VILNIUS UNIVERSITY

MEDICAL FACULTY

The Final thesis

Overview of Clinical Validation Processes for Artificial Intelligence Applications in Pathology

Margaryta Lyzogub, VI year, 2018, 1st group

Institute of Biomedical Sciences;

Department of Pathology and Forensic Medicine

Supervisor	Prof. dr. Arvydas Laurinavičius
Head of Department/Clinic	Prof. dr. Arvydas Laurinavičius

2024

margaryta.lyzogub@stud.mf.vu.lt

CONTENT

SUMMARY

The paper aims to provide a general understanding of the impact of artificial intelligence on the current pathology and explore the aspects that a medical student or a clinician without a background in the computational pathology need to be aware of to evaluate the performance of the algorithm.

It starts with the overview of the history of computational pathology development emphasizing the speed and the effort needed for the novel trends' introduction into research and clinical practice. In then explores the most common types of algorithms deployed in the industry and cases of their application. An idealistic model of fully symbiotic machine-pathologist workload is depicted. Next, information about the process of artificial intelligence models' development is introduced, focusing on the steps of analytical and clinical validation. The approaches to the data distribution for training, tuning, and testing are discussed and the most commonly used statistical measures are explained. The work concludes with the ideas on the steps needed to take today in terms of medical education, to provide the healthcare specialists of tomorrow with the relevant knowledge to face the era or artificial intelligence prepared.

In conclusion, it takes a combination of pathology, data science, medical statistics, medical law and many more areas of knowledge to accurately assess the novel model, thus providing the healthcare professionals with basic concepts in diverse expertise domains during medical education using stratified approach based on their future role in AI development is crucial.

KEYWORDS

Artificial intelligence, machine learning, deep learning, supervised learning, computational pathology, validation, Ki-67 enumeration algorithm

ABBREVIATIONS

AI	Artificial intelligence
CNN	Convolutional neural network
CPU	Central processing unit
CV	Cross-validation
DL	Deep learning
DP	Digital pathology
EMA	European Medical Agency
FDA	Food and Drugs Agency
FFPE	Formalin-fixed paraffin-embedded blocks
FN	False negative values
FP	False positive values
GAN	Generative adversarial network
GNN	Graph neural network
GPU	Graphics processing unit
H&E	Hematoxylin and eosin
IHC	Immunohistochemistry
MCC	Matthew’s correlation coefficient
ML	Machine learning
MIL	Multiple instance learning
MSI	Microsatellite instability
ROI	Region of interest
ROC	Receiver operating characteristic curve
TN	True negative values

TP	True positive values
WSI	Whole-slide image

1.BACKGROUND

Pathology is one of the specialties benefitting the most from artificial intelligence (AI) algorithms becoming increasingly better, performing such tasks as interpreting complex histopathological images and detecting the patterns suggestive of disease. The potential gains from adapting the digital tools into the workflow include reducing the time per case, and improving the accuracy of assessment, resulting in higher efficiency and scalability in practice. However, before allowing the machine to make decisions impacting human health, rigorous validation is needed. The process aims to provide the practitioners with a scientific and statistical basement demonstrating the reliability of the model while deployed over diverse patient datasets.

This thesis is dedicated to unraveling the existing practices in clinical validation within pathology. By diving into the methodologies and statistical measures employed in this process, it aims to enlight the path toward enjoying the full potential of AI while safeguarding the integrity of diagnostic decisions in clinical settings, providing the illustration of the validation of the Ki67 enumeration tools.

2. LITERATURE SEARCH STRATEGY

Literature search was performed from January 1st 2023, to May 10th 2024, with keywords "Computational pathology OR Artificial intelligence in pathology AND Validation„ "Statistical measures for validation OR cross validation," using PubMed. Only original research articles written in English were selected. Full texts of the relevant articles were extracted after being screened for titles and abstracts.

3. HOW DID WE GET WHERE WE ARE NOW

Pathology has always been the discipline connecting clinical practice and science. It took only 50 years for the journey of artificial intelligence from being a newly invented term in computer sciences to becoming a part of clinical routine.

In digital diagnostics, it is considered that the start took place in 1965 as the year that brought to us the invention of computerized image analysis and the introduction to computer vision for microscopic fields of blood smears. It continued as convolutional neural networks were developed in 1988, as a kind of deep learning algorithm for automated recognition and classification of histopathological images [1]. In 1990 the development of whole-slide scanners was started and in 1994 the first commercial solution called BLISS went through FDA approval successfully. Ever since, multiple commercial scanners with fluorescent and brightfield capabilities have come to the market.

In 1995 the first commercial digital tool for Pap smears also received the green light from the FDA. The digital progress continued as in 2017 another large approval happened in 2017 with Philips and their Intellisite. Moreover, 1 year later AI-based diagnostic tool for diabetic retinopathy was also approved. This set the precedent of AI-based tool registration for healthcare, posing more questions on how to assess the efficacy for diverse groups of patients and resulting in the need for better regulation of the field of digital healthcare solutions. That's why in 2019 FDA released 'AI/ML based software as medical device'. Up to date, EMA has not released any AI-related regulations with the main document regulating the field still being 2017's 'In Vitro Diagnostic Medical Devices Regulation (2017/746)'. However, Office of the European Union has developed and published 'Ethics guidelines for trustworthy AI' in 2019 and 2 years later, in 2021 'The AI Act' was developed and is still going through the revision.

However, pathology is not limited to the digital diagnostics aspect, but also includes computational pathology aspect, such as automatic cytological smear screening becoming possible in 1959, statistical analysis of cell morphology performed in 1965, survival predictions based on nuclear analysis (2009) and morphological features (2011). The more recent advances include microsatellite instability (MSI) prediction from WSI, cancer detection using MIL, HoverNet nucleus segmentation and classification (2019), WSI image search development, prediction of metastasis development, PAIGE.AI, Proscia, DeepLens, PathAI and Inspirata start-ups (2020-2021) and many more. The field is booming with many commercial and academic attempts to produce new more exciting tools.

However, it quite early became evident that to advance quicker, much public collaboration is needed. The success of The Human Genome Project was evident, that's why the idea to create a similar database for pathologists was shared by many minds of the industry. That's how TCGA – public archives of WSI were started in 2005. Another milestone in public efforts is considered to be the CAMELYON challenge, which in 2017 provided the world with algorithms for metastasis detection in the lymph nodes. In 2019 BACH initiative brought us the chance to subtype the breast cancer of regions of interest (ROIs). In 2021 EU Big Picture project was started as a repository of WSIs, including 3 million slides collected and stored with respect to patients' privacy and data confidentiality [2].

The most significant milestones are depicted in Figure 1.[1,2]

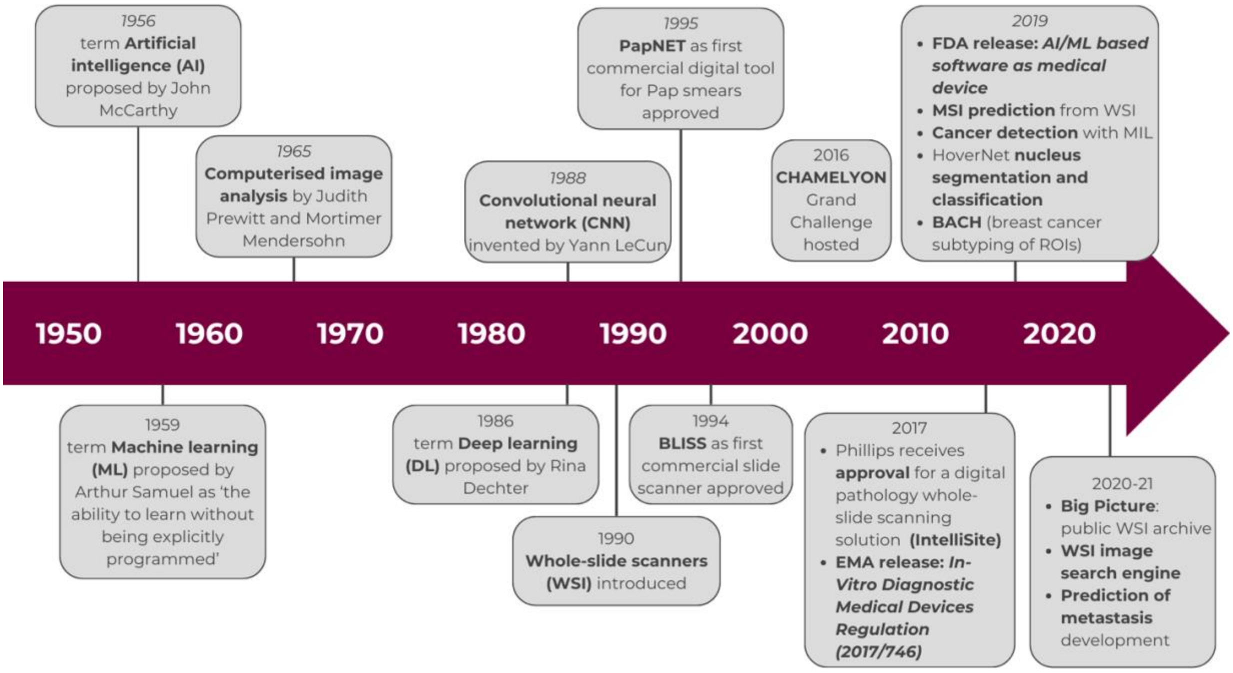

1956
term Artificial intelligence (AI) proposed by John McCarthy
1965
Computerised image analysis by Judith Prewitt and Mortimer Mendersohn
1988
Convolutional neural network (CNN) invented by Yann LeCun
1995
PapNET as first commercial digital tool for Pap smears approved
2016
CHAMELYON Grand Challenge hosted
2019
• FDA release: AI/ML based software as medical device
• MSI prediction from WSI
• Cancer detection with MIL
• HoverNet nucleus segmentation and classification
• BACH (breast cancer subtyping of ROIs)
1950
1960
1970
1980
1990
2000
2010
2020
1959
term Machine learning (ML) proposed by Arthur Samuel as 'the ability to learn without being explicitly programmed'
1986
term Deep learning (DL) proposed by Rina Dechter
1990
Whole-slide scanners (WSI) introduced
1994
BLISS as first commercial slide scanner approved
2017
• Phillips receives approval for a digital pathology whole-slide scanning solution (IntelliSite)
• EMA release: In-Vitro Diagnostic Medical Devices Regulation (2017/746)
2020-21
• Big Picture: public WSI archive
• WSI image search engine
• Prediction of metastasis development

Fig. 1. *The timeline of the most significant events in the development of computational pathology (modified on Bera K. et al. 2019; A.H. Song et al. 2023)*

4. THE DIGITALIZATION OF PATHOLOGY

4.1 The algorithms applied in pathology

Before starting the discussion on all the possible applications of AI, let's review the technologies and methods involved.

