Article

Not peer-reviewed version

Prioritizing Data Quality Governance for AI in Prostate Cancer: A Methodological Proof-of-Concept Using Neural Networks for Risk Stratification

Vanessa Talavera-Cobo , Jose Enrique Robles-Garcia , Francisco Guillen-Grima * , Andres Calva-Lopez , Mario Tapia-Tapia , Luis Labairu-Huerta , Francisco Javier Ancizu-Marckert , Laura Guillen-Aguinaga , Daniel Sanchez-Zalabardo , Bernardino Miñana-Lopez

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1207.v1

Keywords: prostate cancer; artificial neural network; D'Amico risk stratification; multilayer perceptron; ISUP grade; Briganti nomogram; data quality governance; FAIR principles; AI-readiness; reproducibility; proof-of-concept

Article

Prioritizing Data Quality Governance for AI in Prostate Cancer: A Methodological Proof-of-Concept Using Neural Networks for Risk Stratification

Vanessa Talavera-Cobo [1], Jose Enrique Robles-Garcia [1], Francisco Guillen-Grima [2,3,4,5,*], Andres Calva-Lopez [1], Mario Tapia-Tapia [1], Luis Labairu-Huerta [1], Francisco Javier Ancizu-Marckert [1], Laura Guillen-Aguinaga [6], Daniel Sanchez-Zalabardo [1] and Bernardino Miñana-Lopez [1]

[1] Department of Urology, Clinica Universidad de Navarra, 31008 Pamplona, Spain.
[2] Department of Preventive Medicine, Clinica Universidad de Navarra, 31008 Pamplona, Spain
[3] Department of Health Sciences, Public University of Navarra; 31008 Pamplona, Spain
[4] Group of Clinical Epidemiology, Area of Epidemiology and Public Health, Healthcare Research Institute of Navarre (IdiSNA), 31008 Pamplona, Spain
[5] CIBER in Epidemiology and Public Health (CIBERESP), Institute of Health Carlos III, 46980 Madrid, Spain
[6] Department of Nursing, Clinica Universidad de Navarra, 28027 Madrid, Spain

* Correspondence: frguillen@unav.es

Abstract

Background: An accurate D'Amico risk stratification is mandatory for prostate cancer (PCa) management. The purpose of this proof-of-concept study was to establish a methodological framework of integrating validated clinical nomograms with strict data-quality governance in order to generate reliable artificial neural networks (ANN), even when the sample is small**. Methods**: We performed a retrospective analysis of a curated cohort of 49 patients from one center. A multilayer perceptron (MLP) was trained using 11 variables, including the ISUP biopsy grade and Briganti nomogram. Model development was guided by a proactive data-quality protocol based on FAIR principles, with stringent checks for accuracy, consistency and validity to ensure data were "AI-ready". A sensitivity analysis was conducted on three data partitioning scenarios (20/80, 34/66 and 39/61).**Results:** From a starting pool of 76 patients, the FAIR-based data governance architecture was applied to create a highly selected cohort of 49 patients. A multilayer perceptron (MLP) trained on this "AI-ready" dataset achieved a mathematically perfect but clinically uninterpretable discrimination (AUC 1.000) for High vs. Intermediate risk groups on a small internal test set (N=9 for the 20/80 split). However, this complete accuracy is a best-case scenario reflecting the high data quality, not proof of generalizable clinical utility, as the large confidence interval (66.4-100%) and the requirement to exclude instances with unusual attributes for model validation (as described in the methods) highlight. **Conclusions**: The main contribution of this proof-of-concept study is the effective illustration of a strict, repeatable data governance approach for producing "AI-ready" urological datasets. Although the MLP demonstrated a robust internal signal for risk discrimination, its flawless accuracy is an ideal, non-generalizable situation. The most important deliverable that needs external validation is the framework, not the model's performance metrics.

Keywords: prostate cancer; artificial neural network; D'Amico risk stratification; multilayer perceptron; ISUP grade; Briganti nomogram; data quality governance; FAIR principles; AI-readiness; reproducibility; proof-of-concept

 doi:10.20944/preprints202603.1207.v1

1. Introduction

As a major cause of cancer incidence and mortality in men worldwide, prostate cancer (PCa) continues to be a major global health concern [1–4] Widespread PSA screening and population aging are two factors contributing to its prevalence, especially in Western nations. However, this high frequency presents a clinical conundrum: the need for early detection must be carefully weighed against the possibility of overtreatment; this balance depends on precise risk classification [5,6]

Despite lower incidence rates, PCa is a growing issue in developing and low-to-middle-income countries, where the mortality-to-incidence ratio is almost five times higher than in high-income countries (0.95 vs. 0.24) because of low health awareness, a lack of early screening, and restricted access to gold-standard treatment [7,8]. In some areas, this leads to lower clinical outcomes and more advanced-stage diagnoses [9–12].

Highly precise diagnostic tools have become essential to increase treatment accuracy and resource allocation in a variety of clinical contexts as this global disparity widens [13–16]. To resolve the clinical management dilemma and avoid overtreatment in modern health facilities, accurate risk classification has been essential [17–22]. However, the inclusion of "AI-ready" data and strong quality governance for clinical trustworthiness have taken importance over the ability of these prediction models [13,23–25].

Established tools such as the D'Amico classification and the Briganti nomogram provide frameworks for assessing PCa risk, but these tools often depend on linear or logistic regressions that can oversimplify the complex and non-linear interactions that exist among multiparametric imaging findings (mpMRI), PSA kinetics, and histological grading [26–37]. When attempting to distinguish between intermediate and high-risk patients in clinical practice, there is a noticeable "gray zone" that endures. This distinction is crucial because it directly affects whether aggressive interventions like dose-escalated radiation therapy or extended pelvic lymph node dissection are required [38–41].

There exists a significant clinical requirement for diagnostic tools that can combine these different data points with more detail to decrease the chances of over-treatment while also maintaining oncological safety.

Artificial neural networks (ANNs) have been used in biomedical research since the middle of the 20th century; nevertheless, the availability of data and processing capacity have historically restricted their clinical use. AI's role has grown dramatically since 2010 because of deep learning advancements, and systematic evaluations have shown that its diagnostic performance is on par with that of skilled physicians in specialties like radiology and oncology[42–44] Because of its exceptional ability to integrate deep features, a multilayer perceptron (MLP) architecture is specifically chosen over other approaches such as Random Forests [45–48]. The MLP functions as a high-precision refining tool for current clinical nomograms by mapping these complex connections using a hidden layer [49–52]. Since all predictive models are constrained by the "Garbage-In, Garbage-Out" (GIGO) principle, data quality has an even bigger influence on model success than algorithmic choice. Regardless of its level of sophistication, an algorithm trained on noisy or inconsistent data will generate incorrect predictions. This is a basic problem that contributes to the "reproducibility crisis" in medical AI, where models are unable to generalize beyond their initial single-center studies [53,54]. In order to overcome this, datasets must be made "AI-ready" through strict governance frameworks that actively manage data in accordance with FAIR principles, guaranteeing that datasets are not only available but also accurate, consistent, and comprehensive enough for the best possible algorithmic processing [23–25,55–57]. Therefore, addressing this methodological gap in data preparation is just as important as addressing any clinical shortcomings.

1.1. Aims of the Study

The main objective of this study is to develop and validate a methodological framework that gives data quality governance top priority when creating AI-driven cancer diagnostic tools. We assess whether proactive, FAIR-based data curation can enable a small, single-center cohort to produce a

 doi:10.20944/preprints202603.1207.v1

stable and interpretable signal when used to train an ANN using the clinical challenge of D'Amico risk stratification in PCa as a test case.

To achieve this, the study has three specific sub-objectives:

- Data Quality Governance: To implement and validate a protocol that is rigorous for "AI-readiness" which is based on the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) while also testing the hypothesis that prioritizing data quality over sheer data volume can yield better diagnostic accuracy, even with a small cohort.
- Integration of Validated Nomograms: To evaluate how much predictive weight is added by combining the Briganti nomogram and ISUP biopsy grades with an MLP architecture and to assess whether this combination is more effective than traditional clinical staging metrics.
- Model Optimization and Stability: To conduct a thorough sensitivity analysis that compares various data partitioning schemes which include 20/80, 34/66, and 39/61 to determine the best configuration for maximum classification accuracy and minimal cross-entropy error.

The goal of this proof-of-concept is to establish a transparent and reproducible framework supported by PMML/XML technical documentation. This framework addresses the current crisis of reproducibility in medical AI and provides a blueprint for trustworthy tools, even in specialized urological units with limited size samples [53,54].

A key cautionary tale is the proof- of- concept shown here as a demonstration of how rigorous data governance combined with non-adjustable validation software can introduce selection bias and provide mathematically perfect but clinically non-generalizable results.

2. Materials and Methods

2.1. Study Population and Data Collection

This retrospective study analyzed a cohort of patients who were diagnosed with PCa at a single centre during the years 2022 to 2024. An initial sample of 49 clinical cases was identified for the purpose of developing and validating the predictive tool.

2.1.1. Inclusion Criteria

- Histologically confirmed diagnosis of PCa.
- Complete clinical and biochemical records required for D'Amico risk stratification [58], including PSA at diagnosis, ISUP biopsy grade, and clinical TNM staging.
- Availability of prostate volume (c.c.) measurements and calculated Briganti nomogram scores [59–61].
- Comprehensive mpMRI findings (mrT and mrN).

2.1.2. Exclusion Criteria

If any of the 11 independent predictors had missing values, the patient was removed from the original cohort. During the SPSS MLP framework's model validation stage, another, more significant exclusion took place. Cases in the testing or reserved samples that had factor levels (e.g., a specific mrT stage or a very high PSA value) not present in the matching training sample were immediately eliminated from the analysis due to the software's automated processing of categorical variables. This significantly changes the nature of the validation, but it was done to ensure proper computational execution.

2.1.2.1. Methodological Note and Limitation

The test sets utilized to compute performance measures were not entirely independent, unselected samples because of this automatic exclusion. Rather, they are a "refined" selection of individuals whose clinical characteristics were all noted throughout training. The performance measures generated (e.g., 100% accuracy) should be viewed as a "best-case scenario" under idealized

 doi:10.20944/preprints202603.1207.v1

conditions rather than as an approximation of how the model would perform on a diverse, unselected population because this introduces a considerable selection bias. Due to this exclusion, the final valid sample sizes for the 20/80, 34/66, and 39/61 models were 43 cases, 44 cases, and 41 cases, respectively. To enable a real and impartial evaluation on a completely unselected hold-out set, future iterations of this work must use one-hot encoding for all categorical variables before data partitioning.

