Journal of
Electronics Technology Exploration
Homepage: htttps://shmpublisher.com/index.php/joetex
p-ISSN 3025-3470 | e-ISSN 3026-1066

Classification of Pancreatic Cancer Diagnosis with CatBoost Using Urine Biomarker Combination

Yulizchia Malica Pinkan Tanga[1], Putri Utami[2], Aditya Yoga Darmawan[3], Jumanto[4]
[1,2,3,4] Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Mar 06, 2025
Revised Apr 22, 2025
Accepted May 21, 2025

Keywords:

Pancreatic cancer
Catboost algorithm
Urinary biomarkers
Machine learning
Cancer prediction

ABSTRACT

Uncontrolled cell growth in the pancreatic gland, is one of the most aggressive types of cancer with a high mortality rate, called pancreatic cancer. This research focuses on improving early diagnosis methods for pancreatic cancer by using CatBoost. Urine biomarker datasets were collected and subjected to pre-processing, including label coding, standardized scaling, and balancing via the Synthetic Minority Oversampling Technique (SMOTE). The CatBoost model achieved an accuracy of 98.89%, specificity of 99.35%, sensitivity of 98.71%, and Area Under the Curve (AUC) of 0.9951. These results show that the CatBoost model significantly outperforms the diagnosis models in previous studies, overcoming the challenges of early detection and classification of pancreatic cancer. This study shows that CatBoost is effective for diagnosing pancreatic cancer and suggests that future research explore other models on larger and more diverse datasets.

Corresponding Author:

Yulizchia Malica Pinkan Tanga,
Department of Computer Science,
Universitas Negeri Semarang,
Semarang, Jawa Tengah, Indonesia,
Email: yulizchiamalica@students.unnes.ac.id

https://doi.org/10.52465/joetex.v4i1.651

1. INTRODUCTION

A disease that can attack various parts of the human body due to abnormalities during the cell turnover process is called cancer. Cancer is known to have the ability to spread rapidly and damage various vital organs, thus becoming one of the leading causes of human death worldwide [1], [2]. There are various types of cancer that can attack the human body, such as skin cancer, breast cancer [3], [4], lung cancer [5], pancreatic cancer, colorectal cancer and many other types of cancer. Of the several types of cancer, there is one of the most aggressive cancers, which is pancreatic cancer [6].

Uncontrolled cell growth in the pancreatic gland leads to a cancerous mass that can spread to other tissues in the body known as Pancreatic Cancer (PC) disease [7]. This disease is very dangerous and has a high mortality rate, especially if not diagnosed early [8], [9]. Predicted survival of pancreatic cancer patients is low, with a survival rate of only five years [10], [11]. About 9% of pancreatic ductal adenocarcinoma (PDAC) patients survive more than five years [8], [12]–[14]. Several factors such as obesity, diabetes, unhealthy diet, genetic predisposition, advanced age, excessive alcohol consumption, smoking, physical inactivity, chronic pancreatitis, and chronic exposure to air and water pollutants contribute to the increasing prevalence of

Pancreatic Cancer (PC) [7]. Recent years have shown that pancreatic cancer can be detected early to improve long-term survival [15], [16]. Currently, there is a need for more accurate methods for early diagnosis of pancreatic cancer, given the low survival rate and difficulty in identifying it at an early stage. Current pancreatic cancer diagnosis methods face challenges such as low accuracy for early detection and difficulty to implement effectively in populations with different demographic characteristics.

Previous research that also used a combination dataset of urine biomarkers for pancreatic cancer was conducted by Irem Acer, et all [12] in 2023. This study compared machine learning classification methods such as Support Vector Machine (SVM), Naive Bayes (NB), k-Nearest Neighbor (kNN), Random Forest (RF), Light Gradient Boosting Machine (LighyGBM), AdaBoost, and Gradient boosting classifier (GBC) to detect Pancreatic Ductal Adenocarcinoma (PDAC) disease classification. By analyzing the classes of healthy controls, pancreatic disorders, and PDAC patients, this study attempts to determine the best machine learning model. CV-10 was the most effective classification method for classifying patients with PDAC and healthy controls, while the GBC model was 92.99% (AUC=0.9761) and the LightGBM model reached 86.37% (AUC=0.9348). Another study was conducted by Malte Heyen [17] in 2023 using gamma regression and multinomial logistic regression models. The results showed that LYVE1, REG1B and TFF1 were increased in PDAC patients. The logistic regression model trained at 50% was able to detect 84.7% of PDAC patients with test data. Research conducted by Oleg Blyuss et al [18] in 2019 compared several algorithms such as logistic regression, neural network, random forest, support vector machine, and neuro fuzzy to build a biomarker-based risk score. The results showed that the logistic regression model achieved specificity of 0.9, sensitivity of 0.82, and AUC of 0.94.