According to Abels et al. (2019), '***Artificial intelligence*** (AI) is a branch of computer science dealing with the simulation of intelligent behavior in computers'.

The branch of it that specifically processes data, makes an intelligent judgement, performs the set task and improves the performance is called ***machine learning (ML)***.

Last, but not least, ***deep learning (DL)*** is the branch of machine learning that imitates logical structure via artificial neural network's multilayer hierarchical algorithm that processes large amounts of data to perform the tasks and self-train [3].

Several AI model types are deployed to solve the tasks posed by pathologists. One of the oldest ones used by pathologists since the last century is the ***convolutional neural network*** (CNN). CNNs are supervised machine learning models that require a large pool of labelled data (typically images) for training. They are hypothesis-driven and consist of one or more convolutional layers, pooling layers and 1 or more fully connected layers. The typical tasks that are deployed by these systems include image classification, object detection and semantic segmentation [4]. In the articles, such examples of CNN use for tumor-infiltrating lymphocytes classification, differentiation of malignant and benign lesions between colorectal polyps and lung nodules, interstitial fibrosis and tubular atrophy detection, Ki67 quantification with piNET calculator and many more [5-8]. The biggest advantage of such a neural network is that it's understandable for pathologists, thus deemed more trustworthy to be deployed into clinical practice. The disadvantages include limitations of the use (can only perform pre-trained tasks) and the need for a large data pool with pathologists' annotations to train the algorithm efficiently (if the model is built from scratch).

A more complex model is a ***graph neural network*** (GNN), which are moderately-supervised models designed to operate not on the simple images, but on graphs where nodes represent entities and edges represent relationships between them [9]. They usually consist of embedding layer, graph pooling layer, graph convolutional layer and last fully connected layer to produce the output. GNNs are capable of classification/regression (of nodes, edges, or graphs), prediction and clustering [10]. The examples of application in pathology include nucleus segmentation and classification, tumor detection and staging for breast cancer, grading of colorectal cancer, more accurate Gleason scoring for prostate cancer and many more [11]. The advantage of this model is in capturing special relationships between nuclei and other elements of histological architecture, while disadvantage is limited explainability of the algorithm's decision-making, disturbing the pathologists considering entrusting real patient data and validating it.

The next group is represented by end-to-end models, including ***multiple instance learning*** (MIL) algorithms. These are weakly supervised, hypothesis seeking models that typically need grouped data for training into so-called bags, when the information about instances is aggregated in each bag to make a prediction on the group's label. The typical structure includes activating fully connected layer, encoding layer, pooling layer producing bag representation, classification, and output layers [12]. In pathology, MIL offers 3 levels of classification – image, instance, and pixel. Moreover, MIL models are capable of object detection and identification leading to computer-assisted diagnostics, clusterisation, bag label prediction, annotations refining and image retrieval for the specific task [13]. In pathology, these algorithms may predict HER2 and BRAF status, microsatellite instability (MSI) based on hematoxylin & eosin (H&E) breast tissue samples, predict subtypes of cancer based on H&E WSIs, etc [14,15]. These models share the 'black box' problem with their predecessors, while allowing to explore even more advanced associations.

Generative adversarial networks (GANs) typically involve 2 competing neural networks: a generator that creates fake images and discriminator that compares generated data with ground truth and filters aside fake data to produce most realistic output. GANs may be used in pathology to generate a pool of data that may be further used for training of other kinds of models, for education purposes, to improve the image quality and remove the artifacts [16,17]. These models pose the biggest validation challenge as the current metrics for quality assessment are insufficient for them.

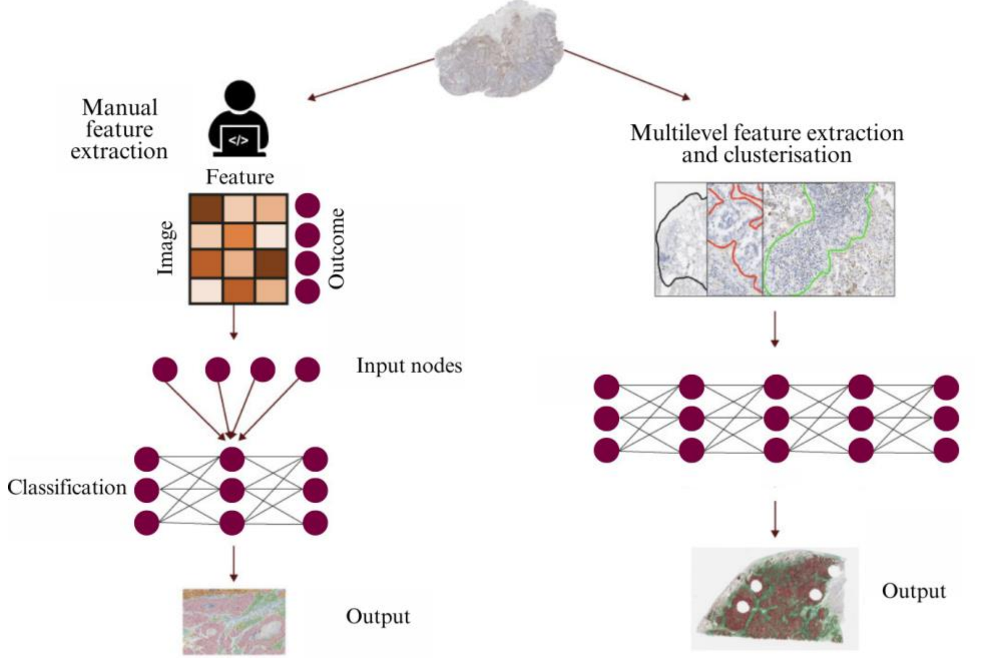
4.2 Supervised learning Vs unsupervised learning

As we have explored in the previous chapter, machine learning models may be supervised to different extent and completely unsupervised. The comparative characteristics of supervised and unsupervised learning is presented in Table 1.

Supervised machine learning is used to cover the current tasks of clinical pathologists. For the training, ground truth annotations are used, where each data item is linked with the expected outcome. It is typically performing such tasks as classification, regression, localisation, and segmentation [18]. The training demands large datasets, and labelling is usually a lengthy process taking away much time from clinicians and researchers. L. Hou et al. specify that providing annotation to 50 WSIs on nuclei segmentation took 120-130 hours of pathologists' workload [19]. This sometimes leads to the limited application and scalability in clinical practice. Another possible disadvantage is related to the challenges of generalisation. As the training dataset is usually quite limited, it might be not diverse enough to represent the target population or variety of condition's forms, thus it might fail to recognise it in unseen data. Unrepresentiveness might be the result of multiple biases, including sampling bias, labelling inaccuracies (for example, immunohistochemical evaluation is often subjected to inter-observer variability), or social factors (including accessibility of healthcare services by various groups). There is also a chance of *overfitting* happening during the training, when the algorithm performs better identifying noise and outliers rather than target features. After overcoming all these potential problems in the process of development, the models are ready to perform very specific tasks with easy to interpret and highly accurate results [20].

Unsupervised machine learning is providing new insight onto the disease diagnostics [3]. They are fed unlabelled data in attempt to discover new useful features and patterns without trying to interpret them. They can solve such problems as clusterisation, dimensionality reduction, and new content generation [18]. These models' validation poses a great challenge. The obvious reason for that is the absence of ground truth. The unsupervised algorithms often produce results that are difficult to comprehend, that's why a lot of effort is now put into producing interpretable AI models, including dimensionality reduction techniques use, deploying initial model-based neural networks for clusters visualisations (such as localisation

***Tab. 1**. Comparative characteristics of supervised and unsupervised learning (modified on Baxi V. et al., 2022; Kim I. et al., 2022)*

	Supervised learning	Unsupervised learning
Scheme of work	 Manual feature extraction Feature Image Outcome Input nodes Classification Output Multilevel feature extraction and clusterisation Output	
Human supervision	Requires intervention to label the data, pathologist is a trainer and a validator	Doesn't need much guidance, pathologist is mostly a validator
Data	Annotated with ground truth	Unlabeled data
Hardware requirements	Mostly CPUs	Mostly GPUs
Training time	Less	Longer
Typical tasks	• Image classification • Object detection and localization • Semantic segmentation • Instance segmentation • Regression • Survival analysis	• Feature extraction and representation learning • Anomaly detection • Data augmentation and enhancement • Transfer learning and domain adaptation • Image generation and synthesis • Treatment effectiveness prediction
Applications in pathology	• Tumor cells identification • Computing of mitotic counts • Immunohistochemistry scoring • Standardized histological scoring criteria application (Gleason score) • Detection of lymph node metastases	• Identifying morphological features (nuclear shape, nuclear orientation, texture, tumor architecture, etc.) to predict recurrence in early-stage non-small cell lung cancer (NSCLC) from H&E slides • Grading prostate cancer • Identifying biomarkers for disease-specific survival in early-stage melanoma • Detection of invasive breast cancer regions on WSIs • Predicting response to chemoradiotherapy in locally advanced rectal cancer

heatmaps depicting the relevance for the model's decision making), explaining outliers and inliers and many more [21,22]. Another problem is the choice of metrics to demonstrate the model's good quality as it is usually quite subjective and depends on the model. For example, to demonstrate internal validity of clustering method, cohesion (intra-cluster metric), separation (inter-cluster metric) or a mixture of both can be used [23]. It is also quite difficult to construct a good training set that would have sufficient variability and minimal noise to ensure algorithm's stability and robustness, as well as validation dataset to be representative. However, these models are more data efficient as they require no annotations, they are capable of extracting hidden features and patterns and are suitable for a variety of tasks. Moreover, they may generate the annotated data to train supervised algorithms faster and more efficiently.