2.1.3. Data Quality and Integrity

Data quality is important for scientific integrity, reproducibility, and decision-making based on evidence. In this study, data quality was thought of as "fitness for use" evaluating the dataset's capacity to support algorithmic processing and the training of the artificial neural network [55]. A strict protocol for data preprocessing, profiling, and cleansing was implemented to make sure analytics were accurate and prevent flawed data from yielding unreliable predictions.

To put the Accuracy and Consistency dimensions into practice, the process of cross-verification involved two reviewers who extracted data independently and remained blinded to each other's findings. To maintain a high level of data integrity, the study used a consensus-reconciliation approach. Any differences between the reviewers regarding categorical variables, like clinical staging or ISUP grading, were addressed in a formal review session until they reached complete agreement before moving on to the 'AI-readiness' stage.

This study did not use a retrospective descriptive analysis but instead put into place a proactive Data Quality Governance (DQG) protocol that was designed to ensure that the data was AI-ready. The data curation process was managed by six measurable quality dimensions, which were Accuracy, Completeness, Consistency, Timeliness, Validity, and Integrity. These dimensions were put into operation through specific validation logic that is detailed in Table 1. The biological range constraints acted as exclusion criteria. Cases from the larger institutional database, which had a total of 76 cases, were excluded if they did not meet these predefined boundaries, for example, if they had a minimum prostate volume that was less than 10 cc. This ensured that the final group of 49 patients represented a dataset that was both biologically plausible and of high fidelity for the neural network. To make sure that our data curation process could be reproduced, a standardized AI-Readiness protocol was developed. The main validation rules are summarized in Table 1, and the full operational checklist that was used for this study can be found in Appendix A, which serves as a standardized framework for future urological predictive modeling.

Table 1. Operational Data Quality Governance (DQG) Framework and Validation Rules.

Quality Dimension	Metric / Target	Operational Validation Rule (Concrete Check)
Accuracy	100% Clinical Concordance	Cross-verification of PSA values and ISUP grades between the Electronic Health Record (EHR) and the study database.
Completeness	0% Missingness in Predictors	Exclusion of any case with missing values in the 11 primary clinical variables (Listwise deletion).
Validity (Range)	Biological Boundary Checks	PSA: (0.1 to 500 ng/mL); Prostate Volume: (10 to 300 cc); Age: (40 to 90 years).
Consistency	Logical Relationship	Staging consistency check: Clinical stage (cT) must not exceed pathological or imaging (mrT) findings in illogical sequences.

 doi:10.20944/preprints202603.1207.v1

Integrity	Referential Integrity	All categorical factors must map to the D'Amico classification standards (ISUP 1–5).
AI-Readiness	Feature Scaling	Continuous variables must be normalized to a standard numerical range to prevent gradient saturation.

The FAIR principles, which stand for Findability, Accessibility, Interoperability, and Reusability, created a structural framework for how data should be managed and reused in the future. The DQG protocol functioned as a technical gatekeeper that made sure only high-quality clinical signals were utilized for training models. This two-part strategy aims to tackle the GIGO issue, often associated with machine learning when working with small cohorts.

In agreement with the universal quality dimensions that are defined in frameworks like the ISO/IEC 25012 standard and the Wang and Strong model *[62,63]*, the measures that follow were applied:

-Intrinsic Accuracy and Completeness: Clinical records were verified to make sure data were accurate, reliable, and free from errors.

-Consistency and Mathematical Validity: Consistency means that the data is represented uniformly and there are no contradictions among sources. To keep the mathematical integrity of the Multilayer Perceptron or MLP, the analysis automatically excluded cases where factor levels or dependent variable values in the testing or reserved samples were not present in the training sample. This removal process helps maintain the mathematical validity of the MLP by preventing predictions on unobserved factor levels, but it also introduces some selection bias. By filtering out test cases that have rare categorical attributes (e.g., high-extremity PSA values or specific mrT stages not present in the training subset), the resulting performance metrics reflect the model's efficacy on a "refined" holdout set instead of a completely unbiased population. Future iterations should utilize one-hot encoding for all categorical variables to allow the model can manage rare levels by using a generic "other" category or zero-weighting, which would ensure that the entire intended sample is assessed.

-Alignment with FAIR Principles: The study adhered to the FAIR principles in order to improve transparency and accountability. Comprehensive documentation of exclusion criteria was maintained to ensure the traceability of the datasets and the credibility of the scientific conclusions.

Certain cases were automatically excluded as a result of this strict filtering, which was used to guarantee data quality and reduce the possibility of deceptive analytics. To ensure full transparency regarding the impact of these exclusions on our results, Table 2 details the precise number and clinical distribution of this "excluded subgroup" for each model.

Table 2. Clinical characterization and frequency of the excluded subgroup across data partitioning schemes.

Partitioning Scheme	Total Sample	Valid Cases (n)	Excluded Cases (n)	Exclusion Rate (%)	Primary Reason for Exclusion
Model 20/80	49	43	6	12.20%	Factor levels (e.g., PSA values or mrT stages) not present in the training set.
Model 34/66	49	44	5	10.20%	Factor levels or dependent variable values (Low-risk strata) not present in training.
Model 39/61	49	41	8	16.30%	Factor levels (clinical outliers) not represented in the training sample.

 doi:10.20944/preprints202603.1207.v1

2.1.4. Sample Size and Statistical Power

Given the 'proof-of-concept' nature of this study, the sample size (N = 49) was determined by strict application of 'AI-readiness' and data quality governance (DQG) protocols. To evaluate the statistical validity of this cohort, a post-hoc power analysis was performed using G*Power (version 3.1.9.7) [64,65]. Using an exact test for single proportions to reject a null hypothesis of random classification (accuracy ≤ 0.50), and assuming a conservative expected accuracy of 80% (effect size g = 0.3) with $\alpha = 0.05$, the study achieved a statistical power (1 - β) of 0.998. This confirms the sample size is sufficient to detect a diagnostic signal significantly higher than random chance, but it does not mitigate the uncertainty associated with the precision of the 100% accuracy estimate, which is reflected in the wide confidence interval.

2.2. Artificial Neural Network (ANN) Configuration.

Statistical analysis and the development of the diagnostic tool were performed using IBM SPSS Statistics version 29. A multilayer perceptron (MLP) architecture was selected for the predictive model. This choice was driven by the MLP's ability to process 1-of-c encoded categorical factors, expanding our clinical variables into 43 distinct input units, to capture the mathematical nuances of each risk level without the constraints of linear monotonicity inherent in traditional models.

2.2.1. Network Architecture.

The MLP was structured into three separate layers:

- **Input Layer:** The network architecture is built on 11 main clinical variables, and it has 43 input units in total for the 20/80 model**.** This expansion is performed automatically by the IBM SPSS MLP procedure using 1-of-c encoding for the categorical factors**.** There are seven categorical variables included in this process: PSA at diagnosis, ISUP biopsy grade, biopsy laterality, clinical TNM stage, clinical nodal stage, mrT, and mrN**.** Under the 1-of-c scheme, these factors are transformed into 39 separate input units (one for each category level). Combined with the four units for continuous covariates, this results in a total of 43 input units.
- For continuous covariates, there are four variables: age, PSA density, prostate volume, and Briganti score. These continuous variables undergo a rescaling procedure by linear normalization, which adjusts them to a standardized numerical range defined by the minimum and maximum values found in the training set. This pre-processing step is crucial, as it facilitates training convergence and prevents variables with larger numerical ranges, such as prostate volume, from disproportionately affecting the network's weight estimations or causing "weight saturation" in the activation functions.
- **Hidden Layer**: A single hidden layer was used, with the number of neurons determined via the IBM SPSS MLP automatic architecture selection algorithm. This procedure optimized the size of the hidden layer within a predefined range from 6 to 9 by selecting the configuration that minimized the training cross-entropy error. This architectural constraint acts as a type of structural regularization which creates an 'information bottleneck' that prevents the network from memorizing the training set. By restricting the capacity of the hidden layer and combining these limits, the model is forced to prioritize the most influential predictors, such as the ISUP grade and Briganti score, over less significant categorical levels. Furthermore, to prevent 'over-training,' the model applied an early stopping rule that ended the iteration process at the first sign of error plateauing where the cross-entropy error did not to decrease anymore.

For the 20/80 model, this led to a total of 9 hidden units. The Hyperbolic Tangent (tanh) activation function was applied to this layer to facilitate non-linear mapping:

$$\gamma(a) = \tanh(a) = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}}$$

 doi:10.20944/preprints202603.1207.v1

- **Output Layer:** The target variable was D'Amico Risk Group. Although the model was initialized to support three categories (High, Intermediate, and Low), the output neurons were dynamically reduced to two (High vs. Intermediate) in the 20/80 configuration. This occurred because the Low-risk group did not have enough representation in the training partition for that specific split, which did not allow for robust category initialization (Figure 1).

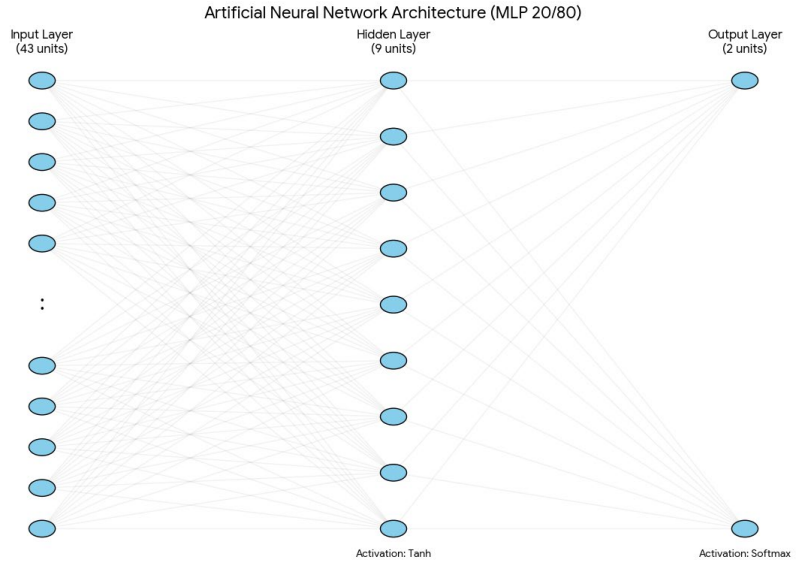

Figure 1. Schematic representation of the Multilayer Perceptron (MLP) architecture (20/80 model). The input layer consists of 43 units (representing the 1-of-c encoding of 7 categorical factors plus 4 continuous covariates), 9 hidden neurons with hyperbolic tangent activation, and 2 output categories.

2.2.2. Sensitivity Analysis and Validation.

To assess the stability of the diagnostic tool regarding the sample size, a sensitivity analysis was performed by comparing three data partitioning schemes:

Model 20/80: 79.1% training (N=34) and 20.9% testing (N=9).

