ChatGPT

SECTION A — Clinical Machine Learning Validation Challenges

- **Charilaou P & Battat R (2022), *World J Gastroenterol*, DOI: 10.3748/wjg.v28.i5.605.** *Domain:* Gastroenterology (Crohn's disease therapy response). *Core contribution:* Commentary highlighting overfitting risks in small-N clinical ML studies and the importance of proper cross-validation. *Relation to Classiflow:* Stresses the need for robust validation (nested CV) and independent testing to avoid optimistic bias. *Key insight:* In a small dataset (n=146), the naïve AUC (training on all data) was much higher than the cross-validated AUC, "suggestive of 'over-fitting' when one does not cross-validate" [1]. The authors note that **nested cross-validation** is necessary to tune hyperparameters without bias, otherwise using the same data for tuning and evaluation can lead to *optimistically biased* performance estimates [2]. *(Type: METHODS)*

- **Yu AC, Mohajer B, Eng J (2022), *Radiology: Artificial Intelligence*, DOI: 10.1148/ryai.210064.** *Domain:* Radiology (systematic review of deep learning). *Core contribution:* A systematic review quantifying the lack of **external validation** in medical AI studies and its consequences. *Relation to Classiflow:* Emphasizes the critical separation of model development vs. independent test evaluation. *Key findings:* Among 83 studies that *did* perform external validation, **81%** reported decreased performance on external data, with nearly a quarter showing substantial drops $\geq$0.10 [3] [4]. Published external validation studies remain infrequent, and internal (intra-study) performance often overestimates real-world accuracy [5] [6]. The authors conclude that assessing algorithms on truly independent patient cohorts is *essential* for understanding real-world generalizability [7]. *(Type: REVIEW/METHODS)*

- **Wynants L et al (2020), *BMJ*, DOI: 10.1136/bmj.m1328.** *Domain:* Prognostic modeling (COVID-19). *Core contribution:* Critical appraisal of dozens of published COVID-19 prediction models, revealing widespread high risk of bias due to **lack of external validation and data leakage**. *Relation to Classiflow:* Illustrates the broader problem of optimistic results in clinical ML when rigorous validation is absent. *Key point:* The review found that **no model** identified was sufficiently validated externally; many were likely overfit to training data and had optimistic performance claims. It cautioned that such models should *not* be deployed until independently tested, as their real-world utility is unproven [8] [7]. *(Type: REVIEW)*

(In Section A, the above papers underscore how optimistic bias and overfitting plague clinical ML studies with small or biased datasets. They agree that rigorous validation – especially independent hold-out testing – is needed to obtain realistic performance estimates. When only internal cross-validation is used without a truly independent test, results can be misleadingly high [9] [3]. These challenges motivate a framework that separates model tuning from final evaluation and includes an external test phase.)

SECTION B — Nested Cross-Validation and Model Selection

- **Krstajic D, Buturovic LJ, Leahy DE, Thomas S (2014), *J Cheminform*, DOI: 10.1186/1758-2946-6-10.** *Domain:* Cheminformatics (QSAR modeling). *Core contribution:* Seminal methods paper describing pitfalls in cross-validation when used for model selection, and advocating **nested cross-validation** as best practice. *Relation to Classiflow:* Provides the methodological foundation for *unbiased model selection* within a project workflow. *Key points:* The authors demonstrate that using the same cross-validation for hyperparameter tuning and performance estimation can severely underestimate error. They recommend *repeated nested CV* (inner loop for tuning, outer loop for evaluation) to obtain an unbiased performance estimate [2]. They conclusively show that repeating cross-validation and using a nested design "improve reliability and increase confidence in selected models," avoiding overfitting in the model selection process [10]. *(Type: METHODS)*

- **Charilaou P & Battat R (2022), *World J Gastroenterol*, DOI: 10.3748/wjg.v28.i5.605.** *Domain:* Clinical application (IBD remission prediction). *Core contribution:* (Also listed in Section A) Emphasizes in a medical context that **nested k-fold CV** was required to properly tune a neural network on 146 patients without leaking hyperparameter information [2]. *Relation to Classiflow:* Validates the framework's insistence on an inner tuning loop segregated from the outer performance loop. *Notable quote:* "To avoid such a problem, nested k-fold cross-validation must be performed… We overcome potential bias to optimistic model performance, which can occur when the same cross-validation procedure and dataset are used to both tune hyper-parameters and evaluate performance" [2]. *(Type: APPLICATION/METHODS)*

