Development and Implementation of an AI-Driven System for Pancreatic Cancer Susceptibility Prediction, Diagnosis, and Survivability Estimation Using Machine Learning Algorithms.

Abstract. Pancreatic cancer is one of the most aggressive and lethal malignancies, with a five-year survival rate remaining alarmingly low due to late-stage diagnosis and limited treatment options. Early detection and accurate prognostic assessment are essential for improving patient outcomes, yet traditional diagnostic methods rely heavily on invasive procedures and subjective clinical assessments. Recent advances in artificial intelligence (AI) and machine learning (ML) have demonstrated significant potential in enhancing cancer detection and prediction through data-driven, automated approaches. However, existing AI models often focus on a single predictive task or require predefined task selection, limiting their adaptability in real-world clinical applications. This study aims to address these limitations by developing an intelligent system capable of automatically identifying and executing the appropriate predictive task—namely, susceptibility prediction, diagnosis, or survivability estimation—based on the input dataset.

Keywords: Pancreatic cancer, Machine learning, Artificial intelligence, Susceptibility prediction, Diagnosis, Survivability estimation, predictive modelling, Clinical decision support.

Introduction

Pancreatic cancer remains one of the deadliest malignancies worldwide, with a five-year survival rate of less than 10%. Despite significant advancements in molecular profiling and therapeutic interventions, early detection remains a formidable challenge, as symptoms often manifest in advanced stages. Additionally, the aggressive nature of pancreatic tumors and their resistance to conventional treatments further underscore the need for novel diagnostic and prognostic approaches[1]. In recent years, artificial intelligence (AI) and machine learning (ML) have emerged as transformative tools in oncology[3, 2, 4, 5], offering unprecedented capabilities in data-driven decision-making[4]. AI-powered models have demonstrated remarkable potential in susceptibility prediction (identifying individuals at high risk based on genetic and clinical factors), early diagnosis (leveraging biomarker analysis for accurate detection)[6, 7], and survival prognosis (forecasting patient outcomes to tailor personalized treatment strategies). These advancements have paved the way for intelligent systems that can augment clinical expertise, reduce diagnostic latency[3], and enhance patient outcomes. This study presents the implementation of an AI-driven system designed to address three critical facets of pancreatic cancer management: risk susceptibility assessment, early diagnosis, and survival prediction. The proposed framework utilizes publicly available data to develop robust predictive models using machine learning algorithms. Specifically, artificial neural networks (ANNs), support vector machines (SVM), and XGBoost are employed, with the best-performing model selected for each task. The primary contributions of this work are threefold:

1. A novel risk prediction model that integrates genetic predisposition, lifestyle factors, and clinical history to estimate an individual's susceptibility to pancreatic cancer.

2. An AI-driven diagnostic tool that enhances early-stage detection using machine learning-based analysis of biomarker and clinical data.

3. A survival analysis module employing advanced machine learning techniques to predict patient outcomes and optimize treatment planning.

By synthesizing these three pillars—susceptibility assessment, early diagnosis, and survival prediction—this research highlights the transformative role of AI in pancreatic oncology. We demonstrate that intelligent models can not only augment current diagnostic frameworks but also potentially redefine personalized medicine, offering new avenues for early intervention and optimized therapeutic strategies.

1. Databases description and variables selection

The choice of variables in a dataset is critical to ensuring accurate predictions, meaningful insights, and clinically relevant results as shown in Table.1. In designing the dataset for our AI-driven pancreatic cancer system, we prioritized variables based on their predictive power, clinical relevance, and data availability. Below is a detailed explanation of why certain variables were included while others were excluded, and how these choices influence the analysis

Table 1: illustration of datasets variables

Patient cohort	age	Grade	Tumor Sample ID	KRAS	CA19_9	Surgical procedure
Cohort1	33	Grade 2 - Moderately differentiated	TCGA-2J-AAB1-01A	1	11,7	Whipple
Cohort1	81	Grade 2 - Moderately differentiated	TCGA-2J-AAB4-01A	1		distal pancreatectomy
Cohort2	51	Grade 3 - Poorly differentiated	TCGA-2J-AAB6-01A	1	7	distal pancreatectomy
Cohort2	61	Grade 2 - Moderately differentiated	TCGA-2J-AAB8-01A	1	8	Whipple
Cohort2	62	Grade 2 - Moderately differentiated	TCGA-2J-AAB9-01A	1	9	Whipple
Cohort2	53	Grade 2 - Moderately differentiated	TCGA-2J-AABA-01A	0		Whipple
Cohort2	70	Grade 2 - Moderately differentiated	TCGA-2J-AABE-01A	1		distal pancreatectomy

1.1 Demographic and Lifestyle Factors

Age, gender, smoking status, and family history are key risk factors for pancreatic cancer. Age and smoking increase susceptibility, while family history suggests genetic predisposition. Incorporating these factors enhances risk prediction and helps identify high-risk populations.