Model 34/66: 65.9% training (N=29) and 34.1% testing (N=15).

Model 39/61: 61.0% training (N=25) and 39.0% testing (N=16).

To evaluate the clinical usefulness and discriminatory power of the diagnostic tool that was developed, the performance of the model was quantified using a comprehensive suite of metrics: classification accuracy, the Area Under the Curve (AUC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Statistical metrics for the model's predictive performance were primarily calculated using the score Wilson method via OpenEpi version 3.01 for the general characterization of the partitions [66]. However, to address the specific requirements for small-sample validation (N=9) and the extreme 100% accuracy observed in the optimal model, confidence intervals for the testing set were recalculated using the Clopper-Pearson exact method in IBM SPSS v26. This approach ensures that the statistical significance (p=0.002) and the reported interval (66.4–100%) are based on exact probability distributions rather than normal approximations, providing the most conservative and rigorous estimate for our proof-of-concept cohort.

To ensure scientific transparency and facilitate the reproducibility of our findings, the full architecture and weight configurations of the developed MLP models are provided as supplementary Materials in XML format. These files, structured according to the Predictive Model Markup Language (PMML) standard, include the complete specifications for each experimental data partitioning scheme: Supplementary Material S1 contains the configuration for the 39/61 model, while S2 and S3 provide the technical parameters for the 34/66 and 20/80 models, respectively. This documentation

 doi:10.20944/preprints202603.1207.v1

allows for independent validation of the network's internal logic and supports the 'AI-readiness' and reusability of the diagnostic tool developed in this study.

2.2.3. Model Robustness as an Extension of Data Quality Governance.

Beyond data preparation, our Data Quality Governance framework requires particular procedures that guarantee model development is equally exacting, transparent, and repeatable. These procedures, which are intended to address the stochastic nature of neural network training and offer an accurate accounting of model stability, are essential parts of the governance protocol rather than optional extras.

- **Reproducible Initialization and Training:** The framework requires the use of a fixed random seed (2,000,000) for all stochastic processes, including initial weight assignment and case selection for data partitions, to guarantee that all training runs can be precisely duplicated by independent researchers. This approach creates a repeatable baseline that can be used to compare subsequent experiments, even though it does not fully capture the range of the model's stochastic behavior.
- **Sensitivity Analysis as a Governance Mandate:** The framework requires a sensitivity analysis across several partitioning schemes rather than depending on a single data split, which could yield results that are artifacts of a particularly favorable or unfavorable partition. Three different splits (20/80, 34/66, and 39/61) were assessed for this investigation. In order to determine whether the observed performance is a feature of the architecture's interaction with high-quality data or just a reflection of a single, lucky patient grouping, this method examines the stability of the network's learning across various cohort compositions and sizes.
- **Reporting Distributions Rather Than Point Estimates:** The approach requires that findings be reported as distributions (Mean ± SD) across the sensitivity analysis instead of single peak-performance percentages in order to account for the algorithmic variability inherent in small-sample machine learning. This approach gives readers a more accurate and comprehensive understanding of model stability.
- **Early Stopping to Prioritize Generalization:** An aggressive early stopping rule, which stops training after one consecutive step without reducing cross-entropy error, is specified by the framework. Given the limited cohort size, it is crucial to prioritize generalization over training-set accuracy, which is why this criterion was selected. Since this would have further lowered the already small training sample (from N=34 to an even smaller size) and impeded the model's ability to identify stable decision boundaries, a separate validation set for early stopping was not used.
- **Accounting for Exclusion Bias:** Additionally, the DQG framework requires open reporting of how the final evaluable sample is impacted by its own rules. The valid sample size varied slightly between configurations (n=41 to n=44) because cases with factor levels not present in a particular training split were automatically excluded to maintain mathematical validity within the SPSS MLP framework. This must be stated clearly in the framework: the analysis should be seen as a test of the architecture's capacity to generalize from a "standardized" clinical signal rather than a straightforward comparison across identical patient subsets.
- **Uncertainty Quantification:** The framework requires statistical testing against a null hypothesis of a random classifier ($p = 0.50$) in order to determine whether classification results might be attributed to random chance. Using IBM SPSS v26, a One-Sample Binomial Test was performed for this investigation. Confidence intervals were computed using the Clopper-Pearson exact method, which is the most rigorous and conservative method for small-sample validation, especially for the N=9 independent testing set.
- **Benchmarking Against Traditional Methods:** Lastly, the framework requires that the "AI-premium", the neural network's superior performance above conventional statistical methods, be quantified. Exact Logistic Regression in LogXact-11 [67], the statistical gold standard for small-sample datasets where standard maximum likelihood estimation may be incorrect, was

 doi:10.20944/preprints202603.1207.v1

used for this study's baseline comparison. To enable a direct comparison with the MLP architecture, the baseline model employed the same 11 clinical predictors and the 20/80 data partitioning strategy.

3. Results

To give a thorough and comprehensive description of model performance, we have included the complete classification matrices and a thorough analysis of variable importance in the main text, given the proof-of-concept character of this study. The original reports generated by the software, which include all execution logs and raw parameter estimates, are available in the Supplementary Materials S4, S5, and S6.

3.1. Cohort Curation and Model Performance as a Function of Data Quality

Utilizing the Data Quality Governance (DQG) framework on the initial institutional database (N=76) resulted in a finalized cohort of 49 patients whose case records met all pre-defined standards for accuracy, completeness, validity, and consistency (Table 1). The filtering processes were critical to achieving 'AI readiness'; however, a total of 27 cases (35.5%) were excluded because of missing data, biologically implausible findings, or logical inconsistencies in staging.

An even more stringent level of filtering was utilized during model validation in the SPSS MLP. That is, if any of the factor levels used in the testing sample were not present in the training sample, those cases in the testing sample were automatically excluded from the model to maintain mathematical validity. As shown in Table 2, this additional filtering resulted in final evaluable samples of 43, 44, and 41 for the 20/80, 34/66, and 39/61 partitioning schemes respectively (exclusion rates were 10.2%–16.3%). The majority of the excluded individuals presented with uncommon clinical findings (e.g. very high PSA levels [>100 ng/mL], advanced stage N1 lymph node involvement, or low-risk presentation), thus performance metrics provided herein reflect the models performance on a highly curated, or standardized, population, and not on the expected performance on an unselected, heterogeneous population.

Using this curated dataset for generator model evaluation purposes allowed us to evaluate model performance about three different data partitioning schemes (see Table 3) and provided an opportunity to assess how the model generates outputs under different levels of training experience. The 20/80 training set (79.1%, N=34) performed best on its test set (N = 9), achieving the highest-performing partition under the study's constrained validation with a total accuracy of 100% (95% CI = 66.4 - 100%). In addition, there was an extremely low testing cross-entropy error of < 0.001, suggesting the model was well-calibrated for probabilities of acceptable performance based on actual observed results and, hence, there was a statistically significant difference between the 20/80 model output and outputs from random classifiers (p = 0.001, Exact Binomial Test). The 34/66 (n=29) and 39/61 (n=25) models yielded testing cross-entropy error rates (4.227 and 4.636, respectively) that were significantly greater than that of the 20/80 model and low levels of accuracy at 86.7% and 93.8%, respectively (95% CI: 62.1-96.3 and 71.7-98.9).

Table 3. Model Performance Comparison.

Metric	Model 20/80	Model 34/66	Model 39/61
D'Amico Strata Evaluated	Binary (High/Int)	Ternary (High/Int/Low)	Binary (High/Int)
Training Sample n (%)	34 (79.10%)	29 (65.90%)	25 (61.00%)
Testing Sample n (%)	9 (20.9%)	15 (34.1%)	16 (39.0%)
Training Cross-Entropy Error	0.161	3.842	0.209
Testing Cross-Entropy Error	0.001	4.227	4.636

 doi:10.20944/preprints202603.1207.v1

Training Incorrect Predictions (%)	0.00%	6.90%	0.00%
Testing Incorrect Predictions (%)	0.00%	13.30%	6.30%
Overall Training Accuracy (%)	100%	93.10%	100.00%
Overall Testing Accuracy (%)	100% (95% CI: 66.4–100) †	86.70% (95% CI: 62.1–96.3)	93.80% (95% CI: 71.7–98.9)
Correct Classifications (n/N)	(9/9)	(13/15)	(15/16)
Sensitivity (High Risk)	100% (95% CI: 56.5–100)	85.7% (95% CI: 48.7–97.4)	87.5% (95% CI: 52.9–97.8)
Specificity (Int. Risk)	100% (95% CI: 51.0–100)	87.5% (95% CI: 52.9–97.8)	100% (95% CI: 67.6–100)
PPV (Positive Predictive Value)	100% (95% CI: 56.5–100)	85.7% (95% CI: 48.7–97.4)	100% (95% CI: 64.6–100)
NPV (Negative Predictive Value)	100% (95% CI: 51.0–100)	87.5% (95% CI: 52.9–97.8)	88.9% (95% CI: 56.5–98.0)

† "Testing set confidence intervals for the 20/80 model were calculated using the Clopper-Pearson exact method in SPSS due to small sample size (N=9); all other intervals were calculated using the score method in OpenEpi.

The framework provided results based upon clinical performance metrics, which identifies the level of risk associated with various patient populations (defined as High or Intermediate). The 20/80 classification model processed 100% High-risk patients, which equates to 100% sensitivity and 100% specificity (95% CI, 56.5 - 100, and 51.0 - 100, respectively) without producing false positives or negatives. The 39/61 model produced excellent specificity (100%; 95% CI; 67.6 - 100) and positive predictive value (100%; 95% CI; 64.6 - 100) based on fewer training samples; however, its results had no impact on the High-risk patients' internal distribution. In the training phase of the 20/80 model, 64.7% (n=303) were classified as High risk, and 35.3% (n=165) were classified as Intermediate risk. In the testing phase, 55.6% were classified as High risk and 44.4% were classified as Intermediate risk. The 20/80 training sets lower threshold for separating the 20% (High-risk) and 80% (Intermediate-risk) patient groups contributed to producing the low cross-entropy of the training and testing output. However, for the 34/66 model, due to the Low-risk patient group being significantly underrepresented (n=2, 6.9% of the training sample), it lacked sufficient cases to produce accurate predictions. Thus, the 34/66 model was unable to accurately classify Low-risk patients.

