APPENDIX FOR THE ECIS 2024 PAPER:

ARE OUR PREDICTIONS HEALTHY? A COMPARATIVE META-ANALYSIS OF MACHINE LEARNING STUDIES IN PREDICTIVE HEALTHCARE

~~*Completed Research Paper*~~ *Appendix*

Kaj Heinrich, ~~Otto-von-Guericke-Universität~~OVGU Magdeburg, Germany, kai.heinrich@ovgu.de

Armin Keshavarzi~~John Doe~~, ~~Otto-von-Guericke-Universität~~OVGU Magdeburg, Germany, -armin.keshavarzi@st.ovgu.de

~~University of Lisbon, Uppsala, Portugal, john.doe@university.edu~~

hat formatiert: Deutsch (Deutschland)

hat formatiert: Deutsch (Deutschland)

hat formatiert: Deutsch (Deutschland)

hat formatiert: Deutsch (Deutschland)

hat formatiert: (keine)

Formatiert: Einzug: Links: 0 cm, Erste Zeile: 0 cm

Abstract

Predictive healthcare in the case of pancreatic neuroendocrine tumors (PNETs) is a crucial operation as treatment challenges arise due to the heterogeneity of the disease. Surgical approaches vary based on aggressiveness, ranging from resection for milder cases to extensive removal for aggressive PNETs. Thus, machine learning (ML) models are crucial for precise prediction and categorizing PNETs for enhanced outcome forecasting. This systematic review sheds light on the practices of ML approaches within a comparative meta-analysis and a quality assessment employing the standardized IJMEDI checklist. The results show that ML studies within the field of predictive healthcare, despite their potential, face challenges like inadequate data preprocessing, unclear model architecture, and limited clinical applicability.

Keywords: Predictive healthcare, systematic review, meta-analysis, quality.

~~1 Introduction~~

~~Pancreatic neuroendocrine tumors (PNETs) are the second-most common type of solid tumors in the pancreas and originate from neuroendocrine cells. The primary approach for treating PNETs is surgery, the sole method for achieving a cure. Various surgical techniques are available based on the tumor's level of aggressiveness. In cases of less aggressive PNETs, surgical options like exenteration or resection can mitigate surgical risks while not impacting patient prognosis. Conversely, aggressive PNENs necessitate extensive removal along with lymph node dissection, and in cases where required, combination with pharmaceutical intervention is advised to lower the chances of recurrence (*Huang J, Xie X.Pdf* n.d.).~~

~~Utilizing machine learning (ML) models to predict PNETs, rare and heterogeneous tumors involving the pancreas, is significant for several reasons. PNETs are uncommon and diverse growths exhibiting varying biological behaviors and prognoses based on their grade, stage, and molecular attributes. ML models can potentially categorize PNETs into subtypes and pinpoint the most pertinent characteristics for forecasting outcomes (Park et al. 2023). Secondly, they often present with no symptoms or nonspecific symptoms, making them difficult to diagnose early. Hence, prediction models can facilitate the detection of PNETs from diverse data sources, such as imaging, blood tests, or biopsies, and provide a more accurate and prompt diagnosis (Murakami et al. 2023).~~

Moreover, PNETs have limited treatment options and frequently develop resistance to conventional therapies. In this regard, personalizing the treatment can be done by Machine learning models and results in selecting the most effective drugs, doses, and combinations for each patient based on their genomic and phenotypic profiles (Chen et al. 2023). They can also aid in monitoring the response and progression of PNETs and adjusting the treatment accordingly by predicting the survival rate in various stages of the patient's life cycle (Jiang et al. 2023). This way, patients' quality of life and survival rates in PNET cases could be improved, and unnecessary, ineffective treatments could be avoided, saving resources and costs (Murakami et al. 2023). Additionally, evaluating and estimating the cost-effectiveness of different PNET management strategies, considering clinical outcomes and patient and medical provider preferences, can assist in the decision-making process. Moreover, these models can provide tools for decision support, guiding clinical decisions, and enhancing patient outcomes. As a result, the quality and efficiency of PNET care could be increased while uncertainties and practice variations are reduced (White et al. 2021).

The number of PNET patients has been increasing in recent years due to advanced diagnostic methods and the growing recognition of endocrine tumors (Han et al. 2022; Li et al. 2023). There has been an increasing effort in constructing ML models to predict PENT in recent years, but to the best of our knowledge, there is no systematic review of current works in this domain; therefore, it is essential to establish an insight by conducting a comparative meta-analysis. This systematic review aimed to 1) find studies that have used Machine Learning technologies to forecast PNET related outcomes, 2) evaluate the quality of the included literature using the IJMEDI checklist, 3) classify the tasks of different forecasting studies and summarize the development process and performance of models, and 4) identify the current gaps and challenges in building predictive ML based models to offer references for the creation of similar predictive tools in the future. This decision was made due to the increasing utilization of machine learning (ML) in forecasting tasks for PNETs. However, it is worth noting that significant variability exists across different studies regarding forecasting objectives, data sources, derivation and validation techniques, and model performance metrics.

2 Related Work

In this section we focus on related meta-analysis studies in predictive healthcare that are related to our work in methodology and result presentation. A recent study by Yin et al. (2023) systematically reviews predictive models for pancreatic diseases based on endoscopic ultrasound (EUS) using deep learning techniques. The authors extensively searched PubMed and Semantic Scholar databases. The selected studies developed deep learning models for pancreatic diseases utilizing EUS and were evaluated based on the IJMEDI checklist for quality assessment. The review encompasses 23 studies grouped by computer vision tasks such as classification, detection, and segmentation. Various neural network architectures were employed in these categories, emphasizing convolutional neural networks. Remarkably, all the models demonstrated robust performance when applied to EUS images, videos, or voice data. The IJMEDI checklist identified six studies as high-quality, surpassing a score of 35 points.

In a separate work, Zhou et al. (2022) present a systematic review of ML models for predicting outcomes in acute pancreatitis (AP). The authors extensively searched PubMed, Web of Science, Scopus, and Embase databases. The study spanned from inception to May 29, 2021. Studies that employed ML for predictive tools related to AP were included, and the IJMEDI checklist was used for quality assessment. Among 2,913 articles, 24 met the criteria, involving 8,327 patients and resulting in 47 models. These studies were categorized into tasks such as predicting severity, complications, mortality, recurrence, and surgery timing. While ML models showed promising accuracy in these tasks, most studies were retrospective, single-center, and database-driven and lacked external validation. According to the IJMEDI checklist and the authors' criteria, two studies were classified as high quality. Nonetheless, most studies revealed biases in data preparation, validation, and deployment aspects.

~~Another study by Liu et al. (2023) presents a systematic review of ML models for predicting survival outcomes in bladder cancer. The authors searched PubMed, Web of Science, and Embase databases. Fifteen articles meeting inclusion criteria were included, and the Quality Assessment of Diagnostic Accuracy Studies 2 tool evaluated these studies. ML models displayed potential in predicting bladder cancer survival outcomes, but most studies shared characteristics such as being retrospective, single-center, database-oriented, and lacking external validation. The authors emphasize the potential of ML in aiding bladder cancer survival predictions while highlighting the necessity for broader validation in diverse patient populations. The prevalent algorithms in the included articles were artificial neural networks and logistic regression. Most studies were of medium quality per the IJMEDI checklist, with room for enhancement in data preparation and deployment descriptions. The study concludes that challenges in data processing, feature selection, and data source quality must be addressed for robust model development.~~

~~Kamel Rahimi et al. (2022) undertake a systematic review of ML models aiming to enhance care for hospitalized adult diabetes patients using electronic medical records (EMR) data. The authors searched four databases for relevant studies between January 2010 and January 2022. The study included research that developed and validated ML models for diabetes management with EMR data, excluding primary care and community care settings. Data extraction and critical appraisal were conducted using the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS). At the same time, the Prediction Model Risk Of Bias Assessment Tool (PROBAST) assessed bias risk and the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guideline evaluated reporting quality. ML model quality was assessed using the IJMEDI checklist, and external validation methodologies were reviewed. Out of 1317 screened studies, twelve met the inclusion criteria. These studies developed ML models for various purposes, such as predicting disglycemic episodes, total insulin dosage, and risk of readmission for hospitalized diabetes patients. The studies varied in ML types, cohort characteristics, input predictors, sample sizes, and performance metrics. Although a couple of studies adhered to the TRIPOD guideline, methodological reporting across studies displayed a high risk of bias. The quality of ML models was generally considered poor. None of the studies underwent robust external validation or practical implementation in clinical settings.~~

