with RF and use AdaBoost to improve the classification accuracy. Training data make up 80% of the dataset, whereas testing data make up 20%. There are three subsections within the section: evaluation metrics, results, and comparison with other results of other approaches. This study employs classification algorithms and assesses their performance using metrics such as Accuracy, Recall, Precision, and the F-1 score.

Let TP (True Positive) represent the outcome in which the model correctly predicts the positive class, TN (True Negative) represents the outcome in which the model correctly predicts the negative class, FP (False Positive) represent the outcome in which the model incorrectly predicts the positive class, and FN represent the outcome in which the model incorrectly predicts the negative class. The above-mentioned performance measurements may thus be described as follows:

1) Accuracy: Accuracy, one of the most straightforward classification metrics, is computed as the ratio of correct predictions to all predictions made.

For calculation:

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

2) Recall: True Positives represent predictions that correctly match the total number of positive cases, whether they were accurately predicted as positive or incorrectly predicted as negative (True Positives and False Negatives). The following is the calculation method for recall:

$$\text{Recall} = \frac{TP}{TP+FN}$$

3) Precision: Precision measures the proportion of positive predictions that were accurate. It can be expressed as the ratio of all correctly predicted positive instances (True Positives and False Positives) to the total number of positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \times 100$$

4) F1-Score: Calculating the F1 Score involves taking the harmonic mean of Precision and Recall, giving equal weight to both metrics. The formula for computing the F1 score is as follows:

$$\text{F1- score} = 2 * \frac{precision * recall}{precision + recall}$$

5) ROC Curve: The classification accuracy also computed by the Receiver-Operating-Characteristic curve. It's constructed using True-Positive (TP) and False-Positive (FP) at different target levels. TP is determined by recall, while FP is established using fallout. This illustrates how a ROC curve plots the fallout and sensitivity of the classifier. Fig. 8 shows ROC analysis of the proposed system.

6) Confusion matrix: A confusion matrix is a common tool used in machine learning and statistics to evaluate the performance of a classification algorithm, particularly in the context of binary classification. It provides a clear and detailed breakdown of how a classifier's predictions align with the actual class labels in a dataset. A confusion matrix (see Fig. 9) is typically represented as a 2x2 matrix, and it contains four key metrics: TP, TN, FP, FN.

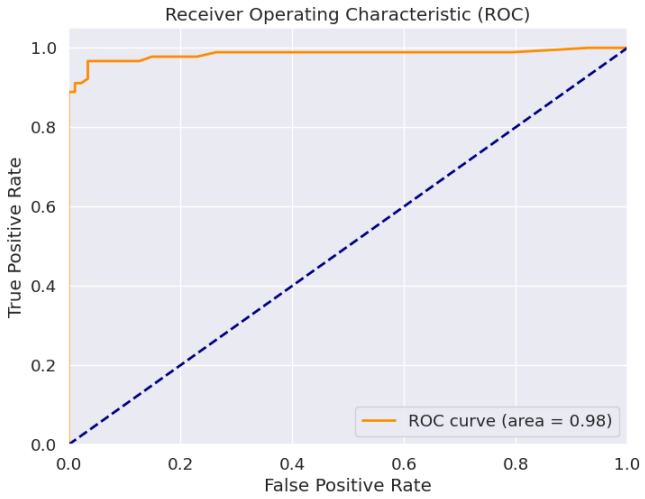

Fig. 8. ROC analysis of the proposed system.

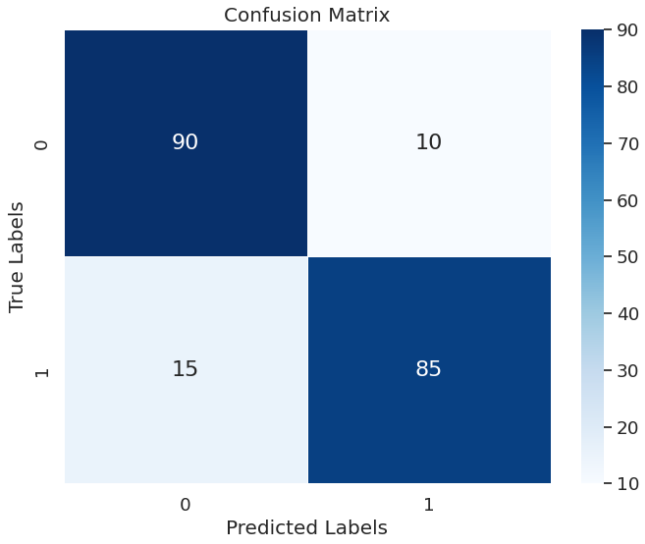

Fig. 9. Confusion matrix of proposed system.