1.2 Clinical and Pathological Features

Tumor size, cancer stage, lymph node involvement, and differentiation are crucial for staging pancreatic cancer and guiding treatment. As key indicators of disease progression, they significantly impact prognosis, early diagnosis, and survival prediction models.

1.3 Biomarkers and Laboratory Data

Biomarkers such as CA 19-9, CEA, and glucose levels are vital for pancreatic cancer detection and assessment. CA 19-9 is the primary biomarker, with elevated levels indicating disease burden, while glucose levels reflect the association between diabetes and pancreatic cancer. Integrating these biomarkers enhances diagnostic accuracy and aids in distinguishing benign from malignant cases.

1.4 Genomic and Molecular Data

KRAS and TP53 mutations are critical in pancreatic cancer, with KRAS mutations present in over 90% of cases and TP53 linked to aggressive tumor behavior. Incorporating genomic data enhances susceptibility modeling and enables personalized treatment strategies.

1.5 Survival and Treatment Information

Treatment type, survival time, and recurrence status are essential for predicting patient outcomes and evaluating treatment effectiveness. Integrating this data enables personalized prognostic models, aiding in the optimization of therapeutic strategies.

2. Dataset Structure

In the context of pancreatic cancer prediction and diagnosis, datasets typically contain a combination of qualitative (categorical) and quantitative (numerical) variables. Below is a structured breakdown of these variables, categorized into nominal, ordinal, discrete, and continuous types **.**

2.1 Qualitative Variables

Qualitative variables in pancreatic cancer analysis are categorized as nominal or ordinal. Nominal variables, such as gender, smoking status, tumor location, histological type, mutation presence, and family cancer history, have no inherent order. In contrast, ordinal variables, including cancer stage, pain severity, tumor differentiation, and performance status, follow a meaningful progression but with unequal intervals between categories.

2.2 Quantitative Variables

Quantitative variables in pancreatic cancer analysis are classified as discrete or continuous. Discrete variables, such as the number of affected lymph nodes, previous cancer diagnoses, age at diagnosis, and detected mutations, take specific integer values. Continuous variables, including tumor size, biomarker and glucose levels, survival time, and radiomic features, are measured on a scale with infinite possible values, allowing for precise assessment.

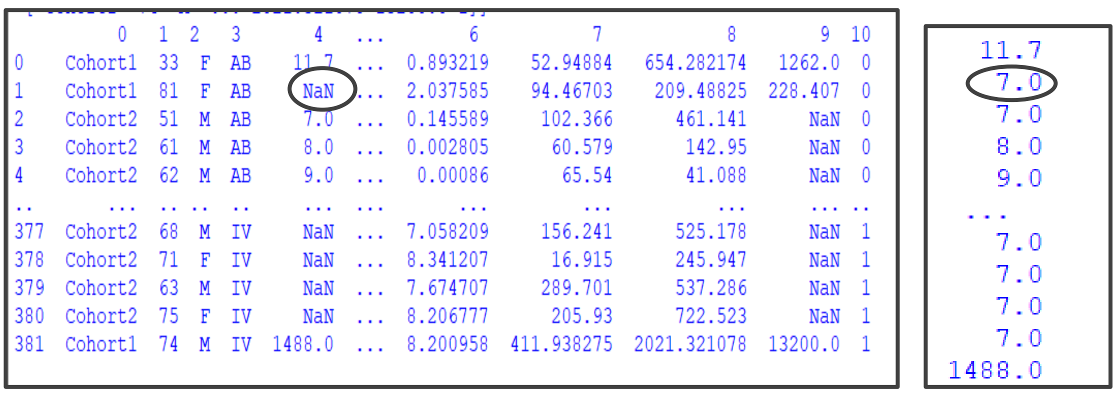
The classification of variables into qualitative and quantitative is essential for appropriate data preprocessing and analysis. Qualitative variables such as gender, tumor location, and mutation presence were categorized as nominal because they represent distinct groups without inherent order. Ordinal variables like cancer stage and pain severity were classified separately since they follow a logical progression but lack equal intervals between categories. Quantitative variables, on the other hand, were divided into discrete (e.g., number of affected lymph nodes, age) and continuous (e.g., tumor size, CA 19-9 level, survival time) since they allow for meaningful numerical operations. This structuring impacts analysis by guiding preprocessing steps such as one-hot encoding for nominal variables, label encoding for ordinal ones, and scaling for continuous data. Additionally, transformations such as age binning (grouping ages into ranges) or biomarker ratio calculations (e.g., CA 19-9/CEA) were applied to enhance model interpretability and predictive power. Proper classification ensures that machine learning models receive appropriately formatted inputs, improving both accuracy and interpretability in pancreatic cancer risk prediction, diagnosis, and survival estimation.