~~In the study by Lai et al. (2020), a systematic review concerning AI-based prognostic models for hepatocellular carcinoma (HCC) is presented. The authors conducted an extensive web-based literature search using keywords related to artificial intelligence, deep learning, and HCC. Nine relevant articles were identified, and their quality was evaluated using the Risk of Bias In Non-randomized Studies of Interventions tool. The studies employed various AI methodologies, including artificial neural networks, support vector machines, artificial plant optimization, and peritumoral radiomics. Six studies focused on artificial neural networks, while the remaining three used different AI techniques. These studies consistently compared the performance of artificial neural networks against traditional statistics, using training cohorts to develop neural networks for subsequent validation. AI models consistently outperformed traditional statistics, exhibiting notably improved predictive performance as indicated by areas under the curve. However, most studies had retrospective, single center designs, relied on database data, and lacked external validation. Despite the demonstrated potential of AI in aiding HCC prognosis decision-making, the authors concluded that further research is necessary to validate these models across larger and more diverse patient populations.~~

~~In Harding Theobald et al. (2021) work, a systematic review assessed the effectiveness of radiomics-based predictive models in diagnosing and predicting the prognosis of HCC. The authors extensively searched PubMed, Web of Science, and Embase databases, identifying and analyzing 54 relevant studies. The radiomic features employed demonstrated strong discriminatory ability, distinguishing HCC from other solid lesions. Several consistent features were highlighted as influential for diagnostic and prognostic radiomic tools, including imaging skewness, peritumoral region assessment, and arterial imaging phase feature extraction. Despite the promising findings, the overall quality of the included studies was observed to be low. Common deficiencies encompassed insufficient internal and~~

external validation, inconsistent standardized imaging segmentation, and a lack of comparisons to a gold standard. The authors stressed the need for further research to validate these models across more prominent, diverse patient populations. The study underscored radiomics' potential in informing HCC management decisions, emphasizing the importance of refining study methodologies for more robust results.

Lastly, W. Xu et al. (2022) present a meta-analysis that investigates the association of mucin family members with the prognosis of pancreatic cancer patients. The search spanned databases such as PubMed, Embase, and Web of Science, with inclusion criteria met by 16 articles. Study quality was assessed using the Newcastle-Ottawa Scale. Ultimately, the authors concluded that mucin family members have the potential to serve as prognostic biomarkers for pancreatic cancer patients.

While several meta-analysis studies in predictive healthcare have been conducted, addressing various diseases and utilizing different methodologies, our work focuses on a unique aspect that has not been extensively explored within these existing studies. Our research addresses several key questions regarding predictive modeling for Pancreatic Neuroendocrine Tumors (PNETs). Firstly, we investigate the current methodologies and performance outcomes of Machine Learning (ML) models employed in forecasting PNET-related outcomes. Additionally, we compare predictive models for PNETs in terms of their problem understanding, data understanding and preparation, modeling and development processes, performance metrics, and validation techniques. We also examine the various aspects of PNETs that these predictive models target, including tumor grade, aggressiveness, diagnosis accuracy, mortality risk estimation, and recurrence probability. Furthermore, we explore how methodologies, performance metrics, and outcomes differ across different categories of predictive models for PNETs, aiming to provide a comprehensive understanding of the predictive modeling landscape for this specific disease entity.

Kommentiert [A1]: A paragraph to discuss the research question and better integration of the literature review with the next section

3 Methodology

3.1 Search strategy

Adhering to the standard of studies mentioned in related work for comparability reasonsAdhering to the standard of studies mentioned in related work for comparability reasons, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (PRISMA) (Moher et al. 2009). A thorough literature search was performed in the PubMed, Web of Science, Scopus, and Springer databases from January 2022 to September 2023 for studies that presented predictive tools for pancreatic diseases derived using ML. The search strategy employed a broad range of search terms for ML to include as many studies as possible. The following were the general search terms used in the conducted search: ((pancreatic OR pancreatitis OR pancreatitis) AND (predict OR predicted OR predicting OR prediction OR predictions OR predictive OR predictivities OR predictivity OR predicts))) AND (2022:2023). This general search approach is implemented for further relevant studies.

3.2 Inclusion and exclusion criteria

Studies were included when they met the following criteria: 1) investigations centered around the application of Machine Learning in the realm of the pancreas; 2) adherence to the World Health Organization's 2017 Classification for the diagnosis of PNETs; 3) utilization of Machine Learning techniques in constructing predictive models for outcomes linked to PNETs; and 4) precise delineation of outcome criteria.

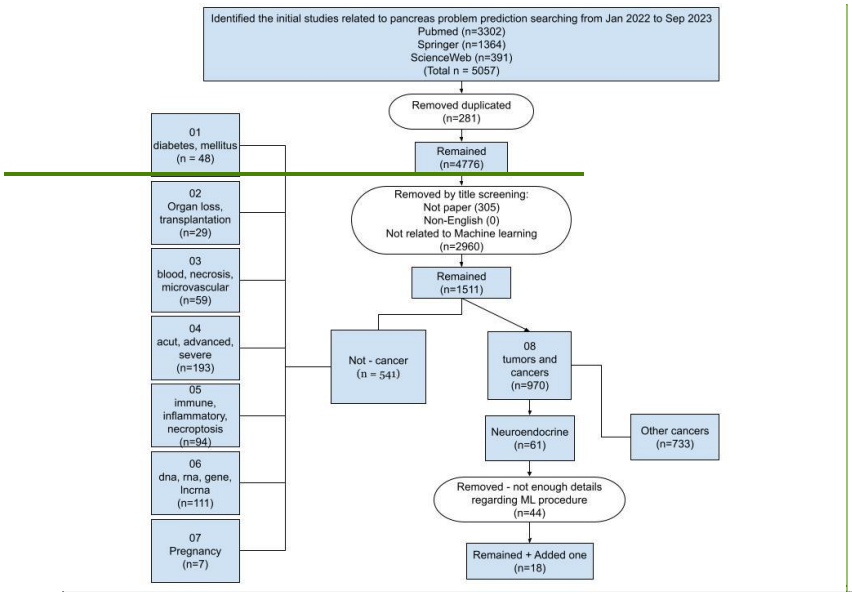
A multi-tiered exclusion process was carried out. Initially, duplicated documents, non-English articles, reviews, abstracts, correspondences, case reports, meeting summaries, studies unavailable in full-text, and non-human studies were eliminated. Subsequently, papers were clustered based on their titles

using the K-means algorithm implemented through the Sklearn package in Python. This aimed to determine the optimal number of clusters while setting this parameter as unknown and identifying the keywords associated with each cluster. This methodology was employed to uncover prevalent topics within this domain and provide insights for future research. Ultimately, only articles pertinent to neuroendocrine aspects were retained, resulting in a compilation of papers directly linked to this study.

Afterward, full-text evaluation was conducted, and the following information was recorded: 1) year of publication; 2) number of cases in each data set; 3) ML classifier; 4) validation methods of models; 5) predictive factors; and 6) model performance indicators: Accuracy, area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, and 95% confidence intervals, which were reported in most of the included studies.

3.3 Assessment of the quality of the studies

The included studies' quality was assessed using the IJMEDI checklist (Cabitza and Campagner 2021). This checklist, proposed by IJMEDI, is a tool for evaluating the quality of medical artificial intelligence studies. It aims to differentiate between comprehensive ML studies and more straightforward medical data mining studies. Researchers employ this checklist to gauge the quality of selected literature, wherein pertinent studies are screened based on study objectives, inclusion criteria, and expert knowledge. The checklist encompasses six dimensions: problem comprehension, data comprehension, data preprocessing, modeling, validation, and implementation. A total of 30 questions are included in the checklist. Respondents can classify each question as either "OK" (satisfactorily addressed), "mR" (satisfactory yet with room for improvement), or "MR" (inadequately addressed). Drawing from prior research, efforts were made to correlate scores with responses for each item (Zhou et al. 2022). For high-priority items, scores of 2, 1, and 0 were respectively assigned to "OK," "mR," and "MR." In low-priority items, scores were halved. The potential maximum score achievable was 50 points. The quality of a study was categorized as low (0–19.5), medium (20–34.5), or high (35–50), based on the achieved score.



