DOI: https://doi.org/10.57233/ihcfs.040

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

1ST INTERNATIONAL HYBRID CONFERENCE OF FACULTY OF SCIENCE

(A PUBLICATION OF FACULTY OF SCIENCE, FEDERAL UNIVERSITY GUSAU, NIGERIA)

EARLY DETECTION OF PANCREATIC CANCER USING URINARY BIOMARKERS AND CLINICAL FEATURES: A BIOSTATISTICAL AND MACHINE LEARNING APPROACH

[1]Nafisat Yusuf, [2]Musa Uba Muhammad and [3]Suraj Mahmud Umar

[1,2]Department of Statistics, Aliko Dangote University of Science and Technology, Wudil, Nigeria.

[3]Department of Statistics, Bayero University, Kano, Nigeria.

Corresponding Author's email: yusuffeenah1139@gmail.com

ABSTRACT

Pancreatic ductal adenocarcinoma (PDAC) remains highly lethal largely because reliable early-detection methods are lacking. Traditional tools, including imaging and serum CA19-9, show poor sensitivity in early disease and limited specificity, contributing to late diagnoses. Emerging evidence suggests urinary biomarkers may offer a non-invasive alternative, though their clinical integration with machine-learning models is still limited. This study evaluated urinary biomarkers (LYVE1, REG1B, TFF1) alongside age, CA19-9, and creatinine for early PDAC detection. Using logistic regression, LASSO, SVM, and Random Forest, predictive models were developed and assessed. Random Forest achieved the best performance (AUC 0.952), showing strong discrimination between PDAC and non-PDAC groups. Findings indicate urinary biomarkers provide complementary diagnostic value to traditional markers, supporting a multimodal approach. Therefore, the study demonstrates that a urinary biomarker–based machine-learning model is a feasible, effective, non-invasive tool for early PDAC detection and future screening applications.

Keywords: Pancreatic Ductal Adenocarcinoma (PDAC), Urinary biomarkers, clinical features, Machine learning.

1.0INTRODUCTION

Pancreatic cancer, particularly Pancreatic Ductal Adenocarcinoma (PDAC), has become one of the most challenging and deadly forms of cancer globally. According to recent global cancer statistics, PDAC ranks among the top five causes of cancer-related mortality in both men and women, despite being less prevalent than other major cancers like lung, breast, and colorectal cancer (Bray *et al*., 2018; Rawla, Sunkara and Gaduputi, 2019). What makes PDAC particularly devastating is its aggressive nature, complex tumour microenvironment, and late symptom onset, all of which contribute to a five-year survival rate of less than 10% (Ferlay *et al*, 2019). Despite accounting for only 3% of all cancers diagnosed in the United States, PDAC is the fourth leading cause of cancer-related death, and is projected to become the second by 2030 due to its increasing incidence and minimal improvement in survival rates (Rahib *et al*., 2014; McGuigan *et al*., 2018). PDAC exhibits a clear age-associated pattern. The majority of cases occur in individuals over the age of 60, with the median age at diagnosis being approximately 70 years (Ilic and Ilic, 2016). Male sex has also been associated with a slightly higher risk of PDAC, possibly due to differences in tobacco use and occupational exposures (Lowenfels and Maisonneuve, 2006). Urine offers several unique advantages over blood: it is non-invasive to collect, poses minimal risk of infection or contamination, can be obtained repeatedly and easily even in low-resource settings, and requires no complex infrastructure or highly trained personnel for acquisition, making it ideal for population-level screening, especially in remote or underserved regions (Dinges *et al*., 2019).

1.1 Aim and Objectives

To develop and evaluate biostatistical models using urinary biomarkers and clinical risk factors for the early detection of pancreatic ductal adenocarcinoma (PDAC), and the objectives are;

i. To preprocess and explore the dataset for quality assurance and data integrity.