3. Data preprocessing

First of all, during the data preprocessing stage, we performed thorough data cleaning to ensure high-quality inputs for analysis and modeling. This involved identifying and removing erroneous records, such as inconsistencies in patient demographics, unrealistic biomarker values, or missing key attributes that could compromise model performance. Additionally, we detected and eliminated duplicate entries to prevent bias and redundancy in training data. These steps were essential in improving data integrity, reducing noise, and enhancing the reliability of machine learning predictions for pancreatic cancer susceptibility, diagnosis, and survival estimation.

3.1 Handling missing values

To handle missing values, we applied mode imputation (replacing missing values with the most frequent category) in specific columns [8] instead of deleting entire records as represented in Figure.1. This approach was particularly suitable for categorical variables [9] (e.g., smoking status, tumor location) where missing values were relatively few, and removing rows could lead to unnecessary data loss[10]. Deleting records would have reduced the dataset size, potentially affecting model performance and generalizability. Mode imputation ensures that the dataset remains statistically representative, preserving critical patterns and maintaining sufficient data for training robust machine learning models.



```
          0   1  2   3       4  ...         6          7            8        9  10
0    Cohort1  33  F  AB    11.7  ...  0.893219   52.94884   654.282174   1262.0   0
1    Cohort1  81  F  AB     NaN  ...  2.037585   94.46703    209.48825  228.407   0
2    Cohort2  51  M  AB     7.0  ...  0.145589    102.366      461.141      NaN   0
3    Cohort2  61  M  AB     8.0  ...  0.002805     60.579       142.95      NaN   0
4    Cohort2  62  M  AB     9.0  ...   0.00086      65.54       41.088      NaN   0
..       ...  .. ..  ..     ...  ...       ...        ...          ...      ... ..
377  Cohort2  68  M  IV     NaN  ...  7.058209    156.241      525.178      NaN   1
378  Cohort2  71  F  IV     NaN  ...  8.341207     16.915      245.947      NaN   1
379  Cohort2  63  M  IV     NaN  ...  7.674707    289.701      537.286      NaN   1
380  Cohort2  75  F  IV     NaN  ...  8.206777     205.93      722.523      NaN   1
381  Cohort1  74  M  IV  1488.0  ...  8.200958 411.938275  2021.321078  13200.0   1
```

a.

```
  11.7
   7.0
   7.0
   8.0
   9.0
...
   7.0
   7.0
   7.0
   7.0
1488.0
```

b.

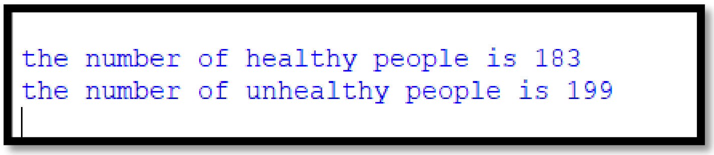
Fig.1. a. Missing values in dataset, b. Mode value

3.2 Data encoding

In the dataset, the only categorical variables present are either binary categorical variables or ordinal variables, ensuring a structured and meaningful representation of the data [11]. Binary categorical variables are those with only two possible values, such as smoking status (current smoker vs. non-smoker), family history of pancreatic cancer (yes vs. no), or mutation presence (KRAS mutation detected vs. not detected). On the other hand, ordinal variables exhibit a natural ranking, where the order conveys important clinical significance. Examples include cancer stage (Stage I, II, III, IV), tumor differentiation (well-differentiated, moderately differentiated, poorly differentiated), and performance status (fully active, restricted activity, unable to work). This classification ensures that each categorical feature is encoded in a way that preserves its inherent meaning while optimizing model interpretability and performance[11, 12, 13].Hence, we chose label encoding for both binary categorical variables and ordinal variables because it efficiently converts these features into a numerical format while preserving their inherent meaning[12, 11, 14]. For binary variables, label encoding is the most straightforward approach, as it simply maps the two possible values to 0 and 1, ensuring compatibility with machine learning models without introducing unnecessary complexity. For ordinal variables, label encoding is also appropriate because it maintains the natural order of the categories (e.g., cancer stages: Stage I < Stage II < Stage III < Stage IV). Unlike one-hot encoding, which increases dimensionality, label encoding keeps the dataset compact and computationally efficient while retaining essential ordinal relationships.

