

Evaluating Hyperparameter Tuned Machine Learning Classifiers for Pancreatic Cancer Detection via Urinary Biomarkers

By Nivriith Ananth Iyer

Table of Contents

Abstract	3
Introduction	4
Related Works	5
Data Sets and Methods	6
Experiments and Discussion	14
Conclusion	19
Acknowledgments	20
References	21

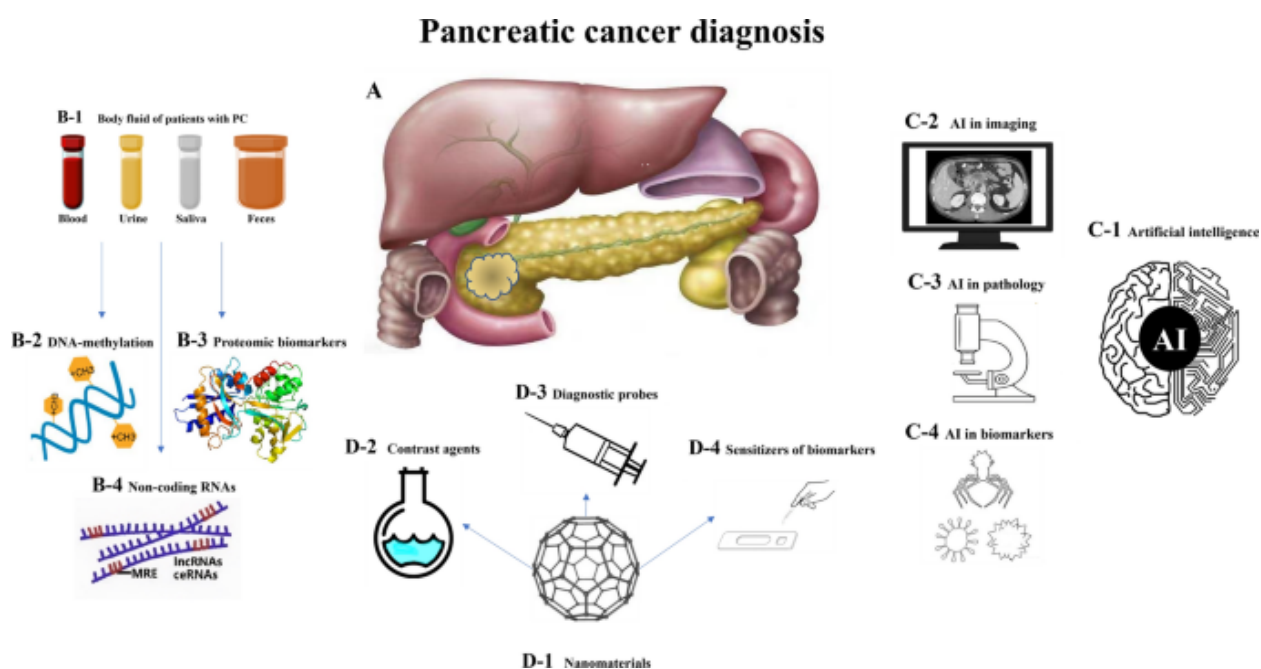
Abstract

Pancreatic cancer, though accounting for only 3% of all cancers in the US, is highly lethal. Early detection is critical for effective treatment, with up to 44% recovery when identified before tumor maturation. Current diagnostic tools, mainly blood tests, often lack sensitivity for early detection, highlighting the need for a new approach to diagnostic tests. This study explores urinary biomarkers as a promising alternative, specifically creatinine, LYVE1, REG1B, and TFF1. Hyperparameter tuning is employed to enhance the accuracy using a dataset from the Spanish National Cancer Research Center. The comparison includes gradient boosting models (XGBoost and LightGBM), a Random Forest Classifier, a support vector machine, and a 1D CNN-LSTM model. Results show that XGBoost and LightGBM perform equally well with an accuracy of 91% in early-stage pancreatic cancer detection. The other models tested achieved up to a 90% accuracy score with the exception of the 1D CNN-LSTM model, with a 78% accuracy. With the promising results of the 1D CNN-LSTM model being highlighted in other papers, the specific parameters and environment needed for the model come into question. Furthermore, extensions delving into the reasoning behind why gradient boosting models work the best can expand the future development of detection accuracy. This study highlights the potential of urinary biomarkers and the importance of model selection in improving pancreatic cancer detection accuracy.

Introduction

Pancreatic cancer, despite comprising about 3% of all cancers in the US, is a very deadly disease [1]. Despite the extreme dangers present with pancreatic cancer, it can be treated if the tumor is detected in its early stages before the tumor starts to mature [2]. Up to 44% of patients who have pancreatic cancer detected before tumors start to develop recover completely [3]. Unfortunately, many cases of pancreatic cancer show no symptoms until the cancer has metastasized, and by then survival will only last for 8-12 months [4]. Furthermore, the current diagnostic tool for pancreatic cancer is blood tests that don't allow for early detection of pancreatic cancer due to lower undetectable levels of certain gradients [5].

A promising alternative for pancreatic cancer detection is urinary biomarkers [6]. Four biomarkers can be found in urine that can be used to detect an abnormality in the pancreas [7]: (i) creatinine, a protein that is often used as an indicator of kidney function; (ii) lymphatic vessel endothelial hyaluronan receptor (YVLE1); (iii) Regenerating Family Member 1 Beta, a protein that is associated with pancreas regeneration; and (iv) a trefoil factor 1 (TFF1), which is related to regeneration and repair of the urinary tract. These four biomarkers have been experimented with to improve the accuracy of early-stage pancreatic cancer detection by up to 90% [8]. The current results can be boosted by performing hyperparameter tuning that could increase the accuracy of detection even further. Here, we examined a dataset from the Spanish National Cancer Research Center and tested out gradient boosting models to strive for an even higher accuracy than any current model of pancreatic cancer detection.



Courtesy of Cancer Cell International - BioMed Central: The diagram above describes how pancreatic cancer can be diagnosed: different protein structures, blood tests, proteomic biomarkers in blood and urine samples, and more. The right of the diagram describes the different ways pancreatic cancer can be diagnosed using AI models, with either imaging, pathology, or biomarkers. This project focuses specifically on training different AI models to detect pancreatic cancer based on the abundance and correlation of different biomarkers in urine samples.

Related Works

A research paper published by Plos Medicine features the dataset used for this analysis. The title of this research paper is, “A Combination of Urinary Biomarker Panel and PancRISK Score for Earlier Detection of Pancreatic Cancer: A Case-control Study,” written by fourteen researchers. This paper delves into the background of the project, explaining the importance of pancreatic cancer detection and the practicality of using urinary biomarkers to detect early-stage pancreatic cancer. With the help of the Spanish National Research Center and various universities, the researchers developed a model based on score collection and generating ROC curves for each biomarker, known as PancRISK. The scientists use the PancRISK model to test each biomarker and its correlation to the early-stage detection of pancreatic cancer. The researchers found out that all of the biomarkers except creatinine possessed an ROC curve higher than 90%, thus expressing the accuracy of the models. It can be noticed that there is promising evidence for the efficiency of urine tests and that extensions can be made upon this basis of information. Numerous research papers have based their information on this sole paper, for it serves as the backbone for urinary biomarker samples. Other research papers have utilized the dataset mentioned in this paper to test their accuracy models, and notable discoveries have been made regarding the increased accuracies in various tests. The research paper concludes with an extension into the precision score of the PancRisk model, which is yet to have a proper article from the same researchers.