DOI: https://doi.org/10.57233/ihcfs.040

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025

WEBSITE: fos.fugusau.edu.ng

ii. To identify the most informative urinary and clinical biomarkers associated with PDAC using feather importance and comparative statistics.
iii. To build and compare multiple predictive models predictive models including logistic regression, LASSO, Random Forest, and Support Vector Machines (SVMs).
iv. To thoroughly evaluate model performances conducted using cross-validation, confusion matrices, and ROC-AUC analysis to ensure model robustness and generalizability.
v. To interpret model outputs and assess clinical applicability through feature importance analysis.

2.0 MATERIALS AND METHODS

2.1 Research Design

The study adopts a retrospective observational design using a case–control dataset. It involves secondary data analysis of the Debernard.csv dataset, originally compiled in a biomarker-based PDAC detection study. The research framework includes data preprocessing, feature selection, predictive model development, performance evaluation, and interpretation. Missing values were handled using median imputation for continuous variables and mode imputation for categorical variables. The target variable was created by classifying diagnosis code 3 as PDAC (positive class) and codes 1-2 as non-PDAC (negative class).

2.2 Model Development

The following models were trained using a 70/30 train-test split:

2.2.1 Logistic Regression:

A logistic regression will model the chance of an outcome based on individual characteristics. Because chance is a ratio, what will be actually modeled is the logarithm of the chance given by (Sperandei, S. 2014):

$$\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_mX_m...(2)$$

where π indicates the probability of an event and β_i are the regression coefficients associated with the reference group and the X_i explanatory variables.

2.2.2 Random Forest Classifier

A random forest is a classifier consisting of a collection of tree-structured classifiers $U\{h(\mathrm{X},\Theta_k),k=1,...\}$ where the Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input X (Breiman, 2001).

2.2.3 Support Vector Machines (SVMs)

Support vector classifiers are based on the class of hyperplanes $(w,x)+b=0$ with $w,x\in\mathbf{R}^n$, $b\in\mathbf{R}$, corresponding to decision functions $f(x)=sign(\langle w,x\rangle+b)$. The points x which lie on the hyperplane satisfy $(w,x)+b=0...(5)$, where w defines a direction perpendicular to the hyperplane, while varying the value of b moves the hyperplane parallel to itself (Alessia, Marco and Nello, 2009).

2.2.4 Least Absolute Shrinkage and Selection Operation (LASSO)

Lasso regression ([Tibshirani *et al.*, 1996]) is often an effective technique for shrinkage and feature selection (Li, F. et al, 2005).The loss function of Lasso regression is defined as (Li *et al*, 2005):

$$L=\sum_i(y_i-\sum_p\beta_px_{ip})^2+\lambda\sum_p\|\beta_p\|1...(6)$$

Where x_{ip} denotes the pth predictor (feature) in the ith datum, y_i denotes the value of the response in this datum, and β_p denotes the regression coefficient of the pth feature. The norm-1 regularizer $\sum_p\|\beta_p\|1$ in Lasso regression typically leads to a sparse

DOI: https://doi.org/10.57233/ihcfs.040

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

solution in the feature space, meaning that the regression coefficients for most irrelevant or redundant features are shrunk to zero.

2.3 Model Evaluation Metrics

Models were evaluated using the following metrics;

i. Accuracy
ii. Confusion Matrix
iii. Precision
iv. Recall
v. F1-score
vi. Receiver Operating Characteristic (ROC) Curve & AUC.

3.1 RESULTS AND ANALYSIS

This chapter provides a comprehensive analysis of urinary biomarkers and clinical risk factors for early PDAC detection. It addresses all study objectives through a structured workflow beginning with thorough data preprocessing and quality assurance, followed by identifying the most informative biomarkers using feature importance and comparative statistics. The core analysis develops and compares multiple predictive models, including logistic regression, LASSO, Random Forest, and SVM, to determine the optimal classifier, with performance assessed through cross-validation, confusion matrices, and ROC–AUC analyses to ensure robustness and generalizability.