~~Figure 1. Flow diagram of study selection~~

~~4 Results~~

~~4.1 Search results~~

~~Figure 1 shows the process of selecting the studies. The initial search yielded 5057 5057 studies. After removing duplicates, non English, and non papers, 1,511 studies remained, which were clustered by the algorithm. The results showed that 541 papers dealt with non-cancer issues in pancreatic diseases and 970 papers focused on cancer and tumor issues, which made up 64% of the total. Of these, 61 papers studied neuroendocrine issues, representing 6% of the papers that worked on cancer-related problems. After screening the title, abstract, methods, and conclusion, 35 papers were excluded because they did not apply machine learning concepts. Then, the full text of 26 studies was thoroughly reviewed, and nine were excluded because they did not use Machine Learning techniques. Finally, 17 papers that developed ML models to predict PNET were included. Finally, two papers were published during our exclusion procedure, in which one used ML in PNET prediction. This resulted in 18 studies being included in the systematic review[1].~~

~~These studies were published from January 2022 to September 2023. The sample data and feature extraction methods used in these studies are summarized in Table 1. Properties of Machine Learning models developed in these studies are summarized in Table 2. The studies had different sample sizes, ranging from 58 to 7750, and different numbers of models, ranging from one to twelve. A total of 66~~

[1] ~~The full datasets and quality assessment criteria with individual items can be found within a digital appendix: https://github.com/kheinovgu/pred_healthcare~~

Kommentiert [A2]: Updated on:
Search result: Jan 2022 to Sep 2023
The text: Removed - not enough details regarding ML procedure

~~models that applied ML algorithms were included in this systematic review. The prediction studies were classified into five categories: Grades, Aggressiveness, Diagnosis, Mortality, and Recurrence.~~

First Author	Sample size	Data source	Input	Feature extraction method
(An et al. 2022)	244	Jiangsu Province Hospital China	Clinical images and information	Radiomic - LASSO regression
(Chiti et al. 2022)	78	Picture Archiving and Communication System	Clinical images and information	Radiomic - LASSO regression
(G. Xu et al. 2022)	1489	SEER database	Clinical data	multivariate Cox regression
(J. Huang et al. 2022)	104	Hospital of Sun Yat-sen University	Clinical images and information	Deep learning CNN
(Javed et al. 2022)	1024	11 centers in China, Italy, USA	Clinical, pathological, and radiological	Not reported
(Jiang et al. 2023)	3239	SEER database	Clinical information	Cox regression
(Liao et al. 2022)	4809	SEER database	Clinical information	LASSO Cox regression
(Liu et al. 2022)	123	Changhai Hospital	Clinical images and information	Radiomic Wilcoxon rank-sum test
(Lu et al. 2022)	7750	SEER database	Clinical data and prognosis differences	Cox regression
(Mori et al. 2022)	101	San Raffaele Institute Italy	Clinical images and information	Mann-Whitney test, RF redundancy limitation
(Murakami et al. 2023)	371	Kyushu insitute	Clinical information	NaN
(Otto et al. 2023)	3474	Charité Universitätsmedizin Berlin	bulk RNA sequencing data	Transcriptomic deconvolution
(Park et al. 2023)	58	Samsung Medical Center	Clinical images and data	Radiomic - LASSO regression
(Thiis-Evensen et al. 2022)	135	VieDoc, Uppsala Sweden	Clinical data and plasma biomarkers and CgA	Not reported
(Wang et al. 2022)	139	West China Sichuan University Hospital	Clinical images and data	Radiomic Mann–Whitney U test and LASSO
(X.-T. Huang et al. 2022a)	2742	Hospital of Sun Yat-sen University	Clinical and radiological features	Not reported
(Yu et al. 2022)	182	Hospital of Sun Yat-sen University	Enhanced computed tomography	univariate and multivariate logistic regression
(Zhu et al. 2022)	187	Beijing, Peking, Qingdao, Kunming Hospitals	Clinical and MRI features	Lasso regression
~~First Author~~	~~Sample size~~	~~Data source~~	~~Input~~	~~Feature extraction method~~
~~(X. T. Huang et al. 2022a)~~	~~2742~~	~~Hospital of Sun Yat-sen University~~	~~Clinical and radiological features~~	~~Not reported~~
~~(Javed et al. 2022)~~	~~1024~~	~~11 centers in China, Italy, USA~~	~~Clinical, pathological, and radiological~~	~~Not reported~~
~~(Zhu et al. 2022)~~	~~187~~	~~Beijing, Peking, Qingdao, Kunming Hospitals~~	~~Clinical and MRI features~~	~~Lasso regression~~
~~(Yu et al. 2022)~~	~~182~~	~~Hospital of Sun Yat-sen University~~	~~Enhanced computed tomography~~	~~univariate and multivariate logistic regression~~
~~(J. Huang et al. 2022)~~	~~104~~	~~Hospital of Sun Yat-sen University~~	~~Clinical images and information~~	~~Deep learning CNN~~
~~(Mori et al. 2022)~~	~~101~~	~~San Raffaele Institute Italy~~	~~Clinical images and~~	~~Mann-Whitney test, RF~~

Feldfunktion geändert
hat formatiert: Englisch (Vereinigte Staaten)
Feldfunktion geändert
hat formatiert: Englisch (Vereinigte Staaten)
hat formatiert: Englisch (Vereinigte Staaten)
Feldfunktion geändert
hat formatiert: Englisch (Vereinigte Staaten)
hat formatiert: Englisch (Vereinigte Staaten)
Feldfunktion geändert
Feldfunktion geändert
hat formatiert: Englisch (Vereinigte Staaten)
hat formatiert: Englisch (Vereinigte Staaten)
Feldfunktion geändert
Feldfunktion geändert
Feldfunktion geändert
Feldfunktion geändert

First Author	Sample size	Data source	Input	Feature extraction method
			information	redundancy limitation
(Thiis-Evensen et al. 2022)	135	VieDoc, Uppsala Sweden	Clinical data and plasma biomarkers and CgA	Not reported
(Otto et al. 2023)	3474	Charité Universitätsmedizin Berlin	bulk RNA sequencing data	Transcriptomic deconvolution
(Wang et al. 2022)	139	West China Sichuan University Hospital	Clinical images and data	Radiomic Mann - Whitney U test and LASSO
(Liu et al. 2022)	123	Changhai Hospital	Clinical images and information	Radiomic Wilcoxon rank-sum test
(Chiti et al. 2022)	78	Picture Archiving and Communication System	Clinical images and information	Radiomic - LASSO regression
(Park et al. 2023)	58	Samsung Medical Center	Clinical images and data	Radiomic - LASSO regression
(Lu et al. 2022)	7750	SEER database	Clinical data and prognosis differences	Cox regression
(Liao et al. 2022)	4809	SEER database	Clinical information	LASSO Cox regression
(Jiang et al. 2023)	3239	SEER database	Clinical information	Cox regression
(G. Xu et al. 2022)	1489	SEER database	Clinical data	multivariate Cox regression
(Murakami et al. 2023)	371	Kyushu insitute	Clinical information	NaN
(An et al. 2022)	244	Jiangsu Province Hospital China	Clinical images and information	Radiomic - LASSO regression

Table A1 Sample sources and features extraction

Grades prediction

Five distinct research investigations have been conducted, each striving to formulate 23 models utilizing ML techniques to classify the grades of PNETs. Three of these studies employed a blend of clinical data and images, utilizing the Radiomic methodology to extract image features. Li et al. (2023) embarked on a study employing 123 samples to develop four models grounded in Linear Discriminant Analysis (LDA) classifiers.

Author	ML model	Validation method	AUROC (95% CI)	Accuracy	Sensitivity	Specificity
		Aggressiveness				
(J. Huang et al. 2022)	MLR - Clinical DL - Clinical DL - combined	70 - 30 random split	0.78 0.81 0.85	0.67 0.79 0.75	0.58 0.75 0.75	0.75 0.83 0.75
(X. T. Huang et al. 2022a)	Nomogram scoring RPA	external validation	0.82 0.81	0.75 0.74	0.71 0.7	0.76 0.75
(Javed et al. 2022)	MLR RF	80 - 20 random split	0.72 0.86	0.71 0.82	0.5 0.75	0.78 0.8

Feldfunktion geändert

Feldfunktion geändert

Feldfunktion geändert

Formatiert: Standard

Kommentiert [A3]: The sorting is alphabetically, and first based on category, and then based on Author