4.3 Value of AI for pathology workflow

Artificial intelligence keeps expanding its abilities to contribute more towards the medical diagnosis. However, the integration of the innovations is quite limited with the centres' abilities to digitalise the workflow, have sufficient funds to invest into purchasing of all necessary hardware and software solutions and have sufficient storage space for the data accumulated, high quality Wi-Fi connection, trained technicians to operate them. Several cost-benefit studies on the digitalisation may be found. For example, Matthew G. Hanna et al. in 2019 has published their estimate of 1.3 million dollars of operational savings over 5 years after integration of digital pathology workflow in the large academic pathology centre [24]. Moreover, College of American Pathologists keeps advocating for the new digital pathology (DP) reimbursement codes, which became effective since January 1st 2023, to ensure sustainable development of the DP programme.

Furthermore, integration of unsupervised ML algorithms is also restrained with ***black box problem*** related to the difficulties to interpret the decision-making of the model. Even the explainable AI algorithms highly depend on their operators and might not meet the needs of the clinical pathologists and physicians as the final users [25].

In the ideal situation, algorithm would assist in every stage of patients' data and material processing, from quality control of formalin-fixed paraffin-embedded (FFPE) blocks to providing binary decisions to the clinicians regarding the possible diagnosis. However, as the options generated might be image processing-based, rather than evidence-based, the final decision should always be made by the pathologist and the algorithm has to go through vigorous validation prior to introduction into the clinical workflow. In this scenario AI symbiosis with clinical pathologist is aiming to increase accuracy and speed of processing each case [20]. The depiction of tasks in which AI may be involved in various parts of pathologists' work is schematically portrayed in Figure 2.

It is also important to consider that the center might be participating in the clinical trials and might additionally benefit from quality controls at every checkpoint of the study and more efficient way to archive, store and retrieve the slides of interest. Moreover, as the system would be able to access the patient meta-data, it might be able to generate new links that may be further investigated by the researchers or generate prognosis or treatment response predictions for individual patients providing additional information for physicians that would contribute to precision and personalization of healthcare services [26].

Another possible application of AI algorithms lies in the educational use. Students and residents may use the AI-selected WSIs with pathologists' annotations to train more efficiently and interactively. The students can zoom through the images, mark ROIs, start communication with the supervisors about areas in question [27].

The opportunity to work flexibly from the comfort of home and provide healthcare services to remote areas more efficiently is the last, but not the least important of the benefits of digitalisation.

Overall, the AI integration provides a lot of opportunities and benefits, especially in the long-term use perspective.

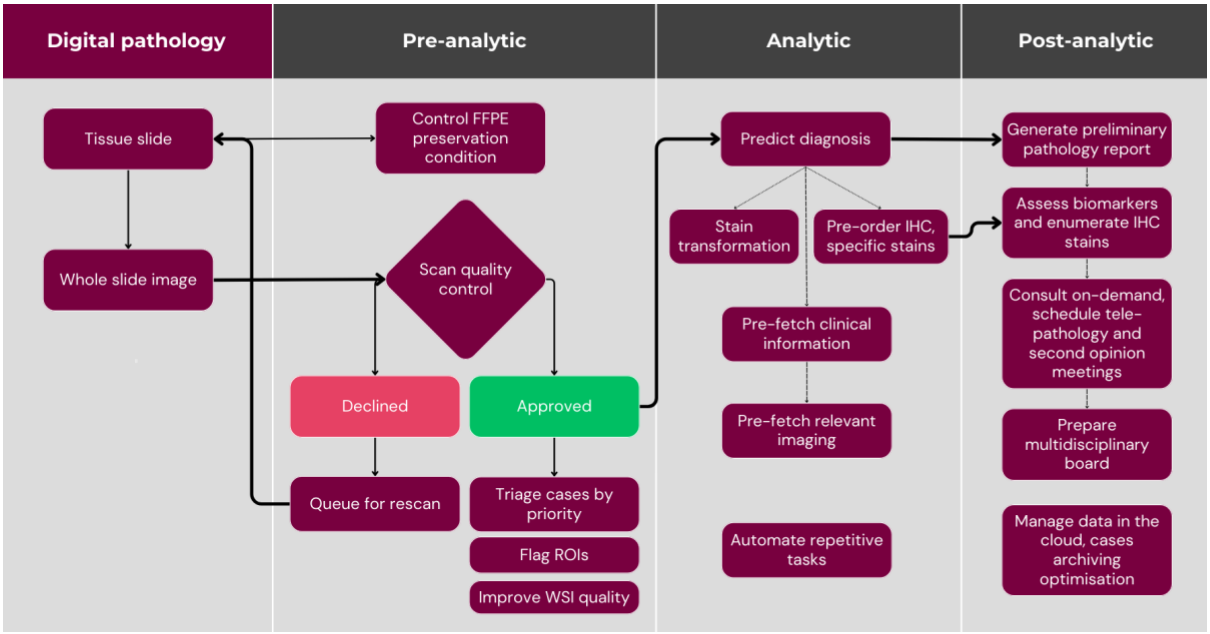

Digital pathology
Pre-analytic
Analytic
Post-analytic
Tissue slide
Whole slide image
Control FFPE preservation condition
Scan quality control
Declined
Approved
Queue for rescan
Triage cases by priority
Flag ROIs
Improve WSI quality
Predict diagnosis
Stain transformation
Pre-order IHC, specific stains
Pre-fetch clinical information
Pre-fetch relevant imaging
Automate repetitive tasks
Generate preliminary pathology report
Assess biomarkers and enumerate IHC stains
Consult on-demand, schedule tele-pathology and second opinion meetings
Prepare multidisciplinary board
Manage data in the cloud, cases archiving optimisation

Fig. 2. *The workflow of pathology center with fully integrated AI ecosystem (modified on Kim I. et al. 2022)*

5. THE LIFECYCLE OF THE ALGORITHM DEVELOPMENT

5.1 Stakeholders and stages

The birth to the artificial intelligence happens in the human brain, when a member of industry, academia or clinician sees the need that may be solved with the technology. They usually perform the research of the current market to look for analogues, potential sources of funding, evidence for such algorithm potential efficacy etc. Quite often the initial stages of research are well protected by intellectual property laws, thus several groups of researchers might be working on the same solution or industry might be developing the application that solves non-existing problem as they are biased with the opinion of expert or group of them that are employed by the company. After the idea is worded, it is submitted to the ***Ethical committee for approval***. It should be noted that it tends to be more challenging to receive approval for the solutions that require redesign of the workflow or substantial funding. Once the permission is received, the research protocol is designed, outlying:

- pre-processing stage (expected output, design of algorithm to achieve it),
- analytical stage (pilot and follow-up patient sample size),
- processing stage (data organisation, storage, statistical analysis measures) [28].

The following stage is ***validation***, which is defined by College of American Pathologists' as 'the process to establish that the performance of a test, tool or instrument is acceptable for its intended purposes' [29]. It is usually divided into 2 stages – analytical and clinical validation, which will be discussed more in the next chapter. Following validation, ***verification*** is executed to prove that the new model performs according to specifications if used according to manufacturer's instructions. Usually, the performance is assessed in multiple centres and by multiple pathologists.

The results of validation and verification are further used for ***regulatory approval***. AI algorithms in computational pathology are considered to be in-vitro diagnostic devices, thus according to current EMA legislation (2017/746 and 2017/745), they are classified according to the potential risk to patient (classes A-D, where WSI are classified as class C) and certified by Notified Bodies. The similar principle is employed by FDA, and the devices that are registered as Class II or III must go through Pre-market notification [30]. The registration process may be avoided if the device is marked as 'For Research Use Only'.

After the approval is received, the model goes through the ***accreditation*** process with regular external quality assessments to keep the accreditation valid. Now it may become the part of ***clinical practice*** and if the performance is satisfactory in long-term run – it becomes part of management guideline. The typical challenges, that need to be overcome at this stage include limited protocols for AI deployment (optimisation of interface and model's output for pathologists' use), limited computing resources of the centre and limited reimbursement models to make the use of AI sustainable [31]. Moreover, any in-house changes to algorithm would demand update to each stage of accreditation.