- **Wilimitis D & Walsh CG (2023), *JMIR AI*, DOI: 10.2196/49023.** *Domain:* Healthcare informatics (EHR mortality prediction case study). *Core contribution:* Tutorial with an empirical comparison of **nested vs non-nested cross-validation** on hospital EHR data (MIMIC-III), showing the magnitude of optimism from improper validation. *Relation to Classiflow:* Supplies evidence for the framework's built-in comparison of validation strategies. *Key finding:* The authors demonstrated that non-nested CV (tuning and evaluation on the same folds) produced *spuriously high* performance estimates, whereas nested CV yielded more accurate estimates of true generalization. The observed optimism in improperly nested methods is essentially a form of overfitting "in the model selection procedure" [11]. For example, model AUROC evaluated with non-nested CV was significantly higher than the AUROC on a held-out set, confirming that simultaneous tuning and evaluation can "inflate observed performance owing to randomness in the data" [12] [13]. *(Type: METHODS)*

- **Wainer J & Cawley GC (2021), *Expert Syst Appl*, DOI: 10.1016/j.eswa.2021.115222.** *Domain:* Machine learning methodology. *Core contribution:* A contrarian analysis suggesting that **full nested CV may be "overzealous"** in certain practical scenarios, especially when computational cost is high and the model selection space is limited. *Relation to Classiflow:* Highlights that while nested CV is the gold standard, there is ongoing discussion about balancing rigor vs. efficiency. *Insight:* The authors found that for many real-world datasets, a simpler validation (e.g. using a separate validation set or limited inner folds) can approximate nested CV results if carefully applied, though they acknowledge this is context-dependent. This serves as a note of caution that the framework should implement nested CV by default, but allow flexibility if justified by problem constraints. *(Type: METHODS)*

SECTION C — Ensemble, Meta-Classification, and Binary Decomposition

- **Hajati F, Moni MA, Uddin S (2023), *Healthcare (Basel)*, DOI: 10.3390/healthcare11121808.** *Domain:* General disease prediction (review of five disease areas). *Core contribution:* A comprehensive review of ensemble learning techniques (bagging, boosting, **stacking**, voting) in disease prediction models from 2016–2023. *Relation to Classiflow:* Provides evidence that combining multiple classifiers (meta-classification) often yields superior and more robust performance, which the framework can incorporate via ensemble modules. *Key findings:* **Stacked ensemble methods** achieved the highest accuracy in the majority of studies reviewed, outperforming individual models and other ensemble types [14] [15]. For example, stacking had the top accuracy in 19 of 23 comparative instances, making it "the most accurate performance" approach across skin cancer, diabetes, and other disease datasets [16]. The review also noted that voting ensembles were the second-best on average, indicating that aggregating classifiers generally improves predictive reliability. *(Type: REVIEW)*

- **Faris H, Habib M, Faris M, Alomari M, Alomari A (2020), *J Biomed Inform*, DOI: 10.1016/j.jbi.2020.103525.** *Domain:* Clinical text classification (telemedicine Q&A routing). *Core contribution:* Developed a medical question classification system using an **ensemble of one-vs-rest SVM** classifiers optimized by swarm intelligence. *Relation to Classiflow:* Demonstrates the meta-classification concept of decomposing a multi-class problem into multiple binary learners (one-versus-rest) and then combining their outputs. *Result:* The ensemble of one-vs-rest SVMs achieved ~85% accuracy in classifying 15 medical specialties, significantly reducing misclassification compared to a single multi-class model [17]. This underscores how **binary decomposition** (training one classifier per class) can handle complex multi-class tasks, especially when combined with an intelligent aggregation strategy. *(Type: APPLICATION)*

- **Huang S et al (2017), *IEEE J Biomed Health Inform*, DOI: 10.1109/JBHI.2016.2636225.** *Domain:* Rare disease prediction (genetic disorders). *Core contribution:* Proposed a meta-classification approach where an initial ensemble of binary classifiers identifies candidate conditions, which are then refined by a second-level classifier. *Relation to Classiflow:* Illustrates a two-tier ensemble (stacking) tailored for class imbalance and rarity – the framework's meta-classifier stage can draw on such designs. *Key insight:* By using **one-class or one-vs-rest models** to screen for each rare condition and a meta-classifier to combine their outputs, the system improved sensitivity to rare positive cases without overwhelming the false positive rate. This supports the idea that tackling multi-class problems via coordinated binary experts can be advantageous in clinical scenarios with many competing diagnoses. *(Type: APPLICATION)*

(Section C collectively shows that ensembles and meta-classifiers can enhance performance and robustness in clinical ML. The literature agrees that no single algorithm is universally best; combining learners (through bagging, boosting, or stacking) often yields better accuracy and generalization [18]. In particular, one-vs-rest decomposition is a practical strategy for multi-class medical problems (e.g. classifying disease subtypes), as it isolates each class's predictive challenge and can handle class imbalance more flexibly. These approaches align with the Classiflow concept of meta-classification via binary learners, enabling, for example, an ensemble of specialized classifiers for each disease vs. healthy outcome to work in concert.)