Formatierte Tabelle

(Mori et al. 2022)	RF for Metastasis	70 - 30 random split	0.697			
	RF for Grade		0.717			
	RF - Radiomics for Metastasis		0.769			
	RF - Radiomics for Grade		0.806			
	RF - Radiomics for Lymphnodes		0.689			
	RF - Radiomics for Microvascular Invasion		0.75			
(X.-T. Huang et al. 2022a)	Nomogram scoring	external validation	0.82	0.75	0.71	0.76
	RPA		0.81	0.74	0.7	0.75
(Yu et al. 2022)	MLR	external validation	0.84	0.8	0.77	0.81
(Zhu et al. 2022)	MLR	35 - 65 random split	0.849	0.86	0.75	
		Diagnosis				
(Thiis-Evensen et al. 2022)	BT*	3-flod cross validation	0.98	0.89	0.95	0.727
	SVM*		0.97	0.87	0.74	0.734
	LDA*		0.96	0.83	0.60	
		Grades				
(Chiti et al. 2022)	Arterial - Model not reported	75 - 25 random split	0.82			
	Venous - Model not reported		0.6813			
(Liu et al. 2022)	LDA - Clinical	70 - 30 random split	0.77	0.76	0.72	
	LDA - MRI		0.83	0.83	0.8	
	LDA - CT		0.75	0.71	0	
	LDA - combined		0.85	0.83	0.84	
(Otto et al. 2023)	softmax MLR	80 - 20 random split		0.85	0.85	0.96
	Deconvolution model			0.81	0.8	0.97
(Park et al. 2023)	ANN - Clinical - Grades	5-fold cross validation	0.705	0.655	0.68	0.89
	ANN - Radiomics - Grades		0.857	0.724	0.64	0.95
	ANN - combined - Grades		0.864	0.776	0.72	0.99
	RF - Clinical - Grades		0.664	0.603	0.48	0.69
	RF - Radiomics - Grades		0.819	0.751	0.64	0.58
	RF- combined - Grades		0.853	0.828	0.8	0.58
	ANN - Clinical - prognosis		0.728	0.672	0.741	0.63
	ANN - Radiomics - prognosis		0.662	0.655	0.667	0.881
	ANN - combined - prognosis		0.83	0.776	0.776	0.636
	RF - Clinical - prognosis		0.72	0.724	0.741	0.788
	RF - Radiomics - prognosis		0.596	0.569	0.481	0.818
	RF- combined- prognosis		0.741	0.707	0.593	0.697
(Wang et al. 2022)	SVM-linear - group 1	60 - 40 random split 5-fold cross	0.84	0.75	0.83	0.79
	SVM-linear - group 2		0.87	0.75	0.83	0.89
	SVM-linear - group 3		0.88	0.78	0.86	
		Mortality				
(Jiang et al. 2023)	Cox proportional hazards	70 - 30 random split 5-fold cross validation	0.87	0.74	0.7501	0.1397
	Neural Multitask Logistic Regression		0.87	0.84	0.7616	0.1418
	DeepSurv		0.9	NaN	0.7882	0.1278
	Random Survival Forest		0.86	NaN	0.7612	0.1432
(Liao et al. 2022)	Cox models, RF	70 - 30 random split	Nan	NaN	0.76	
(Lu et al. 2022)	Cox models, Bayesian network	70 - 30 random split	Nan	NaN	0.82	
(G. Xu et al. 2022)	Cox models	bootstrapping and external	0.822	Nan	0.826	

hat formatiert: Englisch (Vereinigte Staaten)

		validation				
		Recurrence				
(An et al. 2022)	Clinical data - regression model	70 - 30 random split	0.786			
	Radiomics - regression model		0.712			
	Combined Radiomics - regression model		0.824			
(Murakami et al. 2023)	RSF - 1 year	70 - 30 random split	0.937	0.97	0.841	0.108
	RSF - 5 year		0.835	0.98	0.841	0.108
	RSF - 10 year		0.911	0.86	0.841	0.108
	Cox models 1 year		0.936	0.97	0.82	0.151
	Cox models 5 year		0.737	0.98	0.82	0.151
	Cox models 10 year		0.81	0.85	0.82	0.151

Table A2. Machine Learning models properties

Abbreviations: MLR: Multivariable logistic regression, RF: Random Forest, ANN: Artificial Neural Network, DL: Deep learning model, RPA: recursion partitioning analysis, BT: Boosted Tree, SVM: Support vector machines, LDA: Linear discriminant analysis, RSF: Random Survival Forest, *: multiple models were averaged

hat formatiert: Schriftart: 9 Pt.

4.21.1 Grades prediction

Five distinct research investigations have been conducted, each striving to formulate 23 models utilizing ML techniques to classify the grades of PNETs. Three of these studies employed a blend of clinical data and images, utilizing the Radiomic methodology to extract image features.

Li et al. (2023) embarked on a study employing 123 samples to develop four models grounded in Linear Discriminant Analysis (LDA) classifiers. These models were crafted using various inputs: solely clinical data, Magnetic Resonance Imaging (MRI) images, Enhanced Computed Tomography (CT) scan data, and a fusion of clinical, MRI, and CT features. Among these, the most promising outcome emerged from the amalgamated model, which achieved a commendable Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.85 and an accuracy of 0.83 during validation.

Likewise, Park et al. (2023) harnessed a dataset encompassing 58 whole body PET and CT images. Their endeavor produced six models, leveraging Neural Network and Random Forest algorithms. Each model was constructed based on distinct inputs – clinical data alone, image data exclusively, and a synthesis of both. The pinnacle of achievement was found in the amalgamation of a neural network, yielding an AUROC of 0.864 and an accuracy of 0.776, a result validated through a 5 fold cross-validation process. With a similar approach, they developed six different models for detecting the prognosis and classifying the disease progression within two years. They also reported that a Neural Network with an Integrated clinic-radiomics model achieved AUROC of 0.83 and an accuracy of 0.776 as the highest performance model. Using a similar method, the researchers also built six other models to predict the prognosis by classifying the disease progression within two years. Among these, the most effective was the one that combined the neural network with information from clinical records and radiology scans. This specific model achieved remarkable results, with an AUROC of 0.83 and a value of 0.776 for prediction accuracy, making it the best-performing model in their study.

Meanwhile, Wang et al. (2022) embarked on the creation of linear Support Vector Machine (SVM) models, drawing upon three distinct collections of combined features derived from a dataset of 139 samples. The optimal AUROC of 0.88, accompanied by an accuracy of 0.78, was attained by amalgamating radiomic signature attributes, which included parameters like T stage, Dilated MPD/BD, clinical TNM stage, and tumor margin.

On a separate trajectory, Otto et al. (2023) embraced a dataset of 3474 bulk RNA sequencing inputs to craft two distinct models. The first model hinged on a multi class SoftMax logistic regression, securing an accuracy of 0.85. The second model, a deconvolution model, showcased an accuracy of

0.81. While the AUROC index was not disclosed, the models exhibited notable sensitivity and specificity values of 0.85, 0.80, and 0.96, 0.97, respectively.

Conversely, Chiti et al. (2022) created a pair of models during the arterial and venous phases, employing radiomic feature extraction alongside LASSO regression. They indicate that the arterial model exhibited a superior accuracy score of 0.82. Nonetheless, they refrained from disclosing the machine learning algorithms utilized and additional performance metrics.

4.3 Aggressiveness prediction

Six research studies conducted analyses on the aggressiveness of PNETs. These studies collectively formulated 15 distinct classification models to predict various facets of tumor aggressiveness.

In a study by J. Huang et al. (2022), an investigation was carried out utilizing Contrast-Enhanced Ultrasound (CEUS) examinations, pathological findings, and clinical information. This research yielded three distinct models. Image features were extracted through the application of a Convolutional Neural Network model. Their approach encompassed the development of a Multivariable Logistic Regression model and a Deep Learning model, both based solely on clinical features. Additionally, a deep learning model was devised by integrating combined features. Notably, this combined model exhibited an Area Under the AUROC value of 0.85 and an accuracy of 0.79 during validation.

In a parallel attempt, X. T. Huang et al. (2022a) employed radiological and clinical attributes to predict the occurrence of Lymph Node Metastasis in nonfunctioning PNET patients. Using a dataset comprising 2742 samples, they created a nomogram scoring system and conducted recursive partitioning analysis. Their former model achieved the highest AUROC of 0.82 and an accuracy of 0.75 during validation, with validation extended to external datasets.

Zhu et al. (2022), also pursuing the same objective, constructed a Multivariate Logistic Regression model. Their model was built upon MRI images and clinical data derived from a sample of 187 patients. Demonstrating substantial efficacy, their model yielded an AUROC of 0.849 and an accuracy of 0.86 during validation.

Another pertinent study by Javed et al. (2022) concentrated on nodal disease prediction in nonfunctional PNETs. This research involved the integration of clinical, pathological, and radiological attributes within their analysis of a dataset encompassing 1024 samples. Their approach yielded two classification models - one employing logistic regression and the other utilizing random forest algorithms. The latter model stood out with an elevated AUROC of 0.86 and an accuracy of 0.82.

Meanwhile, Mori et al. (2022) explored diverse dimensions of tumor aggressiveness, encompassing Metastasis, Lymph Node involvement, and Microvascular Invasion. Radiological variables and radiomic features were incorporated to construct four distinct models, all employing the Random Forest algorithm. The models yielded AUROC values of 0.697, 0.769, 0.689, and 0.75 for Metastasis with Radiomics features, Metastasis and Lymph Node involvement with combined features, and Microvascular Invasion, respectively.

Lastly, Yu et al. (2022) devised a multivariate logistic regression model to predict a Ki-67 index lower than 5%. CT features were harnessed, utilizing a sample size of 182. Impressively, their model achieved an AUROC of 0.84 and an accuracy of 0.80 when validated against an external dataset.

4.4 Diagnosis prediction

Solely one scholarly article, specifically the work conducted by Thiis-Evensen et al. (2022), is dedicated to diagnosing Pancreatic Neuroendocrine Tumors. The central objective of their research was to identify and differentiate this particular tumor type from intestinal neuroendocrine tumors. This endeavor was accomplished by integrating plasma protein biomarkers and clinical data as features

based on a sample size of 135. The study encompassed the development of three distinct classifiers: Boosted Tree (BT), LDA, and SVM. These classifiers collectively yielded twelve distinct classification models due to their incorporated features and prediction goal variations.