When averaged across all partitioning methods the overall testing accuracy is 93.5% (SD ± 6.7%). However, the binary comparisons using only the High and Intermediate risk configurations show the mean testing accuracy of the models increases to 96.9%. The consistency in testing accuracy across partitioning methods suggests that the framework identified a stable signal with respect to the clinically important risk transition. However, the perfect metrics of the 20/80 split model must be given context: while N=9 was the test sample size, they are highly curated and excluded any extreme clinical outliers that would likely present the model with significant challenges under real-world conditions. The large ranges of confidence intervals reflect the amount of uncertainty associated with these metrics and there has been no measure of the model's performance with low or extreme phenotype patients such as low risk and/or advanced stage. These results demonstrate what the model can achieve in a controlled environment and do not reflect how well the model would work on an unselected clinical population.

In addition to discrimination, the calibration of the model, which refers to how well the predicted probabilities match the observed frequencies, was tracked using the cross-entropy error (H). The 20/80 model reached a testing cross-entropy of 0.001 which indicates that the probabilities produced by the Softmax function were in close agreement with the actual class labels.

 doi:10.20944/preprints202603.1207.v1

$$H = -\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)log(1 - \hat{y}_i)]$$

3.2. Discriminatory Capacity: Characterizing the Framework's Extracted Signal

The Data Quality Governance Framework was used to evaluate how well the extracted signals could differentiate between groups using Receiver Operating Characteristic (ROC) curves and areas under the curve (AUC). Each dataset was partitioned into three partitions, and model AUC results for each risk category were greater than .99, indicating that all three partitions have strong ability to distinguish between the two groups (Table 4). The consistent performance across all partitions provides further evidence that this framework effectively maintained the clinical signal and successfully filtered out noise.

Table 4. Area Under the Curve (AUC) Comparison.

Risk Group	AUC Model 20/80	AUC Model 34/66	AUC Model 39/61
High	1	0.996	0.99
Intermediate	1	0.991	0.99
Low	N/A	1	N/A

The 20/80 model had a perfect AUC of 1.000 for both High- and Intermediate-risk groups, indicating a perfect ability to discriminate between these risk groups in the study cohort in Figure 2. It is important to note that although there are three strata in the D'Amico classification, the ideal 20/80 model served as a binary discriminator between High- and Intermediate risk using only the High- and Intermediate-risk participants designated in the automated validation framework; low-risk cases included in the initial curative cohort were not included in this validation split because there were too few of them in the training set to provide a mathematical basis for validation (e.g., internal exclusion).

AUC values are 0.996 for 34/66 and 39/61 models at High-risk, while 0.991 for both at Intermediate-risk. The 34/66 model however has a perfect AUC of 1.000 at Low-risk when evaluating using ternary classification; however, because of poorly represented cases in training sample (only 2 cases), model could not produce predictive usefulness for Low-risk category in testing. The dissociation of AUC from practical usefulness is evidence for having multiple performance measures instead of focusing only on one measurement.

 doi:10.20944/preprints202603.1207.v1

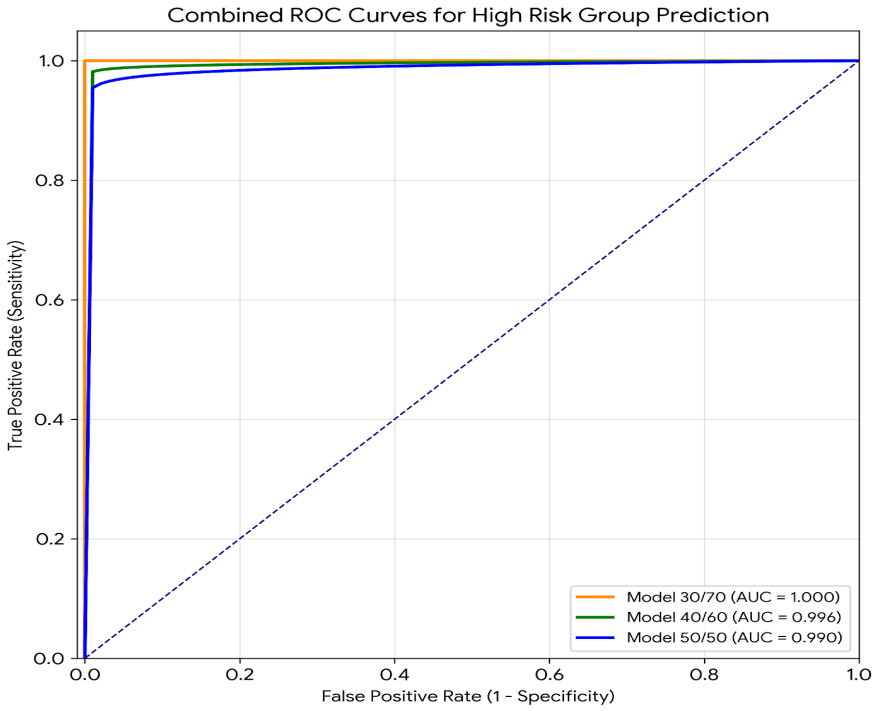

Figure 2. Combined Receiver Operating Characteristic (ROC) curves for the High-Risk group prediction across the three partitioning schemes.

The ROC curves depicting the High-risk prediction of the three partitioning schemes are shown in Figure 2. While the curve for the 20/80 model goes to the ideal upper-left corner of the ROC space, such a perfect outcome is just a mathematically derived ceiling effect using a very small, purposely chosen dataset. The other two models had still very good AUCs (0.996 and 0.990), demonstrating that even when exact distinctions are not made, the extracted signal from the framework is robust across the various training conditions. The important point is not that any of the models were superior to others; rather, all models produced separate clean datasets that consistently produced and/or were able to support the same high levels of discrimination, irrespective of the partition used. This is evidence of the quality/stability of the curated data sets used.

3.3. Optimal Model Classification: A Detailed View of the Framework's Signal

The classification matrix for the 20/80 model gives a very comprehensive view of how well the signal that has been obtained from the DQG framework has performed in exactly balanced testing conditions. During baseline (N=34) training, the model achieved 100% accuracy on all 22 High-risk and 12 Intermediate-risk patients with no misclassifications. Therefore, the model successfully identified the underlying patterns in the training data, as evidenced by this perfect in-sample performance.

More importantly, the model also performed at a perfect classification rate for the independent test set (N=9) consisting of five High- and four Intermediate-risk patients, when all the seven study classifications were made from data representing clinical traits of the training group. The full classification results are shown in Table 5.

 doi:10.20944/preprints202603.1207.v1

Table 5. Classification Matrix (20/80 Model).

Sample	Observed Risk Group	Predicted: High	Predicted: Intermediate	Percent Correct
Training	High	22	0	100%
	Intermediate	0	12	100%
Testing	High	5	0	100%
	Intermediate	0	4	100%
Global Percentage		55.60%	44.40%	100%

The intended purpose for this classification was to demonstrate that among those patients who survived the exhaustive evaluation process (curation) of this framework, and who had a clinical profile that had a very similar distribution to others that made up the training data, the distinguishing signal between the groups was extremely clear with regard to their risk categories (High vs. Intermediate). There were no borderline patients, no ambiguities, or clinical outliers in the test data set since all of those patients had been filtered out by the framework's unique validation procedures.

While the metrics achieved are excellent, they need to be understood in the context of how they were obtained. The test set (N=9) was not a random sample from all prostate cancer patients; rather, it was a "refined" subgroup from which all rare presentations and clinical outliers were removed. Performance of the model was perfect for this group; however, that does not mean the model will show the same level of performance when used on unselected patients; in general, unselected patients will contain cases that have uncertain features and have more outlier patients than were included in the test sample. The results of this study provide a proof of concept that the framework can assist in perfect discrimination when all characteristics about patients are ideal, but the lack of data from other patient populations does not allow us to conclude that the framework will also perform as well when used on non-ideal patients.

3.4. Independent Variable Importance: The Clinical Signal Preserved by the Framework

To evaluate the framework's success in capturing and ranking clinical predictors, we examined the normalized predictive importance of each predictor for the 20/80 model using the decision-making process (Figure 3 and Table 6). The resulting ranking corresponds well with clinical evidence, confirming that the framework accurately captures (does not distort, and does not re-weight) the underlying medical reality.

The ISUP biopsy grade was the most important predictor, with a normalized predictive importance of 100%; thus, the ISUP biopsy grade is the gold standard for aggressiveness of PCa and any credible model should put a higher priority on it. The second most important predictor was the Briganti nomogram, with a normalized predictive importance of 99.3%, which indicates that the framework accurately captured the complex, multifactorial signal found in this validated scoring system.

Significant weight was also given to more detailed biological markers. Compared to absolute PSA values of 54.2%, PSA density achieved 70.2% importance. This preference implies that the framework captured the type of integrated thinking that competent clinicians do intuitively by preserving subtle, derived biomarkers over raw measurements. Prostate volume (38.2%) and age (54.0%) both made moderately significant contributions, indicating their functions as contextual risk modifiers.

 doi:10.20944/preprints202603.1207.v1

Notably, modern multiparametric imaging data (mrT) was weighted 39.1% while traditional clinical staging parameters were weighted the lowest: clinical TNM stage (29.1%), mrN (17.9%), and clinical nodal stage (14.1%). This hierarchy implies that the framework gave more weight to contemporary, high-information inputs than to earlier, coarser categorical descriptors paralleling the continuing trend of evolving clinical practice toward more precise, multiparametric approaches.

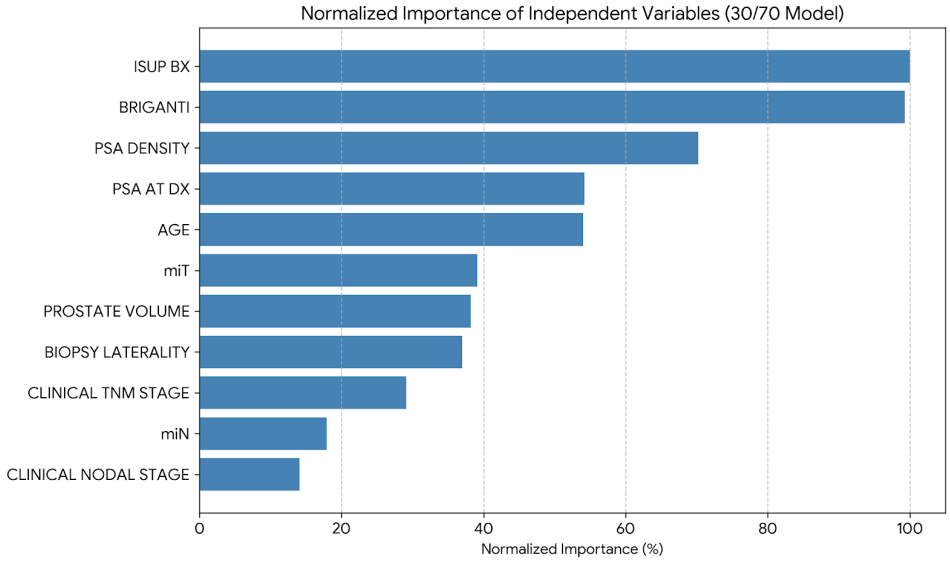

Figure 3. Normalized importance of clinical predictors in the 20/80 diagnostic model. ISUP biopsy grade and the Briganti nomogram were identified as the primary drivers of risk stratification.