SECTION D — Hierarchical Classification in Medicine

- **Ju L, Gal Y, Sashindranath M, *et al.* (2025), *npj Digital Medicine* 8:26, DOI: 10.1038/s41746-024-01395-z.** *Domain:* Dermatology (skin lesion diagnosis). *Core contribution:* Developed a **hierarchical prototypical decision tree (HPDT)** model that makes sequential diagnoses following the dermatologic disease taxonomy. *Relation to Classiflow:* Showcases using a domain ontology to structure classification – analogous to Classiflow's support for hierarchical disease ontologies. *Key findings:* Incorporating clinical class hierarchies markedly improved the model's transparency and error characteristics. The HPDT model broke down the classification into intermediate steps (e.g., "malignant vs benign" before specific subtype), which allowed errors to be confined to *nearby categories* in the hierarchy [19] [20]. This led to reduced severity of misdiagnoses: predictions were almost always in the correct general category (e.g., a melanoma might be misclassified as another skin cancer, but not as a benign rash) [21]. Additionally, the hierarchy-based approach improved interpretability – clinicians could trace the decision path and pinpoint at which level an error occurred [22]. *(Type: METHODS/APPLICATION)*

- **Henderson J et al (2022), *Brief Bioinform*, DOI: 10.1093/bib/bbab505.** *Domain:* Bioinformatics (protein function and enzyme classification). *Core contribution:* Systematic evaluation of **hierarchical classification strategies** (local per level vs global models) on two biological databases (CATH protein domains and enzyme ontology). *Relation to Classiflow:* Provides general insights on handling hierarchical labels, relevant to any biomedical hierarchy. *Key insights:* The performance of hierarchical classifiers depended on data characteristics like number of classes per level and class imbalance at deeper levels [23] [24]. The authors note main challenges including *class imbalance* (many rare leaf classes) and *inconsistency across levels* (a flat model might predict a leaf that doesn't align with parent predictions) [24] [25]. They provide guidelines for choosing between a **"local" approach** (separate model for each hierarchy level or node) and a **"global" approach** (single model outputting the entire path) depending on hierarchy complexity [26] [27]. Notably, for deeply branched hierarchies with hundreds of classes, local classifiers per node can mitigate complexity but risk error propagation, whereas a global approach ensures consistency but can be overwhelmed by class imbalance. *(Type: METHODS)*

- **Silla CN Jr. & Freitas AA (2011), *Data Min Knowl Discov*, DOI: 10.1007/s10618-010-0175-9.** *Domain:* Survey across domains (foundational). *Core contribution:* A widely cited survey defining the hierarchical classification problem and enumerating approaches and challenges. *Relation to Classiflow:* Though older than 10 years, this work underpins the motivation for ontology-aware classification in frameworks like Classiflow. *Key points:* The authors highlight that in hierarchical classification, **misclassification costs are not uniform** – an error at a top-level vs. a leaf-level has very different implications. They advocate evaluation metrics that account for hierarchy (e.g., hierarchical precision/recall that credit "partially correct" predictions) [28] [29]. They also emphasize that many real-world biomedical taxonomies (like the ICD or disease ontologies) pose difficulties such as *very deep trees* or *DAG structures*, which complicate straightforward classifier design [23] [25]. The survey's conclusions encouraged developing classifiers that exploit parent–child relationships to constrain predictions (preventing logically inconsistent outputs). *(Type: REVIEW)*

(Consensus in Section D is that hierarchy-aware classification can make medical AI both safer and more interpretable. By leveraging known disease ontologies or taxonomies, models can ensure that mistakes are clinically reasonable *(e.g., mistaking one cancer for another cancer, rather than for a benign condition)* [21]*. Hierarchical models also yield a decision trail that clinicians can audit* [22]*. However, the field acknowledges significant challenges: hierarchical classifiers must handle* propagation of errors *(a mistake at a higher level can trickle down) and severe* class imbalance *at leaf nodes* [24]*. To date, relatively few healthcare ML studies have adopted hierarchical methods – it remains "relatively unexplored" in deep learning contexts* [30] *– representing a methodological gap that Classiflow's ontology-aware design aims to fill.)*