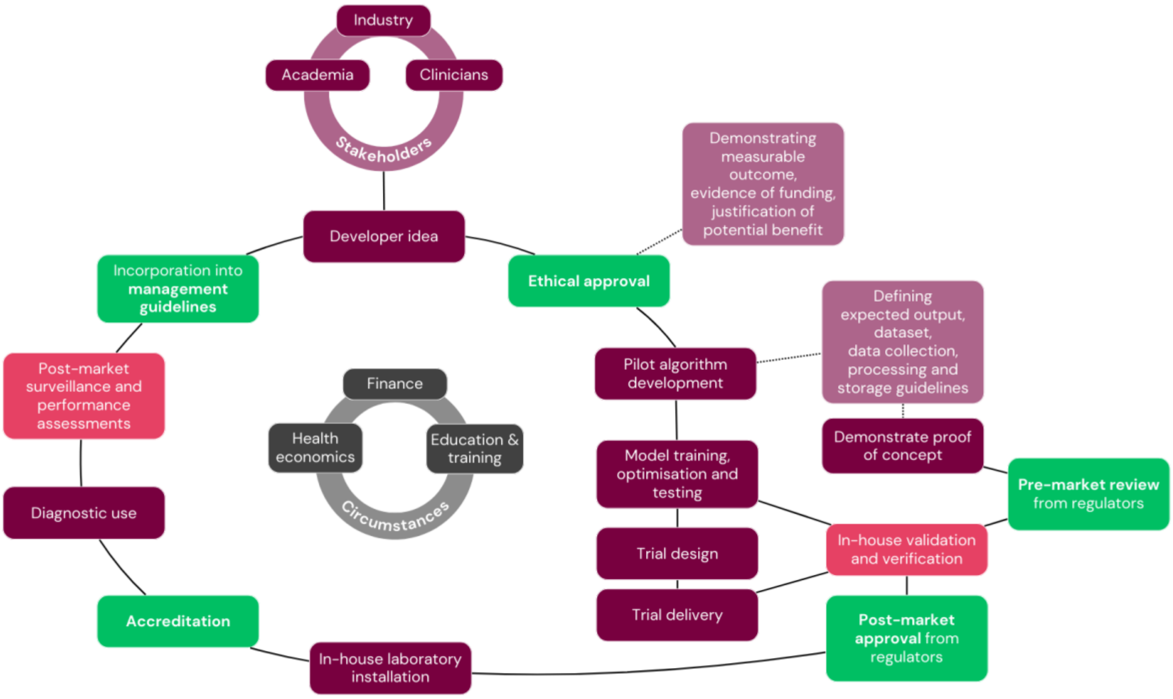

Industry
Academia
Clinicians
Stakeholders
Developer idea
Demonstrating measurable outcome, evidence of funding, justification of potential benefit
Ethical approval
Incorporation into management guidelines
Defining expected output, dataset, data collection, processing and storage guidelines
Pilot algorithm development
Post-market surveillance and performance assessments
Finance
Health economics
Education & training
Circumstances
Demonstrate proof of concept
Model training, optimisation and testing
Pre-market review from regulators
Diagnostic use
In-house validation and verification
Trial design
Trial delivery
Accreditation
Post-market approval from regulators
In-house laboratory installation

Fig. 3. *Roadmap of AI tool development. In green marked the stages that require external assessment and approval (by Ethical committee, EMA/FDA/national regulators, international professional society/expert groups), in red – stages of performance assessments (modified on Colling R. et al. 2019)*

5.2 Analytical and clinical performance stages

In last chapter, the concepts of analytical and clinical validation were introduced.

Analytical validation is focused on the ability of the algorithm to process and interpret the data as intended. This stage usually includes the development of the study protocol that would include the algorithm of data collection and specimens handling, choice of markers to test performance, limits of detection and quantification, cut-offs and other. It is important to note that these markers are relying on the ground truth annotations and may be impacted by inter-observer and intra-observer variability, thus opinions of multiple pathologists are needed.

Different levels of validation are used at this stage. It usually includes internal and external validation. *Internal validation* is usually performed on the smaller unseen portion of dataset that was used for training. *External validation* utilizes external datasets to evaluate generalizability and robustness of the model. It is important to blind the developers if the data from multiple centers is used to avoid bias. The summary of data processing for analytical validation is provided at Figure 4 [32,33].

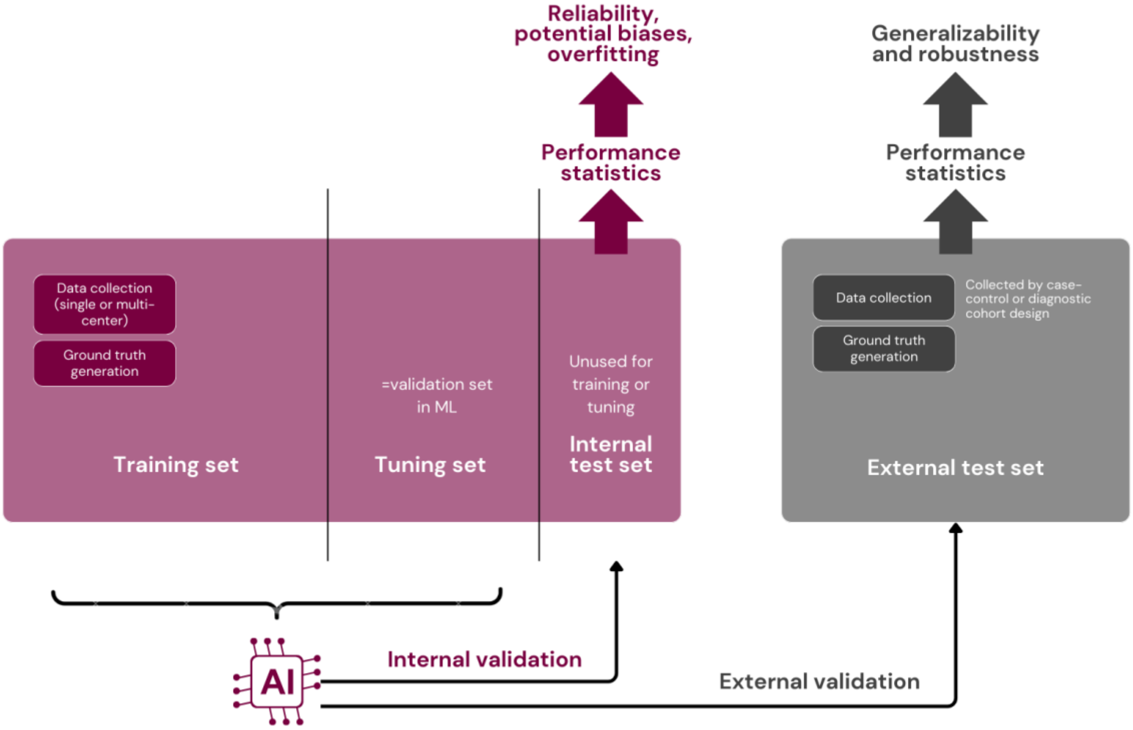


Fig. 4. Datasets for development and testing of AI models according to hold-out cross-validation design (modified on Park S.H. et al., 2021; van der Laak J. et al., 2021)

In case of unsupervised algorithms, the choice of metrics is even more challenging, as for internal validation cohesion, separation, proximity matrix or hierarchical methods may be used, while for external validation classification accuracy coefficient, purity, precision-recall, F-measure, Jaccard coefficient, entropy or other methods may be used [23,34].

Nakagawa K. et al. propose to develop standardized protocols for performance assessments in pathology, that would have to include demographic report and publicly available code, performance on data from independent datasets and on images from different scanners [31].

As a final product of analytical validation, internal documentation, noting down the population metrics, ground truth or reference selection, explanation on the choice of statistics and its calculations; regulatory submission; journal article should be produced [35].

Clinical validation sets its goal on assessing the benefit to the patient and clinical need in such medical device typically as a clinical trial performed on cohorts of patients with and without the feature of interest. It must assess the diagnostic accuracy via sensitivity and specificity, likelihood rations and expected values), ensure diagnostic reproducibility. It also compares the method to gold standard in the industry (for example, AI output with the output of several experienced pathologists) and assesses the potential clinical benefits and losses. The benefits might be advances in treatment selection and response prediction, more accurate prognostication, changes in treatment outcomes for patients, time optimizations. It may include exploration of the peculiarities of new model's application in clinics and financial calculations [28]. It is typically designed as independent crossover or sequential study. *Independent crossover design* has 2 arms: pathologist evaluation with and without AI intervention, which is usually achieved by substantial time between pathologist's evaluations. *Sequential study design* has 1 arm marking the pathologist evaluation prior to model application and post-AI intervention [36].

As the result, all the data is collected and submitted to the Regulatory Body for approval.

5.3 Analytical validation data strategies

As it was stated before, scarcity of annotated data is usually one of the biggest challenges in the algorithm development. The centre may decide to produce its own database; however, it will require much financial and time effort from the pathologists. For this reason, several strategies of data splitting were developed to mitigate this challenge and provide the ability to use smaller datasets.

Train/test split, suggests initially randomly dividing the data into the training and unseen validation datasets prior to the development of the model to assess the performance of the algorithm. Usually the seen sets include 70-90% of images, while test set includes 10-30% of cases [37]. It is a classic technique deployed if sufficient data is available and the method effectively represents the model's performance if deployed in real clinical practice [38].

Cross-validation (CV) is the typical solution to the limited dataset issue, it is a statistical approach of data resampling from same core dataset aiming to evaluate generalization and prevent overfitting.

- ***Hold-out CV*** also depicted on Fig. 4, advised by several authors, and is the simplest of cross-validation methods as it required only 1 run of data. It involves splitting into 3 datasets, using a separate portion of training set for fine-tuning (validation or tuning set) and a separate testing set. The final performance is evaluated when using the testing set. Its application is however constrained if the size of the set is small or it's imbalanced. There is also a novel MuSCID method, proposing to use WSIs of off-target organ for calibration of algorithms to prevent data leakage between training and testing sets [39].
- ***K-fold CV*** involves splitting the data into k equal portions randomly. One section of data is excluded from the training set and further used for validation. Then a different

fold is left out and used for testing until each sub-sample has been used for validation and all the data is fed to the model. The performance is calculated as the mean of assessment results. This method does not ensure that training and testing data are separate, therefore might provide opportunity for overfitting (if approach is used for tuning) and overly positive estimates [40].

- ***Stratified K-fold CV*** is the K-fold CV variation, where consistent representation of each class is ensured in every fold, marking the fixed proportion in a fold. This approach is more favoured over K-fold CV as it is more efficient for problems of classification with unequal class distribution.
- ***Leave-P-out CV*** is a technique that suggests keeping away P samples as test set and use all the remaining data for training. The process is repeated until all the data points combinations in P group were achieved. It is important to note, that the higher the number of P is, the more combinations are possible [41]. Due to this factor a simplified version called ***leave-one-out CV*** exist, where P=1. Some authors consider it being a type of K-fold validation, where number of folds is equal to the sample size [42].
- ***Nested CV*** consists of 2 stages. In the primary stage or outer layer, the principle of K-fold CV is employed to split data into seen and testing sets. In the internal layer, data is split into a training and tuning sets according to the same principle. This provides an opportunity to calibrate hyperparameters. Performance is calculated as the mean of outer layer results and the results are considered to be more reliable, especially for complex models [43].
- ***Partially nested or hierarchical CV*** combines the test-split simplicity and nested CV computational efficiency. At first, the data is randomly split into training and testing sets. Next, the data in the training set is arranged into folds. The model is trained on all folds except one used for intermediate evaluation, just like in K-fold CV. If hyperparameter tuning is needed, the data may be split according to outer and inner layer principle of nested CV. After completing the training, the final evaluation is performed on the initially separated training set [38].
- ***Monte Carlo CV or random sub-sampling*** performs random selection of portion of data for training and further using all the remaining data for testing several times. The training-testing ratio is the same, however a sample might be selected several or no times for validation. The average of performance indicators is used [44].