Research by Mohamed Esmail Karar et al [14] in 2023 proposed Convolutional Neural Network (CNN) and Long short-term memory (LSTM) for the classification of urine biomarkers for pancreatic cancer diagnosis. The results showed that the CNN model combined with LSTM achieved the best accuracy of 97%. CNN and LSTM models were successfully developed for pancreatic cancer diagnosis with four proteomic biomarkers creatinine, LYVE1, REG1B, and TFF1. Previous research utilizing clinical data from 6 million patients (24,000 pancreatic cancer cases) in Denmark (Danish National Patient Registry (DNPR)) and 3 million patients (3,900 cases) in the United States (US Veteran Affairs (US-VA)) was conducted by Davide Placido et al [11] in 2022. The cross-validation performance of the Danish model to the US-Va data was lower (AUROC = 0.71), and retraining was required to improve it to AUROC = 0.78 and AUROC (3m) = 0.76. These results may improve the ability to make realistic surveillance in high-risk patients. This may help and prolong quality of life by detecting aggravated cancers quickly.

Previous studies have used various models to detect pancreatic cancer. However, there are still limitations in terms of accuracy and early detection, especially in the early stages of the disease. Therefore, this research focuses on further development in pancreatic cancer diagnosis using the CatBoost algorithm. This is driven by the limitations of diagnosis methods as well as the urgency to overcome the challenges in early detection of pancreatic cancer. The CatBoost algorithm was chosen because of its ability to work well in processing large and complex data [19]. This algorithm is able to provide more stable prediction results and reduce overfitting [20], so it is expected to provide more accurate diagnosis results and can be widely applied.

2. METHOD

This research uses the CatBoost model for pancreatic cancer disease classification. The stages in this research include data collection, data preprocessing, and model training. The flowchart of this research is shown in Figure 1.

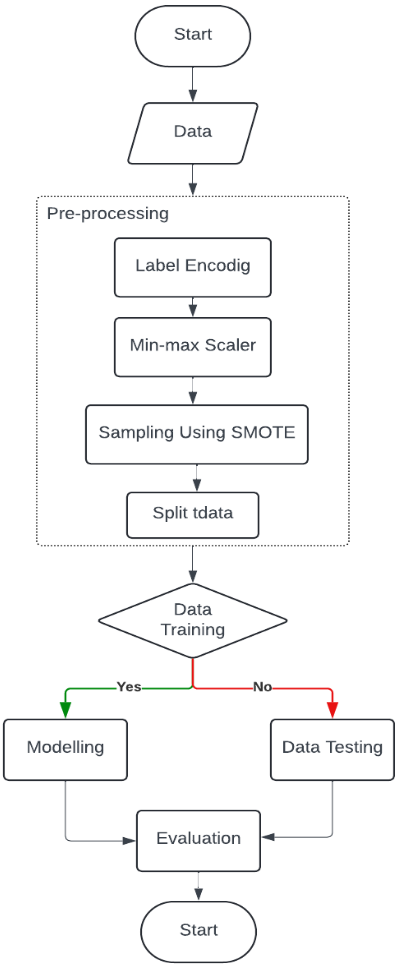

Figure 1. Proposed method

1.1 Data Collecting

The dataset used was Unrinary Biomarkers for Pancreatic Cancer [8]. The data was retrieved through the kaggle platform which can be accessed through Dataset Kaggle unirary bimarkers for pancreatic cancer. The dataset consists of 590 data and 14 columns. Features in the dataset include sample_id, patient_cohort, sample_origin, age, sex, diagnosis, stage, benign_sample_diagnosis, plasma_CA19_9, creatinine, LYVE1, REG1B, TFF1, and REG1A. This dataset contains three categories: healthy controls, patients with noncancerous pancreatic disorders such as chronic pancreatitis, and patients with pancreatic ductal adenocarcinoma. The distribution of the data samples used is shown in Table 1.