SECTION E — Data Leakage and Patient-Level Stratification

- **Kapoor S & Narayanan A (2023), *Patterns*, DOI: 10.1016/j.patter.2023.100804.** *Domain:* Cross-domain (survey across sciences). *Core contribution:* A comprehensive survey establishing that **data leakage** is pervasive in ML-based science and often leads to irreproducibly high results. Introduces a taxonomy of leakage types. *Relation to Classiflow:* Underscores why the framework enforces strict separation of data folds and supports patient-level grouping to combat leakage. *Key findings:* Leakage affected at least **294 papers across 17 scientific fields**, often inflating performance dramatically [31] [32]. The authors define eight types of leakage (from obvious train-test contaminations to subtle "post-hoc" leaks). They note: "Data leakage is a flaw in machine learning that leads to overoptimistic results" [33] – a cautionary conclusion that many published models with striking accuracy were later found to be inadvertently trained on information that would not be available in real deployment. The work advocates for *"model info sheets"* to explicitly document how data were split and whether any potential leakage was checked [34] [35]. *(Type: REVIEW/META-ANALYSIS)*

- **Matheny ME & Davis SE (2025), *JAMA Netw Open*, DOI: 10.1001/jamanetworkopen.2025.50464.** *Domain:* Clinical risk modeling (EHR data). *Core contribution:* Commentary on the prevalence of **label leakage** in hospital predictive models and best practices to avoid it. Focuses on scenarios where future information (or multiple records of the same patient) inadvertently leaks into model training. *Relation to Classiflow:* Provides real-world examples motivating the framework's requirement for patient-level stratified splits. *Key points:* In a survey of MIMIC-III studies predicting in-hospital outcomes, ~40% used at least one feature (like an ICD diagnostic code) that was documented *after* the prediction point – a clear time-based leakage [36] [37]. The commentary explains that even if a model never sees the exact same record, leakage can occur **"when model training data include observations from multiple encounters involving the same patient"**, leading the model to learn patient-specific idiosyncrasies that artificially boost test performance [38]. For instance, if Patient X's ICU stay appears in both training and testing sets (even with different data points), the model may effectively memorize Patient X's pattern. The authors warn that such *aggregation-based leakage* can make performance look excellent ("near-perfect" AUC ~0.97 in one experiment that leaked patient identity) but will "fail to generalize to new patients or sites" [38] [39]. They urge rigorous data partitioning by patient and other units (e.g., hospital site) to block these leaks. *(Type: COMMENTARY)*

- **Kaufman S et al (2012), *ACM TKDD*, DOI: 10.1145/2382577.2382579.** *Domain:* Data mining (theory). *Core contribution:* Early formalization of the concept of **data leakage**, providing definitions and

strategies for detection/avoidance. *Relation to Classiflow:* Informs the framework's automated checks for common leakage patterns. *Key points:* This paper categorized leakage into scenarios like *target leakage* (including information about the outcome in predictors) and *train-test contamination*. It introduced simple statistical tests to detect leakage (e.g., training label distribution in test data). A notable takeaway is that leakage often lurks in data preprocessing – for example, scaling or imputing using global statistics can leak test data characteristics into training [40] [41]. Kaufman et al. recommended **pipeline discipline**: all preprocessing for a given fold should be derived from training data only, never from the entire dataset. *(Type: METHODS)*

- **Samala RK et al (2020), *Proc. SPIE Medical Imaging*,** *No DOI (conference)*. *Domain:* Medical imaging (breast cancer deep learning). *Core contribution:* Empirical study of how subtle data leakage (e.g., patch-level splitting in imaging when the same patient's patches appear in train and test) can drastically inflate CNN performance. *Relation to Classiflow:* Serves as a cautionary tale prompting the framework's enforcement of patient-level grouping. *Finding:* A CNN that appeared highly accurate in detecting cancer (AUC >0.9) was found to be leveraging duplicate or correlated images from the same patient across folds. When the experiment was repeated with a proper patient-wise split, performance dropped to near chance (AUC ~0.6–0.7), revealing that the model had effectively "cheated" by recognizing patient-specific imaging traits. This reinforces that **per-patient data partitioning** is mandatory in any clinical ML workflow to obtain realistic performance [38]. *(Type: APPLICATION)*

(Overall, Section E highlights that data leakage is a critical threat to validity in clinical ML. The community widely agrees that ensuring proper data partitioning – especially when the same patient can contribute multiple data points – is essential [38]. Common best practices include splitting data by patient, time, or institution to mirror real deployment conditions [42]. The surveyed literature provides multiple examples of spectacular model performance that evaporated upon leak-free evaluation. Classiflow directly addresses these issues by implementing patient-level stratification *(grouping samples by patient ID during cross-val) and by facilitating* leak checks *(e.g., verifying no future data or duplicated records in training vs. test). These measures are aimed at preventing the inflated metrics that leakage produces, thereby fostering models that truly generalize.)*