Bootstrapping validation involves the random sampling of training dataset and the remaining data items are used for validation. However, afterwards a re-sampling is performed, and the repeated between runs data-points are replaced, causing duplications in each following ‘strap’. The bootstrap estimate is calculated as the mean of all runs. The result is highly dependent on representativeness and size of the initial dataset and the process may be time-consuming. This approach is considered to improve stability of algorithms [45].

The schematic illustration of all methods is provided on Figure 5.

5.4 Performance indicators

There are a lot of statistical measures used to illustrate the performance of AI models. Let’s review the ones that are more frequently seen in the pathology articles.

Let’s start with building the confusion matrix. First the output must be converted into binary format (yes-no). At this stage, a threshold has to be set as it affects the final result. Next, the results need to be summarised into ***diagnostic cross-table***, looking as follows:

Tab. 2. *Confusion matrix: TP = true positive, FP = false positive, FN = false negative, TN = true negative*

		The feature / disease	
		Present	***Absent***
Algorithm output	***Present***	TP	FP
	Absent	FN	TN

Sensitivity or recall is a probability of the algorithm’s output being positive, when the feature / disease are actually present. It represents the true values that were correctly classified and is calculated via formula:

$$\mathit{Sensitivity} = TP / (TP+FN)$$

It should be interpreted cautiously in the imbalanced datasets.

Specificity is a chance of the output being negative when the feature or condition are actually absent. It depicts the ratio of false values being correctly classified and is evaluated by:

$$\mathit{Specificity} = TN / (TN+FP)$$

Accuracy is denoting how well does the model provide true values, is a measure of systematic errors and is measured as the ratio of correctly identified cases to all data points and is calculated as:

$$\mathit{Accuracy} = (TP+TN) / (TP+FP+TN+FN)$$

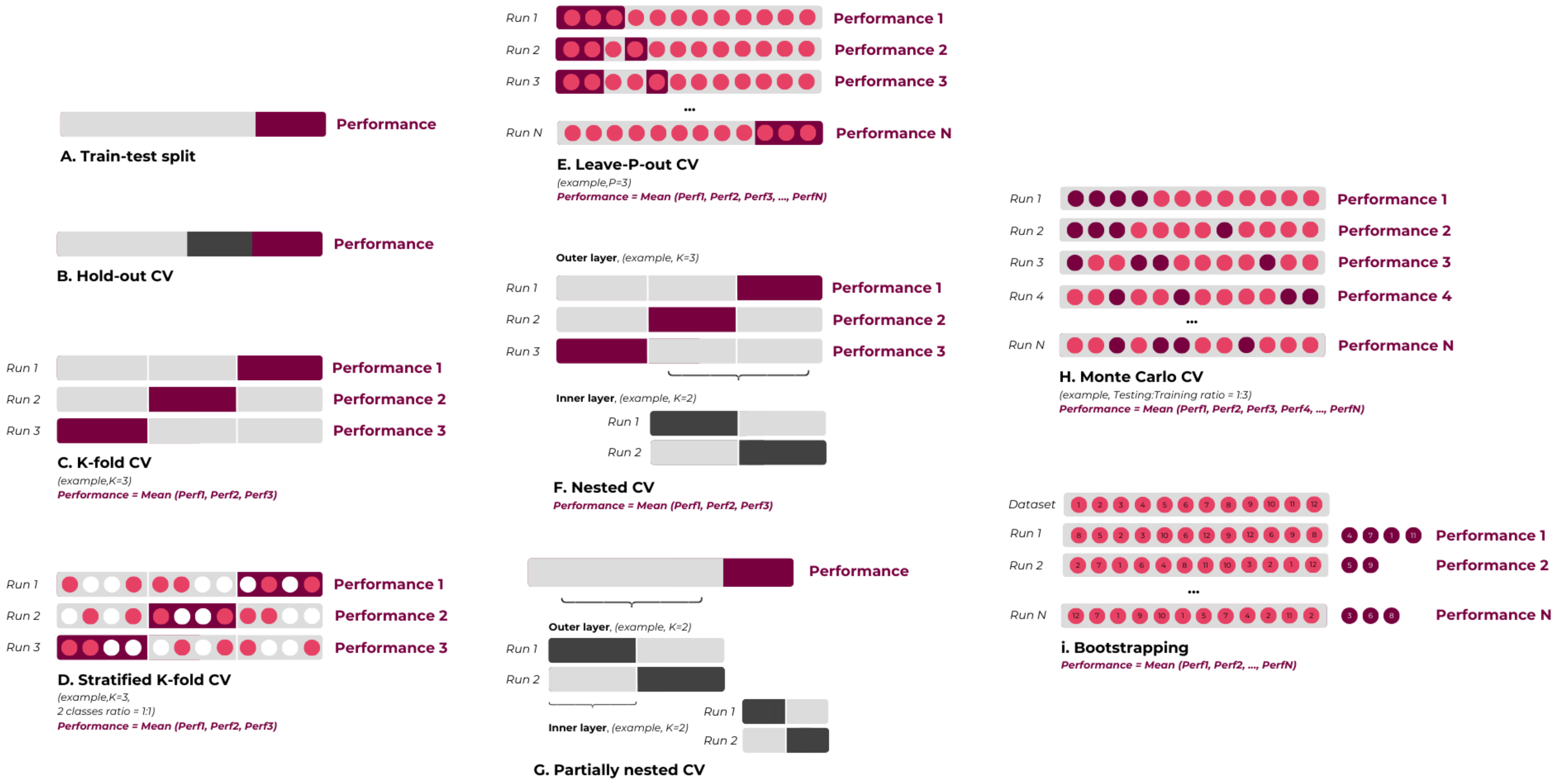


Fig 5. *The methods used for splitting the data during analytical validation (light grey depicts training set, dark grey – tuning set, burgundy – testing set; each circle depicts 1 data piece, red and white are used to distinguish 2 classes)*

Precision is calculated as a proportion of successfully identified cases to all outputs within the positive or negative class. It is counted as follows:

Precision = TP / (TP+FP) or Precision = TN / (TN+FN)

Often the first formula's result is referred to as positive predictive value, while second – as negative predictive value. It is important to note that all these values should lie within [0;1] span [32, 41, 46].

The dependences between these values may be illustrated via 2 common visualisations that may be prepared only with the direct access to the model, therefore are more difficult to check.

Receiver operating characteristic curve (ROC curve) is a demonstration of dependency between specificity and sensitivity. To plot it, specificity values are noted on X-axis, while sensitivity – on Y-axis. This allows to identify area under ROC curve, which is the same for all threshold values and can be 1 at maximum. It is important to note that in clinical practice the algorithm is still functioning with a certain threshold, thus sensitivity and specificity counted according to it are the main true measures of the algorithm's performance. It might provide an evaluation too optimistic to the real picture if the dataset is imbalanced as FP might be too small, when in reality it is more significant.

A variation of ROC curve is a ***free-response ROC curve***, where sensitivity is kept on Y-axis, while mean number of FP is depicted on X-axis. It is used to assess the performance of algorithms for computer-aided detection.

Precision-recall curve shows the relation between recall depicted on X-axis and precision plotted on Y-axis. It demonstrated the performance only in the positive class. Similarly to ROC curve, area under precision-recall curve is used as the metric of the graph, for a random classifier it is equal to TP ratio in the whole set, therefore it might deal with the imbalanced data more realistically.

For demonstration, ROC and precision-recall curves from deep-learning model for predicting of multiple sclerosis-associated RNA by Xiaoping Sun et al. (2022) is added as Figure 6 [47].

Using the confusion matrix and previously mentioned statistical measures, more advanced calculation may be performed.

F1-score is used for unbalanced datasets and is a harmonic mean of precision and recall. The harmonic mean promotes similar measures of precision and recall, thus the larger difference between precision and recall is, the worse F1-score will be got [48]. It is estimated as:

F1-score = 2 (Precision*Recall) / (Precision+Recall) = TP / (TP + ½(FP+FN))*

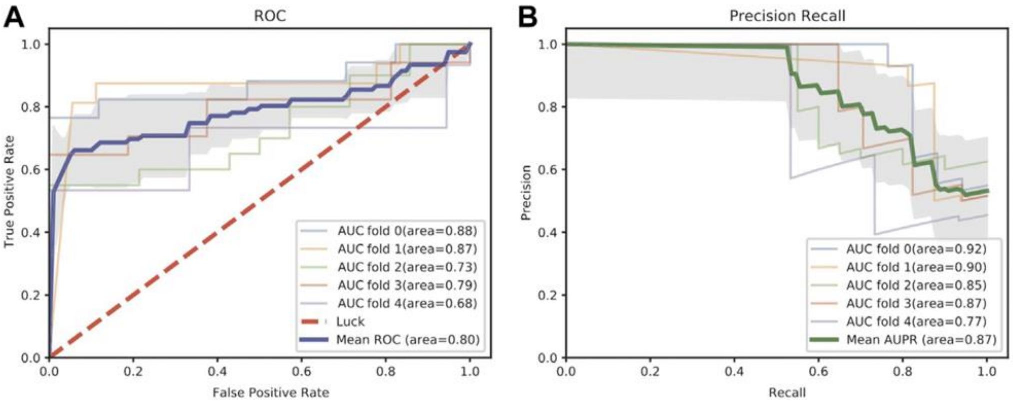


Fig. 6. *ROC curve (A) and precision-recall curve (B) in five-fold cross validation of the miRNA-multiple sclerosis association prediction. Standard deviation is presented as a shade. From Xiaoping Sun et al., 2022.*

Kappa is a measure that demonstrates inter-rater reliability as the extent of agreement between 2 raters, in our case – truthing pathologist and AI, during classification.

$$\kappa = (Po\text{-}Pe) / (1\text{-}Pe),$$

where Po – observed agreement between raters, Pe – expected agreement between raters be chance. For binary systems the formula may be modified:

$$\kappa = 2 * (TP*TN - FN*FP) / ((TP+FP) * (FP + TN) + (TP + FN) * (FN + TN))$$

The values are in the range [-1;1]. To interpret it, following references are used: κ = 1 corresponds to perfect agreement between raters, κ = 0 means that agreement is equivalent tp chance, κ < 0 is interpreted as agreement worse than chance [46].