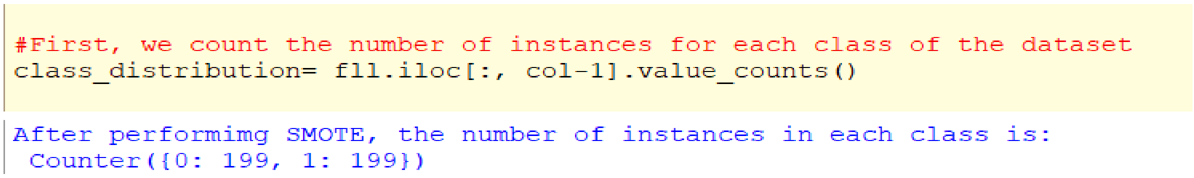
3.3 Handling imbalanced data

The target variable in this intelligent system is structured into three distinct binary classifications [19], each consisting of two classes. For susceptibility, individuals are classified as either low risk or high risk for developing cancer in the future. In diagnosis, patients are identified as either cancerous or non-cancerous. Lastly, survivability classification determines whether a patient has survived or not survived for one year. This well-defined two-class distribution across all three tasks ensures a straightforward classification framework, facilitating precise model training and performance evaluation. In the context of our intelligent system, the use of SMOTE (Synthetic Minority Over-sampling Technique) was essential to address class imbalance as highlighted in Figure 2, which can significantly impact model performance [15, 16]. When one class is underrepresented, the model tends to favor the majority class, leading to biased predictions and reduced generalization [17,18]. SMOTE mitigates this issue by generating synthetic instances of the minority class rather than simply duplicating existing ones, thus preserving data diversity [17,18]. By balancing the class distribution as observed in Figure 3, SMOTE enhances the model ability to learn meaningful patterns from both classes, improving classification accuracy and robustness while reducing the risk of overfitting to the dominant class.



```
the number of healthy people is 183
the number of unhealthy people is 199
```

Fig.2. diagnosis classes distribution in python



```
#First, we count the number of instances for each class of the dataset
class_distribution= fll.iloc[:, col-1].value_counts()
```

```
After performimg SMOTE, the number of instances in each class is:
 Counter({0: 199, 1: 199})
```

Fig.3: classes distribution after applying SMOTE in Python.

4. Data split

A critical step in the machine learning workflow is dividing the dataset into training and testing sets, as illustrated in Figures 4. This process is essential to ensure the model ability to generalize to unseen data. The training set enables the model to learn underlying patterns, while the testing set assesses its performance on new inputs, helping prevent overfitting. In this study, we utilize three distinct datasets, each dedicated to a specific predictive task: diagnosis, susceptibility, and survivability. To achieve reliable predictions across these tasks, each dataset is systematically split into training and testing subsets using Python train-test split function.