A research paper following the last one’s research was published soon after. It can be found on Springer Link: “Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks.” This paper explores the accuracy performances of a proposed 1-D CNN + LSTM model which analyzes the dataset obtained from the Spanish National Research Center. The paper utilizes 183 healthy pancreas samples, 208 benign hepatobiliary disease samples, and 199 PDAC(Pancreatic Ductal Adenocarcinoma) samples. Compared to other prevalent models such as the KNN model, logistic regression, and gradient boosting models, this research paper believes in using a CNN model with a one-dimensional layer. The paper later details the accuracy score of the proposed CNN model: a 97% best accuracy score along with an AUC curve score of 98%. Compared to other research papers around the same topic and dataset, this paper utilized a deep-learning classifier with a one-dimensional layer, allowing such accuracy to be achieved. The LSTM, one of the most

popular architectures of recurrent neural networks, was utilized to manipulate the sequential data. The research paper uploaded goes into more context with the LSTM model, but by using a one-dimensional layer, it was able to achieve such a high accuracy with different layers constantly analyzing the training data to produce a highly accurate testing data result. This paper will analyze this model using a newly developed form to attempt validation of this research paper's original method with a novel system based on the 1D CNN-LSTM Model. Other papers have trained different models based on the categorical dataset, with most of them achieving a maximum of 90%. The most prevalent models that were used were either XGBoost models or LightGBM models. By using gradient-boosting artificial intelligence, these models were able to achieve on average higher accuracies [9]. The exact reason for this is not known but will be explored later on.

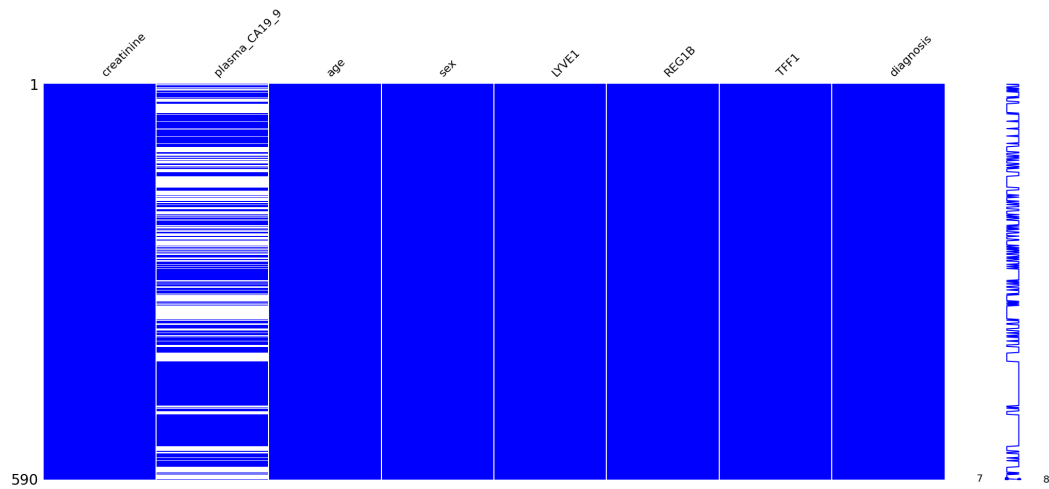
Since the code is open source, this file will be explored later in this study to confirm its validation along with different iterations of the code and their extensions.

Data Sets and Methods

The dataset used for this analysis was obtained from the Spanish National Cancer Research Center and uploaded onto Kaggle, the world's largest data science community, by John Davis [10]. The dataset contains three groups: healthy controls, patients with non-cancerous pancreatic conditions, and patients with pancreatic ductal adenocarcinoma. In addition, the dataset contains age, sex, and diagnosis groups, which can be used to show the demographics that tend to be linked with pancreatic cancer more often. Finally, the dataset contains the levels of each biomarker for their respective patients, cancerous or not. With 590 samples, 14 columns contain each patient's data in relation to the disease.

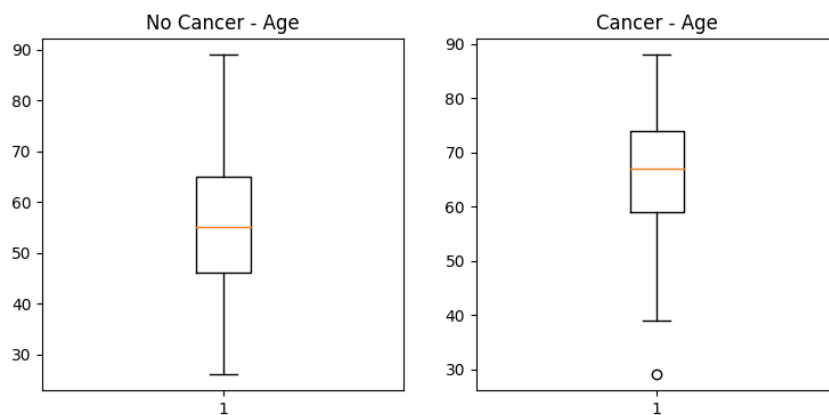
To examine the relationships in the dataset, a data visualization file was created in Google Colaboratory [11]. Initiating data comparisons will indicate any non-separable data. If too many of the false and true diagnosis points collide with each other, then a trustworthy accuracy will not be produced unless the data is manually separated.

A matrix was first developed to examine the abundance of each biomarker per patient.

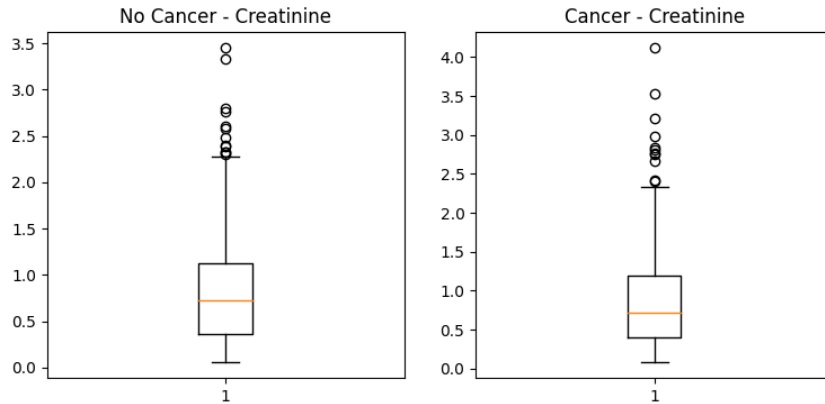


Plasma_CA19_9, a tumor marker, was the only group to not appear for every single patient, which makes sense since there are only a certain number of cancer patients who have tumors.

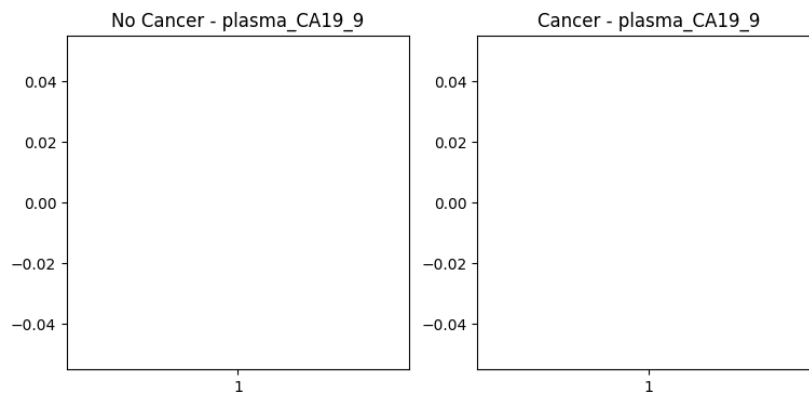
Sequentially, boxplots were created, differentiating each patient into either a cancer or no cancer group followed by their respective demographic.



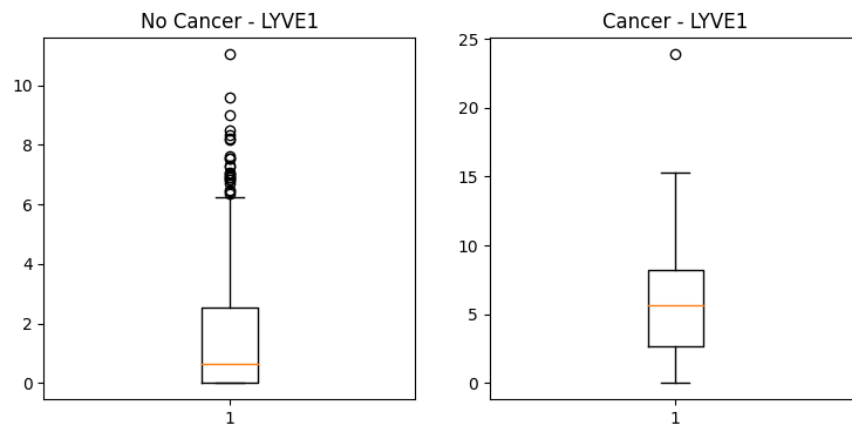
This comparison indicates that older people tend to be associated with pancreatic cancer more often. However, the extensions of both box plots suggest that such a difference is not proven to generally be true.