Matthew's correlation coefficient (MCC) is a balanced measure of classification performance used for imbalanced datasets. It is considered to be more reliable measure as it reflects the performance in all 4 confusion matrix categories and is equally influenced by both positives and negatives [49]. It is calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

There are many more statistical measures, used less often, such as balanced accuracy, bookmaker informedness, markedness, diagnostic odds ratio, confusion entropy, Brier score, log-loss, Jaccard coefficient, etc. The choice is made based on the type of task algorithm is trying to solve, importance of false positives and false negatives, data distribution and many other factors.

6. EXAMPLE OF Ki67 ENUMERATION ALGORITHMS AND THEIR VALIDATION

Ki67 is cellular proliferation marker, the higher values of which were linked to the worse survival in breast cancer patients. Various guidelines recommend assessing from 300 cells (100 cells in 3 ROIs) to 1000 cells per case. The patients are usually stratified into the risk groups with the most common cut-offs of 14% or 20%. The full-scale assessment is time-consuming, thus eyeballing technique with approximate estimation is used in clinical practice. This creates the task, that might be efficiently solved with artificial intelligence model. Let's illustrate the theory discussed with three articles on clinical validation experience of various Ki67 enumeration algorithms and the variety of statistical representations.

Feng M. et al. (2020) introduced a concept of machine-pathologist competion assessing the accuracy of in-house algorithm compared to the work of 10 pathologists [50]. The database included 1017 sections with the split to 677 sections in training set, 153 in verification set and 187 in test set. 28 labelling workers in 3 groups performed annotations on over 200 000 cells per person with a senior pathologist providing the feedback on annotations and 2 attending pathologists conducting the quality check during more than 80 hours. The competion was performed on unseen 10 slides. The accuracy of the algorithm developed was 99.4% Vs 90% accuracy of participating doctors [50].

Van den Berg E.J. et al. (2021) assessed correlation between eyeballing, counting by 2 pathologists and ImmunoRatio and interobserver agreement between 2 pathologists on 204 breast carcinoma core biopsies. To illustrate the validation results, Lin concordance correlation coefficient, $\chi 2$ test and Kohen's kappa (with cut-offs 14% and 20%). were used. The highest correlation assessed by Lin's concordance correlation coefficient was observed between the counts by 2 pathologists (0.965), next between ImmunoRatio and Pathologist 1 (0.790), ImmunoRatio-eyeballing (0.716), eyeballing-Pathologist 1 (0.698). To interpret, one should know that the closer the value of Lin's concordance correlation coefficient is to 1, the higher the agreement is. Kappa was considered to be more accurate than $\chi 2$ test, moderate agreement between Pathologist 1 and Immunoratio was demonstrated both with cut-off 14% and 20% with the values of $\kappa=0.428$ ($P<0.001$) and $\kappa=0.560$ ($P<0.001$) respectivelly [51].

Zehra T. et al. (2023) assessed concordance between Mindpeak software and manual assessment on 60 retrospectively collected cases for 1,5 months. The groups presented mean values as 36.40±25.7 and 38.35±24.7 for manual and AI-based enumeration. Paired t-test was used to demonstrate high concordance and the value was 0.00 with p<0.001 demonstrating statistical significance [52].

7. THE FUTURE OF MEDICAL PROFESSION IN THE AI CONTEXT

As we have discussed in the history chapter, the development of computational pathology and AI application in medicine overall is getting more and more intense. Modern healthcare students are already well aware of GPT as large language model, Grammarly as an AI-based writing assistant, Otter.ai for spoken text to transcript and many more. They might be less aware of Med-PaLM as a tool able to generate artificial clinician's responses, AlphaFold in drug discovery, numerous models already approved for healthcare. There are many challenges in the integration of AI into diagnostic and treatment processes, one of which being training of healthcare professionals. The idea to split clinicians into 3 groups of developers, translators and consumers based on the level of interaction with AI is proposed by Faye Yu Ci Ng et al. (2023) [53]. The healthcare workers are facing information overload, thus the depth of technical knowledge should be stratified [54]. Developers require much technical competency; they are computer and medical sciences qualified. Translators are responsible for formulating the diagnostic questions for engineers, actively participate in the model development and validation. Lastly, customers are majority of healthcare specialists uninvolved in any of the development stages, however they still need technical knowledge to correctly choose and deploy the models.

The goal of medical education is to release the specialists in every category mentioned, therefore several universities in America (United States, Canada, Mexico) and Asia (Korea, China, Singapore) started various initiatives to improve the knowledge of healthcare students in the realm of artificial intelligence. Majority are several months long (1-6 months), elective, many are integrated into informatics courses. The proposed curriculum to be studied, pre-stratified based on the 3 groups of specialists outlines the need in knowledge regarding technicalities, validation, ethics, and appraisal. The concepts advised to study are presented in Figure 7.

The earlier healthcare professionals will be exposed to the stratified AI-related content, the more prepared the medical community will be tomorrow.

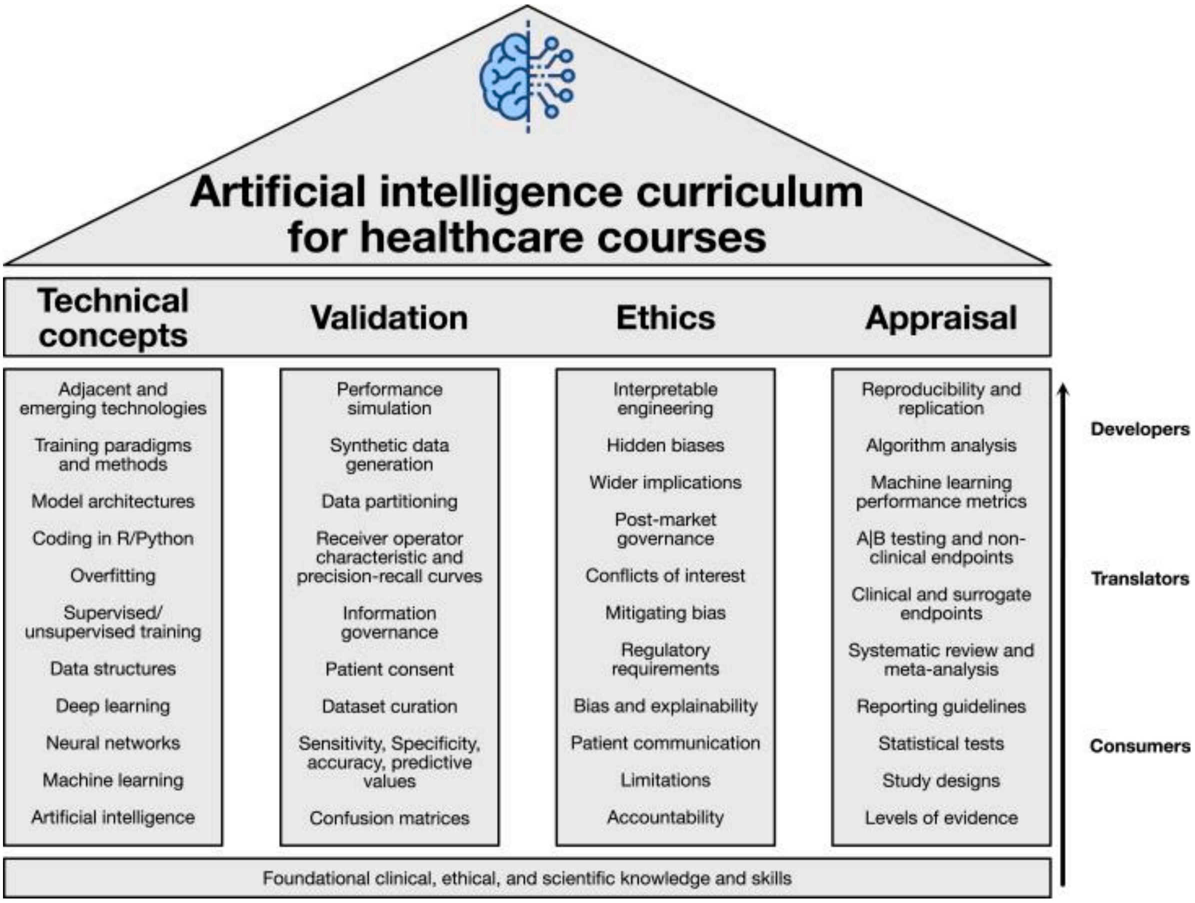


Fig. *7. The map of AI curriculum (from Faye Yu Ci Ng et al., 2023 [50])*

8. CONCLUSION

This paper provides a narrative literature review of artificial intelligence application in pathology and its validation peculiarities. The lifecycle of the algorithm is discussed in most detail, outlining stakeholders, stages of development and assessment, validation data strategies and most common statistical measures. Advantages and disadvantages of various algorithms and validation approaches are outlined.

Looking ahead, the technological advancements present many exciting opportunities to the future healthcare specialists and ensuring their basic computational education according to their plan to be involved into the development and clinical integration process.