Table 1. Number of samples

	Healthy Controls	Pancreatic patients	Pancreatic ductal adenocarcinoma
Patients	183	208	199
Avg. Age	56.33	54.70	66.18
Female	155	101	83
Male	68	107	116

1.2 Pre-processing

1.2.1 Label Encoding

Label encoding is used to convert categorical values in the data into numerical values [22]. This technique is used to convert categorical data types into numerical columns. Some columns have a categorical data type, so they need to be converted to a numerical format in order to be used by machine learning algorithms.

1.2.2 Min-Max Scaler

Min-max scaler is used to normalise or standardise data by changing the size of the data into the range of 0 and 1 [23]. This is important to do so that the machine learning model is not biased towards features with a larger value scale. Formula 1 explain how min-max scaler works [24].

$$X_{scaled} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \tag{1}$$

Where, X is the original value, X_{min} is the minimum featurem, X_{max} is the maximum feature value, and X_{scaled} is the normalised data value.

1.2.3 SMOTE (Synthetic Minority Oversampling Technique)

SMOTE is used to synthetize the minority data by generate a random instance between 2 data, a minority instance under consideration and one of its k closest neighbours from the same minority class, selected at random [25]. With this principle, SMOTE will handle the imbalanced class problem on the dataset by synthetize the class with less data (minority class) until it balanced [26].

1.3 Split Data

Datasets that have gone through pre-processing will be divided into two parts, namely training data and testing data. The data will be divided into 80% as training data, while the remaining 20% is used for testing data. Data splitting is done using train_test_split from scikit-learn, with the parameter random_state=42.

1.4 Modelling

At this stage, the CatBoost algorithm is used to build a model that will predict pancreatic cancer diagnosis. CatBoost is a gradient boosting learning algorithm designed to handle data categories and prevent overfitting [27]. CatBoost works by building a series of consecutive decision trees, where each tree aims to correct the error of the previous tree [28]. The boosting process combines the predictions from multiple trees to generate better predictions [29]. CatBoost works by building a series of consecutive decision trees, where each tree aims to correct the error of the previous tree. The boosting process combines the predictions from multiple trees to generate better predictions. Table 2 below are the Hyperparameter setting for CatBoost.

Table 2. Catboost hyperparameter

Hyperparameter	
iterations	1000
learning_rate	0.075
l2_leaf_reg	0.5
random_state	42

1.5 Evaluation Model

Evaluation model for Pancreatic Cancer Diagnosis Classification using confusion matrix, which determines accuracy, precision, recall, F1-score. These metrics provide an overall picture of the model's ability to correctly classify data and identify prediction errors [30]. The accuracy of the proposed model in comparison to previous studies is the focus of the confusion matrix performance analysis. The confusion matrix formula is written in equations (2), (3), (4), and (5).

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN) \times 100} \tag{2}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4}$$

$$F1\ Score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \tag{5}$$

Description:
True Positive (TP) : number of positive cases that are classified as positive.
False Positive (FP): the number of negative cases that are classified as positive.
True Negative (TN): the number of negative cases that are classified as negative.
False Negative (FN): the number of positive cases that are classified as negative.

3. RESULTS AND DISCUSSIONS

1. Result

3.1.1. Result of Data Preprocessing

.

Data from kaggle contain 590 data. However, the data have some missing value, N/A and NULL value inside. Data Preprocessing is needed to clean the data so only data with correct value will be passed to next process. Table 3 the summary of null value of each column.

Table 3. Data value before preprocessing

Data	
sample_id	0
patient_cohort	0
sample_origin	0
age	0
sex	0
diagnosis	0
stage	391
benign_sample_diagnosis	382
plasma_CA19_9	240
creatinine	0
LYVE1	0
REG1B	0
TFF1	0
REG1A	284

First, 'sample_id', 'patient_cohort', 'sample_origin' was dropped since it contain information that is not important, while 'stage' and 'benign_sample_diagnosis' was also dropped due to the colomn contain missing value since 'stage' and 'benign_sample_diagnosis' column will have no value if the patient's diagnosis is control or no pancreatic disease. Data values after column drop are shown in Table 4.

Table 4. Data value after drop column

Data	
age	0
sex	0
diagnosis	0
plasma_CA19_9	240
creatinine	0
LYVE1	0
REG1B	0
TFF1	0
REG1A	284

After the feature selection process, a number of empty values were still found in several variables. Empty values can interfere with the model training process because they can reduce the quality of learning and cause bias in the prediction results. Therefore, data rows containing empty values were removed from the dataset to ensure that the model was trained using complete, consistent data that was more representative of existing patterns. Table 5 below are the data after dropping N/A value from the data.