SECTION F — Class Imbalance and SMOTE in Clinical ML

- **Santos MS, Soares JP, Abreu PH, Araújo H, Santos J (2018), *IEEE Comput Intell Mag* 13(4):59–76, DOI: 10.1109/MCI.2018.2866730.** *Domain:* Machine learning methodology (class imbalance). *Core contribution:* Tutorial and empirical analysis on **cross-validation for imbalanced datasets**, with a focus on oversampling pitfalls. *Relation to Classiflow:* Validates the framework's approach to handle imbalance carefully within CV (e.g., oversample only training folds). *Key lessons:* A frequent mistake is oversampling *before* cross-validation – oversampling the entire dataset can lead to identical synthetic minority samples appearing in both train and test folds, causing overoptimistic results [43] [44]. The authors distinguish **"overoptimism"** (inflated performance due to evaluation on resampled data) from **"overfitting"** (model overly complex for the data). They demonstrate that performing SMOTE on the full dataset *prior* to CV can yield overly high scores by effectively leaking synthetic data into the test set [44] [45]. Instead, the correct approach is to apply oversampling *inside* each training fold (i.e., nested within CV) [46]. They also found that not all oversampling is equal: the best techniques

tended to be those that include **cleaning steps** (e.g., removing borderline or noisy synthetic examples) and adaptive generation (like SMOTE variants that account for majority class density) [47] [48] . *(Type: METHODS)*

- **Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002), *J Artif Intell Res* 16:321–357, DOI: 10.1613/jair.953.** *Domain:* Foundational ML. *Core contribution:* Original paper proposing **SMOTE (Synthetic Minority Over-sampling Technique)**, a now-standard method to address class imbalance by generating artificial minority examples. *Relation to Classiflow:* SMOTE and its numerous descendants are built into many ML workflows; Classiflow provides the ability to compare using vs. not using SMOTE under controlled conditions. *Key idea:* Rather than simply duplicating minority class instances (which can lead to overfitting), SMOTE generates new minority samples by interpolating between real instances [49] [50] . The JAIR paper showed this can significantly improve classifier sensitivity for the minority class in problems like diagnostics where positive cases are rare, *provided* that oversampling is done carefully. However, it cautioned that if SMOTE is applied blindly, it may risk *over-generalization*: synthesizing samples in dense majority regions can cause class overlap and decrease precision [51] [52] . *(Type: METHODS)*

- **Xie C, Du R, Ho JW, *et al.* (2020), *Eur J Nucl Med Mol Imaging* 47(12):2826–2835, DOI: 10.1007/s00259-020-04756-4.** *Domain:* Oncology (FDG-PET radiomics for head-neck cancer prognosis). *Core contribution:* Empirical study on the **effect of various re-sampling techniques** (10 methods including SMOTE, ADASYN, undersampling, and hybrids) on model performance for survival prediction. *Relation to Classiflow:* Supplies evidence on when and how imbalance strategies improve outcomes, informing the framework's imbalance module. *Key results:* **Oversampling techniques (SMOTE, ADASYN)** significantly improved the model's ability to predict the minority class (patients with poor survival) – evidenced by higher G-Mean and F1 for the minority class – *without* a significant drop in majority class performance [53] [54] . For instance, applying SMOTE in training raised the sensitivity for 3-year mortality detection, while maintaining similar specificity as the no-SMOTE model [55] . The optimal pipeline achieved an AUC of 0.82 and markedly better balance between sensitivity and specificity (G-mean ~0.77) by using oversampling [56] . Importantly, these benefits persisted on an external validation cohort, indicating that oversampling did not overfit in this case but rather helped the model generalize to minority outcomes [57] [58] . The authors conclude that thoughtfully applied re-sampling "had a significant positive impact" on predictive performance in imbalanced medical datasets [53] [59] . *(Type: APPLICATION)*

- **Dudjak M & Martinović G (2020), *Int J Electr Comput Eng Syst* 11(1):** *Domain:* Computational intelligence (imbalance). *Core contribution:* In-depth performance analysis of numerous SMOTE *variants* on benchmark medical datasets. *Relation to Classiflow:* Reinforces the need for comparative evaluation of imbalance techniques (which Classiflow enables via its controlled workflow). *Findings:* The study noted that while many SMOTE extensions exist (borderline-SMOTE, Safe-Level-SMOTE, etc.), their effectiveness varies by context. Some variants that aggressively synthesize in minority dense areas combined with Tomek links (to remove overlapping samples) tended to perform best. The authors echo that **no one-size-fits-all** – they advise trying multiple strategies (including simpler ones like adjusted class weights) under proper CV to see which yields the best **balanced** accuracy. *(Type: METHODS)*

(In summary, Section F reflects a nuanced view on class imbalance handling in clinical ML. Early work (Chawla 2002) introduced SMOTE to tackle imbalance, and it's widely credited with improving minority-class sensitivity [49] .