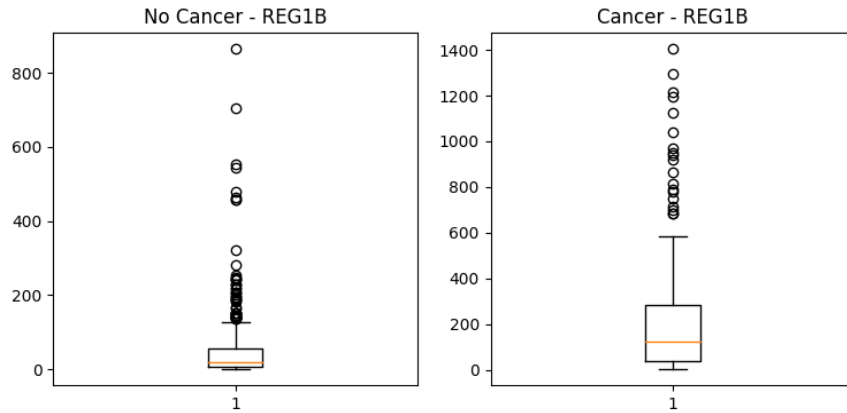
This comparison shows that patients with pancreatic cancer tend to have a generally lower source of creatinine, suggesting that pancreatic cancer patients have lower energy to complete daily tasks.



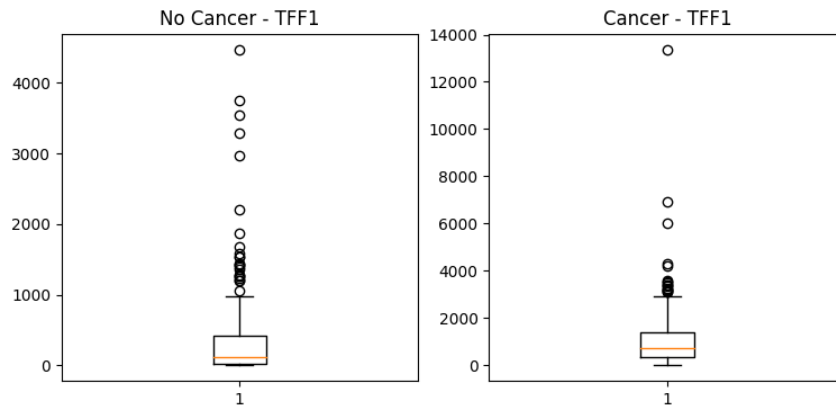
There is no boxplot drawn for either group because of a result of missing values in the dataset. In this data visualization notebook, the missing values are filled after this plot is created.



This comparison shows that pancreatic cancer patients typically have a higher level of LYVE1 than healthy patients. However, the boxplots are close in value to each other, so this can not be proven entirely.

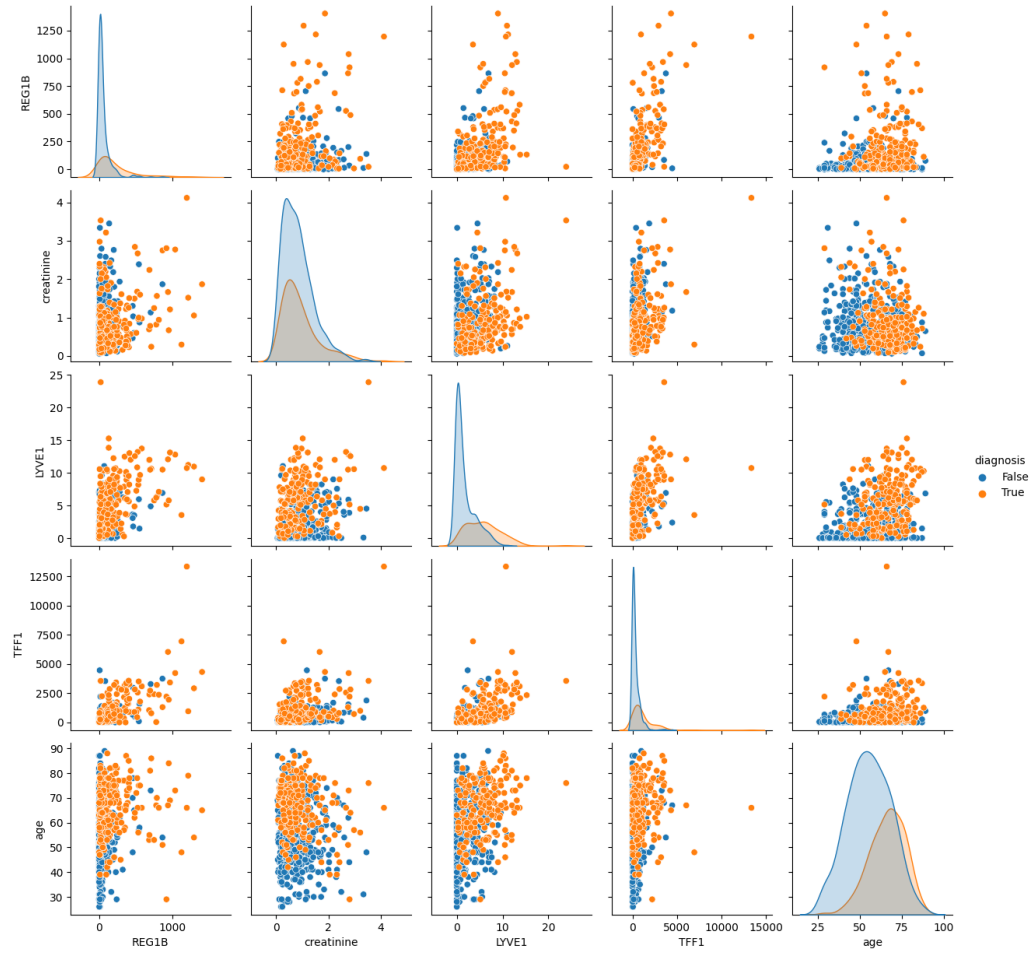


This comparison shows that pancreatic cancer patients tend to have higher levels of REG1B than healthy patients. However, the boxplots are very similar to each other, so this can not be generalized.