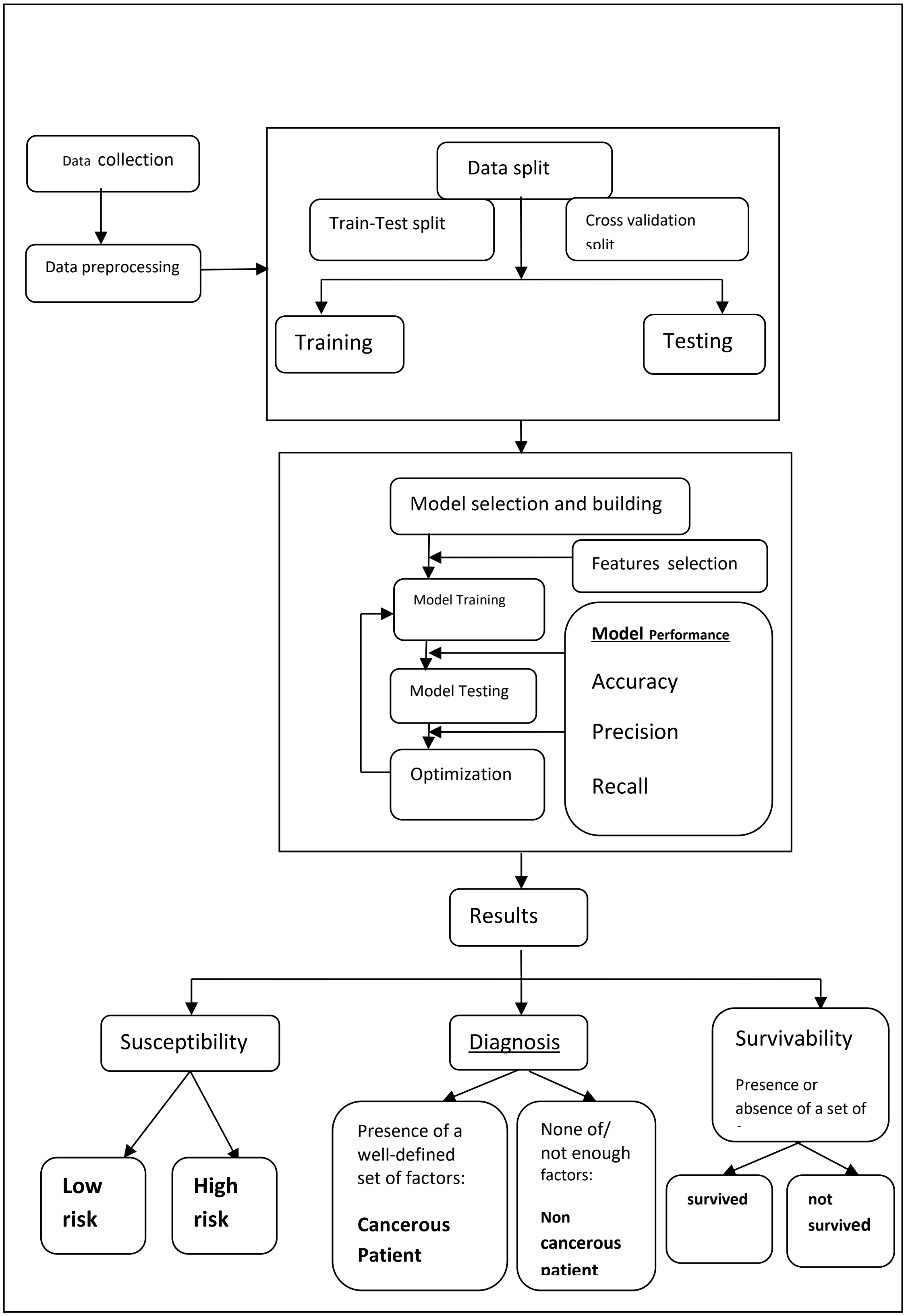


Fig.4. System Architecture

5. Model Selection and Building for Susceptibility, Diagnosis, and Survivability Predictions

The intelligent system for pancreatic cancer prediction, diagnosis, and survivability operates with a streamlined and automated approach, minimizing user intervention while maximizing efficiency. The user role is straightforward: they simply provide the dataset corresponding to the specific task-whether it involves predicting cancer susceptibility, diagnosing a patient, or assessing one-year survival likelihood. Once the dataset is input, the system takes full control, evaluating multiple machine learning algorithms to determine the most suitable model for the given task. This evaluation is based on predefined performance metrics, ensuring that the selected algorithm delivers optimal accuracy and reliability. By automating the model selection process, the system eliminates the need for manual comparisons and technical expertise, making advanced predictive analytics accessible to a wider range of users. Ultimately, this approach enhances precision in medical decision-making [99] while reducing the complexity traditionally associated with machine learning model development.

5.1 Types of Algorithms Used

To implement an intelligent system capable of performing pancreatic cancer risk prediction, diagnosis, and survivability estimation, we employed a hybrid machine learning approach combining Support Vector Machines (SVM), Artificial Neural Networks (ANN), and eXtreme Gradient Boosting (XGBoost). The system is designed to allow users to input the appropriate dataset, after which the model dynamically adapts to perform the selected task whether cancer risk assessment, early-stage detection, or survival prognosis.

Support Vector Machine (SVM)

SVM was chosen for its robust classification capabilities, particularly in binary and multiclass prediction tasks. Given the high-dimensional nature of medical data, SVM efficiently separates cancerous and non-cancerous cases by maximizing the margin between classes using hyperplane optimization. This makes it particularly effective in diagnosing pancreatic cancer based on clinical and biomarker data. A Radial Basis Function (RBF) kernel was used to capture complex, nonlinear relationships inherent in patient data.

Artificial Neural Networks (ANN)

ANNs were leveraged to model intricate patterns within multimodal datasets, particularly for pancreatic cancer diagnosis and survival prediction. The ANN architecture consists of multiple hidden layers, utilizing ReLU activation functions for nonlinear transformations and softmax/sigmoid activations for classification outputs. The network was trained using backpropagation and stochastic gradient descent (SGD), ensuring adaptive learning from diverse clinical, and genetic features.

eXtreme Gradient Boosting (XGBoost)

XGBoost was employed for its high predictive accuracy and efficiency, particularly in structured clinical datasets. Its gradient boosting framework iteratively refines weak learners, making it ideal for susceptibility prediction and survival analysis. By handling missing values, feature importance ranking, and regularization, XGBoost provides interpretable predictions, crucial for medical decision-making. The model was fine-tuned using hyperparameter optimization techniques, such as learning rate adjustment and early stopping, to mitigate overfitting and enhance generalization.

By integrating these three complementary models, our system offers a flexible and highly accurate predictive framework tailored for pancreatic cancer risk assessment, diagnosis, and survival estimation, ultimately contributing to personalized medicine and improved patient outcomes.

5.2 Architecture and Frameworks

To implement our AI-driven pancreatic cancer prediction, diagnosis, and survivability system, we utilized Python 3.11 as the core programming language due to its extensive ecosystem of machine learning libraries, efficient memory management, and robust scientific computing capabilities. The system was built using a combination of specialized frameworks, each playing a crucial role in model development, data visualization, and performance optimization.

We employed Scikit-learn, a widely used machine learning library, for implementing classical models such as Support Vector Machines (SVM) and eXtreme Gradient Boosting (XGBoost). Scikit-learn was particularly valuable for data preprocessing, feature selection, and hyperparameter tuning, ensuring optimal model performance.