Lately, machine learning techniques have seen substantial adoption in healthcare, particularly in the realm of cancer diagnosis. Utilizing hybrid ensembled classifiers, creatinine, LYVE1, REG1B, and TFF1 urine biomarkers have been effectively analyzed to detect pancreatic cancer. The proposed hybrid ensembled model surpasses other basic models with the maximum accuracy score of 97% in prior studies, as shown in Table III's assessment findings. Conventional ML models, such as LR, RF, and SVM, performed inadequately to correctly detect conditions related to pancreatic cancer. Therefore, hybrid ensembled classification model has been developed to predict pancreatic cancer based on urine biomarkers. As described above, the advantageous of gridsearchCV have the capability to ignore unnecessary feature and tune the feature which are very efficient to detect pancreatic cancer from the urine biomarkers.

7) Comparative analysis: The Table III offers an extensive comparison of the proposed model with established methods, conducting a thorough analysis of classification outcomes. Utilizing state-of-the-art classifiers, the proposed system is comprehensively evaluated in terms of accuracy, recall, precision, and the F1 score. Fig. 10 shows the comparative analysis of proposed system upon different classifiers based on urinary biomarkers.

TABLE III. A COMPARATIVE EVALUATION OF THE PROPOSED APPROACH WITH DIFFERENT CLASSIFIERS FOR PANCREATIC CANCER CLASSIFICATION USING URINARY BIOMARKERS

Classifier	Recall	Precision	F1-score	Accuracy
Random Forest	87	79	82	75
LR	76	64	89	74
KNN	64	65	-	64
GBC	77	73	-	72
1D CNN	100	90	95	93
LSTM-1D CNN	100	96	98	97
GridsearchCV+ RF+Ada Boost	**99.98**	**99.98**	**99.98**	**99.98**

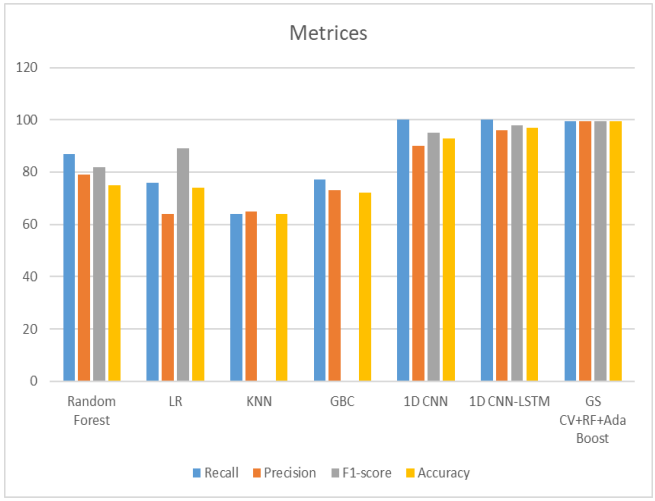

Fig. 10. Comparative analysis of proposed system upon different classifiers based on urinary biomarkers.

V. CONCLUSION

The scarcity of publicly available medical datasets is a significant challenge to the training of supervised learning models, particularly with regard to pancreatic cancer, as a result of which the only available training model performs poorly in terms of classification accuracy. A new ensemble strategy that uses urinary biomarkers gathered from Kaggle, all efficient features from dataset are tuning based on gridsearchCV then to classifies the pancreatic cancer and non-pancreatic cancer uses random forest, to improve performance of classification boosting algorithm Ada Boost was proposed. Metrics used to gauge the proposed model's performance Precision, F-1 score, Accuracy, and Recall all scored 99.98.

In future the machine learning models obtained optimum performances to detect pancreatic cancer at early stages automatically which are helpful for both clinicians and patients to save their life. This type of systems may assure prominent results in real time medical scenarios.

REFERENCES

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72(1):7–33.

[2] Chang YH, Margolin A, Madin O, et al. Deep learning-based nucleus classification in pancreas histological images. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2017, p. 672–675.

[3] Gupta, Anish, Apeksha Koul, and Yogesh Kumar. "Pancreatic cancer detection using machine and deep learning techniques." *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*. Vol. 2. IEEE, 2022.

[4] S. Fukushige and A. Horii, ''Road to early detection of pancreatic cancer: Attempts to utilize epigenetic biomarkers,'' Cancer Lett., vol. 342, no. 2, pp. 231–237, Jan. 2014.

[5] M. Kalubowilage et al., ''Early detection of pancreatic cancers in liquid biopsies by ultrasensitive fluorescence nanobiosensors,'' Nanomedicine, Nanotechnol., Biol. Med., vol. 14, no. 6, pp. 1823–1832, Aug. 2018.

[6] S. Boeck, D. P. Ankerst, and V. Heinemann, ''The role of adjuvant chemotherapy for patients with resected pancreatic cancer: Systematic review of randomized controlled trials and meta-analysis,'' Oncology, vol. 72, nos. 5–6, pp. 314–321, 2007.

[7] Malhotra A, Rachet B, Bonaventure A, Pereira SP, Woods LM. Can we screen for pancreatic cancer? Identifying a sub-population of patients at high risk of subsequent diagnosis using machine learning techniques applied to primary care data. PLoS One. 2021;16(6):e0251876.