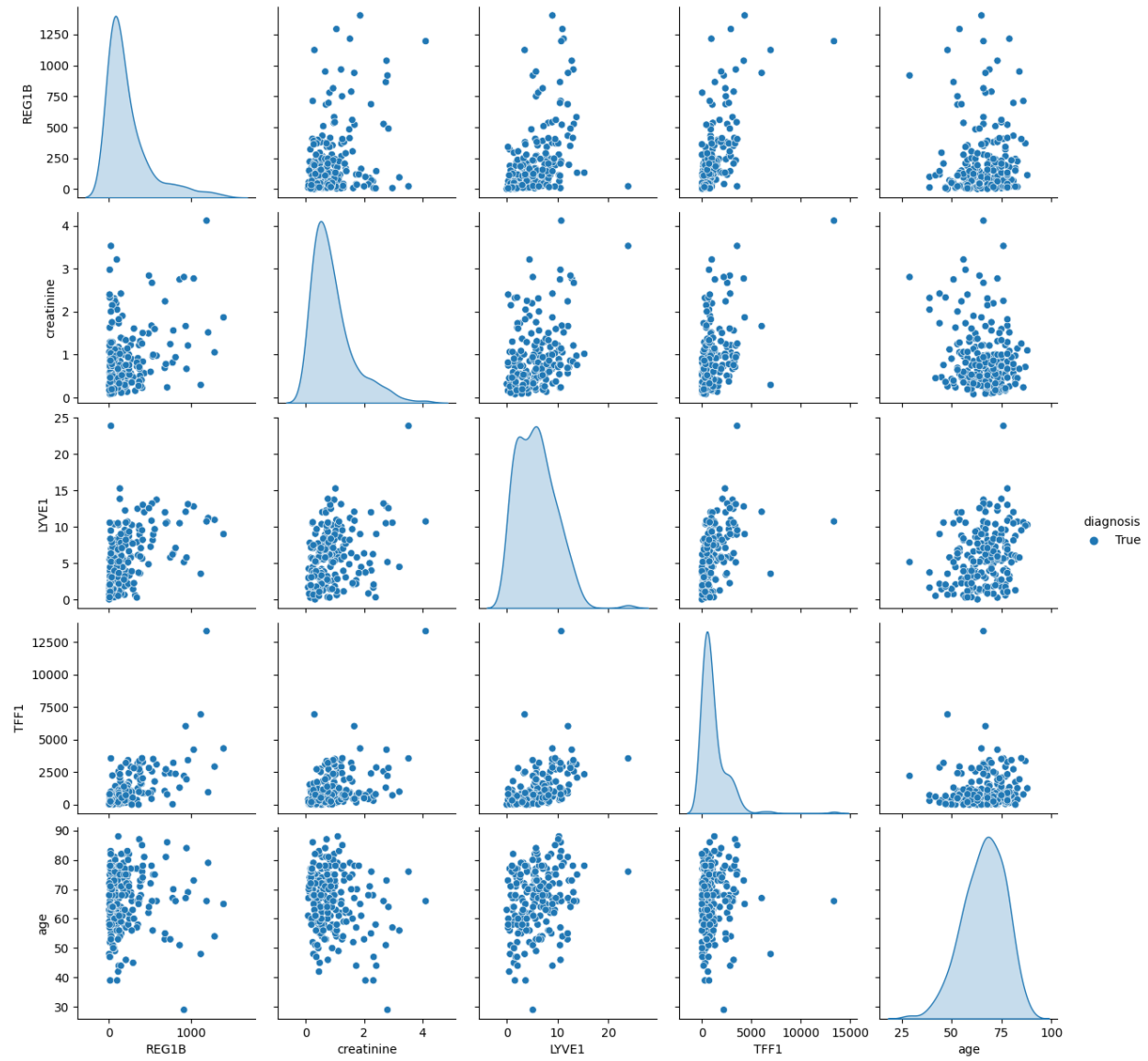


Regardless of having pancreatic cancer or not, all groups of patients seem to have the same levels of TFF1.

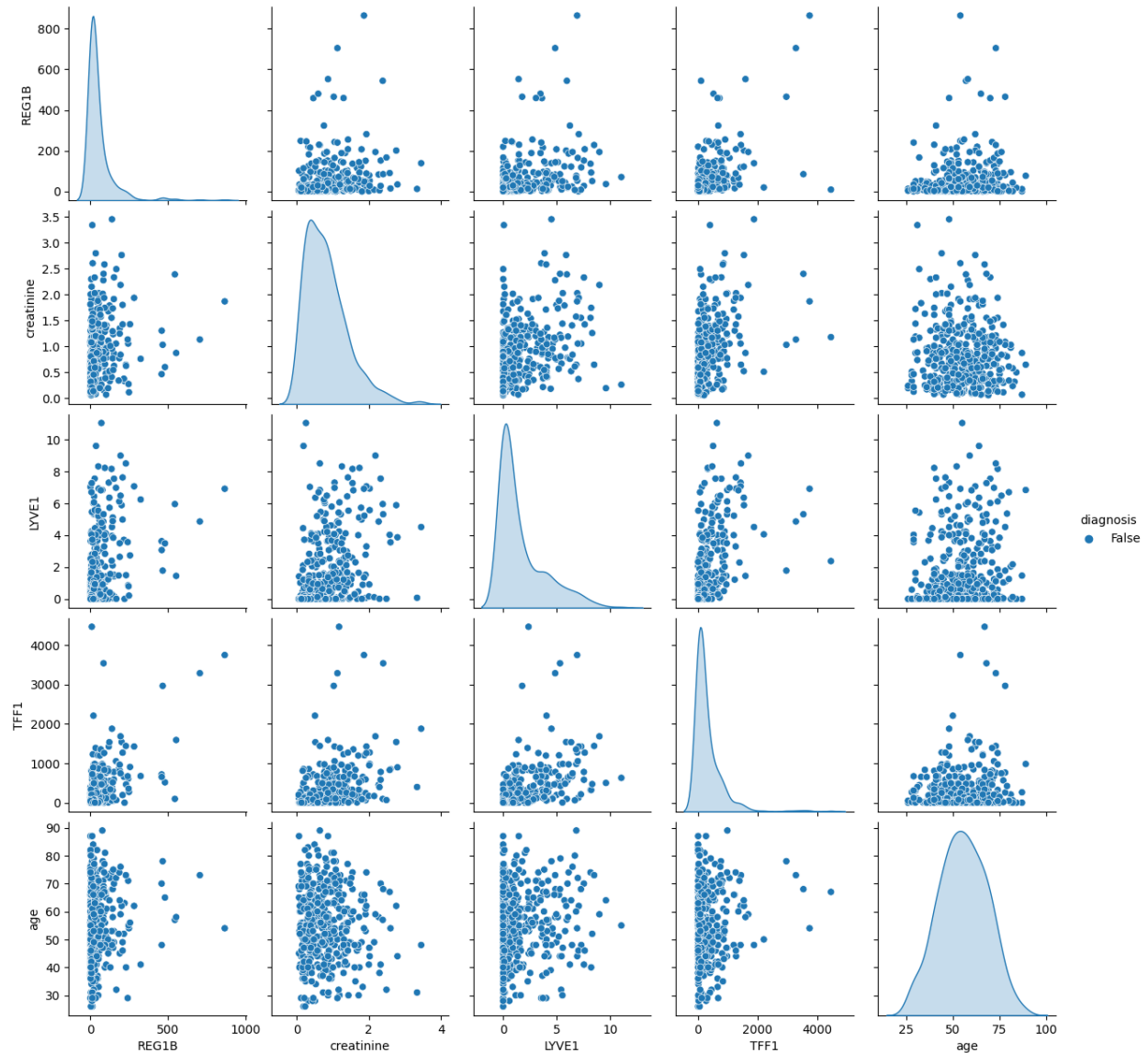
The boxplot comparisons seem to suggest that the data points, differentiated between true and false, seem to land at around the same point in each graph. This means that there is no separability between cancer and no-cancer data points for each group in the dataset. This is further confirmed with a pair plot:



Each graph in this pair plot seems to have both false and true diagnosis data points lying in the same spots. This means there initially seems to be no differentiation in the levels of the biomarkers tested for both cancer and non-cancer patients. Two pair plots only containing true and false data points respectively hint at this occurrence as well.



This pair plot shows all the general patterns of biomarker levels for cancer patients in more detail.



This pair plot shows all of the general patterns of biomarker levels for non-cancer patients in more detail.

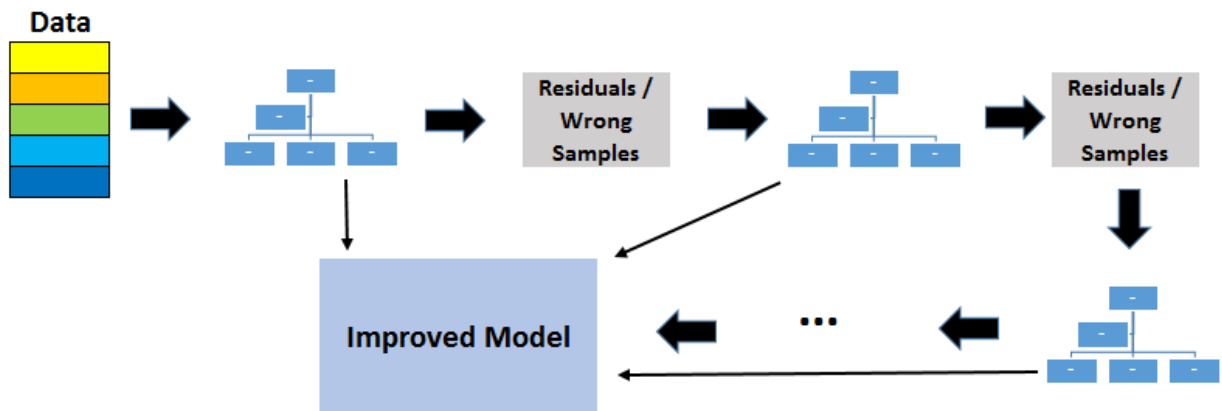
It is clear that the data must be separated into differentiable groups so that a proper accuracy score can be generated. This is because if models are tested on the dataset without any modification, then they will give automatic high scores since there is no clear difference between non-cancerous and cancerous patients, thus invalidating the accuracy score. The models will only base their results on the data points, which may cause overfitting in the tested models since the model will generalize true and false points to hold the same value when in reality, biomarkers associated with pancreatic cancer have more different levels than healthy patients [12].