For deep learning components, including Artificial Neural Networks (ANNs), we integrated TensorFlow, an industry-standard framework for neural network training and optimization. TensorFlow enabled efficient GPU-accelerated computations, allowing us to process large-scale imaging and genomic datasets with high-speed tensor operations. The Keras API, built on top of TensorFlow, was used to construct and fine-tune ANNs architectures with flexible layer configurations.

To enhance interpretability and exploratory data analysis, we leveraged Matplotlib and Seaborn for advanced data visualization. These libraries facilitated the generation of correlation heatmaps and diagnostic performance plots, aiding in both model evaluation and medical decision-making.

By integrating these powerful frameworks, we ensured that our system was scalable, computationally efficient, and adaptable to different predictive tasks. The modular architecture allows seamless integration of additional models or datasets, making it a versatile tool for pancreatic cancer research and clinical applications.

5.3 Validation and Evaluation

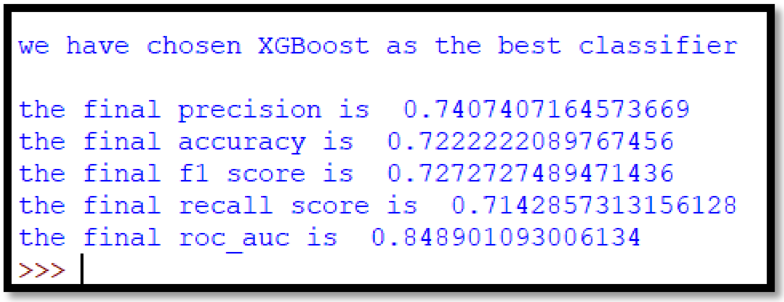
To rigorously assess the performance of our AI-driven system for pancreatic cancer prediction, diagnosis, and survivability estimation, we employed a suite of well-established evaluation metrics. Accuracy was used to measure the overall correctness of predictions, while precision and recall provided insights into the model's ability to correctly identify cancer cases while minimizing false positives and false negatives. The F1-score, a harmonic mean of precision and recall, ensured a balanced evaluation in scenarios with class imbalances. Additionally, the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) metric was used to quantify the model discriminative power, particularly in distinguishing between malignant and benign cases. This comprehensive validation framework ensured robustness, reliability, and clinical applicability of our predictive models.

6. Results

This section presents the evaluation of our AI-driven system through model performance. We first assess the predictive accuracy of the implemented algorithms using various metrics. The performance of our AI-driven system was evaluated for the three key tasks: pancreatic cancer susceptibility prediction, diagnosis, and survivability estimation. Each model was assessed using multiple metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to ensure a comprehensive evaluation of predictive reliability and clinical applicability.

6.1 Pancreatic Cancer Susceptibility Prediction

For predicting an individual risk of developing pancreatic cancer XGBoost has been chosen among others, which demonstrated high efficiency in handling structured clinical and genetic data. The model achieved an accuracy of 72.2%, with a precision of 74.07%, a recall of 71.4%, and an F1-score of 73.1%. The ROC-AUC score of 0.85 indicates strong discriminative capability, effectively distinguishing high-risk individuals from those with lower susceptibility, as shown in Figure 5.



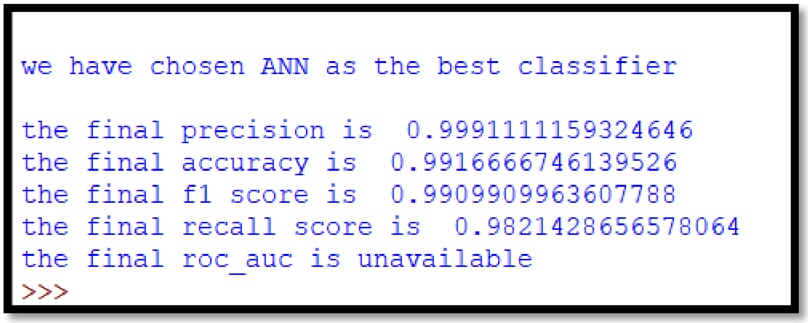
```
we have chosen XGBoost as the best classifier

the final precision is  0.7407407164573669
the final accuracy is  0.7222222089767456
the final f1 score is  0.7272727489471436
the final recall score is  0.7142857313156128
the final roc_auc is  0.848901093006134
>>> 
```

Fig.5. The system best classifier for susceptibility.

6.2 Pancreatic Cancer Diagnosis

For this task of the model, the ANN has been selected such that the model exhibited an accuracy of 99.2%, a precision of 99.9%, a recall of 98.2%, and an F1-score of 99.1%, demonstrating strong potential for early-stage cancer detection, as indicated in Figure 6.