[8] Karar ME, Alotaibi B, Alotaibi M. Intelligent medical IoT-enabled automated microscopic image diagnosis of acute blood cancers. 2022;22(6):2348

[9] Lee J, Lee HS, Park SB, Kim C, Kim K, Jung DE, Song SY: Identification of circulating serum miRNAs as novel biomarkers in pancreatic cancer using a penalized algorithm. Int J Mol Sci. 2021; 22:1007.

[10] Reddy, Santosh, and M. Chandrasekar. "PAD: A Pancreatic Cancer Detection based on Extracted Medical Data through Ensemble Methods in Machine Learning." *International Journal of Advanced Computer Science and Applications* 13.2 (2022).

[11] Lepowsky E, Ghaderinezhad F, Knowlton S, Tasoglu S. Paper-based assays for urine analysis Biomicrofuidics. 2017;11(5): 051501

[12] for urine analysis Biomicrofuidics. 2017;11(5): 051501. 17. Radon TP, Massat NJ, Jones R, Alrawashdeh W, Dumartin L, Ennis D, Dufy SW, Kocher HM, Pereira SP, Guarner L, et al. Identifcation of a three-bio- marker panel in urine for early detection of pancreatic adenocarcinoma. Clin Cancer Res. 2015;21(15):3512–21.

[13] Debernardi S, O'Brien H, Algahmdi AS, Malats N, Stewart GD, PlješaErcegovac M, Costello E, Greenhalf W, Saad A, Roberts R, et al. A combina- tion of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case–control study. PLoS Med. 2020;17(12): e1003489.

[14] Lee J, Lee HS, Park SB, Kim C, Kim K, Jung DE, Song SY: Identification of circulating serum miRNAs as novel biomarkers in pancreatic cancer using a penalized algorithm. Int J Mol Sci. 2021; 22:1007.

[15] Long NP, Jung KH, Anh NH, Yan HH, Nghi TD, Park S, Yoon SJ, Min JE, Kim HM, Lim JH et al: An integrative data mining and omics-based translational model for the identification and validation of oncogenic biomarkers of pancreatic cancer. Cancers. 2019; 11:155.

[16] Debernardi S, Massat NJ, Radon TP, Sangaralingam A, Banissi A, Ennis DP, Dowe T, Chelala C, Pereira SP, Kocher HM, et al. Non-invasive urinary miRNA biomarkers for early detection of pancreatic adenocarcinoma. Am J Cancer Res. 2015;5(11):3455–66

[17] Ko J, Bhagwat N, Yee SS, Ortiz N, Sahmoud A, Black T, Aiello NM, McKen- zie L, O'Hara M, Redlinger C, et al. Combining machine learning and nanofluidic technology to diagnose pancreatic cancer using exosomes. ACS Nano. 2017;11(11):11182–93

[18] Lee H-A, Chen K-W, Hsu C-Y: Prediction model for pancreatic cancer-A population-based study from NHIRD. Cancers. 2022; 14:882.

[19] Agarwal, Deepesh, et al. "Early detection of pancreatic cancers using liquid biopsies and hierarchical decision structure." *IEEE Journal of Translational Engineering in Health and Medicine* 10 (2022): 1-8.

[20] Thanya, T., and Wilfred Franklin S. "Novel computer aided diagnostic system using hybrid neural network for early detection of pancreatic cancer." *Automatika* 64.4 (2023): 816-827.

[21] Karar, Mohamed Esmail, Nawal El-Fishawy, and Marwa Radad. "Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks." *Journal of Biological Engineering* 17.1 (2023): 28.

[22] Blyuss O, Zaikin A, Cherepanova V, Munblit D, Kiseleva EM, Prytomanova OM, Dufy SW, Crnogorac -Jurcevic T. Development of PancRISK, a urine biomarker -based risk score for stratifed screening of pancreatic cancer patients. Br J Cancer. 2020;122(5):692–6.

[23] Jiao, Yan, et al. "ITGA3 serves as a diagnostic and prognostic biomarker for pancreatic cancer." *OncoTargets and therapy* 12 (2019): 4141.

[24] Liu, Yawen, et al. "Circulating lncRNA ABHD11-AS1 serves as a biomarker for early pancreatic cancer diagnosis." *Journal of Cancer* 10.16 (2019): 3746.

[25] Iwano, Tomohiko, et al. "High-performance collective biomarker from liquid biopsy for diagnosis of pancreatic cancer based on mass spectrometry and machine learning." *Journal of Cancer* 12.24 (2021): 7477.

[26] Takahashi, Kenji, et al. "Circulating extracellular vesicle-encapsulated HULC is a potential biomarker for human pancreatic cancer." *Cancer science* 111.1 (2020): 98-111.

[27] Brancaccio, Mariarita, et al. "Cell-free DNA methylation: the new frontiers of pancreatic cancer biomarkers' discovery." *Genes* 11.1 (2019): 14.

[28] Hasan, Syed, et al. "Advances in pancreatic cancer biomarkers." *Oncology reviews* 13.1 (2019).

[29] Chen, Qiangda, et al. "CD73 acts as a prognostic biomarker and promotes progression and immune escape in pancreatic cancer." *Journal of cellular and molecular medicine* 24.15 (2020): 8674-8686.