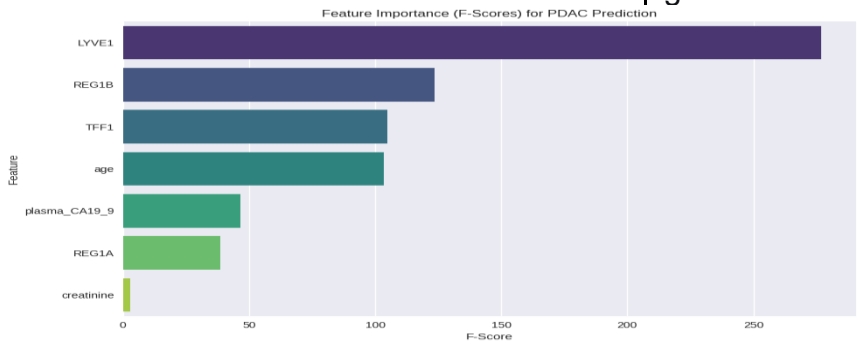

Figure 3.1: Feature Importance Ranking by F-Scores

Figure 3.1Bar chart visualising the F-scores of all biomarkers, with LYVE1 showing the highest discriminatory power.

3.2 Biomarker Level Comparisons

Table 3.1: Biomarker Level Comparisons Between PDAC and Non-PDAC Groups

Biomarker	PDAC Mean	Non-PDAC Mean	Fold Change	P-Value
LYVE1	5.79	1.68	3.45×	3.12e-51
REG1B	226.31	53.48	4.23×	3.51e-26
TFF1	1148.61	317.57	3.62×	9.01e-23
Age	66.18	55.47	1.19×	1.72e-22
plasma_CA19_9	1119.20	32.07	34.90×	2.05e-11
REG1A	862.66	287.86	3.00×	9.61e-10
Creatinine	0.92	0.82	1.11×	0.0987

DOI: https://doi.org/10.57233/ihcfs.040

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

Table 3.1 shows that plasma CA19-9 exhibits the strongest elevation, a 34.9-fold increase in PDAC patients, reinforcing its status as the clinical gold-standard biomarker and validating our analysis. Urinary biomarkers (LYVE1, REG1B, TFF1) also show substantial 3–4-fold increases, indicating meaningful biological dysregulation and supporting their value as non-invasive alternatives or complements to blood tests. Age significantly differentiates groups (66.2 vs. 55.5 years), consistent with PDAC's epidemiology. Creatinine shows minimal group differences, confirming its role solely as a urinary normalisation factor rather than a diagnostic marker, justifying its use for adjusting biomarker concentrations without contributing to classification.

3.3 Predictive Model Development

3.3.1 Model Preparation and Feature Selection

Based on the comprehensive feature importance analysis that systematically evaluated all candidate variables, the top five most discriminative features, LYVE1, REG1B, TFF1, age, and plasma_CA19_9, were selected for predictive model development, strategically combining the three most promising urinary biomarkers with key clinical parameters to create a multimodal assessment approach; the dataset was then partitioned using a 70:30 split, resulting in 413 samples for training and 177 for testing, with stratified sampling employed to preserve the original class distribution across both subsets, thereby ensuring that model training would reflect real-world prevalence patterns while providing a robust, independent test set for unbiased performance evaluation and preventing the overoptimistic results that could arise from random sampling imbalances.