Essentially, if the model isn't able to differentiate key characteristics that define whether a patient most likely has pancreatic cancer, then its accuracy score will ultimately not have much significance since the machine will base its results only on the data and not common patterns that happen in daily life.

Despite there being a lack of separability, this does not invalidate the Spanish National Cancer Research Center's results since their dataset values were properly evaluated during tests. The dataset being used for this analysis has to be separated so that the models used can differentiate cancer and non-cancer points more efficiently, and thus generate accuracy scores that are more valid. Teaching the machine which characteristics are more common in cancerous patients will boost its accuracy score and give us a more reliable percentage.

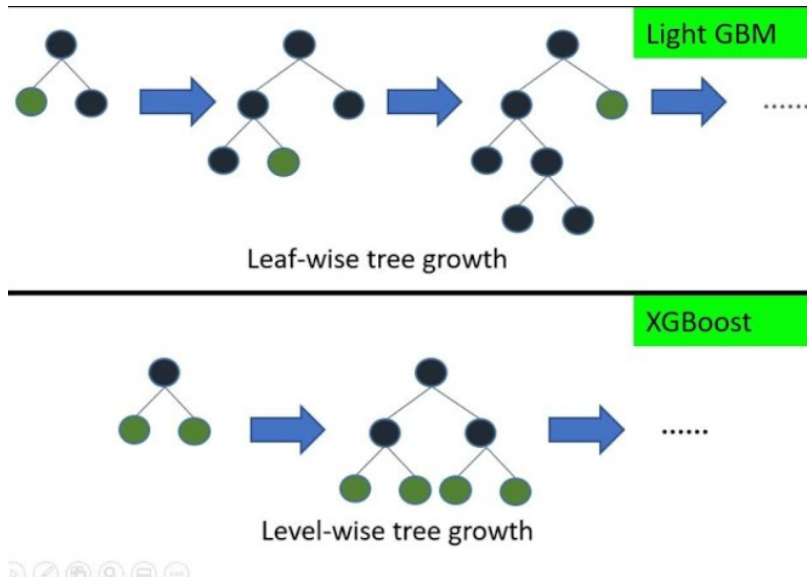
In a model training notebook from Google Colaboratory [13], many models were tested, such as the Knn model, logistic regression models, XGboost models, and the LightGBM Model. Each model will be explored in this notebook to try and generate a higher accuracy. However, due to the previous results of different organizations, it can be predicted that gradient boosting models and long short-term memory network models will be most efficient. However, which model will provide the best accuracy?

LightGBM and XGBoost are similar in many ways since they are gradient-boosting frameworks, but they have key differences that define their implementations, features, and performances. LightGBM [14] uses a leaf-wise tree growth strategy, meaning that it grows the tree by splitting nodes that provide maximum reduction in losing data. LightGBM tends to have higher results on larger datasets, but there is no specified number for this generalization. XGBoost [15] uses a depth-wise tree growth strategy, splitting all nodes at the current depth before moving on to the next depth. XGBoost typically performs better on smaller datasets, but such a number can not be specified exactly. Both LightGBM and XGBoost are robust to outliers, making them more reliable than other models but also making them take longer to output results. Because these models examine each bit of data, they can sometimes overanalyze useless parts of a dataset that can then interfere with the specified targets of a dataset and their results.



Courtesy of Towards Data Science - A diagram of how the LightGBM Model Works

LightGBM has a more complex procedure than XGBoost, which generally makes it faster and more applicable to larger datasets.



Courtesy of DigitalSreeni - The difference in decision trees between Light GBM and XGBoost is shown.

The LSTM addresses the main problems of gradient boosting models by ignoring the useless information present in long datasets. An LSTM layer contains three gates, which each play an important role in enabling the model to precisely depict connections between resultant combinations in training data: the forget gate, the input gate, and the output gate. The forget gate either passes or ignores dataflow from a dataset, as defined by the following equation:

$$F_t = \sigma(W_F \times [X_t, h_{t-1}] + b_F)$$

Variable Names: Output of Forget Gate = Sigmoid Activation Function multiplied by the product of the weight matrix and the current timestamp input and hidden timestamp state when added to the bias coefficient.

The forget gate also deals with presenting the updating and current timestamps of cell states, with C_t and C_{t-1} respectively.

$$C_{t-1} \times F_t = \begin{cases} 0, & F_t = 0 \\ C_{t-1}, & F_t = 1 \end{cases}$$

The input gate selects certain information to be updated using a sigmoid function, I_t , and then compresses the given input pattern using a hyperbolic tangent(tanh) function, C_t , to add an immediate state of long-term impact.

$$I_t = \sigma (W_I \times [x_t, h_{t-1}] b_I)$$

$$\tilde{C}_t = \tanh(W_c \times [X_t, h_{t-1}] + b_c)$$

The output gate figures out the long-term effect and updates the outputs of the current cell state, C_t , and the hidden state(h_t) using sigmoid and tanh functions via the following equations:

$$O_t = \sigma(W_O \times [X_t, h_{t-1}] + b_O)$$

$$h_t = O_t \times \tanh(C_t)$$

The LSTM method, LightGBM, XGBoost, and other classification and regression models will be analyzed in terms of precision, recall, F1-score, and accuracy to determine which demographic of models produce the highest results for the future of pancreatic cancer diagnoses via urinary biomarkers.

Each model tested is also hyperparameter-tuned to its fullest capability. Parameters that were tuned for the gradient boosting models include the following: the boosting type, the objective, the metric, the number of leaves in one tree, the number of boosting rounds, the learning rate, the feature fraction, the bagging fraction, the bagging frequency, and the level of verbosity. Out of these parameters, hyperparameter tuning the number of leaves per tree and the learning rate benefitted both the LightGBM model and the XGBoost model the most, with an increased accuracy of 2%. In addition, the number of boosting rounds also contributed to increasing the accuracy. Separately, a Random Forest model was tuned to specifically look out for the key groups that correlate the most with cancer patients.

Experiments and Discussion

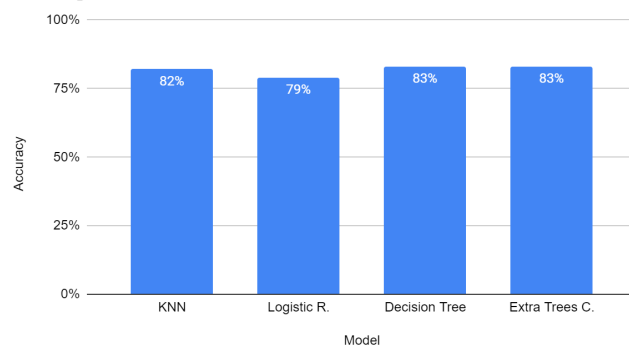
The dataset was downloaded via Google Drive so that Google Colaboratory could access it. As previously mentioned, a model training notebook was created to store each model's results. Before any models were tested, the dataset was converted into a binary classification set so that the models tested would produce better accuracies in correlation with the set. After that, the data was separated into two categories: cancer and no cancer. This was to ensure that the results produced were based on genuine cancer correlations rather than being based solely on the dataset's unique patterns. Following the separation, the missing plasma_CA19_9 level values

were replaced by the average value of valid plasma_CA19_9 levels. After this, the dataset was split into groups, one group being the training data for each tested model. The following models were tested following this split: a KNN model, a logistic regression model, and a decision tree classifier model. Multiple files and techniques were imported from “sklearn,” specifically the accuracy_score technique, which was used to find the accuracy of each model. Later in the file, the extra trees classifier model is also tested for its accuracy. After being matched with testing data, each model came out to produce the following accuracies:

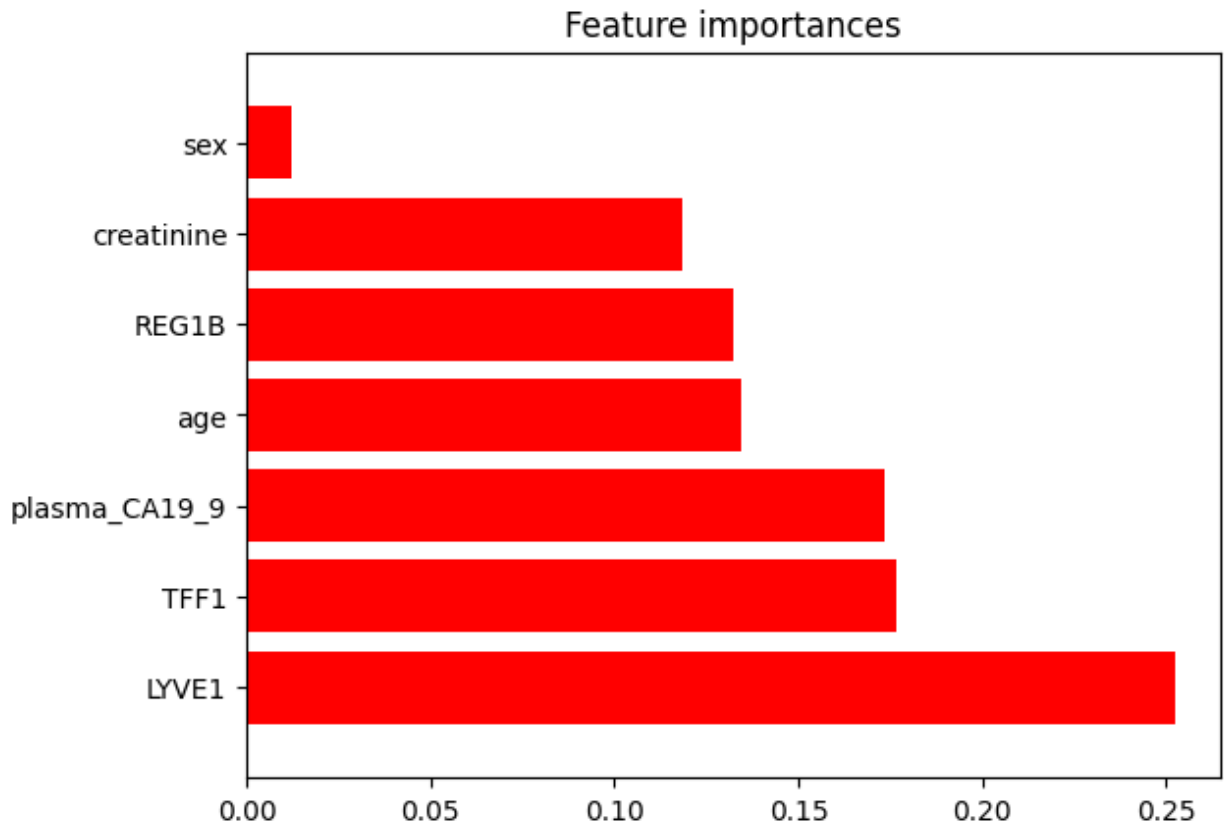
Model	Accuracy
KNN	82%
Logistic R.	79%
Decision Tree	83%
Extra Trees C.	83%

From the data, it can be concluded that these models do not produce high enough accuracies to be considered valuable assets for this analysis. Thus, they will be known as external models for this analysis.

A Comparison of External Model Accuracies



Following the initial model training, a feature importance graph was created with the help of the RandomForest Classifier model, which indicates the features most important for models to use when detecting pancreatic cancer.

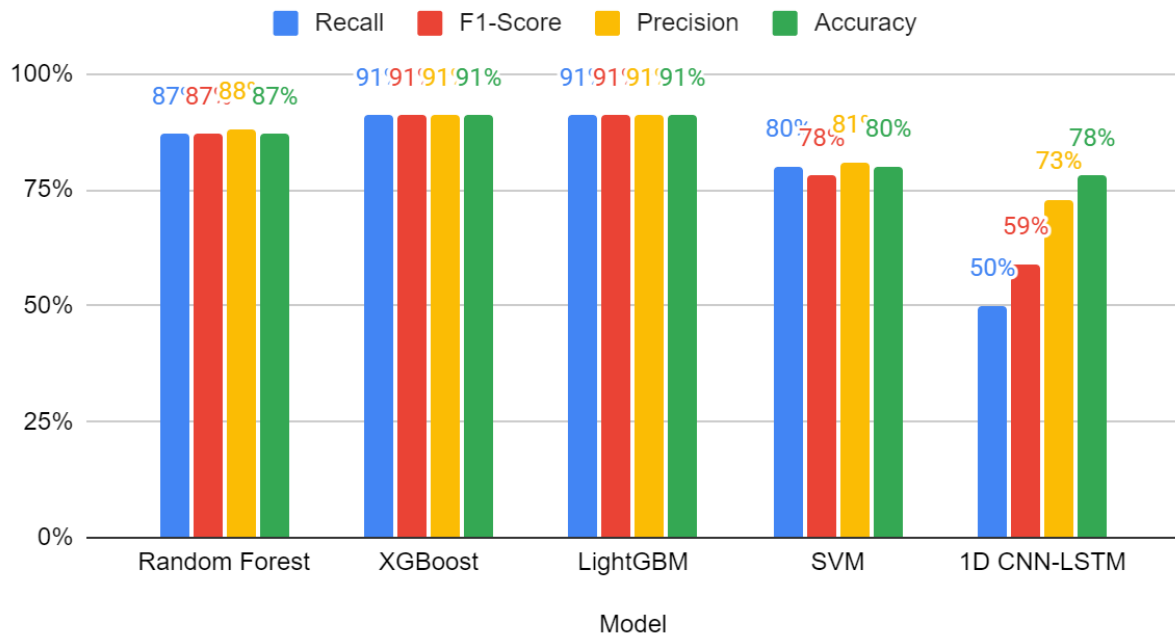


From the bar graph, it can be inferred that urinary biomarkers mostly help a model indicate whether a patient has pancreatic cancer or not.

After the feature importance graph was created, multiple models were tested. Along with the gradient-boosting models and the LSTM model, other models such as the Random Forest Classifier and the support vector machine were tested in terms of recall, F1-Score, precision, and accuracy. Note that the listed models were tested after each model's parameters were manually manipulated to produce the best results. Each model's hyperparameter tuning boosted its accuracy by around 2% on average, and the final results are listed below.

Model	Recall	F1-Score	Precision	Accuracy
Random Forest	87%	87%	88%	87%
XGBoost	91%	91%	91%	91%
LightGBM	91%	91%	91%	91%
SVM	80%	78%	81%	80%
1D CNN-LSTM	50%	59%	73%	78%