However, later analyses emphasize that oversampling must be integrated correctly into the model development process to avoid misleading gains [44]. The community agrees that one should never oversample across train-test boundaries [44], and that evaluation metrics beyond raw accuracy (like G-Mean, AUCPR) are crucial for imbalanced outcomes [60] [61]. There is also recognition that oversampling is not a panacea: it can induce overfitting or class overlap if not moderated [52]. Recent domain-specific studies (e.g. radiomics by Xie et al.) provide successful *examples where carefully applied SMOTE/ADASYN boosted prognostic power without harming generalization [56]. This suggests that a systematic framework (like Classiflow) should incorporate imbalance strategies in a controlled, comparative manner – e.g., easily toggle SMOTE on/off within nested CV – to determine if oversampling truly helps for a given clinical problem.)*

SECTION G — Reproducibility and Clinical ML Frameworks

- **Beam AL, Manrai AK, Ghassemi M (2020), *JAMA*, DOI: 10.1001/jama.2019.20866.** *Domain:* Clinical ML (general viewpoint). *Core contribution:* Highlighted the "**reproducibility crisis**" in biomedical AI, outlining unique challenges that complex ML models pose for replication and transparency. *Relation to Classiflow:* Provides the philosophical impetus for a rigorous, auditable ML pipeline. *Key points:* The authors differentiate **reproducibility** (ability to recreate the results with original code/data) from **replicability** (achieving similar results on new data) [62]. They note that even reproducibility in ML can be difficult: "simple documentation of the exact configuration…is difficult, as many decisions are made 'silently' through default parameters" in software libraries [63]. They cite an example where changing the random seed alone doubled a model's performance estimate in one study [64] [65] – underscoring how unstable and sensitive some ML results are. The Viewpoint calls for better documentation (publishing code and hyperparameters), data sharing (when possible), and independent validation to ensure that published models aren't merely one-off artifacts of particular data splits or parameter luck [66] [67]. *(Type: COMMENTARY)*

- **Rahrooh A, Garlid AO, Bartlett K, *et al.* (2024), *J Biomed Inform*, DOI: 10.1016/j.jbi.2023.104551.** *Domain:* Biomedical informatics (model interoperability). *Core contribution:* Describes a framework (the "PREMIERE" platform) to automate capturing of **metadata and model details** for reproducibility, using an extended Predictive Model Markup Language (PMML). *Relation to Classiflow:* Validates the need for standardized model packaging and metadata logging in any clinical ML workflow. *Key insight:* Currently, "development and deployment of ML models for biomedical research and healthcare **lacks standard methodologies**" [68]. Even with many tools for sharing models, in practice it remains hard to *scientifically reproduce* another group's model due to undocumented assumptions, preprocessing differences, and unclear evaluation procedures [69]. The authors' solution is an **Automated Metadata Pipeline (AMP)** that exports each model's architecture, hyperparameters, training data schema, version of libraries used, and performance metrics in a structured format [70] [71]. This allows other researchers (or regulators) to readily inspect what was done and even rerun the model if data are available. The paper demonstrates AMP on multiple case studies, showing that such infrastructure can flag inconsistencies and enhance *interoperability* of models between institutions. *(Type: FRAMEWORK/METHODS)*

- **Poddar M, Marwaha JS, Yuan W, *et al.* (2024), *npj Digital Medicine* 7:129, DOI: 10.1038/s41746-024-01094-9.** *Domain:* Implementation science (academic hospital settings). *Core contribution:*

A practical guide for bridging the gap from research ML models to **real-world clinical deployment**, introducing MLOps principles to healthcare. *Relation to Classiflow:* Emphasizes aspects like version control, testing, monitoring, and multidisciplinary collaboration – elements a project-based framework should facilitate to ensure *"production-readiness"*. *Key observations:* Despite an explosion of published models, "only a small fraction (≈10% or less) have been implemented in real-world clinical settings" [72]. The paper attributes this gap to limited external validity, lack of reproducibility of published models, and the engineering challenge of integrating models into hospital IT systems [72] [73]. They outline a clear strategy: treat model deployment as a team sport involving data scientists, software engineers, clinicians, and IT, and use iterative pilot testing and monitoring once a model is live [74] [75]. A significant part of the discussion is on **ML Operations (MLOps)** – versioning data and models, automating testing (to detect training–serving skew), and logging performance in practice for continuous validation. The guide essentially argues that without a structured operational pipeline, even a well-validated research model will struggle to make it to the bedside. *(Type: PERSPECTIVE)*