In the pursuit of their objective, the researchers designed various model configurations. Model 1 targeted PNETs and controls, while Model 2 aimed to differentiate between PNETs and Small Intestine Neuroendocrine Tumors. Within this framework, Models 1A and 2A encompassed a comprehensive range of features, integrating 92 plasma protein biomarkers alongside chromogranin A (CgA). In contrast, Models 1B and 2B excluded CgA from their analysis. The study reported notable results through the rigorous implementation of a three-fold cross-validation process.

4.5 Mortality prediction

Four studies aimed to predict survival rates by analyzing mortality using clinical features from the SEER database. These studies used the Cox regression algorithm for feature selection. Jiang et al. (2023) developed four models using standard Cox proportional hazards, neural multitask logistic regression, DeepSurv, and random survival forest. These models were built on a dataset of 3239 samples, utilizing patient clinical data. The optimization of model hyperparameters was achieved through a 5-fold cross-validation approach. The most proficient model's performance was assessed using the internally validated 70-30 percent random split method, resulting in an impressive AUROC of 0.9 and a concordance index (C-index) of 0.7501. An external validation of their model was also conducted to ensure the robustness of their findings.

Similarly, Liao et al. (2022) used a Random Survival Forest model for survival prediction. This approach was based on a dataset comprising 4809 samples, integrating clinical attributes and prognosis-related differences. The resulting C-index achieved a value of 0.76, showcasing the model's predictive capability.

In contrast, Lu et al. (2022) utilized a Bayesian network dynamic nomogram model for survival rate prediction. This sophisticated methodology was implemented using a dataset tailored to their purpose and yielded a C-index of 0.820, signifying substantial accuracy in prognostication.

Likewise, G. Xu et al. (2022) predicted survival rates using the Cox proportional hazards regression model. Their analysis was based on a dataset containing 1489 instances of clinical features. The derived C-index, amounting to 0.826, underscored the model's proficiency in assessing survival outcomes.

4.6 Recurrence prediction

Two studies tried to predict the reoccurrence of PNET and Gastrointestinal Pancreatic Neuroendocrine Neoplasms (GP-NENs), which together developed nine different ML models based on clinical images and data from 2622 samples. An et al. (2022) used Radiomics methods to extract features from CT scanning and used univariate and multivariant regression analysis to select and show the importance of the feature. They developed a regression model based on clinical data, radiomic features, and a combined model. They reported that the combined model with AUROC of 0.824 showed the best performance. On the other hand, Murakami et al. (2023) used random survival forest (RSF) and Cox models to predict the recurrence rate in 1, 2, and 10 years of disease-free survival. They used clinical data of 371 patients and concluded that RSF and Cox models perform similarly in 1-year prediction with AUROC of about 0.93 and specificity of 0.97. However, RSF resulted in better AUROC in 5- and 10-year recurrence prediction with values of 0.835 and 0.911, respectively.

Two distinct research studies aimed to forecast the recurrence of PNETs and GP-NENs. These studies produced nine diverse ML models in collaboration, drawing upon clinical data and images.

In the study conducted by An et al. (2022), Radiomics techniques were employed to extract features from CT scans of 244 patients. Their analysis was further enriched by employing univariate and

multivariate regression analyses to determine feature importance. They designed a regression model grounded in clinical data, radiomic features, and an amalgamated approach within this framework. Their findings highlighted the combined model's exceptional performance, evidenced by an AUROC of 0.824.

Conversely, Murakami et al. (2023) adopted a distinct approach by employing RSF and Cox models to predict recurrence rates over 1, 2, and 10 years within disease-free survival. Their investigation involved clinical data from 371 patients. Their observations indicated that RSF and Cox models exhibited comparable efficacy in the one-year prediction, yielding AUROC values of approximately 0.93 and a high specificity of 0.97. Notably, RSF showcased superior predictive performance in the context of 5- and 10-year recurrence forecasts, attaining AUROC values of 0.835 and 0.911, respectively.

Author	Problem Understanding Max = 10	Data Understanding Max = 6	Data Preparation Max = 8	Modeling Max = 6	Validation Max = 12	Deployment Max = 8	Total Max = 50
(Park et al. 2023)	10	4	8	6	8	2	38
(Otto et al. 2023)	10	6	7	5	4.5	4	36.5
(Wang et al. 2022)	10	4	6	6	6	2	34
(Jiang et al. 2023)	10	6	1	6	7.5	3	33.5
(Thiis-Evensen et al. 2022)	10	3	6	6	6	2	33
(J. Huang et al. 2022)	10	6	2	6	6	1.5	31.5
(G. Xu et al. 2022)	10	6	1	3	9	2	31
(Lu et al. 2022)	10	2	2	4	8	2	28
(Zhu et al. 2022)	10	3	4	6	2	2.5	27.5
(Javed et al. 2022)	10	4	1	4	7	1	27
(Liao et al. 2022)	10	5	1	4	6	1	27
(Murakami et al. 2023)	10	5	2	4	4	1	26
(Yu et al. 2022)	8	3	3	3	6.5	2.5	26
(Chiti et al. 2022)	9	4	2	4	4	1.5	24.5
(Liu et al. 2022)	9	6	0	4	4	0.5	23.5
(Mori et al. 2022)	10	2	0	5	5.5	1	23.5
(X. T. Huang et al. 2022b)	10	4	0	4	2	1	21
(An et al. 2022)	9	3	0	4	0	2.5	18.5

Table 3. Quality assessment scores of the 18 studies according to the IJMEDI checklist

5 Discussion

This is the first systematic review of studies on ML-based models for pancreatic neuroendocrine tumors using the IJMEDI checklist for quality assessment and the first study to associate scores with this checklist. Table 3 summarizes the individual dimension scores and the overall scores across the various studies. The mean score among the studies incorporated was 28.3, spanning from 18.5 to 38. Predominantly, the studies exhibited a moderate quality level, except for two studies, Park et al. (2023) and Otto et al. (2023), that stood out for their high quality standards. In High priority items, studies achieved 61% of total scores on average, while for low priority items, they achieved only 38% on average of total scores.

While the majority of the studies excelled in comprehending the issues at hand and effectively in understanding the problem by conducting experiments, collecting data, and extracting features, there

~~was a conspicuous bias observed in several aspects of data understanding, data preprocessing, modeling, validation, and deployment dimensions. This bias compromised the overall quality in these dimensions. Visualized in Figure 2 is the distribution of responses across each item within the respective dimensions.~~

~~Many studies exhibited an impressive grasp of the problem, achieving an average score of 97% in this domain. This indicates that across all studies, they either achieved the total score or came very close to it in every aspect. These aspects encompassed the meticulous description of the study's subject population, including explicit delineation of inclusion and exclusion criteria, articulation of the study's design and the data's source, and clear explication of the medical task and the methodology behind data collection.~~

~~Table 4 provides a comprehensive summary of predictive features utilized by models that achieved the highest AUORC within each category sharing the same output. It details the specific features employed by each model, along with their corresponding authors and AUORC scores. Table 5 presents an analysis of biases identified within each paper based on the IJMEDI checklist. It outlines the key areas of bias observed in the reviewed studies, including the lack of gold standard reporting, missing feature descriptions, inadequate data preparation techniques, unclear model specifications, validation issues, and deficiencies in deployment strategies.~~

Category	Output	Model (with best AUROC)	AUORC	Predictive Features	Author
Aggressiveness	Aggressive and Non-aggressive prediction	DL - combined	0.85	Sex, Age, Functional, CEA, CA125, CA 19–9, Tumor location, Body, Tail, Tumor size, Texture, Tumor shape, Tumor margin, Echogenicity, DMPD, CDFI, AE, VE, ED	(J. Huang et al. 2022)
	Nodal disease prediction in nonfunctional PNET	RF	0.86	Tumor size, Tumor location, arterial enhancement, portal venous enhancement, and 15 radiomic features	(Javed et al. 2022)
	Grades, Metastasis, Lymph nodes, Microvascular aggressiveness	RF	0.806	Age, Gender, Necrosis, Cystic morphology, Pancreas atrophy, Arterial invasion, Venous invasion, Contiguous organs invasion, Grade (G1 vs. G2/G3), Liver metastasis (M+), Microvascular invasion (VI+), Metastatic lymph nodes (N+)	(Mori et al. 2022)
	Lymph node metastasis (LNM) prediction	Nomogram scoring	0.82	Age, sex, tumor size, tumor location, serum CgA level, serum NSE level, and Ki-67	(X.-T. Huang et al. 2022a)
	Ki-67 index prediction of being less than 5%	MLR	0.84	Tumor size, arterial phase enhancement, portal venous phase enhancement, and arterial phase enhancement pattern	(Yu et al. 2022)