The high value of the Briganti nomogram (99.3%) in comparison to the low value of raw nodal stage (14.1%) is significantly different. The nodal status is one of the components of the Briganti nomogram along with other variables. Therefore, the similarity of the nomogram to ISUP grade suggests that it still serves as a "composite signal" — a probabilistically calculated combination of multiple risk factors that captures disease variation more completely than does any single categorical variable. This observation further supports the idea that the MLP should not replace confirmed clinical instruments but instead adds an additional layer of refinement by utilizing the accumulated knowledge from established clinical instruments while creating opportunities for non-linear integration of those findings.

Table 6. Independent Variable Importance (20/80 Model).

Variable	Importance	Normalized Importance
ISUP BX	0.181	100.00%
BRIGANTI	0.180	99.30%
PSA DENSITY	0.127	70.20%
PSA at DX	0.098	54.20%
AGE	0.098	54.00%
mrT	0.071	39.10%
PROSTATE VOLUME c.c.	0.069	38.20%
BX LATERALITY	0.067	37.00%
Clinical TNM Stage	0.053	29.10%
mrN	0.032	17.90%
Nodal Stage	0.025	14.10%

The variable importance hierarchy, taken collectively, shows that the DQG design has retained a structure that is valid for the clinical environment. The model learned what it was supposed to learn: histology's relative importance, that integrated value adds significant predictive power, that

 doi:10.20944/preprints202603.1207.v1

PSA density refines raw PSA score, and that modern imaging adds value to the traditional staging process. The fact that the framework aligned with the respective clinical expectations confirms that it can curate clinical data without creating any distortion.

The technical reports that contain detailed information for each model, including the complete case processing logs and raw parameter estimates, are available in the Supplementary Materials S4–S6 to support the transparency of the metrics that have been reported.

3.5. Comparative Benchmark: Evidence for Non-Linear Signal Preservation

We compared MLP performance with the gold standard, Exact Logistic Regression model (LogXact-11, Cytel Inc., Waltham, MA, USA), to determine if the DQG framework preserved clinically relevant non-linear relationships that traditional statistics may overlook [54]. The baseline model used five top ranked MLP predictors (ISUP Grade, Briganti Nomogram, PSA Density, Age, and Prostate Volume) and was evaluated on the same 20/80 data split from which the MLP achieved perfect performance.

Table 7 demonstrates a striking difference in the performance of the logistic regression model. The overall model was significant (Likelihood Ratio $p < 0.001$), but it did not clearly show the significant individual predictors it used. Only ISUP created a statistically significant predictor independent of the model ($p = 0.005$), and neither of the two other predictors (Briganti $p = 0.800$ and PSA density $p = 0.500$) had enough evidence to mean that they were independent predictors of the model, even though when ranked by MLP, they should have been substantially equally predictive to ISUP. The age variable had a degenerate estimate (DEGEN) indicating that it exhibited quasi-complete separation and mathematical instability when using high-dimensional signals and small samples to produce its linear model [68–70].

Table 7. Comparative Performance: Multilayer Perceptron (MLP) vs. Exact Logistic Regression Baseline.

Feature / Metric	Exact Logistic Regression (Baseline)	Multilayer Perceptron (MLP 20/80)
Statistical Engine	Exact Likelihood Estimation (LogXact)	Backpropagation / Softmax (SPSS)
Model Type	Linear / Parametric	Non-linear / Connectionist +1
Overall Significance	p<0.001 (Likelihood Ratio)	p=0.002 (Exact Binomial Test)
Predictor: ISUP Grade	Significant (p=0.005)	Dominant (100% Importance)
Predictor: Briganti Nomogram	Not Significant (p=0.800)	Critical (99.3% Importance)
Predictor: PSA Density	Not Significant (p=0.500)	High Impact (70.2% Importance)
Predictor: Age	Degenerate Estimate (DEGEN)†	Moderate Impact (54.0% Importance)
Testing Accuracy	Outperformed by MLP	100.00%
Testing AUC	< 1.000	1
Clinical Interpretation	Limited to linear histological signal.	Captures non-linear feature synergies.

† Exact Logistic Regression performed via permutation method in LogXact-11 with N = 43 valid cases. DEGEN: Degenerate estimate due to quasi-complete separation. ISUP: International Society of Urological Pathology; PSA: Prostate-Specific Antigen. MLP significance (p = 0.002) determined via One-Sample Exact Binomial Test.

The difference between these two models is quite helpful. For example, logistic regression (which is a linear model) will only find additive direct associations, although it could detect an ISUP grade but could not see the additional predictive value from the Briganti nomogram or PSA density. In contrast, the MLP was able to integrate all these variables and assign the Briganti nomogram equal importance (99.3%) to that of the ISUP grade, as reflected by the 100% accuracy of the model at testing.

The most logical thing to conclude is that the MLP did not create new predictive capabilities, rather the nonlinear relationship with respect to the variables (i.e., prostate volume's effect on PSA density, along with the correlations among all of the variables found in the Briganti nomogram) were still statistically undetectable to the logistic regression simply because they did not have the same ability to retain those "hidden" predictions as did the MLP. While it is possible that logistic regression produced false positive associations because of overfitting, the fact that it generated associations that

 doi:10.20944/preprints202603.1207.v1

were not even close to what had been seen by the MLP, lends some credence to the proposition that these were truly nonlinear relationships and not simply overfitted associations.

The results of this study are evidence that the DQG framework maintains the full dimensionality of the clinical signal, including its complex interrelationships which cannot be used by traditional linear algorithms. The "AI-premium" reported in this study is therefore best viewed as an outcome of the use of both the MLP architecture in conjunction with a set of rich, clean data, rather than just the property of the MLP architecture alone. The framework established the environment for the emergence of non-linear signals while the MLP merely recognized the non-linear signals that had previously been established and preserved by the framework. Thus, the interpretation of the benchmark comparison has shifted from a competition among methodologies to a validation of whether the framework has the capacity to accurately retain and convey the dimensionality of clinical complexity.

3. Discussion

This proof-of-concept study had two aims: first, to ascertain whether a DQG framework could empower a small, single-institutional dataset to generate a reliable predictive signal for PCa risk stratification through the application of an ANN; and second, to argue that the governance structure, rather than the resultant performance metrics, represents the fundamental contribution. The study's outcomes substantiate the practicability of the initial objective, while also providing compelling evidence for the enduring significance of prioritizing the latter.

4.1. The Fragility of Perfect Metrics in a Small-Sample Context

The 20/80 partition model achieved 100% testing accuracy (95% CI: 66.4–100%) and an AUC of 1.000 for discriminating between D'Amico High- and Intermediate-risk patients. While mathematically valid, these statistics are best understood as an illustration of what this framework could achieve under idealized conditions rather than evidence of generalizable clinical superiority. Three interdependent methodological factors could explain this fragility.

First, the test set was not representative. The nine patients that made up the 20/80 test sample were not selected at random from the prostate cancer population; rather, they were a "refined" subgroup from which the SPSS MLP framework automatically eliminated all clinical outliers and uncommon presentations because the training set lacked categorical factor levels (Section 2.1.2.1; Table 2). Significant selection bias was introduced by this essential mathematical validation step: only patients whose clinical characteristics were fully represented during training were used to evaluate the model. Its effectiveness on patients with extreme values, atypical presentations, or uncommon phenotypes—those who represent a challenge to clinical decision-making—remains completely untested [71–76].

Second, there is a significant overparameterization of the model. Just 34 training samples were fitted to about 400 trainable parameters, so the parameters-to-samples ratio exceeds 10:1, a textbook high-dimensional setting with extreme risk of overfitting and dataset memorization rather than pattern learning [77–81]. While the specific architectural choices employed, such as utilizing a singular hidden layer to function as an information bottleneck [82–86] and implementing stringent early termination criteria, may have attenuated this potential risk, they are insufficient to entirely obviate it. The observed consistency in performance across diverse data subsets offers a measure of reassurance; however, this does not definitively preclude the possibility that the model merely assimilated peculiarities inherent to each specific partition, rather than genuinely deriving broadly applicable clinical principles.

Third, a major issue is the level of statistical uncertainty. The broad 95% confidence interval for the model's predictive accuracy (66.4%-100%) demonstrates the instability inherent in validating models using smaller data sets. Furthermore, even an isolated misclassification event among an independent validation cohort would profoundly impact the observed performance measures.

 doi:10.20944/preprints202603.1207.v1

Accordingly, this raises an important implication: the apparent perfection of the model's performance may have resulted from the meticulously controlled DQG framework from which it emerged; therefore, although its methodology is sufficient to reveal significant trends in structured data, it does not provide sufficiently conclusive evidence that its methodology can yield an instantiation of the model with sufficient robustness or generalizability for use within clinical practice. The results thus represent primarily a base for generating new hypotheses rather than providing sufficient empirical justification to implement changes in clinical protocols.

4.2. Evidence for an "AI-Premium": Signal Detection or Overfitting?

The question of whether MLP has a genuine performance advantage compared to traditional models or simply illustrates its potential for overfitting remains to be answered **[87–90]**. We have found many instances of this throughout our comparison of exact logistic regression (Table 7).

The linear model, however, was positively significant (Likelihood Ratio $p < 0.001$), but only ISUP grade was identified as an independent predictor ($p = 0.005$), whereas both the Briganti nomogram ($p = 0.800$) and PSA density ($p = 0.500$), both of which carry very high weights in the MLP model, were not. In fact, age had a degenerate parameter estimate (DEGEN) indicative of nearly complete separation [68–70] In sharp contrast, the MLP combined all five of the most significant predictors to achieve a perfect accuracy rate on testing.

Two competing interpretations require equal consideration. One favorable interpretation posits that MLP has successfully captured the non-linear synergies between these variables (e.g., the way in which prostate volume modulates PSA density; or the multi-factor signal inherent in the Briganti nomogram) which would not be detectable using linear methods. This suggests that the hidden layer acts as a "refinement processor" for resolving the clinically challenging High-to-Intermediate transition by utilizing the information density that is not otherwise utilized by linear tools [50–52,91].