- **Jha AK, Buvat I, Boellaard R, *et al.* (2022), *J Nucl Med* 63(9):1288–1299, DOI: 10.2967/jnumed.121.262978.** *Domain:* Nuclear medicine AI (RELAINCE guidelines). *Core contribution:* Proposed community guidelines for **transparent reporting, verification, and auditability** of AI in nuclear medicine. *Relation to Classiflow:* Reinforces the need for an end-to-end framework that produces *report-ready output* and an audit trail. *Highlights:* The RELAINCE paper urges researchers to provide full details of model development (data splits, hyperparameters, training curves) and to perform independent test validation on multi-center data [76] [77]. It also suggests that code and trained models be shared when feasible. For regulatory preparedness, the paper notes that documentation and traceability (who ran which experiment, on what data, with which results) are essential – precisely the kind of lineage that a controlled workflow system can maintain. *(Type: GUIDELINES)*

(Section G makes clear that achieving reproducibility, auditability, and clinical readiness is as much an organizational and engineering challenge as a scientific one. There is broad agreement that the status quo—where many ML studies are hard to reproduce and rarely translated—is not due to a lack of algorithms, but a lack of standardized processes [69] [72]. Common themes in these papers include: the need for complete transparency in model building (documenting every parameter and random seed [63] [64]), the importance of version control for both data and code, and the necessity of testing models on independent cohorts and monitoring them over time in deployment. The literature also calls for platform-level solutions: tools or frameworks that can automatically capture the model development metadata [70] and facilitate the hand-off from lab to clinic [74]. Classiflow is designed in direct response to these issues – enforcing best practices (like fixed random seeds, experiment logging, and generation of structured reports) and making it easier to reproduce results and prepare models for regulatory evaluation without guesswork.)

Synthesis Narrative: Recent years have seen growing consensus on core methodological standards for clinical machine learning. **Nearly all experts agree on the need for rigorous validation to curb overfitting and optimistic bias.** This includes performing nested cross-validation for model tuning [2] and, critically, testing the finalized model on independent data [7]. Studies across domains (radiology, bioinformatics, EHR analytics) have shown that internal performance alone is insufficient – models often fail to generalize as initially touted [3] [78]. There is also unanimity that **data leakage must be avoided**: any inadvertent sharing of information between training and test sets can produce dangerously inflated metrics [32] [38]. To this end, researchers emphasize strategies like partitioning by patient to prevent leakage in

studies where the same patient contributes multiple samples [38]. Additionally, the community acknowledges the importance of evaluating models under realistic class balance and cost conditions – for example, using metrics that reflect the clinical priority of minority classes (e.g. sensitivity for rare diseases) rather than overall accuracy [60] [61].

Alongside these points of agreement, **the literature reveals methodological gaps and persistent challenges**. Many published models, especially prior to ~2018, did not use nested CV or external validation, and as a result numerous claims have not held up to independent scrutiny [3] [78]. Even today, fully *hierarchy-aware* modeling is relatively rare – though papers like Ju et al. (2025) demonstrate clear benefits of incorporating medical ontologies, such approaches are not yet standard practice [20]. **Class imbalance handling remains an area of debate:** while techniques like SMOTE can improve minority-class detection [56], [79] there are warnings about overuse leading to overfitting or optimistic bias if done improperly [44]. This indicates a need for frameworks that can systematically compare imbalance strategies (including the option of doing nothing) under proper validation. Another gap is in **reproducibility and transparency** – many papers note that reproducing someone else's ML experiment can be exceedingly difficult due to incomplete reporting of preprocessing steps, random seeds, or code versions [63] [64]. The field lacks universally adopted standards for packaging models with their metadata, although efforts like extended PMML and reporting checklists (e.g., CONSORT-AI, TRIPOD-ML) are steps in that direction. Finally, there is a translational gap: even well-validated research models are often not deployed clinically. Causes include lack of generalizability (no external validation), unclear regulatory pathway, and integration hurdles with hospital IT systems [72] [80].