	lymph node metastasis prediction	MLR	0.849	Gender, Age, BMI, Symptom, NLR, TB, ALT, AST, FBG, CEA, CA199, CA724, NSE, Lymph node metastasis, Tumor location, SI on T2WI, Maximum diameter of the tumor, Tumor margin, Exophytic growth, MPDD or CBDD, Hyperenhancement at arterial phase, Homogeneity, Vascular and adjacent tissue involvement, Synchronous liver metastases, Long axis of the largest lymph node, Short axis of the largest lymph node, Ratio of the long/short axis of the largest lymph node, Abnormal Shape of the largest lymph node, Number of the lymph nodes with the short axis>5 mm, Number of the lymph nodes with the short axis>10 mm, ADCmean, ADCmax, ADCmin, Tumor volume	(Zhu et al. 2022)
Diagnosis	Diagnosis prediction	BT*	0.98	Age, sex, Distant metastasis, Lymph node metastasis, Ki-67, NET Grade	(Thiis-Evensen et al. 2022)
Grades	Prognosis prediction	ANN - combined - prognosis	0.83	Tumor grade prediction model: Clinical features (1) Primary tumor size (2) Hepatic metastasis (3) Extrahepatic metastasis (4) Tumor grade Radiomic features (1) Morphological surface to volume ratio (2) Morphological center of mass shift (3) Intensity-based kurtosis (4) Intensity-histogram minimum histogram gradient grey level (5) GLCM correlation (6) GLRLM long runs emphasis (7) NGTDM strength (8) GLSZM small zone emphasis (9) GLSZM large zone emphasis (10) GLSZM large zone low grey level emphasis (11) GLSZM grey level non-uniformity; Prognosis prediction model: Clinical features (1) Sex (2) Clinical stage (3) Vascular invasion Radiomic features (1) Morphological spherical disproportion (2) Morphological sphericity (3) Intensity-histogram kurtosis (4) Intensity-histogram quartile coefficient of dispersion (5) Intensity-histogram maximum histogram gradient (6) GLSZM normalized grey level non-uniformity	(Park et al. 2023)
	Grades classes	ANN - combined - - Grades	0.864		
Mortality	Survival rate prediction	DeepSurv	0.9	Age, Gender, Marital status, Race, Primary site, Stage, Grade, Surgery, Radiotherapy, Chemotherapy, Tumor size, Number of tumors, Tumor extension, Distant metastasis, Survival months, Status	(Jiang et al. 2023)
Re	Prognosis	Combined	0.824	Gender, History of hypertension,	(An et al. 2022)

	prediction of GP-NENs	Radiomics - regression model		Smoking history, Drinking history, Age, Tumor pathological type, Primary tumor site, Ki-67, TNM stage, Lymph node metastasis, Distant metastasis, History of diabetes, Radscore 1, Radscore 2, Radscore 3	
	1-year recurrence prediction	RSF	0.937	Age, Sex, Hereditary syndrome, Symptom, Tumor location, Tumor number, Contrast pattern, Cystic component, Calcification, Main pancreatic duct obstruction, Surgical method, Surgical procedure, Lymphadenectomy, Clavien-Dindo classification, Tumor grade, Tumor size, Lymph node metastasis, Stage, Residual tumor, Lymphovascular invasion, Perineural invasion	(Murakami et al. 2023)
	5-year recurrence prediction	RSF	0.835		
	10-year recurrence prediction	RSF	0.911		

Table 4A3. summary of the predictive features and models with highest AUROC among each category and similar ourput.

~~Category~~	~~priority~~	~~bias~~	~~Papers~~
~~Data Understanding~~	~~High~~	~~The gold standard has not been reported.~~	~~(An et al. 2022), (Chiti et al. 2022), (Mori et al. 2022, (Yu et al. 2022), (Zhu et al. 2022)~~
	~~High~~	~~The features have not been described.~~	~~(Lu et al. 2022), (Mori et al. 2022, (Park et al. 2023), (Thiis-Evensen et al. 2022)~~
	~~High~~	~~Outlier detection has not been performed.~~	~~(Lu et al. 2022)~~
~~Data Preparation~~	~~High~~	~~Missing value management has not been described.~~	~~(An et al. 2022), (Chiti et al. 2022), (G. Xu et al. 2022, (J. Huang et al. 2022), (Jiang et al. 2023), (Liu et al. 2022), (Lu et al. 2022), (Mori et al. 2022, (Murakami et al. 2023), (Wang et al. 2022), (X. T. Huang et al. 2022a)~~
	~~High~~	~~How the feature pre-processing was described has not been mentioned.~~	~~(An et al. 2022), (Chiti et al. 2022), (G. Xu et al. 2022, (J. Huang et al. 2022), (Javed et al. 2022), (Jiang et al. 2023), (Liao et al. 2022), (Liu et al. 2022), (Lu et al. 2022), (Mori et al. 2022, (Thiis-Evensen et al. 2022), (X. T. Huang et al. 2022a), (Yu et al. 2022), (Zhu et al. 2022)~~
	~~High~~	~~The imbalance in data distribution, especially when clinical data is used, has not been analyzed or reported.~~	~~(An et al. 2022), (Javed et al. 2022), (Liao et al. 2022), (Liu et al. 2022), (Mori et al. 2022, (X. T. Huang et al. 2022a), (Yu et al. 2022), (Zhu et al. 2022)~~
~~Modeling~~	~~High~~	~~The model output task has not been specified clearly or is vaguely mentioned.~~	~~(An et al. 2022), (Chiti et al. 2022), (G. Xu et al. 2022, (J. Huang et al. 2022), (Javed et al. 2022), (Jiang et al. 2023), (Liao et al. 2022), (Liu et al. 2022), (Lu et al. 2022), (Mori et al. 2022, (Murakami et al. 2023), (X. T. Huang et al. 2022a)~~
	~~High~~	~~The model structure,~~	~~(G. Xu et al. 2022, (Javed et al. 2022), (Liao et al. 2022), (Lu~~

		assumptions, and modifications regarding the model architecture have not been specified clearly or are vaguely mentioned.	et al. 2022), (Yu et al. 2022)
Validation	High	The data splitting procedure has not been described.	(An et al. 2022), (Chiti et al. 2022), (Liu et al. 2022), (Murakami et al. 2023), (X. T. Huang et al. 2022a)
	High	Hyper-parameter optimization and selection method have not been described.	(An et al. 2022)
	High	Model calibration on classification models has not been described, and the Brier score or calibration plot has not been reported.	(An et al. 2022), (Chiti et al. 2022), (G. Xu et al. 2022, (J. Huang et al. 2022), (Jiang et al. 2023), (Liao et al. 2022), (Liu et al. 2022), (Mori et al. 2022, (Thiis-Evensen et al. 2022), (Wang et al. 2022), (X. T. Huang et al. 2022a), (Yu et al. 2022), (Zhu et al. 2022)
	High	The model internal/external validation procedure has not been described.	(An et al. 2022), (Chiti et al. 2022), (Liu et al. 2022), (Mori et al. 2022, (Murakami et al. 2023), (Otto et al. 2023), (Park et al. 2023), (Thiis-Evensen et al. 2022), (Wang et al. 2022), (X. T. Huang et al. 2022a), (Zhu et al. 2022)
	High	No external validation or the characteristics of the external validation set(s) have been described.	(An et al. 2022), (Chiti et al. 2022), (J. Huang et al. 2022), (Javed et al. 2022), (Liao et al. 2022), (Liu et al. 2022), (Lu et al. 2022), (Otto et al. 2023), (X. T. Huang et al. 2022a), (Yu et al. 2022), (Zhu et al. 2022)
	Low	Main error-based metrics have not been fully reported.	(An et al. 2022), (Chiti et al. 2022), (J. Huang et al. 2022), (Javed et al. 2022), (Liao et al. 2022), (Liu et al. 2022), (Lu et al. 2022), (Murakami et al. 2023), (Otto et al. 2023), (Park et al. 2023), (Thiis-Evensen et al. 2022), (Wang et al. 2022), (X. T. Huang et al. 2022a), (Zhu et al. 2022)
	High	Some relevant errors, like noteworthy classification errors or cases for which the regression prediction was much higher (> 2x) than the MAE, have not been reported.	(An et al. 2022), (Zhu et al. 2022)
Deployment	Low	The target users have not been indicated.	(An et al. 2022), (Chiti et al. 2022), (G. Xu et al. 2022, (J. Huang et al. 2022), (Javed et al. 2022), (Jiang et al. 2023), (Liao et al. 2022), (Liu et al. 2022), (Lu et al. 2022), (Mori et al. 2022, (Murakami et al. 2023), (Park et al. 2023), (Thiis-Evensen et al. 2022), (Wang et al. 2022), (X. T. Huang et al. 2022a), (Yu et al. 2022), (Zhu et al. 2022)
	Low	The utility of the model has not been discussed.	(Liu et al. 2022), (Lu et al. 2022), (Murakami et al. 2023), (Park et al. 2023)
	Low	There is no information regarding model interpretability and explainability.	(An et al. 2022), (Chiti et al. 2022), (J. Huang et al. 2022), (Javed et al. 2022), (Jiang et al. 2023), (Liao et al. 2022), (Liu et al. 2022), (Mori et al. 2022, (Murakami et al. 2023), (Thiis-Evensen et al. 2022), (Wang et al. 2022), (X. T. Huang et al. 2022a)
	Low	There is no discussion regarding model fairness, ethical concerns, or risks of bias.	(G. Xu et al. 2022, (Javed et al. 2022), (Liao et al. 2022), (Mori et al. 2022, (Thiis-Evensen et al. 2022), (X. T. Huang et al. 2022a)

hat formatiert: Deutsch (Deutschland)