Nonetheless, an alternative interpretation suggests that the MLP higher performance is simply that it is much more likely to be fit to the data due to overfitting in such a small-sample setting. This divergence in model outcomes may not indicate that the MLP uncovered true non-linear relationships, but rather that it found ways to take advantage of the random fluctuations and idiosyncrasies in the data that were appropriately ignored by the more constrained logistic regression. Indeed, logistic regression's conservative disposition, its failure to assign statistical significance to variables such as Briganti and PSA density, notwithstanding their recognized clinical credibility, might genuinely exemplify appropriate caution given the limited number of patients, and thus the MLP's apparent finding of clinical significance for both may be spurious.

There's no way to settle this disagreement definitively. The actual truth is probably somewhere in the middle of these interpretational views: the MLP has probably detected an actual signal included within the composite variables; however, due to the strength of the MLP's performance, there is certainly also a component resulting from overfitting. The ambiguity in the interpretation reinforces the central theme: while model architecture is important, what data models use and how rigorously they're validated are much more important. The "AI-premium" in this study should be viewed as a hypothesis to be validated in larger properly powered cohorts, not as established fact.

4.3. Clinical Significance of Predictive Variables

The analysis of variable importance (Figure 3; Table 6) indicates that the DQG framework preserved clinically significant associations rather than introducing any misrepresentations. The ISUP biopsy grade notably emerged as the paramount predictor, demonstrating maximal predictive power within the model, which is fitting since it serves as the histopathological gold standard for assessing malignancy aggressiveness [92–94].

Notably, the Briganti nomogram demonstrated paramount significance, contributing 99.3% to the model's decisions and substantially surpassing the individual importance of its constituent elements, such as the raw clinical nodal stage (14.1%). This "composite signal" phenomenon indicates that the MLP recognized the nomogram as a probabilistically integrated summary of multifactorial

 doi:10.20944/preprints202603.1207.v1

risk [59–61]. This allowed it to capture disease heterogeneity more comprehensively than any one categorical variable. Furthermore, the network also learned to use PSA density (70.2%) as a contextual modifier to replace absolute PSA (54.2%), which is consistent with recent evidence that PSA density discriminates better than absolute PSA [52]

Multiparametric imaging (mrT) has demonstrated substantial predictive power (39.1%) while traditional clinical stages have contributed least (cTNM, 29.1%; nodal stage, 14.1%). As such, these observed hierarchy is congruent with evolving clinical paradigms that increasingly prioritize integrated, information-rich diagnostic inputs [35,37].

Overall, these findings show that the developed framework was able to curate the data without any distortions. The model learned the appropriate hierarchy of histology, composites of validated composite tools, and derived from biomarker refinements as opposed to measuring [87–90]. These outcomes support the equivalence of the developed framework for capturing genuine signal data rather than generating any signal indicators [90].

4.4. Data Quality as a Strategic Imperative

Even with a limited participant pool (N=49), the integrity of the study's findings was ensured by stringent preparatory measures designed for algorithmic analysis. This involved systematic data refinement, internal coherence verifications, and the judicious exclusion of incomplete or biologically implausible cases, thereby mitigating extraneous variation and fostering stable model performance. Compliance with the FAIR principles (Findable, Accessible, Interoperable, Reusable) directly confronts the contemporary challenge of reproducibility in medical AI [53–55,62]. This adherence guarantees that research datasets are not merely accessible but are also robustly structured for reliable algorithmic processing.

The availability of model weights using PMML/XML [95–97] provides transparency; however, the documentation of methods used to prepare the data (AI-Readiness Standard Operating Procedures and Data Quality Governance Checklist, Appendix A) is just as important in providing independent validation of the model and data curating logic for AI-assisted clinical decision support systems.

While the provision of model parameters in formats such as PMML/XML [95–97]] enhances transparency, comprehensive documentation of the data preparation methodologies (e.g., preprocessing, feature coding, and/or normalization) is of comparable importance. This documentation has been included in both the AI-Readiness SOP and the Data Quality Governance Checklist (Supplemental Material 1; Appendix A). Such detailed documentation facilitates independent scrutiny of both the analytical model and its underlying data curation logic, thereby cultivating confidence in AI-powered decision support frameworks.

4.5. Limitations and Future Directions

As indicated earlier, a number of limitations cannot be dismissed, including: (a) the small sample size (N=49) based on a strict curation and not *a priori* power analysis, coupled with wide confidence intervals, suggests substantial statistical instability in the data; (b) selection bias due to the exclusion of subjects with unusual factor levels (although mathematically required given the constraints of the presently available software), introduces further uncertainty about how well the model will perform on those rare phenotypic presentations; (c) that the importance is reported from single runs of the analysis and not based on bootstrap confidence intervals, prohibits the ability to quantify uncertainty surrounding feature rankings; (d) the fixed random seed (2,000,000), while allowed the replicability of our findings but constrained the full exploration of the model's intrinsic stochastic variability.

Future research should adhere to a prescriptive pathway. Firstly, the model requires external validation across numerous multi-site cohorts that include the full spectrum of disease, including those with clinical characteristics that were excluded from this study [71–76] Secondly, all categorical variables should be converted into “one-hot-encoded” format before partitioning, so that all hold-out samples could be evaluated against a sample from which they were selected. Thirdly, an objective

 doi:10.20944/preprints202603.1207.v1

evaluation of how the model predicted outcomes relative to logistic regression models using just ISUP and Briganti should be conducted. Fourthly, the recruitment of additional subjects with Low-risk could provide opportunities to validate the findings of the D'Amico classification model for different levels of risk. Fifthly, methods to quantify uncertainty around how each predictor impacts prediction should be developed using either bootstrap resampling methods or SHAP-based methods. Lastly, multiple random seed-based analyses should be conducted to provide confidence intervals around the performance metrics.

4.6. Clinical Translation: A Cautionary Framework, Not a Deployable Tool

The restrictions mentioned above show that this model isn't currently prepared for, and may never be prepared for, direct patient contact. The methodology developed in this research provides a guide to developing responsible approaches to moving AI technologies from proof-of-concept to clinical application. This guide also serves as a formal benchmark for assessing subsequent research in this area.

Before any application of clinical information could be made, a stable governance framework with one or more models must have been validated by large cohorts from multiple institutions reflecting the whole spectrum of the disease including phenotypes that were systematically excluded as well as clinical outliers [71–76]. Such validation requires either prospective studies or carefully designed retrospective studies that are designed using a variety of electronic health record (EHR) systems and that have pre-registered analyses. In addition, performance deteriorating across subgroups should be reported transparently [72,73].

Should its validation proceed as anticipated, the MLP is a candidate for use as a deployment model. When generalizability has been established through future work, the MLP could serve as a "digital second opinion" during multi-disciplinary tumor case review meetings; meaning it would be complementary to, not a replacement for, clinical judgment. Specifically, the MLP would be able to provide guidance in those clinical situations when there is uncertainty surrounding appropriate management of patients (e.g., "the gray zone" of clinical decision-making, as described by D'Amico and Briganti) [92–94]. The clinician would enter 11 variables routinely collected at the time of diagnosis into the MLP; in response, the MLP would generate a probability-based risk classification that may, for instance, identify an apparent high-risk patient whose non-linear signal indicates a biologically less aggressive cancer (thus avoiding overtreatment) or an apparent intermediate-risk patient with occult high-risk cancer features (thus triggering further staging with PSMA-PET). The attached PMML/XML documentation [95–97] provides the means to hypothetically integrate the MLP into current decision support systems without the need for proprietary software.

Given the importance of these aspects in future deployments, the additional protection elements that must be included are: 1) on-going monitoring for data drift and spectrum bias [76]; 2) including explicit confidence intervals associated with each prediction so uncertainty is communicated; 3) inclusion of a human oversight requirement such that the model is only an advisory tool; and, 4) regular re-evaluation as the population and clinical practices continue to change.

There is no current evidence to suggest that this model meets any of the above elements. The only usefulness of this work is to show how to create and govern datasets from which we may someday derive clinically useful models when we have a sufficient total number of satisfactory samples and when those samples represent our population. Only the framework for developing and supporting the datasets will require external validation.

3. Conclusions

This study shows that strict, FAIR-based data quality governance can produce a clear, comprehensible clinical signal from a small cohort, allowing algorithmic development. But it also serves as a sobering warning: the same governance that generates this clean signal can also produce a validation sample that is so carefully selected that performance metrics (like the 100% accuracy reported here) become clinically meaningless but mathematically perfect due to necessary but strict

 doi:10.20944/preprints202603.1207.v1

exclusion criteria. Therefore, this work's main contribution is a transparent and repeatable methodological framework rather than a deployable model. This framework's performance on large, unselected, multi-institutional populations that represent the complete heterogeneity of clinical practice will be the real test of any model it generates.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/doi/s1,. File S1: Artificial Neural Network architecture in PMML/XML format for the 39/61 data partitioning scheme. File S2: Artificial Neural Network architecture in PMML/XML format for the 34/66 data partitioning scheme. File S3: Artificial Neural Network architecture in PMML/XML format for the 20/80 data partitioning scheme. File S4: Full SPSS Output Report for the 20/80 Model, containing the original execution logs, comprehensive classification matrices, and raw neural weight estimates. File S5: Full SPSS Output Report for the 34/66 Model, documenting the multi-category performance and factor level exclusions. File S6: Full SPSS Output Report for the 39/61 Model, providing the complete model summary and variable importance data from the software execution.

Author Contributions: Conceptualization, V.T.C., J.E.R.G. and F.G.G.; methodology, J.E.R.G. and F.G.G.; software, F.G.G. and L.G.A.; validation, V.T.C., J.E.R.G. and F.G.G.; formal analysis, V.T.C., J.E.R.G., F.G.G., A.C.L., M.T.T., L.L.H., F.J.A.M., L.G.A., D.S.Z. and B.M.L.; investigation, V.T.C., J.E.R.G., F.G.G., A.C.L., M.T.T., L.L.H., F.J.A.M., L.G.A., D.S.Z. and B.M.L.; writing—original draft preparation, V.T.C., J.E.R.G. and F.G.G.; writing—review and editing, V.T.C., J.E.R.G., F.G.G., A.C.L., M.T.T., L.L.H., F.J.A.M., L.G.A., D.S.Z. and B.M.L.; visualization, F.G.G. and L.G.A.; project administration, J.E.R.G.; supervision, J.E.R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was carried out following the Declaration of Helsinki and it received official approval from the Ethics Committee of the Clínica Universidad de Navarra on January 22, 2025 (project number: 2025.006, Approval Date: 2 February 2025).

Informed Consent Statement: Patient informed consent was waived since it was not required by the Ethics Committee of Clínica Universidad de Navarra, due to the retrospective nature of this study. All data were anonymized completely before any analysis took place.