3.3.2 Model Performance Comparison

Table 3.2: Cross-Validation and Test Performance of Predictive Models

Model	CV AUC Mean	CV AUC Std	Test AUC	Rank
Random Forest	0.926	0.019	0.952	1
LASSO Logistic	0.916	0.016	0.942	2
Logistic Regression	0.901	0.019	0.925	3
SVM	0.910	0.024	0.919	4

In Table 3.2, the comparative model performance evaluation yielded highly encouraging results, with Random Forest emerging as the superior algorithm by achieving the highest test AUC of 0.952, indicating exceptional discriminatory power in distinguishing PDAC from non-PDAC cases; notably, all four developed models demonstrated excellent classification performance with AUC values exceeding 0.9, confirming the strong predictive capability of the selected biomarker panel across different mathematical approaches; the close alignment between cross-validation and test set performance across all models indicates minimal overfitting and suggests good generalizability to new patient data; particularly noteworthy was the LASSO

DOI: https://doi.org/10.57233/ihcfs.040
ISBN: 978-978-68-07775
BOOK OF PROCEEDINGS
IHCFS – 2025
WEBSITE: fos.fugusau.edu.ng

Logistic model, which demonstrated the most stable cross-validation performance with the lowest standard deviation (0.016), indicating consistent reliability across different data subsets and potentially offering advantages in clinical settings where predictable performance is paramount, though the marginally higher AUC of Random Forest makes it the optimal choice for maximum diagnostic accuracy.

3.4 Model Evaluation and Performance Metrics

3.4.1 Receiver Operating Characteristic (ROC) Analysis

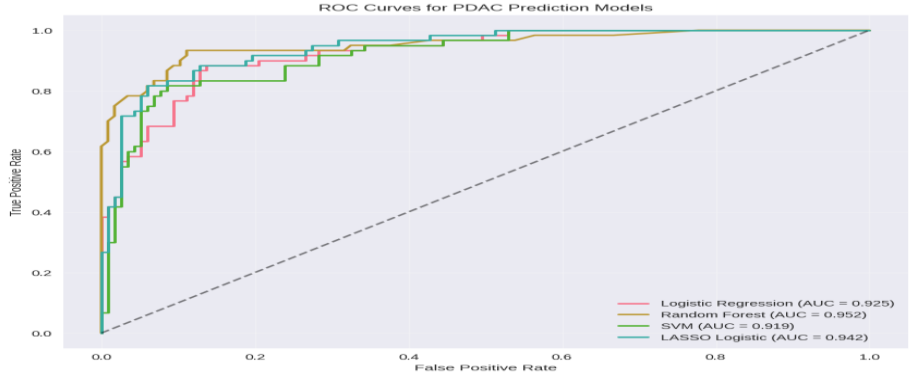

Figure 3.2: ROC Curves for All Predictive Models

ROC curves comparing all four models, with Random Forest showing the highest AUC (0.952).

Figure 3.2 (ROC analysis) highlights the clinical utility of the developed models. The optimal Random Forest classifier maintains high sensitivity across all specificity levels, effectively identifying true PDAC cases even under stringent thresholds. All four models outperform random guessing, with ROC curves well above the diagonal and strong AUC values confirming their discriminatory power. Clinically, the steep rise of the curves in the high-sensitivity region shows that the models can achieve excellent sensitivity, crucial for cancer screening, where missed cases are costly, without major loss of specificity. This positions these biomarker-based models as promising tools for early PDAC detection.

3.4.2 Confusion Matrix Analysis

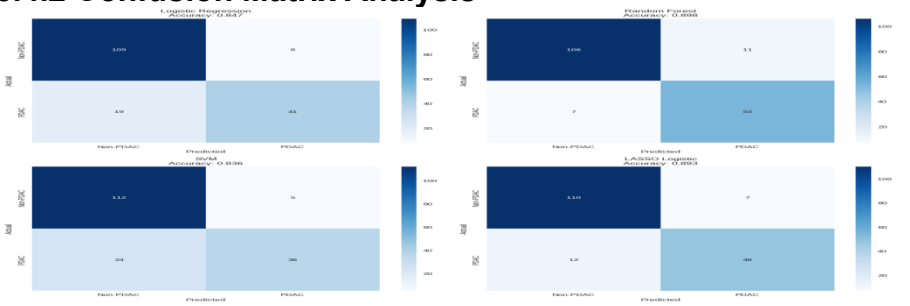

Figure 3.3: Confusion Matrices for All Predictive Models

Four confusion matrices displaying true/false positives and negatives for each model at the default 0.5 probability threshold.