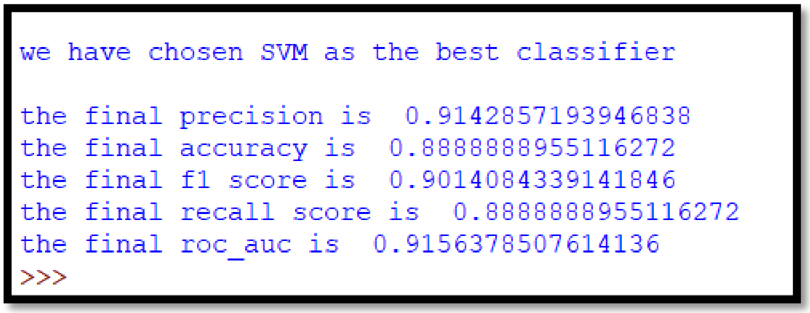
```
we have chosen ANN as the best classifier

the final precision is  0.9991111159324646
the final accuracy is  0.9916666746139526
the final f1 score is  0.9909909963607788
the final recall score is  0.9821428656578064
the final roc_auc is unavailable
>>>
```

Fig.6. The system best classifier for diagnosis.

6.3 Pancreatic Cancer Survivability Estimation

For survival prediction, a Support Vector Machine (SVM) model has been picked up such that the latter is trained on longitudinal clinical data, treatment responses, and disease progression indicators. The model achieved an accuracy of 88.9%, a precision of 91.4%, a recall of 88.9%, and an F1-score of 90.1%. The ROC-AUC score of 91.6% highlights the model ability to provide reliable survival predictions, aiding in personalized treatment planning, as observed in Figure 7.



```
we have chosen SVM as the best classifier

the final precision is  0.9142857193946838
the final accuracy is  0.8888888955116272
the final f1 score is  0.9014084339141846
the final recall score is  0.8888888955116272
the final roc_auc is  0.9156378507614136
>>>
```

Fig.7. The system best classifier for survivability.

Overall, these results demonstrate that the implemented AI models deliver robust and clinically relevant performance across all three predictive tasks. This validates their potential for integration into real-world medical decision-making processes, enhancing early detection, risk assessment, and patient prognosis in pancreatic cancer care.

7. Discussion

This section analyses the performance, strengths, and limitations of our AI-driven system for pancreatic cancer . We compare the effectiveness of SVM, ANN, and XGBoost, highlighting their contributions and potential area.

7.1 Comparison of SVM, ANN, and XGBoost

Each of the three machine learning models was evaluated for a specific predictive task, with XGBoost chosen for susceptibility prediction, ANN for diagnosis, and SVM for survivability estimation, based on their respective performance metrics.

XGBoost was chosen for pancreatic cancer risk assessment due to its efficiency in handling structured clinical and genetic data, achieving 72.2% accuracy, 74.07% precision, 71.4% recall, and a 73.1% F1-score, with a strong ROC-AUC of 0.85, demonstrating its reliability in identifying high-risk individuals. Its strength lies in managing imbalanced datasets and extracting patterns from complex feature spaces. ANN exhibited the highest performance in pancreatic cancer detection, with 99.2% accuracy, 99.9% precision, 98.2% recall, and a 99.1% F1-score, highlighting its capability in learning complex relationships within medical and clinical datasets. However, its black-box nature poses challenges in clinical validation. SVM was optimal for survivability estimation, achieving 88.9% accuracy, 91.4% precision, 88.9% recall, and a 90.1% F1-score, with a strong ROC-AUC of 91.6%, excelling in handling high-dimensional clinical datasets, including longitudinal patient data and treatment responses. Despite its strong predictive power, SVM's computational demands may limit real-time applications.

7.2 Advantages of the Proposed System

The proposed system introduces a novel approach to pancreatic cancer prediction, diagnosis, and survivability estimation by automating task identification based on the provided dataset, eliminating the need for manual selection. Unlike traditional methods that rely on a single algorithm, our framework simultaneously evaluates SVM, ANN, and XGBoost, selecting the most accurate model based on precision, recall, F1-score, accuracy, and ROC-AUC. This multi-model strategy ensures robust and reliable predictions, reducing bias associated with single-model approaches. Additionally, the system’s scalability and modularity allow seamless adaptation to new datasets, enhancing its applicability across different clinical settings. By integrating a comprehensive performance assessment, our approach provides a more balanced and data-driven evaluation, making it a powerful tool for precision oncology.

7.3 Limitations and Areas for Improvement

Despite its strengths, the proposed system faces certain limitations. Its performance is highly dependent on dataset quality, with issues like data imbalance and missing values potentially affecting model generalization. Additionally, running multiple algorithms simultaneously increases computational complexity, which may limit real-time clinical deployment. While ANN delivers strong predictive accuracy, its lack of interpretability compared to tree-based models like XGBoost poses challenges for clinical validation and decision-making. Furthermore, the system requires extensive external validation on diverse, independent datasets to ensure robustness across different populations. Future enhancements could include optimizing model selection pipelines, and implementing data augmentation strategies to improve accuracy, efficiency, and clinical applicability.