	Low	There is no point made about the environmental sustainability of the model.	(An et al. 2022), (Chiti et al. 2022), (G. Xu et al. 2022, (J. Huang et al. 2022), (Javed et al. 2022), (Jiang et al. 2023), (Liao et al. 2022), (Liu et al. 2022), (Lu et al. 2022), (Mori et al. 2022, (Murakami et al. 2023), (Otto et al. 2023), (Park et al. 2023), (Thiis-Evensen et al. 2022), (Wang et al. 2022), (X.-T. Huang et al. 2022a), (Yu et al. 2022), (Zhu et al. 2022)
	Low	The code or data has not been shared with the community, and no reason has been given when it's not.	(An et al. 2022), (Chiti et al. 2022), (G. Xu et al. 2022, (J. Huang et al. 2022), (Javed et al. 2022), (Jiang et al. 2023), (Liao et al. 2022), (Liu et al. 2022), (Lu et al. 2022), (Mori et al. 2022, (Murakami et al. 2023), (Otto et al. 2023), (Park et al. 2023), (Thiis-Evensen et al. 2022), (Wang et al. 2022), (X.-T. Huang et al. 2022a), (Yu et al. 2022), (Zhu et al. 2022)
	High	The system is not already adopted in daily practice.	(Chiti et al. 2022), (G. Xu et al. 2022, (J. Huang et al. 2022), (Javed et al. 2022), (Jiang et al. 2023), (Liao et al. 2022), (Liu et al. 2022), (Lu et al. 2022), (Mori et al. 2022, (Murakami et al. 2023), (Park et al. 2023), (Wang et al. 2022), (X. T. Huang et al. 2022a), (Yu et al. 2022), (Zhu et al. 2022)
	Low	The gold standard has not been reported.	(An et al. 2022), (Chiti et al. 2022), (G. Xu et al. 2022, (J. Huang et al. 2022), (Javed et al. 2022), (Liao et al. 2022), (Liu et al. 2022), (Lu et al. 2022), (Mori et al. 2022, (Murakami et al. 2023), (Otto et al. 2023), (Park et al. 2023), (Thiis-Evensen et al. 2022), (Wang et al. 2022), (X. T. Huang et al. 2022a), (Yu et al. 2022), (Zhu et al. 2022)

Table 5. summary of biases observed among the papers.

Formatiert: Beschriftung, Links

Formatiert: Beschriftung

Formatierte Tabelle

~~Figure 2. Achieved scores for~~in each item of each category ~~in the IJMEDI checklist~~

~~In data understanding, studies generally demonstrated competence, attaining an average score of 70%. While a majority of the papers proficiently detailed the demographic composition of their data and showcased adeptness in the realm of feature selection, as well as the role of these features in the ultimate prediction, a deficiency existed in terms of a comprehensive definition of the classification framework and the processes by which labeling was undertaken.~~

~~A pronounced weakness emerged in the area of data preparation. A substantial portion of the studies lacked aspects such as performing outlier detection, implementing feature preprocessing, addressing missing values, and handling imbalanced data. On average, studies achieved only 32% proficiency in this domain. Excluding noteworthy exceptions such as the work by Park et al. (2023), Otto et al. (2023), Wang et al. (2022), and Jiang et al. (2023) who exhibited commendable performance in this category the majority of studies demonstrated subpar performance.~~

Regarding constructing machine learning models, studies achieved a commendable average score of 78%. These studies tried to elucidate the model's intended task and corresponding output. Nevertheless, many of these studies failed to provide sufficiently detailed explanations, particularly in clarifying the model's outputs' exact nature and practical utility. Furthermore, while many articles reported utilizing specific machine learning models, only a fraction provided in-depth insight into the model's architecture; a substantial number were content with merely naming the model.

Validation emerged as another area where studies achieved less than moderate scores, averaging 48%. While many studies employed the random split method, and a few opted for k-fold cross-validation as an internal validation strategy, the reporting of external validation was limited and inadequately comprehensive. The explanation of model training procedures and the conduct of hyperparameter optimization displayed notable gaps in the literature. Although specific reports mentioned aspects such as model calibration and the Brier score, there remains a need for further development in this sphere.

Deployment emerged as a fragile point in the studies, with a compliance rate of only 23%. Merely three studies openly shared their code with the broader community, while others did not mention their source code's availability. Similarly, only Jiang et al. (2023) successfully integrated their work into practical applications. Notably absent were discussions on the ethical implications or bias risks associated with the models, as well as considerations of environmental sustainability. However, studies briefly touched on target user demographics and the model's utility.

6 Conclusion

In conclusion, the rise in patients with there is a notable surge in the development of machine learning models to predict and comprehend various dimensions of PNETs is paralleled by a notable surge in the development of machine learning models to predict and comprehend various dimensions of this ailment. These encompass predicting Grades, Aggressiveness, Diagnosis, Mortality, and Recurrence. However, this domain's current state of models calls for further refinement to achieve enhanced performance metrics. Furthermore, the quality of existing studies remains relatively modest, marked by issues such as inadequate data samples, insufficient detailing of data preprocessing, suboptimal validation procedures, ambiguous model architecture, absence of well-defined training procedures, and lack of hyperparameter optimization, along with limited clinical applicability. Looking ahead, researchers and medical professionals must address and rectify these challenges. This involves conducting comprehensive studies to assess ML systems' real-world impact and model performance's comparability. The ultimate goal is to craft high-caliber predictive ML-based models that can be seamlessly integrated into clinical practice, thus necessitating a concerted effort to enhance the overall landscape of healthcare with cutting-edge technology.

7 References

An, P., Zhang, J., Li, M., Duan, P., He, Z., Wang, Z., Feng, G., Guo, H., Li, X., and Qin, P. 2022. "Clinical Data CT Radiomics Based Model for Predicting Prognosis of Patients with Gastrointestinal Pancreatic Neuroendocrine Neoplasms (GP-NENs)," *Computational and Mathematical Methods in Medicine* (2022), (E. Hincal, ed.), pp. 1–9.

Cabitza, F., and Campagner, A. 2021. "The Need to Separate the Wheat from the Chaff in Medical Informatics: Introducing a Comprehensive Checklist for the (Self)-Assessment of Medical AI Studies," *International Journal of Medical Informatics* (153), p. 104510.

Chen, L., Wang, W., Jin, K., Yuan, B., Tan, H., Sun, J., Guo, Y., Luo, Y., Feng, S., Yu, X., Chen, M., and Chen, J. 2023. "Special Issue 'The Advance of Solid Tumor Research in China': Prediction of Sunitinib Efficacy Using Computed Tomography in Patients with Pancreatic Neuroendocrine Tumors," *International Journal of Cancer* (152:1), pp. 90–99.

Chiti, G., Grazzini, G., Flammia, F., Matteuzzi, B., Tortoli, P., Bettarini, S., Pasqualini, E., Granata, V., Busoni, S., Messserini, L., Pradella, S., Massi, D., and Miele, V. 2022. "Gastroenteropancreatic Neuroendocrine Neoplasms (GEP-NENs): A Radiomic Model to Predict Tumor Grade," *La Radiologia Medica* (127:9), pp. 928–938.

Han, S., Kim, J. H., Yoo, J., and Jang, S. 2022. "Prediction of Recurrence after Surgery Based on Preoperative MRI Features in Patients with Pancreatic Neuroendocrine Tumors," *European Radiology* (32:4), pp. 2506–2517.

Harding-Theobald, E., Louissaint, J., Maraj, B., Cuaresma, E., Townsend, W., Mendiratta-Lala, M., Singal, A. G., Su, G. L., Lok, A. S., and Parikh, N. D. 2021. "Systematic Review: Radiomics for the Diagnosis and Prognosis of Hepatocellular Carcinoma," *Alimentary Pharmacology & Therapeutics* (54:7), pp. 890–901.

Huang, J., Xie, Xiaohua, Wu, H., Zhang, X., Zheng, Y., Xie, Xiaoyan, Wang, Y., and Xu, M. 2022. "Development and Validation of a Combined Nomogram Model Based on Deep Learning Contrast-Enhanced Ultrasound and Clinical Factors to Predict Preoperative Aggressiveness in Pancreatic Neuroendocrine Neoplasms," *European Radiology* (32:11), pp. 7965–7975.

Huang J, Xie X.Pdf. (n.d.).