LITERATURE

1. Bera, K., Schalper, K.A., Rimm, D.L. et al. **Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology**. Nat Rev Clin Oncol 16, 703–715 (2019). https://doi.org/10.1038/s41571-019-0252-y
2. Song, A.H., Jaume, G., Williamson, D.F.K. et al. **Artificial intelligence for digital and computational pathology**. Nat Rev Bioeng 1, 930–949 (2023). https://doi.org/10.1038/s44222-023-00096-8

3. Abels E, Pantanowitz L, Aeffner F, Zarella MD, van der Laak J, Bui MM, Vemuri VN, Parwani AV, Gibbs J, Agosto-Arroyo E, Beck AH, Kozlowski C: **Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association.** J Pathol 2019. pp. 286-94 doi:10.1002/path.5331 https://www.ncbi.nlm.nih.gov/pubmed/31355445
4. Cheng J. **Convolutional neural networks**. PathologyOutlines.com website. https://www.pathologyoutlines.com/topic/informaticsconvnet.html. Accessed May 2nd, 2024.
5. Meirelles ALS, Kurc T, Kong J, Ferreira R, Saltz JH, Teodoro G. **Building Efficient CNN Architectures for Histopathology Images Analysis: A Case-Study in Tumor-Infiltrating Lymphocytes Classification.** Front Med (Lausanne). 2022 May 31;9:894430. doi: 10.3389/fmed.2022.894430. PMID: 35712087; PMCID: PMC9197439. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9197439/
6. Zhang S, Han F, Liang Z, Tan J, Cao W, Gao Y, Pomeroy M, Ng K, Hou W. **An investigation of CNN models for differentiating malignant from benign lesions using small pathologically proven datasets.** Comput Med Imaging Graph. 2019 Oct;77:101645. doi: 10.1016/j.compmedimag.2019.101645. Epub 2019 Aug 11. PMID: 31454710; PMCID: PMC6800808.
7. Ginley B, Jen KY, Han SS, Rodrigues L, Jain S, Fogo AB, Zuckerman J, Walavalkar V, Miecznikowski JC, Wen Y, Yen F, Yun D, Moon KC, Rosenberg A, Parikh C, Sarder P. **Automated Computational Detection of Interstitial Fibrosis, Tubular Atrophy, and Glomerulosclerosis.** J Am Soc Nephrol. 2021 Apr;32(4):837-850. doi: 10.1681/ASN.2020050652. Epub 2021 Feb 23. PMID: 33622976; PMCID: PMC8017538.
8. Geread RS, Sivanandarajah A, Brouwer ER, Wood GA, Androutsos D, Faragalla H, Khademi A. **piNET-An Automated Proliferation Index Calculator Framework for Ki67 Breast Cancer Images**. Cancers (Basel). 2020 Dec 22;13(1):11. doi: 10.3390/cancers13010011. PMID: 33375043; PMCID: PMC7792768.
9. Ahmedt-Aristizabal D, Armin MA, Denman S, Fookes C, Petersson L. **Graph-Based Deep Learning for Medical Diagnosis and Analysis: Past, Present and Future. Sensors** (Basel). 2021 Jul 12;21(14):4758. doi: 10.3390/s21144758. PMID: 34300498; PMCID: PMC8309939.
10. La Rosa M, Fiannaca A, La Paglia L, Urso A. A Graph Neural Network **Approach for the Analysis of siRNA-Target Biological Networks.** Int J Mol Sci. 2022 Nov 17;23(22):14211. doi: 10.3390/ijms232214211. PMID: 36430688; PMCID: PMC9696923.
11. David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, Lars Petersson. **A survey on graph-based deep learning for computational histopathology**, Computerized Medical Imaging and Graphics, Volume 95, 2022, 102027, ISSN 0895-6111, https://doi.org/10.1016/j.compmedimag.2021.102027.
12. Fatima S, Ali S, Kim H-C. A **Comprehensive Review on Multiple Instance Learning.** *Electronics*. 23; 12(20):4323. https://doi.org/10.3390/electronics12204323
13. Wang Z, Saoud C, Wangsiricharoen S, James AW, Popel AS, Sulam J. **Label Cleaning Multiple Instance Learning: Refining Coarse Annotations on Single Whole-Slide Images.** IEEE Trans Med Imaging. 2022 Dec;41(12):3952-3968. doi: 10.1109/TMI.2022.3202759. Epub 2022 Dec 2. PMID: 36037454; PMCID: PMC9825360.
14. Couture HD. **Deep Learning-Based Prediction of Molecular Tumor Biomarkers from H&E: A Practical Review.** J Pers Med. 2022 Dec 7;12(12):2022. doi: 10.3390/jpm12122022. PMID: 36556243; PMCID: PMC9784641.
15. Pablo Meseguer, Rocío del Amor, Valery Naranjo. **MICIL: Multiple-Instance Class-Incremental Learning for skin cancer whole slide images**. Artificial Intelligence in Medicine, Volume 152, 2024, 102870, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2024.102870.
16. Dolezal JM, Wolk R, Hieromnimon HM, Howard FM, Srisuwananukorn A, Karpeyev D, Ramesh S, Kochanny S, Kwon JW, Agni M, Simon RC, Desai C, Kherallah R, Nguyen TD, Schulte JJ, Cole K, Khramtsova G, Garassino MC, Husain AN, Li H, Grossman R, Cipriani NA, Pearson AT. **Deep learning generates synthetic cancer histology for explainability and education.**

NPJ Precis Oncol. 2023 May 29;7(1):49. doi: 10.1038/s41698-023-00399-4. PMID: 37248379; PMCID: PMC10227067.

17. Rong R, Wang S, Zhang X, Wen Z, Cheng X, Jia L, Yang DM, Xie Y, Zhan X, Xiao G. **Enhanced Pathology Image Quality with Restore-Generative Adversarial Network.** Am J Pathol. 2023 Apr;193(4):404-416. doi: 10.1016/j.ajpath.2022.12.011. Epub 2023 Jan 18. PMID: 36669682; PMCID: PMC10123520.
18. Baxi, V., Edwards, R., Montalto, M. *et al.* **Digital pathology and artificial intelligence in translational medicine and clinical practice**. *Mod Pathol* **35**, 23–32 (2022). https://doi.org/10.1038/s41379-021-00919-2
19. L. Hou, A. Agarwal, D. Samaras, T. M. Kurc, R. R. Gupta and J. H. Saltz, **Robust Histopathology Image Analysis: To Label or to Synthesize?**," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 8525-8534, doi: 10.1109/CVPR.2019.00873.
20. Kim I, Kang K, Song Y, Kim T-J. **Application of Artificial Intelligence in Pathology: Trends and Challenges.** *Diagnostics*. 2022; 12(11):2794. https://doi.org/10.3390/diagnostics12112794
21. Zhang, Z., Chen, X., Tang, R. *et al.* **Interpretable unsupervised learning enables accurate clustering with high-throughput imaging flow cytometry**. *Sci Rep* **13**, 20533 (2023). https://doi.org/10.1038/s41598-023-46782-w
22. Montavon, G., Kauffmann, J., Samek, W., Müller, KR. (2022). **Explaining the Predictions of Unsupervised Learning Models.** In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (eds) xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science(), vol 13200. Springer, Cham. https://doi.org/10.1007/978-3-031-04083-2_7
23. Palacio Niño, Julio. (2019). **Evaluation Metrics for Unsupervised Learning Algorithms**. Preprint. Accessed via https://www.researchgate.net/publication/333102602_Evaluation_Metrics_for_Unsupervised_Learning_Algorithms on May, 4th, 2024
24. Hanna MG, Reuter VE, Samboy J, England C, Corsale L, Fine SW, Agaram NP, Stamelos E, Yagi Y, Hameed M, Klimstra DS, Sirintrapun SJ. **Implementation of Digital Pathology Offers Clinical and Operational Increase in Efficiency and Cost Savings.** Arch Pathol Lab Med. 2019 Dec;143(12):1545-1555. doi: 10.5858/arpa.2018-0514-OA. Epub 2019 Jun 11. PMID: 31173528; PMCID: PMC7448534.
25. Mohammad Hossein Jarrahi, Vahid Davoudi, Mohammad Haeri. **The key to an effective AI-powered digital pathology: Establishing a symbiotic workflow between pathologists and machine.** Journal of Pathology Informatics, Volume 13, 2022, 100156, ISSN 2153-3539, https://doi.org/10.1016/j.jpi.2022.100156.
26. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, Brat DJ, Cooper LAD. **Predicting cancer outcomes from histology and genomics using convolutional networks.** Proc Natl Acad Sci U S A. 2018 Mar 27;115(13):E2970-E2979. doi: 10.1073/pnas.1717139115. Epub 2018 Mar 12. PMID: 29531073; PMCID: PMC5879673. doi: 10.1073/pnas.1717139115
27. Lujan GM, et al. **Digital pathology initiatives and experience of a large academic institution during the coronavirus disease 2019 (COVID-19) pandemic**. Volume 145. Archives of Pathology & Laboratory Medicine; 2021. pp. 1051–61. 9.
28. Colling R, Pitman H, Oien K, Rajpoot N, Macklin P, Group CM-PAiHW, Snead D, Sackville T, Verrill C: **Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice.**J Pathol 2019. pp. 143-50 10.1002/path.5310 https://www.ncbi.nlm.nih.gov/pubmed/31144302https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/path.5310?download=true
29. Goldsmith JD, Troxell ML, Roy-Chowdhuri S, et al. **Principles of analytic validation of immunohistochemical assays: guideline update.** *Arch Pathol Lab Med.* Published online February 23, 2024. doi: 10.5858/arpa.2023-0483-CP
30. Abels E,Pantanowitz L. **Current state of the regulatory trajectory for whole slide imaging devices in the USA.** J Pathol Inform 2017; 8:23.