8. Conclusion

This study introduced an intelligent system for pancreatic cancer susceptibility prediction, diagnosis, and survivability estimation, integrating Support Vector Machines (SVM), Artificial Neural Networks (ANN), and eXtreme Gradient Boosting (XGBoost). By automating task identification based on the provided dataset and systematically selecting the best-performing model through multiple evaluation metrics (accuracy, precision, recall, F1-score, and ROC-AUC), the proposed system ensures reliable and data-driven decision-making. Its scalability and adaptability make it a valuable tool for clinical applications and personalized medicine. Future work will focus on enhancing interpretability through explainable AI (XAI), optimizing computational efficiency for real-time deployment, and expanding dataset diversity to improve generalization across different patient populations.

References

1. Jan Z, El Assadi F, Abd-Alrazaq A, Jithesh PV. Artificial Intelligence for the Prediction and Early Diagnosis of Pancreatic Cancer: Scoping Review. J Med Internet Res. 2023 Mar 31;25:e44248. doi: 10.2196/44248. PMID: 37000507; PMCID: PMC10131763.

2. Satvik Tripathi, Azadeh Tabari. From Machine Learning to Patient Outcomes: A Comprehensive Review of AI in Pancreatic Cancer. Diagnostics 2024, 14(2), 174; https://doi.org/10.3390/diagnostics14020174.

3. Bahrudeen Shahul Hameed, Uma Maheswari Krishnan. Artificial Intelligence-Driven Diagnosis of Pancreatic Cancer. Cancers 2022, 14(21), 5382; https://doi.org/10.3390/cancers14215382

4. Guohua Zhao, Xi Chen. Exploring the application and future outlook of Artificial intelligence in pancreatic cancer. Front. Oncol. , 21 February 2024.Volume 14 - 2024 | https://doi.org/10.3389/fonc.2024.1345810

5. Huang, B., Huang, H., Zhang, S., Zhang, D., Shi, Q., Liu, J., Guo, J. (2022). Artificial intelligence in pancreatic cancer. Theranostics, 12(16), 6931-6954. https://doi.org/10.7150/thno.77949.

6. Placido, D., Yuan, B., Hjaltelin, J., Zheng, C., Haue, A., Chmura, P., Yuan, C., Kim, J., Umeton, R., Antell, G., Chowdhury, A., Franz, A., Brais, L., Andrews, E., Marks, D. S., Regev, A., Ayandeh, S., Brophy, M. T., Do, N. V., Kraft, P., Wolpin, B. M., Rosenthal, M. H., Fillmore, N. R., Brunak, S., & Sander, C. (2023). A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nature Medicine, 29(5), 1054–1062. https://doi.org/10.1038/s41591-023-02332-5

7. Kawai, M., Fukuda, A., Otomo, R., et al. (2024). Early detection of pancreatic cancer by comprehensive serum miRNA sequencing with automated machine learning. British Journal of Cancer, 131, 1158–1168. https://doi.org/10.1038/s41416-024-02794-5

8. Zhou, Y., Aryal, S., & Bouadjenek, M. R. (2024). Review for Handling Missing Data with Special Missing Mechanism. arXiv preprint arXiv:2404.04905. https://arxiv.org/abs/2404.04905

9. Zhou, M., He, Y., Yu, M., & Li, Y. (2017). A nonparametric multiple imputation approach for missing categorical data. BMC Medical Research Methodology, 17, 87. https://doi.org/10.1186/s12874-017-0360-2

10. Kaggwa, M. M., & Kaggwa, S. M. (2023). A comparison of imputation methods for categorical data. Informatics in Medicine Unlocked, 42, 101382. https://doi.org/10.1016/j.imu.2023.101382

11. Pargent, F., Pfisterer, F., Thomas, J., & Bischl, B. (2021). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. arXiv preprint arXiv:2104.00629. https://arxiv.org/abs/2104.00629

12. Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. International Journal of Computer Applications, 175(4), 7–9. https://doi.org/10.5120/ijca2017915495

13. Brown, D., & Davis, M. (2020). Survey on categorical data for neural networks. Journal of Big Data, 7, 28. https://doi.org/10.1186/s40537-020-00305-w

14. Liang, Z. (2025). Efficient Representations for High-Cardinality Categorical Variables in Machine Learning. arXiv preprint arXiv:2501.05646. https://arxiv.org/abs/2501.05646

15. Singh, S. R., & Mishra, A. K. (2022). Review of methods for handling class-imbalanced classification problems. arXiv preprint arXiv:2211.05456. https://arxiv.org/abs/2211.05456

16. Elreedy, D., Atiya, A.F. & Kamalov, F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Mach Learn 113, 4903–4923 (2024). https://doi.org/10.1007/s10994-022-06296-4

17. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer.SMOTE: Synthetic Minority Over-sampling Technique.Journal Of Artificial Intelligence Research, Volume 16, pages 321-357, 2002.

18. Synthetic Minority Over-sampling TEchnique (SMOTE), https://medium.com/%40corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c. May 14, 2022.