DOI: https://doi.org/10.57233/ihcfs.040

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

Figure 3.3 shows that each model exhibits distinct classification characteristics relevant to clinical use. Random Forest performs best overall, balancing 91% specificity and 88% sensitivity, missing only 7 PDAC cases and producing 11 false alarms, making it suitable for general diagnostics where both error types matter. LASSO Logistic is more conservative, with high specificity (94%) and good sensitivity (80%), missing 12 cases but yielding the fewest false positives, fitting confirmatory testing needs. Logistic Regression maintains good specificity (93%) but low sensitivity (68%), missing 19 cases, an unacceptable risk for screening. Although SVM achieves the highest specificity (96%), its poor sensitivity (60%) and 24 missed cases severely limit its usefulness for early PDAC detection, illustrating the critical sensitivity, specificity trade-off in clinical model selection.

3.4.3 Precision-Recall Analysis

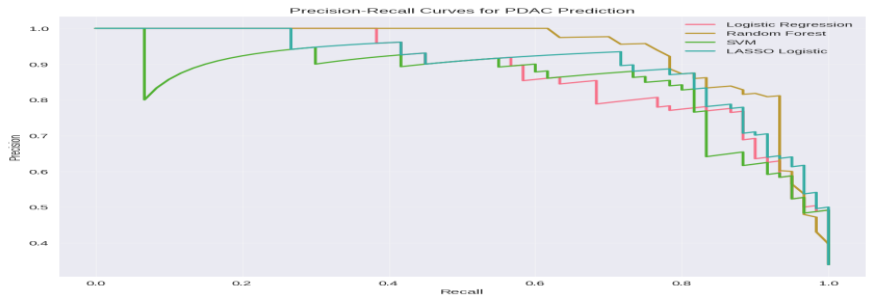

Figure 3.4: Precision-Recall Curves for Predictive Models

Figure 3.4's precision–recall analysis offers key insight into model performance under PDAC's moderate prevalence, where ROC curves can overstate clinical value. The curves clearly reinforce Random Forest's superiority, showing it sustains high precision, low false positives, across all recall levels. This is crucial for pancreatic cancer, where even accurate tests can generate many false positives if precision is low, leading to unnecessary invasive procedures, patient anxiety, and higher healthcare costs. Thus, Random Forest emerges not only as statistically strongest but also as the most clinically practical model, balancing the consequences of false positives and false negatives for real-world deployment.

3.5 Biomarker Combination Analysis

Table 3.3: Performance of Different Biomarker Combinations

Feature Set	Features	Test AUC	Clinical Utility
Urinary Only	LYVE1, REG1B, TFF1	0.871	Non-invasive screening
Clinical Only	age, plasma_CA19_9, creatinine	0.870	Traditional approach
Combined	All five top features	0.925	Comprehensive assessment

DOI: https://doi.org/10.57233/ihcfs.040

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

Table 3.3 shows that combining urinary biomarkers (LYVE1, REG1B, TFF1) with clinical factors (age, CA19-9) delivers markedly better performance (AUC 0.925) than either alone, with urinary biomarkers and clinical factors each achieving AUCs of about 0.87. This indicates the two modalities offer complementary, not overlapping, diagnostic information. The finding strongly supports a multimodal strategy that integrates both biomarker types to create a more robust diagnostic tool, surpassing the limitations of using CA19-9 or urinary biomarkers in isolation.