Data Availability Statement: The original contributions presented in this study are in in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: During preparation of this manuscript, the author(s) used Grammarly for the purpose of correcting grammar and improving the flow. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
AUC	Area Under the Curve
cT	Clinical tumor stage
DQG	Data Quality Governance
EHR	Electronic Health Record
FAIR	Findability, Accessibility, Interoperability, and Reusability
ISUP	International Society of Urological Pathology

 doi:10.20944/preprints202603.1207.v1

mrN	Multiparametric imaging nodal stage (or Imaging nodal stage)
mrT	Multiparametric imaging tumor stage (or Imaging tumor stage)
MLP	Multilayer Perceptron
mpMRI	Multiparametric Magnetic Resonance Imaging
NPV	Negative Predictive Value
PCa	Prostate Cancer
PMML	Predictive Model Markup Language
PPV	Positive Predictive Value
PSA	Prostate-Specific Antigen
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SHAP	SHapley Additive exPlanations
SOP	Standard Operating Procedure
tanh	Hyperbolic Tangent
XML	Extensible Markup Language

Appendix A: AI-Readiness and Data Quality Governance (DQG) Checklist Corresponding to the Study

Title: Standard Operating Procedure (SOP) for Clinical Data Curation in Artificial Neural Network Development.

This checklist outlines the minimum requirements used in this study to ensure that the dataset meets the criteria for "AI-Readiness" before model training.

1. Data Integrity and Source Governance

[ ] Source Verification (Blinded): Every data point (PSA, ISUP, mrTNM) was cross-referenced between the Electronic Health Record (EHR) and the study database by two independent reviewers blinded to each other's assessments.

[ ] Consensus Reconciliation: A formal reconciliation process was used to resolve any inter-rater discrepancies, ensuring 100% clinical concordance for all primary predictors before model training.

[ ] FAIR Compliance: Data fields are mapped to standard clinical vocabularies (e.g., TNM 8th Ed, ISUP 2014) to ensure interoperability and future reuse.

[ ] Anonymization: All identifiers have been removed in accordance with GDPR/Institutional Review Board standards prior to algorithmic processing.

2. Operational Data Quality Rules (The "Technical Gatekeeper")

[ ] Range Constraint Validation: Continuous variables must fall within biological plausibility limits (e.g., PSA >0.1; Age 40–90). Values outside these ranges are flagged for manual re-verification or exclusion.

[ ] Logical Consistency Checks: Staging hierarchies are preserved (e.g., a patient cannot be miN1 without clinical evidence of nodal involvement or suspicious imaging).

[ ] Missingness Protocol: A strict "complete-case analysis" was applied. Any record with missing data in any of the 11 primary clinical predictors was excluded to prevent "noise" in the backpropagation process.

3. Pre-Modeling "AI-Readiness" Transitions

[ ] Feature Encoding: All categorical variables (Factors) must use 1-of-c encoding (one-hot encoding) to ensure the network can mathematically distinguish between distinct clinical strata.

 doi:10.20944/preprints202603.1207.v1

[ ] Scaling & Normalization: Continuous covariates are rescaled using Linear Normalization (Min-Max) to prevent gradient vanishing or weight saturation in the hidden layer's activation functions (tanh).

[ ] Outlier Management: "Clinical outliers" (rare factor combinations) are documented and reported transparently, even if excluded by the software's automatic validation protocol.

4. Evaluation & Stability Governance

[ ] Random Seed Fixation: A fixed random seed (e.g., 2,000,000) is used to ensure all training runs are reproducible by third parties.

[ ] Sensitivity Analysis: Model performance is evaluated across multiple data partitioning schemes (e.g., 80/20, 50/50) to ensure the reported accuracy is not an artifact of a single "lucky" split.

[ ] Uncertainty Reporting: Results are reported as distributions (Mean ± SD) rather than single peak-performance percentages.

References

1. Schafer, E. J.; Laversanne, M.; Sung, H.; Soerjomataram, I.; Briganti, A.; Dahut, W.; Bray, F.; Jemal, A. Recent Patterns and Trends in Global Prostate Cancer Incidence and Mortality: An Update. *Eur. Urol.* **2025**, *87* (3), 302–313. https://doi.org/10.1016/j.eururo.2024.11.013.
2. Wang, L.; Lu, B.; He, M.; Wang, Y.; Wang, Z.; Du, L. Prostate Cancer Incidence and Mortality: Global Status and Temporal Trends in 89 Countries From 2000 to 2019. *Front. Public Health* **2022**, *10*, 811044. https://doi.org/10.3389/fpubh.2022.811044.
3. Chu, F.; Chen, L.; Guan, Q.; Chen, Z.; Ji, Q.; Ma, Y.; Ji, J.; Sun, M.; Huang, T.; Song, H.; Zhou, H.; Lin, X.; Zheng, Y. Global Burden of Prostate Cancer: Age-Period-Cohort Analysis from 1990 to 2021 and Projections until 2040. *World J. Surg. Oncol.* **2025**, *23* (1), 98. https://doi.org/10.1186/s12957-025-03733-1.
4. Kratzer, T. B.; Mazzitelli, N.; Star, J.; Dahut, W. L.; Jemal, A.; Siegel, R. L. Prostate Cancer Statistics, 2025. *CA Cancer J. Clin.* **2025**, *75* (6), 485–497. https://doi.org/10.3322/caac.70028.
5. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R. L.; Soerjomataram, I.; Jemal, A. Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2024**, *74* (3), 229–263. https://doi.org/10.3322/caac.21834.
6. Cornford, P.; Tilki, D.; van den Berghn, R.; Eberlin, D. *Prostate Cancer. EAU Guidelines*; EAU Guidelines Office: Arnhem, 2025.
7. Hassanipour-Azgomi, S.; Mohammadian-Hafshejani, A.; Ghoncheh, M.; Towhidi, F.; Jamehshorani, S.; Salehiniya, H. Incidence and Mortality of Prostate Cancer and Their Relationship with the Human Development Index Worldwide. *Prostate Int.* **2016**, *4* (3), 118–124. https://doi.org/10.1016/j.prnil.2016.07.001.
8. Dudipala, H.; Jani, C. T.; Gurnani, S. D.; Morgenstern-Kaplan, D.; Tran, E.; Edwards, K.; Shalhoub, J.; McKay, R. R. Evolving Global Trends in Prostate Cancer: Disparities Across Income Levels and Geographic Regions (1990-2019). *JCO Glob. Oncol.* **2025**, No. 11. https://doi.org/10.1200/GO-25-00249.
9. Huang, J.; Ssentongo, P.; Sharma, R. Editorial: Cancer Burden, Prevention and Treatment in Developing Countries. *Front. Public Health* **2023**, *10*. https://doi.org/10.3389/fpubh.2022.1124473.
10. Brand, N. R.; Qu, L. G.; Chao, A.; Ilbawi, A. M. Delays and Barriers to Cancer Care in Low- and Middle-Income Countries: A Systematic Review. *Oncologist* **2019**, *24* (12), e1371–e1380. https://doi.org/10.1634/theoncologist.2019-0057.
11. Daniels, J.; Mosadi, L. E.; Nyantakyi, A. Y.; Ayabilah, E. A.; Tackie, J. N. O.; Kyei, K. A. Metastatic Breast Cancer in Resource-Limited Settings: Insights from a Retrospective Cross-Sectional Study at a Radiotherapy Centre in Sub-Saharan Africa. *Ecancermedicalscience* **2025**, *19*. https://doi.org/10.3332/ecancer.2025.1955.
12. Cazap, E.; Magrath, I.; Kingham, T. P.; Elzawawy, A. Structural Barriers to Diagnosis and Treatment of Cancer in Low- and Middle-Income Countries: The Urgent Need for Scaling Up. *Journal of Clinical Oncology* **2016**, *34* (1), 14–19. https://doi.org/10.1200/JCO.2015.61.9189.
13. Morsi, M. H.; Elawfi, B.; ALsaad, S. A.; Nazar, A.; Mostafa, H. A.; Awwad, S. A.; Abdelwahab, M. M.; Tarakhan, H.; Baghagho, E. Unveiling the Disparities in the Field of Precision Medicine: A Perspective. *Health Sci. Rep.* **2025**, *8* (8). https://doi.org/10.1002/hsr2.71102.