Table 5. Data value after dropping row with N/A value

Column	Count
age	0
sex	0
diagnosis	0
plasma_CA19_9	0
creatinine	0
LYVE1	0
REG1B	0
TFF1	0
REG1A	0

All columns in the dataset no longer have empty values after the data cleaning process, which removes rows containing Null or N/A values. The number of samples was reduced from 509 to 209 through this process, ensuring that only complete and consistent data is used in model training. Although the amount of data was significantly reduced, this step was essential to improve the quality of the dataset, reduce the possibility of bias caused by incomplete data, and ensure that the analysis results were more accurate and reliable.

3.1.2. Result of Applying Label Encoding

Label Encoding is performed by utilizing LabelEncoder from sklearn library. The target column is 'sex' column since it contain object type value. Table 6 below are the comparation before and after encoding

Table 6. Comparation before and after label encoding

Before Encoding	After Encoding
F	0
M	1

Table 6 shows the results of applying Label Encoding to the sex feature, where categorical values are converted into numerical representations to help machine learning algorithms process information more easily. In this process, the female category “F” is encoded as 0 and the male category “M” is encoded as 1, so that the data can be used in numerical form without changing the meaning of the categories. In addition to ensuring data compatibility with machine learning models, this transformation aims to maintain consistent input formats during model training and evaluation.

3.1.3. Result of Applying Min-Max Scaler

Min-max Scaler is performed by utilizing MinMaxScaler from sklearn library. Target column for scaling are every column with float or integer data except ‘age’ column. Table 7 below are example of data before and after scaling.

Table 7. Data before and after scaling

Column	Before Scaling	After Scaling
creatinine	1.8322	0.508197
LYVE1	0.8932	0.03737
REG1B	52.9488	0.03737
TFF1	654.282	0.10841

Table 7 shows a comparison of feature values before and after applying the Min–Max Scaler, which aims to transform numerical data into a range of values between 0 and 1. This normalization process helps to equalize the scale between features that have significantly different value ranges, thereby preventing features with large values from dominating the model learning process. By applying the Min–Max Scaler, the data distribution becomes more uniform, which has the potential to improve model training stability and enhance the performance of algorithms that are sensitive to feature scale.

3.1.4. Result of Applying SMOTE

After data preprocessing, some of the data are loss, which affect the balance of the data. Oversampling is needed to make the data balanced by using SMOTE (Synthetic Minority Over-sampling Technique). Table 8 below are the comparation before and after using SMOTE to balance the data.

Table 8. Data before and after SMOTE

Class	Before SMOTE	After SMOTE
0	134	134
1	48	134
2	27	134

Table 8 shows the distribution of sample numbers in each class before and after applying the Synthetic Minority Over-sampling Technique (SMOTE). Before SMOTE was applied, the dataset experienced class imbalance, where the minority class had far fewer samples than the majority class. After the SMOTE process, the number of samples in each class became balanced, namely 134 samples per class.

3.1.5. Model Analysis

Categorical Boosting (CatBoost) was used to classify the pancreatic disease from the Pancreatic patients dataset. Before training, a Hyperparameter for CatBoost need to be set.

Result from the study will be in form of confusion matrix and classification report. Figure 2 shows the Confusion Matrix of the model.

.

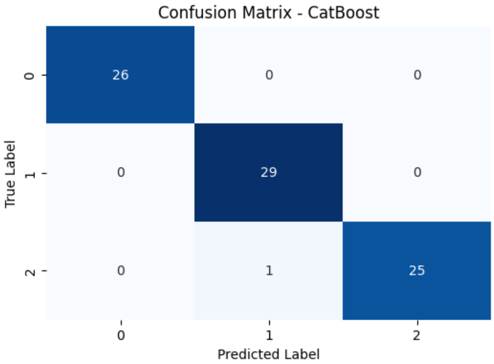

Figure 2. Confusion matrix of the catboost model

CatBoost model sucessfully classify almost all of the test data with only 1 misclassification where 1 ‘pancreatic_cancer’ data are classified as ‘pancreatitis’. The Classification Report of the model are shown on Table 9.