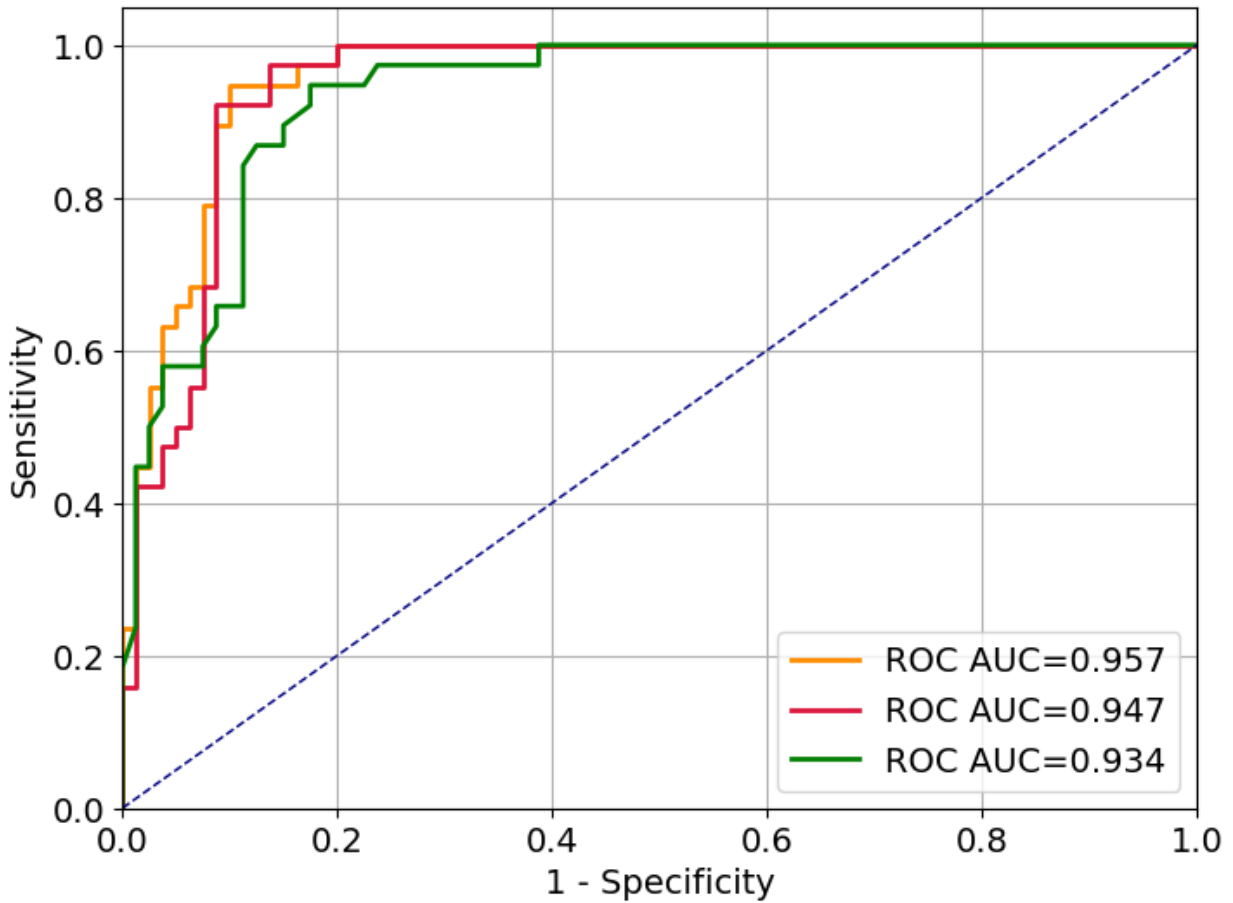
Recall, F1-Score, Precision and Accuracy



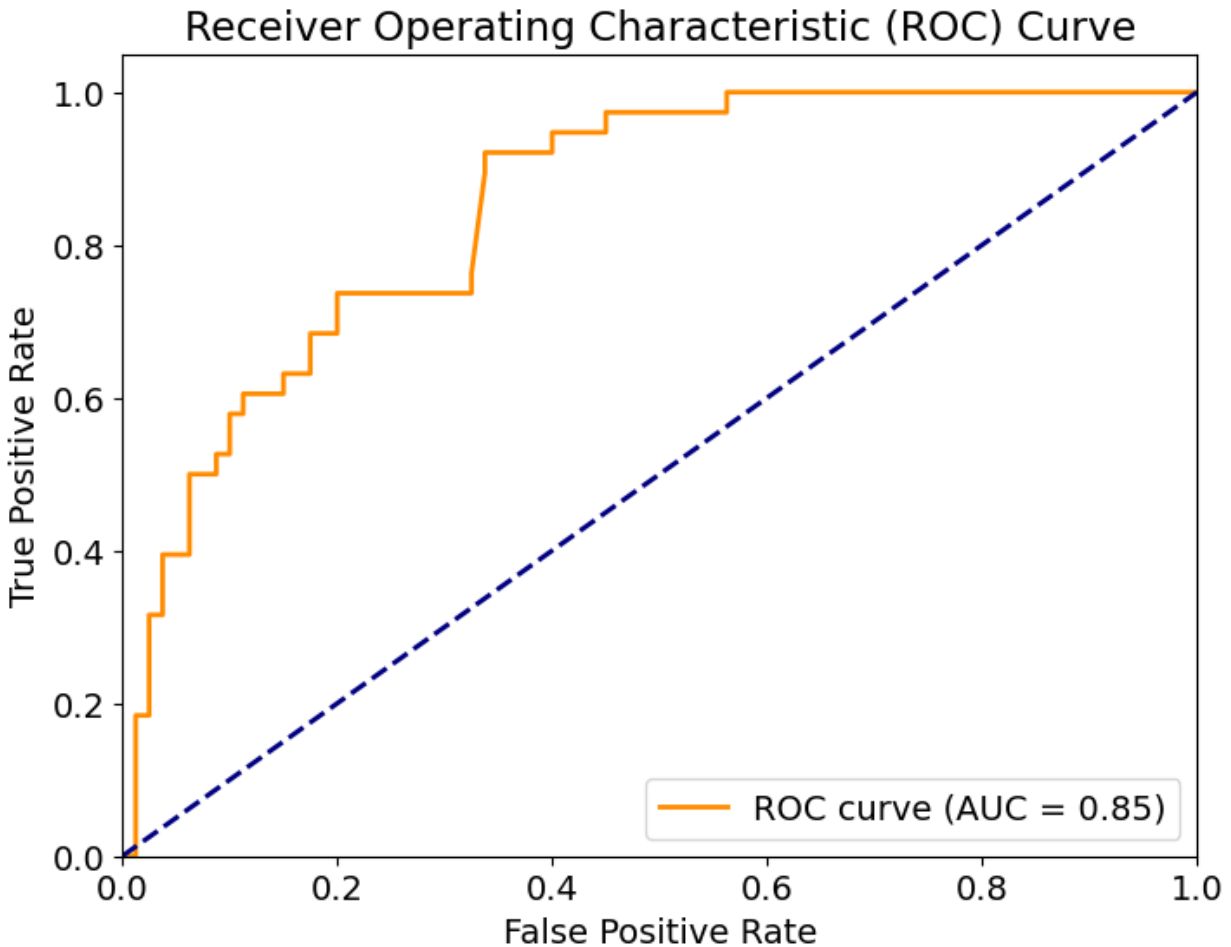
The results show both expected and unexpected results compared to previously discussed results from research papers. The Random Forest Classifier produced an accuracy score of 87%, underperforming compared to XGBoost and LightGBM. XGBoost and LightGBM have scored the same percentages for each group, thus making them equivalent in efficiency at the moment. The support vector machine underperformed with an accuracy score of 80%, which may be because target classes are overlapping. This should've been partially solved when separating the data, but outliers would have affected the accuracy score of this model [16]. The most surprising result was the performance of the 1D CNN-LSTM model. In each category, this model placed last, thus deeming it the most inefficient model given the size of our dataset and the model's given parameters.

After each model's classification report was created, the ROC curves of each significant model were examined. These significant models include the Random Forest Classifier model, the LightGBM model, the XGBoost model, and the 1D CNN-LSTM model even though it underperformed massively given prior expectations.

Model	ROC Curve
Random Forest	0.934
XGBoost	0.947
LightGBM	0.957
1D CNN-LSTM	0.85



The orange line shows LightGBM's ROC performance, the red line shows XGBoost's ROC performance and the green line shows the Random Forest Classifier's ROC performance. From the data, it can be concluded that by a margin of 0.01%, the LightGBM model outperforms the XGBoost model with a slightly higher chance of producing truthful accuracy. However, this is a very minimal difference, thus proving that in general, XGBoost and LightGBM produce the highest accuracy scores when compared to the results of other models.



The following ROC curve indicates that the 1D CNN-LSTM model has a moderate accuracy of around 85% in terms of truthfulness. The graph shows that the model's results are most likely accurate to what it produces, which is a 78% accuracy score.

From the data above, it can be concluded that in general, gradient boosting models produce higher accuracies than any other type of model, including the 1D CNN-LSTM model for this specific dataset. However, upon checking the validity between the dataset used for this analysis and the dataset used for the research paper dealing with the 1D CNN-LSTM model, there are some notable differences that may have affected the model's results. Upon closer inspection, it can be seen that the research paper has a dataset of 590 samples for a specific group that its model is being trained. The dataset used for this analysis has 590 samples in total, but only uses around 200 samples for the model to base itself on. The research paper also defined its LSTM model with many more parameters, taking into account variables and model patterns. The LSTM model used for this dataset was a proposed model technique since the specific code used for the research paper's LSTM model was never publicly given out. If the model used for our dataset had more parameters and a higher data set like the research paper, then perhaps the LSTM model

used for this analysis would increase its accuracy, precision, F1-score, and recall to just as well an amount.