31. Keisuke Nakagawa, Lama Moukheiber, Leo A. Celi, Malhar Patel, Faisal Mahmood, Dibson Gondim, Michael Hogarth, Richard Levenson. **AI in Pathology: What could possibly go wrong?**, Seminars in Diagnostic Pathology, Volume 40, Issue 2, 2023, Pages 100-108, ISSN 0740-2570, https://doi.org/10.1053/j.semdp.2023.02.006.
32. Park SH, Choi J, Byeon JS. **Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence.** Korean J Radiol. 2021 Mar;22(3):442-453.
33. van der Laak, J., Litjens, G. & Ciompi, F. **Deep learning in histopathology: the path to the clinic**. *Nat Med* **27**, 775–784 (2021). https://doi.org/10.1038/s41591-021-01343-4
34. M. Cord and P. Cunningham, eds**., Machine learning techniques for multimedia: case studies on organization and retrieval ; with 20 tables.** Cognitive technologies, Springer, 2008. OCLC: 244009065.
35. Goldsack, Jennifer & Coravos, Andrea & Bakker, Jessie & Bent, Brinnae & Dowling, Ariel & Fitzer-Attas, Cheryl & Godfrey, Alan & Godino, Job & Gujar, Ninad & Izmailova, Elena & Manta, Christina & Peterson, Barry & Vandendriessche, Benjamin & Wood, William & Wang, Ke & Dunn, Jessilyn. (2020). **Verification, Analytical Validation, and Clinical Validation (V3): The Foundation of Determining Fit-for-Purpose for Biometric Monitoring Technologies (BioMeTs)**. 10.1038/s41746-020-0260-4.
36. Dudgeon SN, Wen S, Hanna MG, Gupta R, Amgad M, Sheth M, Marble H, Huang R, Herrmann MD, Szu CH, Tong D, Werness B, Szu E, Larsimont D, Madabhushi A, Hytopoulos E, Chen W, Singh R, Hart SN, Sharma A, Saltz J, Salgado R, Gallas BD. **A Pathologist-Annotated Dataset for Validating Artificial Intelligence: A Project Description and Pilot Study**. J Pathol Inform. 2021 Nov 15;12:45. doi: 10.4103/jpi.jpi_83_20. PMID: 34881099; PMCID: PMC8609287.
37. Lionel C. Briand, Jürgen Wüst. **Empirical Studies of Quality Models in Object-Oriented Systems**. Advances in Computers, Elsevier, Volume 56, 2002, Pages 97-166, ISSN 0065-2458, ISBN 9780120121564, https://doi.org/10.1016/S0065-2458(02)80005-5.
38. Vabalas A, Gowen E, Poliakoff E, Casson AJ. **Machine learning algorithm validation with a limited sample size**. PLoS One. 2019 Nov 7;14(11):e0224365. doi: 10.1371/journal.pone.0224365. PMID: 31697686; PMCID: PMC6837442.
39. Zhou Y, Koyuncu C, Lu C, Grobholz R, Katz I, Madabhushi A, Janowczyk A: **Multi-site cross-organ calibrated deep learning (MuSClD): Automated diagnosis of non-melanoma skin cancer.**Medical image analysis 2023. p. 102702 doi:10.1016/j.media.2022.102702 https://www.ncbi.nlm.nih.gov/pubmed/36516556
40. Varma S, Simon R. **Bias in error estimation when using cross-validation for model selection.** BMC Bioinformatics. 2006 Feb 23;7:91. doi: 10.1186/1471-2105-7-91. PMID: 16504092; PMCID: PMC1397873.
41. Kaliappan J, Bagepalli AR, Almal S, Mishra R, Hu YC, Srinivasan K. **Impact of Cross-Validation on Machine Learning Models for Early Detection of Intrauterine Fetal Demise.** Diagnostics (Basel). 2023 May 10;13(10):1692. doi: 10.3390/diagnostics13101692. PMID: 37238178; PMCID: PMC10217387.
42. Magnusson M., Vehtari A., Jonasson J., Andersen M. **Leave-one-out cross-validation for Bayesian model comparison in large data;** Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020; Palermo, Italy. 26–28 August 2020; pp. 341–351.
43. Jacques Wainer, Gavin Cawley. **Nested cross-validation when selecting classifiers is overzealous for most practical applications**. Expert Systems with Applications, Volume 182, 2021, 115222, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.115222.
44. Mariusz Rafało. **Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis**, ICT Express, Volume 8, Issue 2, 2022, Pages 183-188, ISSN 2405-9595, https://doi.org/10.1016/j.icte.2021.05.001.
45. Topuz, Kazim & Davazdahemami, Behrooz & Delen, Dursun. (2023). **A Bayesian belief network-based analytics methodology for early-stage risk detection of novel diseases**. Annals of Operations Research. 10.1007/s10479-023-05377-4.

46. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, Parasa S. **On evaluation metrics for medical applications of artificial intelligence.** Sci Rep. 2022 Apr 8;12(1):5979. doi: 10.1038/s41598-022-09954-8. PMID: 35395867; PMCID: PMC8993826.
47. Sun, Xiaoping & Ren, Xingshuai & Zhang, Jie & Nie, Yunzhi & Hu, Shan & Yang, Xiao & Jiang, Shoufeng. (2022). **Discovering miRNAs Associated With Multiple Sclerosis Based on Network Representation Learning and Deep Learning Methods.** Frontiers in Genetics. 13. 899340. 10.3389/fgene.2022.899340.
48. Manning, C. D., Raghavan, P., & Schütze, H. (2008). **Introduction to information retrieval.** Cambridge University Press.
49. Chicco D, Jurman G. **The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification**. BioData Min. 2023 Feb 17;16(1):4. doi: 10.1186/s13040-023-00322-4. PMID: 36800973; PMCID: PMC9938573.
50. Feng, M., Deng, Y., Yang, L. *et al.* **Automated quantitative analysis of Ki-67 staining and HE images recognition and registration based on whole tissue sections in breast carcinoma.** Diagn Pathol 15, 65 (2020). https://doi.org/10.1186/s13000-020-00957-5
51. van den Berg Eunice J.; Duarte Raquel; Dickens Caroline; Joffe Maureen; Mohanlal Reena. **Ki67 Immunohistochemistry Quantification in Breast Carcinoma: A Comparison of Visual Estimation, Counting, and ImmunoRatio.** Applied Immunohistochemistry & Molecular Morphology 29(2):p 105-111, February 2021. | DOI: 10.1097/PAI.0000000000000864
52. Zehra T, Shams M, Ahmad Z, Chundriger Q, Ahmed A, Jaffar N. **Ki-67 Quantification in Breast Cancer by Digital Imaging AI Software and its Concordance with Manual Method.** J Coll Physicians Surg Pak. 2023 May;33(5):544-547. doi: 10.29271/jcpsp.2023.05.544. PMID: 37190690.
53. Ng FYC, Thirunavukarasu AJ, Cheng H, Tan TF, Gutierrez L, Lan Y, Ong JCL, Chong YS, Ngiam KY, Ho D, Wong TY, Kwek K, Doshi-Velez F, Lucey C, Coffman T, Ting DSW. **Artificial intelligence education: An evidence-based medicine approach for consumers, translators, and developers.** Cell Rep Med. 2023 Oct 17;4(10):101230. doi: 10.1016/j.xcrm.2023.101230. PMID: 37852174; PMCID: PMC10591047.
54. Wartman SA, Combs CD. **Reimagining Medical Education in the Age of AI.** AMA J Ethics. 2019 Feb 1;21(2):E146-152. doi: 10.1001/amajethics.2019.146. PMID: 30794124.
55. N. Tsip, J. Napolska, M. Lyzogub. **Primary prevention of cervical cancer: factors of success.** Int J Gynecol Cancer. 2017; 27, Suppl.4:2034
56. Tsip N, Lyzogub M, Zaviryukha V. EP1143 **Psychological portrait of women after hydatidiform mole evacuation.** International Journal of Gynecologic Cancer 2019;29:A594.
57. N Tsip, N Khranovska, O Skachkova, M Inomistova, M Lyzogub and O Kolesnik. **EP1005 Spectrum and frequency of hereditary BRCA 1/2 mutations in ovarian cancer patients in Ukraine.** International Journal of Gynecologic Cancer 2019;29:A530.
58. Bobiński M, Hoptyana O, Rasoul-Pelińska K, Lyzogub M, Rychlik A and Pletnev A. **War in Ukraine: the opportunities for oncogynecologic patients in Poland.** International Journal of Gynecologic Cancer Published Online First: 13 April 2022. doi: 10.1136/ijgc-2022-003604
59. Margaryta Lyzogub, Yugrinov Oleg, Kondratiuk Vadym, Vakulenko Galyna, Nataliya Tsip. **1248 Experience of using uterine artery embolization in the treatment of gestational trophoblastic neoplasia accompanied by uterine arteriovenous malformations.** International Journal of Gynecologic Cancer Mar 2024, 34 (Suppl 1) A523-A524; DOI: 10.1136/ijgc-2024-ESGO.1025

WARRANTY

Vilniaus universiteto studijuojančiojo, teikiančio baigiamąjį darbą, GARANTIJA

Vardas, pavardė:Margaryta Lyzogub
Padalinys: Medicinos fakultetas
Studijų programa:Medicina
Darbo pavadinimas: Apžvalga klinikinio patvirtinimo procesų dirbtinio intelekto taikymams patologijoje
Darbo tipas: Naratyvinė literatūros apžvalga

Garantuoju, kad mano baigiamasis darbas yra parengtas sąžiningai ir savarankiškai, kitų asmenų indėlio į parengtą darbą nėra. Jokių neteisėtų mokėjimų už šį darbą niekam nesu mokėjęs.

Šiame darbe tiesiogiai ar netiesiogiai panaudotos kitų šaltinių citatos yra pažymėtos literatūros nuorodose.

WARRANTY of Vilnius University Student Thesis

Name, Surname: Margaryta Lyzogub
Faculty: Faculty of Medicine
Study programme: Medicine
Thesis topic: Overview of Clinical Validation Pocesses for Artificial Intelligence Applications in Pathology
Thesis type: Narrative literature review

I guarantee that my thesis is prepared in good faith and independently, there is no contribution to this work from other individuals. I have not made any illegal payments related to this work. Quotes from other sources used in this thesis, directly or indirectly, are indicated in literature references.

Aš, Margaryta Lyzogub, patvirtinu (pažymėti) ☑
I, Margaryta Lyzogub, confirm (check)

Patvirtinu, kad baigiamasis darbas yra pateiktas į Vilniaus universiteto studijų informacinę sistemą.
I declare that this thesis is submitted to the Vilnius University Study Information System.

Margaryta Lyzogub		***10/05/2024***
(vardas, pavardė / *name, surname*)	(parašas / *signature*)	(data / *date*)