Huang, X. T., Xie, J. Z., Huang, C. S., Li, J. H., Chen, W., Liang, L. J., and Yin, X. Y. 2022a. "Development and Validation of Nomogram to Predict Lymph Node Metastasis Preoperatively in Patients with Pancreatic Neuroendocrine Tumor," *HPB* (24:12), pp. 2112–2118.

Huang, X. T., Xie, J. Z., Huang, C. S., Li, J. H., Chen, W., Liang, L. J., and Yin, X. Y. 2022b. "Development and Validation of Nomogram to Predict Lymph Node Metastasis Preoperatively in Patients with Pancreatic Neuroendocrine Tumor," *HPB* (24:12), pp. 2112–2118.

Javed, A. A., Pulvirenti, A., Zheng, J., Michelakos, T., Sekigami, Y., Razi, S., McIntyre, C. A., Thompson, E., Klimstra, D. S., Deshpande, V., Singhi, A. D., Weiss, M. J., Wolfgang, C. L., Cameron, J. L., Wei, A. C., Zureikat, A. H., Ferrone, C. R., He, J., and Pancreatic Neuroendocrine Disease Alliance (PANDA). 2022. "A Novel Tool to Predict Nodal Metastasis in Small Pancreatic Neuroendocrine Tumors: A Multicenter Study," *surgery* (172:6), pp. 1800–1806.

Jiang, C., Wang, K., Yan, L., Yao, H., Shi, H., and Lin, R. 2023. "Predicting the Survival of Patients with Pancreatic Neuroendocrine Neoplasms Using Deep Learning: A Study Based on Surveillance, Epidemiology, and End Results Database," *Cancer Medicine* (12:11), pp. 12413–12424.

Kamel Rahimi, A., Canfell, O. J., Chan, W., Sly, B., Pole, J. D., Sullivan, C., and Shrapnel, S. 2022. "Machine Learning Models for Diabetes Management in Acute Care Using Electronic Medical Records: A Systematic Review," *International Journal of Medical Informatics* (162), p. 104758.

Lai, Q., Spoletini, G., Mennini, G., Laureiro, Z. L., Tsilimigras, D. I., Pawlik, T. M., and Rossi, M. 2020. "Prognostic Role of Artificial Intelligence among Patients with Hepatocellular Cancer: A Systematic Review," *World Journal of Gastroenterology* (26:42), pp. 6679–6688.

Li, J., Huang, L., Liao, C., Liu, G., Tian, Y., and Chen, S. 2023. "Two Machine Learning Based Nomogram to Predict Risk and Prognostic Factors for Liver Metastasis from Pancreatic Neuroendocrine Tumors: A Multicenter Study," *BMC Cancer* (23:1), p. 529.

Liao, T., Su, T., Huang, L., Li, B., and Feng, L. H. 2022. "Development and Validation of a Novel Nomogram for Predicting Survival Rate in Pancreatic Neuroendocrine Neoplasms," *Scandinavian Journal of Gastroenterology* (57:1), pp. 85–90. (https://doi.org/10.1080/00365521.2021.1984571).

Liu, C., Bian, Y., Meng, Y., Liu, F., Cao, K., Zhang, H., Fang, X., Li, J., Yu, J., Feng, X., Ma, C., Lu, J., Xu, J., and Shao, C. 2022. "Preoperative Prediction of G1 and G2/3 Grades in Patients With Nonfunctional Pancreatic Neuroendocrine Tumors Using Multimodality Imaging," *Academic Radiology* (29:4), pp. e49–e60.

Liu, Y. S., Thaliffdeen, R., Han, S., and Park, C. 2023. "Use of Machine Learning to Predict Bladder Cancer Survival Outcomes: A Systematic Literature Review," *Expert Review of Pharmacoeconomics & Outcomes Research* (23:7), pp. 761–771.

Lu, Z., Li, T., Liu, C., Zheng, Y., and Song, J. 2022. "Development and Validation of a Survival Prediction Model and Risk Stratification for Pancreatic Neuroendocrine Neoplasms," *Journal of Endocrinological Investigation* (46:5), pp. 927–937.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and for the PRISMA Group. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement," *BMJ*

Mori, M., Palumbo, D., Muffatti, F., Partelli, S., Mushtaq, J., Andreasi, V., Prato, F., Ubeira, M. G., Palazzo, G., Falconi, M., Fiorino, C., and De Cobelli, F. 2022. "Prediction of the Characteristics of Aggressiveness of Pancreatic Neuroendocrine Neoplasms (PanNENs) Based on CT Radiomic Features," *European Radiology* (33:6), pp. 4412–4421.

Murakami, M. et al. 2023. "Machine Learning Based Model for Prediction and Feature Analysis of Recurrence in Pancreatic Neuroendocrine Tumors G1/G2," *Journal of Gastroenterology* (58:6), pp. 586–597.

Otto, R., Detjen, K. M., Riemer, P., Fattohi, M., Grötzinger, C., Rindi, G., Wiedenmann, B., Sers, C., and Leser, U. 2023. "Transcriptomic Deconvolution of Neuroendocrine Neoplasms Predicts Clinically Relevant Characteristics," *Cancers* (15:3), p. 936. (https://doi.org/10.3390/cancers15030936).

Park, Y. J., Park, Y. S., Kim, S. T., and Hyun, S. H. 2023. "A Machine Learning Approach Using [18F]FDG PET Based Radiomics for Prediction of Tumor Grade and Prognosis in Pancreatic Neuroendocrine Tumor," *Molecular Imaging and Biology*.

Thiis-Evensen, E., Kjellman, M., Knigge, U., Gronbaek, H., Schalin Jäntti, C., Welin, S., Sorbye, H., Del Pilar Schneider, M., Belusa, R., and The Nordic NET Biomarker Group. 2022. "Plasma Protein Biomarkers for the Detection of Pancreatic Neuroendocrine Tumors and Differentiation from Small Intestinal Neuroendocrine Tumors," *Journal of Neuroendocrinology* (34:7).

Wang, X., Qiu, J. J., Tan, C. L., Chen, Y. H., Tan, Q. Q., Ren, S. J., Yang, F., Yao, W. Q., Cao, D., Ke, N. W., and Liu, X. B. 2022. "Development and Validation of a Novel Radiomics-Based Nomogram With Machine Learning to Preoperatively Predict Histologic Grade in Pancreatic Neuroendocrine Tumors," *Frontiers in Oncology* (12), p. 843376.

White, B. E., Mujica-Mota, R., Snowsill, T., Gamper, E. M., Srirajaskanthan, R., and Ramage, J. K. 2021. "Evaluating Cost-Effectiveness in the Management of Neuroendocrine Neoplasms," *Reviews in Endocrine and Metabolic Disorders* (22:3), pp. 647–663.

Xu, G., Xiao, Y., Hu, H., Jin, B., Wu, X., Wan, X., Zheng, Y., Xu, H., Lu, X., Sang, X., Ge, P., Mao, Y., Cai, J., Zhao, H., and Du, S. 2022. "A Nomogram to Predict Individual Survival of Patients with Liver-Limited Metastases from Gastroenteropancreatic Neuroendocrine Neoplasms: A US Population-Based Cohort Analysis and Chinese Multicenter Cohort Validation Study," *Neuroendocrinology* (112:3), pp. 263–275.

Xu, W., Zhang, M., Liu, L., Yin, M., Xu, C., and Weng, Z. 2022. "Association of Mucin Family Members with Prognostic Significance in Pancreatic Cancer Patients: A Meta-Analysis," *PloS One* (17:6), p. e0269612.

Yin, M., Liu, L., Gao, J., Lin, J., Qu, S., Xu, W., Liu, X., Xu, C., and Zhu, J. 2023. "Deep Learning for Pancreatic Diseases Based on Endoscopic Ultrasound: A Systematic Review," *International Journal of Medical Informatics* (174), p. 105044. (https://doi.org/10.1016/j.ijmedinf.2023.105044).

Yu, H., Li, M., Cao, D., Wang, Y., Zeng, N., Cheng, Y., Huang, Z., and Song, B. 2022. "Enhanced Computed Tomography Features Predict Pancreatic Neuroendocrine Neoplasm with Ki-67 Index Less than 5," *European Journal of Radiology* (147), p. 110100.

Zhou, Y., Ge, Y., Shi, X., Wu, K., Chen, W., Ding, Y., Xiao, W., Wang, D., Lu, G., and Hu, L. 2022. "Machine Learning Predictive Models for Acute Pancreatitis: A Systematic Review," *International Journal of Medical Informatics* (157), p. 104641. (https://doi.org/10.1016/j.ijmedinf.2021.104641).

Zhu, H., Nie, P., Jiang, L., Hu, J., Zhang, X. Y., Li, X. T., Lu, M., and Sun, Y. S. 2022. "Preoperative Prediction of Lymph Node Metastasis in Nonfunctioning Pancreatic Neuroendocrine Tumors from Clinical and MRI Features: A Multicenter Study," *Insights into Imaging* (13:1), p. 162.