Table 9. Classification catboost

	Precision	Recall	F1-Score	Support
0	1.0000	1.0000	1.0000	26
1	0.9667	1.0000	0.9831	29
2	1.0000	0.9615	0.9804	26
Accuracy			0.9877	81
Macro Avg	0.9889	0.9872	0.9878	81
Weighted Avg	0.9881	0.9877	0.9876	81

From the classification report, CatBoost model perform very well detecting each class on the test dataset with 98.77% Accuracy, 98.89% Precision, 98.72% Recall, and 98.78% F1-Score. The weighted average precision, recall, and f1-score is also close with macro average at 98,81% Precision, 98.77% recall, and 98.76% F1-Score shown that the model can perform well on slightly unbalanced dataset, although the difference is just 3 data on ‘pancreatitis’ class with other class.

2. Discussion

The results obtained in this research using the CatBoost model will be compared with previous research models. The model comparison is shown in Table 10.

Table 10. Comparison with previous research

Author	Methode	Accuracy	Specificity	Sensitifity	AUC
[18]	Logistic regression,	-	0.9	0.81	0.94
[8]	Logistic regression	-	96.30	96.70	0.992
[12]	Gradien boosting	72.91	-	-	0.8733
Proposed Method	**CatBoost**	**98.89**	**99.35**	**98.71**	**0.9951**

In research [18] and [8] using the same dataset but with different methods. In this study using CatBoost with the results achieved an accuracy of 98.89, specificity 99.35, sensitivity 98.71, and AUC 0.9951. The accuracy results obtained show that this CatBoost model has better prediction capabilities compared to models in previous studies.

4. CONCLUSION

This study focuses on the early diagnosis of pancreatic cancer due to its aggressive nature and high mortality rate. In some previous studies, there are still challenges in the accuracy and effectiveness of early detection. The use of the CatBoost model for pancreatic cancer classification was successfully demonstrated in this study by achieving accuracy of 98.98, specificity of 99.35, sensitivity of 98.71, and AUC of 0.9951. Compared to previous studies this model achieves better performance and can predict better on complex data in classifying pancreatic cancer diagnosis. For further development, it is recommended that future researchers try to use other models and test them on larger and more diverse datasets.

REFERENCES

[1] A. Tayanloo-Beik *et al.*, “OMICS insights into cancer histology; Metabolomics and proteomics approach,” *Clin. Biochem.*, vol. 84, no. June, pp. 13–20, 2020.

[2] Rofik, F. Artiyani, D. Ananda, and A. Pertiwi, “Journal of Information System Breast Cancer Diagnosis Utilizing Artificial Neural Network (ANN) Algorithm for Integrating Multi-Omics Data and Clinical Features,” vol. 2, no. 2, pp. 61–68, 2024.

[3] Jumanto, M. F. Mardiansyah, R. Pratama, M. F. Al Hakim, and B. Rawat, “Optimization of breast cancer classification using feature selection on neural network,” pp. 105–110, 2022.

[4] A. U. Dullah and F. N. Apsari, “Ensemble learning technique to improve breast cancer classification model,” pp. 107–113, 2023.

[5] N. P. Damayanti, M. Nabiel, D. Ananda, and F. W. Nugraha, “Lung cancer classification using convolutional neural network and DenseNet,” pp. 133–141, 2023.

[6] Z. Y. Zhao and W. Liu, “Pancreatic Cancer: A Review of Risk Factors, Diagnosis, and Treatment,” *Technol. Cancer Res. Treat.*, vol. 19, pp. 1–13, 2020.

[7] K. C. Ravi, R. R. Dixit, T. Indhumathi, S. Singh, A. Gopatoti, and A. S. Yadav, “AI-Powered Pancreas Navigator: Delving into the Depths of Early Pancreatic Cancer Diagnosis using Advanced Deep Learning Techniques,” *IEEE 9th Int. Conf. Smart Struct. Syst. ICSSS 2023*, pp. 1–6, 2023.

[8] S. Debernardi *et al.*, “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. 1–23, 2020.

[9] A. Chakraborty and C. P. Tsokos, “An AI-driven Predictive Model for Pancreatic Cancer Patients Using Extreme Gradient Boosting,” *J. Stat. Theory Appl.*, vol. 22, no. 4, pp. 262–282, 2023.

[10] A. Gupta, A. Koul, and Y. Kumar, “Pancreatic Cancer Detection using Machine and Deep Learning Techniques,” *Proc. 2nd Int. Conf. Innov. Pract. Technol. Manag. ICIPTM 2022*, vol. 2, pp. 151–155, 2022.