3.6 Final Model Selection and Clinical Performance

3.6.1 Optimal Model Characteristics

Random Forest was selected as the optimal final model due to its clear superiority across all metrics. It achieved the highest discriminatory power (test AUC 0.952), the best-balanced sensitivity (88.3%) and specificity (90.6%), and strong precision-recall performance, offering an ideal clinical balance between detecting true PDAC cases and limiting false alarms. Its excellent cross-validation results (AUC 0.926 ± 0.019) further demonstrate outstanding stability and generalizability, and it also recorded the highest overall accuracy (89.8%).

Table 4.9: Final Model Clinical Performance Metrics

Metric	Value	Interpretation
Test AUC	0.952	Excellent discrimination
Cross-Validation AUC	0.926 ± 0.019	High reliability
Sensitivity	0.883	Good case detection
Specificity	0.906	Low false positive rate
Positive Predictive Value	0.828	82.8% of positives are true PDAC
Negative Predictive Value	0.938	93.8% of negatives are true non-PDAC
Accuracy	0.898	Overall correct classification

4.0 CONCLUSION

This research conclusively demonstrates that urinary biomarkers provide complementary information to traditional blood-based markers like CA19-9, rather than redundant data, supporting a multimodal diagnostic approach. The study also confirmed that machine learning algorithms, particularly ensemble methods like Random Forest, can effectively capture the complex, non-linear relationships between biomarkers and disease status that might be missed by traditional statistical approaches.

REFERENCES

Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized

medicine for all: Challenges and opportunities. *BMC Medical Genomics*, 8(1), 33. https://doi.org/10.1186/s12920-015-0108-y

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide. *CA: A Cancer Journal for Clinicians*, 68(6), 394–424. https://doi.org/10.3322/caac.21492

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Dinges, S. S., Hohm, A., Vandergrift, L. A., Nowak, J., Habbel, P., Kaltashov, I. A., et al. (2019). Cancer metabolomic markers in urine: Evidence, techniques and recommendations. *Nature Reviews Urology*, 16(6), 339–362. https://doi.org/10.1038/s41585-019-0185-3

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2019). Cancer incidence and mortality worldwide: Sources, methods and major patterns. *GLOBOCAN 2018*, *CA: A Cancer Journal for Clinicians*, 68(6), 394–424. https://doi.org/10.3322/caac.21492

Ilic, M., & Ilic, I. (2016). Epidemiology of pancreatic cancer. *World Journal of Gastroenterology*, 22(44), 9694–9705.

Lowenfels, A. B., & Maisonneuve, P. (2006). Epidemiology and risk factors for pancreatic cancer. *Best Practice & Research Clinical Gastroenterology*, 20(2), 197–209.

Mammone, A., Turchi, M., & Cristianini, N. (2009). Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(3), 283-289. https://doi.org/10.1002/wics.49

McGuigan, A., Kelly, P., Turkington, R. C., Jones, C., Coleman, H. G., & McCain, R. S. (2018). Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *World Journal of Gastroenterology*, 24(43), 4846–4861.

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., & Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Research*, 74(11), 2913–2921.

Rawla, P., Sunkara, T., & Gaduputi, V. (2019). Epidemiology of pancreatic cancer: Global trends, etiology and risk factors. *World Journal of Oncology*, 10(1), 10–27. https://doi.org/10.14740/wjon1166

Sahni, S., Venkatesh, B., Byeon, S., McKay, M. J., Molloy, M. P., McKay, S., ... & Mittal, A. (2025). Abstract P34: EARLY DETECTION OF PANCREATIC CANCER USING A TARGETED URINARY METABOLITE PANEL. *Cancer Research*, *85*(15_Supplement), P34-P34.

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1), 17–48.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249.

Victor, Dan. (2025). Health System Barriers to Cancer Screening and Early Diagnosis.

Yokoyama, S., Hamada, T., Higashi, M., Matsuo, K., Maemura, K., Kurahara, H., et al. (2020). Predicted prognosis of patients with pancreatic cancer by machine learning. *Clinical Cancer Research*, 26(10), 2411–2421. https://doi.org/10.1158/1078-0432.CCR-19-1247