However, this does not affect the validity and reliability of every other model's performance. The high accuracies that the LightGBM and XGBoost models produced correlate with the results that other files published on Kaggle[9] along with the results of the research papers mentioned in the related works section. After hyperparameter tuning each machine learning classifier to its fullest achievable potential, the Random Forest Classifier model also achieved a high accuracy score compared to other models. This may be because it is robust to outliers, deals with missing values, and can handle both classification and regression. By using the feature importance function, the Random Forest Classifier model is able to accurately match patterns in a dataset, thus generating a high and reliable accuracy score [17]. When explaining why the LightGBM model might have barely outperformed the XGBoost model, there is no apparent answer. The dataset must just be large enough that LightGBM is able to process it with a better understanding than XGBoost. However, given that each model has been hyperparameter-tuned to its fullest potential, there is no significant difference in performance between the two models. It is not exactly known why light gradient boosting models work specifically well in this scenario, but it may just be due to the unique patterns in the dataset that happen to fit better with light gradient boosting model procedures.

Conclusion

In conclusion, this study delves into the critical realm of pancreatic cancer detection using urinary biomarkers, with a specific approach of enhancing model accuracy through hyperparameter-tuned machine learning classifiers. Pancreatic cancer's deadly nature only emphasizes the urgency for successful early detections, and blood tests often provide inefficient results. Urinary biomarkers, including creatinine, LYVE1, REG1B, and TFF1, emerge as promising candidates, demonstrating potential accuracy of up to 90% in early-stage detection. This analysis utilized a dataset from the Spanish National Cancer Research Center. The models examined include gradient boosting frameworks, a Random Forest Classifier, a 1D CNN-LSTM model, and a support vector machine. In the end, the gradient boosting models, XGBoost and LightGBM, exhibited the highest accuracy scores, outperforming every other tested model. The unexpected performance of the 1D CNN-LSTM model raises questions about its sustainability for this specific dataset. Further investigation into the model's parameters and dataset characteristics may provide extensions towards which kinds of datasets the model performs well and poorly in. As seen from other research papers, the 1D CNN-LSTM model has great potential, but further investigation into the environments in which it works best will provide great benefits for future detection standards. The findings contribute invaluable insights for future research aimed at refining early decision-making learning models progressing the fight against pancreatic cancer, a disease with alarming prognoses that demand innovative solutions.

Acknowledgments

I would like to thank my mentor Kahye Song for taking the time to teach me about AI and machine learning classifiers. Throughout each stage of my project, I was guided in all aspects, from obtaining a proper dataset to evaluating hyperparameter-tuned learning classifiers properly. For her immense dedication to my project, I sincerely thank her.

I would also like to thank my mentor Jorge Alfaro for giving his consideration towards my project and research paper. With his guidance, I was inspired to begin pursuing machine learning. Thanks to his help, I was motivated to begin learning more and more about artificial intelligence, helping me to build connections with others.



I would like to thank my father's friend, Nanda, for giving his thoughts on my research project and informing me of one key aspect that helped me to dive deeper into the topics presented in the paper. I am grateful for the insights that he provided which helped me better validate my results in the end,

I would like to sincerely thank the Spanish National Cancer Research Center for supplying the dataset used for this analysis, John Davis for uploading the dataset on Kaggle, and many others for their research papers concerning different learning classifiers and their effects on the dataset.

I would like to give a big thanks to my family for supporting me throughout the entire research process. Their support helped motivate me to complete my project and learn more about the future of AI.

References

- [1]“Key Statistics for Pancreatic Cancer.” American Cancer Society,
www.cancer.org/cancer/types/pancreatic-cancer/about/key-statistics.html#:~:text=Pancreatic%20cancer%20accounts%20for%20about.
- [2]Tan, Daniel J., et al. “Intron Retention Is a Robust Marker of Intertumoral Heterogeneity in Pancreatic Ductal Adenocarcinoma.” *Npj Genomic Medicine*, vol. 5, no. 1, Dec. 2020,
<https://doi.org/10.1038/s41525-020-00159-4>.
- [3]“Pancreatic Cancer - Statistics.” Cancer.Net, 4 May 2023,
www.cancer.net/cancer-types/pancreatic-cancer/statistics#:~:text=If%20the%20cancer%20is%20detected,relative%20survival%20rate%20is%2015%25.
- [4]“UpToDate.” UpToDate,
www.uptodate.com/contents/supportive-care-of-the-patient-with-locally-advanced-or-metastatic-exocrine-pancreatic-cancer/print#:~:text=The%20median%20survival%20for%20patients,syste mic%20chemotherapy%20can%20improve%20survival.
- [5]Hoffman, Matthew, MD. “Pancreatic Cancer Diagnosis and Early Detection.” WebMD, 6 Mar. 2009,
www.webmd.com/cancer/pancreatic-cancer/pancreatic-cancer-diagnosis#:~:text=Certain%20sub stances%2C%20such%20as%20carcinoembryonic,is%20advanced%2C%20if%20at%20all.
- [6]New Tests for Early Detection of Pancreatic Cancer Offer Significant Hope - Queen Mary University of London.
www.qmul.ac.uk/research/featured-research/new-tests-for-early-detection-of-pancreatic-cancer-offer-significant-hope/#:~:text=Currently%20pancreatic%20cancer%20is%20identified,to%20per form%20in%20any%20laboratory.
- [7]Debernardi, Silvana, et al. “A Combination of Urinary Biomarker Panel and PancRISK Score for Earlier Detection of Pancreatic Cancer: A Case–control Study.” *PLOS Medicine*, vol. 17, no. 12, Dec. 2020, p. e1003489. <https://doi.org/10.1371/journal.pmed.1003489>.
- [8]Tatjana Crnogorac-Jurcevic, Professor of Molecular Pathology and Biomarkers
www.qmul.ac.uk/research/featured-research/new-tests-for-early-detection-of-pancreatic-cancer-offer-significant-hope/#:~:text=The%20biomarkers%20appeared%20to%20be,potential%20for%20future%20clinical%20application.

- [9]“Urinary Biomarkers for Pancreatic Cancer.” Kaggle, 10 Dec. 2020,
www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer/code.
- [10]“Urinary Biomarkers for Pancreatic Cancer.” Kaggle, 10 Dec. 2020,
www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer.
- [11]  NIVRITH ANANTH IYER - data_visulation_pancreatic_cancer.ipynb
- [12]O’Neill, Robert S, and Alina Stoita. “Biomarkers in the Diagnosis of Pancreatic Cancer: Are We Closer to Finding the Golden Ticket?” World Journal of Gastroenterology, U.S. National Library of Medicine, 14 July 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8311531/.
- [13]  NIVRITH ANANTH IYER - model_training_pancreatic_cancer.ipynb
- [14]GeeksforGeeks. “LightGBM Light Gradient Boosting Machine.” GeeksforGeeks, 23 May 2023, www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/#.
- [15]“What Is XGBoost?” NVIDIA Data Science Glossary,
www.nvidia.com/en-us/glossary/data-science/xgboost.
- [16]Raj, Ashwin. “Everything About Support Vector Classification — Above and Beyond.” Medium, 6 Aug. 2022,
<https://towardsdatascience.com/everything-about-svm-classification-above-and-beyond-cc665bfd993e#:~:text=SVM%20does%20not%20perform%20very,samples%2C%20the%20SVM%20will%20underperform>.
- [17]Team, Cfi. “Random Forest.” Corporate Finance Institute, 22 Nov. 2023,
<https://corporatefinanceinstitute.com/resources/data-science/random-forest/#:~:text=Among%20all%20the%20available%20classification,other%20classes%20in%20the%20data>.
- [18] NivritthA. “GitHub - NivritthA478/Pancreatic-Cancer-Detection-Via-Urinary-Biomarkers: Pancreatic Cancer Detection With Urinary Biomarkers Data Visualization.” GitHub,
<https://github.com/NivritthA478>