[11] D. Placido *et al.*, “A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories,” *Nat. Med.*, vol. 29, no. 5, pp. 1113–1122, 2023.

[12] İ. Acer, F. Orhanbulucu, S. Içer, and F. Latifoglu, “Early diagnosis of pancreatic cancer by machine learning methods using urine biomarker combinations,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 31, no. 1, pp. 112–125, 2023.

[13] K. Cao *et al.*, “Large-scale pancreatic cancer detection via non-contrast CT and deep learning,” *Nat. Med.*, vol. 29, no. 12, pp. 3033–3043, 2023.

[14] M. E. Karar, N. El-Fishawy, and M. Radad, “Automated classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks,” *J. Biol. Eng.*, vol. 17, no. 1, pp. 1–12, 2023.

[15] M. Dbouk *et al.*, “The Multicenter Cancer of Pancreas Screening Study: Impact on Stage and Survival,” *J. Clin. Oncol.*, vol. 4, no. 28, 2022.

[16] M. Goggins, “The role of biomarkers in the early detection of pancreatic cancer,” *Fam. Cancer*, vol. 23, no. 3, pp. 309–322, 2024.

[17] M. Heyen, “Diagnosing Pancreatic Cancer via Urinary Biomarkers,” no. 11, 2023.

[18] O. Blyuss *et al.*, “Development of PancRISK, a urine biomarker-based risk score for stratified screening of pancreatic cancer patients,” *Br. J. Cancer*, vol. 122, no. 5, pp. 692–696, 2020.

[19] C. S. Kulkarni, “Advancing Gradient Boosting: A Comprehensive Evaluation of the CatBoost Algorithm for Predictive Modeling,” *J. Artif. Intell. Mach. Learn. Data Sci.*, vol. 1, no. 5, pp. 54–57, 2022.

[20] S. Ben Jabeur, C. Gharib, S. Mefteh-Wali, and W. Ben Arfi, “CatBoost model and artificial intelligence techniques for corporate failure prediction,” *Technol. Forecast. Soc. Change*, vol. 166, no. January, p. 120658, 2021.

[21] A. M. Wyglinski, D. P. Orofino, M. N. Ettus, and T. W. Rondeau, “Revolutionizing software defined radio: case studies in hardware, software, and education,” *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 68–75, Jan. 2016.

[22] J. T. Hancock and T. M. Khoshgoftaar, “Survey on categorical data for neural networks,” *J. Big Data*, vol. 7, no. 1, 2020.

[23] W. A. Arifin, I. Ariawan, A. A. Rosalia, L. Lukman, and N. Tufailah, “Data scaling performance on various machine learning algorithms to identify abalone sex,” *J. Teknol. dan Sist. Komput.*, vol. 10, no. 1, pp. 26–31, 2022.

[24] B. Deepa and K. Ramesh, “Epileptic seizure detection using deep learning through min max scaler normalization,” *Int. J. Health Sci. (Qassim).*, vol. 6, no. April, pp. 10981–10996, 2022.

[25] E. Boonchieng, V. Chouvatut, and D. Lipsky, “Article FLEX-SMOTE : Synthetic over-sampling technique that flexibly adjusts to different minority class distributions FLEX-SMOTE : Synthetic over-sampling technique that flexibly adjusts to different minority class distributions,” *Patterns*, p. 101073, 2024.

[26] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, “SMOTE for Handling

Imbalanced Data Problem : A Review," *2021 6th Int. Conf. Informatics Comput. ICIC 2021*, pp. 1–8, 2021.

[27] H. V Nerlekar, K. Moholkar, D. Wagh, and S. L. Bangare, "Integrating AI with Health Informatics for Early Detection of Pancreatic Cancer Frontiers in Health Informatics," vol. 13, no. 2, pp. 204–217, 2024.

[28] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost : gradient boosting with categorical features support," pp. 1–7.

[29] N. Khodadadi, H. Roghani, F. De Caso, E. M. El-kenawy, Y. Yesha, and A. Nanni, "Thin-Walled Structures Data-driven PSO-CatBoost machine learning model to predict the compressive strength of CFRP- confined circular concrete specimens," *Thin-Walled Struct.*, vol. 198, no. February, p. 111763, 2024.

[30] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," pp. 1–17, 2020.