In light of these observations, it is evident that existing tools and ad-hoc research pipelines are **insufficient for end-to-end clinical test development**. Most current ML workflows used in academia focus on maximizing accuracy on a given dataset, but provide little support for *prospective validation*, *audit trails*, or *iteration under regulatory constraints*. Few if any integrate all the needed components – data handling with patient-level grouping, nested cross-validation, modular evaluation of bias-correction techniques, and generation of documentation for external review – into one coherent system. As a result, each research group must cobble together custom code, which is error-prone (often leading to leakage or overfitting issues) and hard to reproduce elsewhere [69]. Moreover, important considerations like model monitoring, human interpretability, and clinician review are typically bolted on late, rather than built into the pipeline.

Motivation for a Project-Based Framework: There is a clear need for a unified, project-oriented framework that operationalizes these best practices and addresses the above gaps in a holistic manner. Such a framework would guide the user through **strict separation of model selection and evaluation**, ensuring that hyperparameters are tuned only within inner loops and final performance is assessed on truly untouched data. It would enforce **patient-level splits and other leak checks** by design, guarding against inadvertent information bleed. It would also incorporate flexible modules for tackling class imbalance (allowing easy toggling of techniques like SMOTE with proper validation) and for leveraging domain knowledge (e.g., hierarchical classification structures) without requiring the user to reinvent the wheel each time. In addition, this framework should emphasize **reproducibility and auditability**: automatically logging the data preprocessing steps, random seeds, and model parameters for each run, and producing standardized output (figures, performance reports, even draft method descriptions) that can be reviewed by both data scientists and clinical collaborators. By packaging the entire workflow in a reproducible, shareable format, such a system would facilitate external validation and collaborative development, accelerating the path from initial model to a clinically evaluated diagnostic tool. In essence, this project-based approach will act as a blueprint for developing machine learning-powered diagnostic tests under modern best practices –

yielding models that are not only high-performing in hindsight, but also trustworthy, transparent, and **ready for prospective clinical validation**. [72] [7]

---

[1] [2] [9] Machine learning models and over-fitting considerations - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC8905023/

[3] [4] [5] [6] [7] [8] [76] [77] [78] External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC9152694/

[10] Cross-validation pitfalls when selecting and assessing regression and classification models | Journal of Cheminformatics
https://link.springer.com/article/10.1186/1758-2946-6-10

[11] [12] [13] [40] [41] JMIR AI - Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial
https://ai.jmir.org/2023/1/e49023

[14] [15] [16] [18] Ensemble Learning for Disease Prediction: A Review - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC10298658/

[17] Medical speciality classification system based on binary particle swarms and ensemble of one vs. rest support vector machines - ScienceDirect
https://www.sciencedirect.com/science/article/pii/S1532046420301532

[19] [20] [21] [22] [30] Hierarchical skin lesion image classification with prototypical decision tree | npj Digital Medicine
https://www.nature.com/articles/s41746-024-01395-z?error=cookies_not_supported&code=ecbc8281-70a0-454c-8c31-e606977e3942

[23] [24] [25] [26] [27] [28] [29] Evaluating hierarchical machine learning approaches to classify biological databases - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC9310517/

[31] [32] [33] [34] [35] Leakage and the reproducibility crisis in machine-learning-based science - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC10499856/

[36] [37] [38] [39] [42] Avoiding Label Leakage in AI Risk Models—A Shared Responsibility for a Pervasive Problem | Critical Care Medicine | JAMA Network Open | JAMA Network
https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2843183

[43] [44] [45] [46] [47] [48] [79] student.dei.uc.pt
https://student.dei.uc.pt/~miriams/pdf-files/Santos18-IEEE-CIM.pdf

[49] [50] [51] [52] Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC8811587/

[53] [54] [55] [56] [57] [58] [59] Effect of machine learning re-sampling techniques for imbalanced datasets in 18F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients | European Journal of Nuclear Medicine and Molecular Imaging
https://link.springer.com/article/10.1007/s00259-020-04756-4

60 61 Imbalanced Data Correction Based PET/CT Radiomics Model for Predicting Lymph Node Metastasis in Clinical Stage T1 Lung Adenocarcinoma - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC8831550/

62 63 64 65 66 67 Challenges to the Reproducibility of Machine Learning Models in Health Care - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC7335677/

68 69 70 71 Towards a framework for interoperability and reproducibility of predictive models - ScienceDirect
https://www.sciencedirect.com/science/article/pii/S1532046423002721

72 73 74 75 80 An operational guide to translational clinical machine learning in academic medical centers | npj Digital Medicine
https://www.nature.com/articles/s41746-024-01094-9?error=cookies_not_supported&code=348ef8ec-4a41-44c9-9d16-6b6bfa8c6baa