 doi:10.20944/preprints202603.1207.v1

14. Zhou, S.; Feng, X.; Hu, Y.; Yang, J.; Chen, Y.; Bastow, J.; Zheng, Z.-J.; Xu, M. Factors Associated with the Utilization of Diagnostic Tools among Countries with Different Income Levels during the COVID-19 Pandemic. *Glob. Health Res. Policy* **2023**, *8* (1), 45. https://doi.org/10.1186/s41256-023-00330-1.
15. Fu, E.; Yager, P.; Floriano, P. N.; Christodoulides, N.; McDevitt, J. T. Perspective on Diagnostics for Global Health. *IEEE Pulse* **2011**, *2* (6), 40–50. https://doi.org/10.1109/MPUL.2011.942766.
16. Hubley, J. H. Barriers to Health Education in Developing Countries. *Health Educ. Res.* **1986**, *1* (4), 233–245. https://doi.org/10.1093/her/1.4.233.
17. Olah, C.; Mairinger, F.; Wessolly, M.; Joniau, S.; Spahn, M.; Kruithof-de Julio, M.; Hadaschik, B.; Soós, A.; Nyirády, P.; Győrffy, B.; Reis, H.; Szarvas, T. Enhancing Risk Stratification Models in Localized Prostate Cancer by Novel Validated Tissue Biomarkers. *Prostate Cancer Prostatic Dis.* **2025**, *28* (3), 773–781. https://doi.org/10.1038/s41391-024-00918-9.
18. D'Amico, AV.; Whittington, R.; Malkowicz, S.; Schulz, D.; Blank, K.; Broderick, G.; Tomaszewski, J.; Renshaw, A.; Kaplan, I.; Beard, C.; Wein, A. D'Amico risk classification for prostate cancer (Version: 1.28) - Evidencio https://www.evidencio.com/models/show/300 (accessed 2026 -02 -10).
19. Hernandez, D. J.; Nielsen, M. E.; Han, M.; Partin, A. W. Contemporary Evaluation of the D'Amico Risk Classification of Prostate Cancer. *Urology* **2007**, *70* (5), 931–935. https://doi.org/10.1016/j.urology.2007.08.055.
20. Cooperberg, M. R. Clinical Risk Stratification for Prostate Cancer: Where Are We, and Where Do We Need to Go? *Canadian Urological Association Journal* **2017**, *11* (3–4), 101. https://doi.org/10.5489/cuaj.4520.
21. Rastinehad, A. R.; Baccala, A. A.; Chung, P. H.; Proano, J. M.; Kruecker, J.; Xu, S.; Locklin, J. K.; Turkbey, B.; Shih, J.; Bratslavsky, G.; Linehan, W. M.; Glossop, N. D.; Yan, P.; Kadoury, S.; Choyke, P. L.; Wood, B. J.; Pinto, P. A. D'Amico Risk Stratification Correlates With Degree of Suspicion of Prostate Cancer on Multiparametric Magnetic Resonance Imaging. *Journal of Urology* **2011**, *185* (3), 815–820. https://doi.org/10.1016/j.juro.2010.10.076.
22. Chierigo, F.; Flammia, R. S.; Sorce, G.; Hoeh, B.; Hohenhorst, L.; Tian, Z.; Saad, F.; Gallucci, M.; Briganti, A.; Montorsi, F.; Chun, F. K. H.; Graefen, M.; Shariat, S. F.; Guano, G.; Mantica, G.; Borghesi, M.; Suardi, N.; Terrone, C.; Karakiewicz, P. I. The Association of the Type and Number of D'Amico High-Risk Criteria with Rates of Pathologically Non-Organ-Confined Prostate Cancer. *Cent. European J. Urol.* **2023**, *76* (2), 104–108. https://doi.org/10.5173/ceju.2023.030.
23. Coetzer, M. How AI Ready Data Strengthens Clinical Analytics and Outcomes Research https://imatsolutions.com/index.php/2025/12/the-role-of-ai-ready-data-in-strengthening-clinical-research-predictive-modeling-and-care-quality/ (accessed 2026 -02 -10).
24. Domagalski, M. J.; Lu, Y.; Pilozzi, A.; Williamson, A.; Chilappagari, P.; Luker, E.; Shelley, C. D.; Dabic, A.; Keller, M. A.; Rodriguez, R. M.; Lawlor, S.; Thangudu, R. R. Preparing Clinical Research Data for Artificial Intelligence Readiness: Insights from the National Institute of Diabetes and Digestive and Kidney Diseases Data Centric Challenge. *Journal of the American Medical Informatics Association* **2025**, *32* (10), 1609–1616. https://doi.org/10.1093/jamia/ocaf114.
25. Bönisch, C.; Schmidt, C.; Kesztyüs, D.; Kestler, H. A.; Kesztyüs, T. Proposal for Using AI to Assess Clinical Data Integrity and Generate Metadata: Algorithm Development and Validation. *JMIR Med. Inform.* **2025**, *13*, e60204–e60204. https://doi.org/10.2196/60204.
26. Wichard, J. D.; Cammann, H.; Stephan, C.; Tolxdorff, T. Classification Models for Early Detection of Prostate Cancer. *Biomed Res. Int.* **2008**, *2008* (1). https://doi.org/10.1155/2008/218097.
27. Chen, S.; Jian, T.; Chi, C.; Liang, Y.; Liang, X.; Yu, Y.; Jiang, F.; Lu, J. Machine Learning-Based Models Enhance the Prediction of Prostate Cancer. *Front. Oncol.* **2022**, *12*. https://doi.org/10.3389/fonc.2022.941349.
28. Elmarakeby, H. A.; Hwang, J.; Arafeh, R.; Crowdis, J.; Gang, S.; Liu, D.; AlDubayan, S. H.; Salari, K.; Kregel, S.; Richter, C.; Arnoff, T. E.; Park, J.; Hahn, W. C.; Van Allen, E. M. Biologically Informed Deep Neural Network for Prostate Cancer Discovery. *Nature* **2021**, *598* (7880), 348–352. https://doi.org/10.1038/s41586-021-03922-4.
29. Shanei, A.; Etehadtavakol, M.; Azizian, M.; Ng, E. Y. K. Comparison of Different Kernels in a Support Vector Machine to Classify Prostate Cancerous Tissues in T2-Weighted Magnetic Resonance Imaging. *Explor. Res. Hypothesis Med.* **2022**, *000* (000), 000–000. https://doi.org/10.14218/ERHM.2022.00013.

 doi:10.20944/preprints202603.1207.v1

30. Chiu, P. K.-F.; Shen, X.; Wang, G.; Ho, C.-L.; Leung, C.-H.; Ng, C.-F.; Choi, K.-S.; Teoh, J. Y.-C. Enhancement of Prostate Cancer Diagnosis by Machine Learning Techniques: An Algorithm Development and Validation Study. *Prostate Cancer Prostatic Dis.* **2022**, *25* (4), 672–676. https://doi.org/10.1038/s41391-021-00429-x.
31. Singh, S.; Pathak, A. K.; Kural, S.; Kumar, L.; Bhardwaj, M. G.; Yadav, M.; Trivedi, S.; Das, P.; Gupta, M.; Jain, G. Integrating MiRNA Profiling and Machine Learning for Improved Prostate Cancer Diagnosis. *Sci. Rep.* **2025**, *15* (1), 30477. https://doi.org/10.1038/s41598-025-99754-7.
32. Jiang, M.; Miao, Z.; Xu, R.; Guo, M.; Li, X.; Li, G.; Luo, P.; Hu, S. Clinical-Radiomics Hybrid Modeling Outperforms Conventional Models: Machine Learning Enhances Stratification of Adverse Prognostic Features in Prostate Cancer. *Front. Oncol.* **2025**, *15*. https://doi.org/10.3389/fonc.2025.1625158.
33. Zamo, F. C. D.; Ebongue, A. N.; Bongue, D.; Ndontchueng, M. M.; Njeh, C. F. Classification of PSQA Outcomes in Prostate VMAT Treatments: A Comparative Study of Machine Learning Models. *Biomedical Engineering Advances* **2026**, *11*, 100206. https://doi.org/10.1016/j.bea.2026.100206.
34. Nieboer, D.; Vergouwe, Y.; Roobol, M. J.; Ankerst, D. P.; Kattan, M. W.; Vickers, A. J.; Steyerberg, E. W. Nonlinear Modeling Was Applied Thoughtfully for Risk Prediction: The Prostate Biopsy Collaborative Group. *J. Clin. Epidemiol.* **2015**, *68* (4), 426–434. https://doi.org/10.1016/j.jclinepi.2014.11.022.
35. Zhao, Y.; Zhang, L.; Zhang, S.; Li, J.; Shi, K.; Yao, D.; Li, Q.; Zhang, T.; Xu, L.; Geng, L.; Sun, Y.; Wan, J. Machine Learning-Based MRI Imaging for Prostate Cancer Diagnosis: Systematic Review and Meta-Analysis. *Prostate Cancer Prostatic Dis.* **2025**. https://doi.org/10.1038/s41391-025-00997-2.
36. Morote, J.; Miró, B.; Hernando, P.; Paesano, N.; Picola, N.; Muñoz-Rodriguez, J.; Ruiz-Plazas, X.; Muñoz-Rivero, M. V.; Celma, A.; García-de Manuel, G.; Servian, P.; Abascal, J. M.; Trilla, E.; Méndez, O. Developing a Predictive Model for Significant Prostate Cancer Detection in Prostatic Biopsies from Seven Clinical Variables: Is Machine Learning Superior to Logistic Regression? *Cancers (Basel).* **2025**, *17* (7), 1101. https://doi.org/10.3390/cancers17071101.
37. Zhang, Y.-D.; Wang, J.; Wu, C.-J.; Bao, M.-L.; Li, H.; Wang, X.-N.; Tao, J.; Shi, H.-B. An Imaging-Based Approach Predicts Clinical Outcomes in Prostate Cancer through a Novel Support Vector Machine Classification. *Oncotarget* **2016**, *7* (47), 78140–78151. https://doi.org/10.18632/oncotarget.11293.
38. Siech, C.; Wenzel, M.; Grosshans, N.; Cano Garcia, C.; Humke, C.; Koll, F. J.; Tian, Z.; Karakiewicz, P. I.; Kluth, L. A.; Chun, F. K. H.; Hoeh, B.; Mandel, P. The Association Between Lymphovascular or Perineural Invasion in Radical Prostatectomy Specimen and Biochemical Recurrence. *Cancers (Basel).* **2024**, *16* (21), 3648. https://doi.org/10.3390/cancers16213648.
39. Chung, D. H.; Han, J. H.; Jeong, S.-H.; Yuk, H. D.; Jeong, C. W.; Ku, J. H.; Kwak, C. Role of Lymphatic Invasion in Predicting Biochemical Recurrence after Radical Prostatectomy. *Front. Oncol.* **2023**, *13*. https://doi.org/10.3389/fonc.2023.1226366.
40. Fajkovic, H.; Mathieu, R.; Lucca, I.; Hiess, M.; Hübner, N.; Al Awamlh, B. A. H.; Lee, R.; Briganti, A.; Karakiewicz, P.; Lotan, Y.; Roupret, M.; Rink, M.; Kluth, L.; Loidl, W.; Seitz, C.; Klatte, T.; Kramer, G.; Susani, M.; Shariat, S. F. Validation of Lymphovascular Invasion Is an Independent Prognostic Factor for Biochemical Recurrence after Radical Prostatectomy. *Urologic Oncology: Seminars and Original Investigations* **2016**, *34* (5), 233.e1-233.e6. https://doi.org/10.1016/j.urolonc.2015.10.013.
41. Karwacki, J.; Stodolak, M.; Dłubak, A.; Nowak, Ł.; Gurwin, A.; Kowalczyk, K.; Kiełb, P.; Holdun, N.; Szlasa, W.; Krajewski, W.; Hałoń, A.; Karwacka, A.; Szydełko, T.; Małkiewicz, B. Association of Lymphovascular Invasion with Biochemical Recurrence and Adverse Pathological Characteristics of Prostate Cancer: A Systematic Review and Meta-Analysis. *Eur. Urol. Open Sci.* **2024**, *69*, 112–126. https://doi.org/10.1016/j.euros.2024.09.007.
42. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **2015**, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.
43. Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J. A. W. M.; van Ginneken, B.; Sánchez, C. I. A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* **2017**, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005.

 doi:10.20944/preprints202603.1207.v1

44. Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D. S. W.; Karthikesalingam, A.; King, D.; Ashrafian, H.; Darzi, A. Diagnostic Accuracy of Deep Learning in Medical Imaging: A Systematic Review and Meta-Analysis. *NPJ Digit. Med.* **2021**, *4* (1), 65. https://doi.org/10.1038/s41746-021-00438-z.
45. Erdem, E.; Bozkurt, F. A Comparison of Various Supervised Machine Learning Techniques for Prostate Cancer Prediction . *European Journal of Science and Technology* **2021**, No. 21, 610–620. https://doi.org/10.31590/ejosat.802810.
46. Shu, X.; Liu, Y.; Qiao, X.; Ai, G.; Liu, L.; Liao, J.; Deng, Z.; He, X. Radiomic-Based Machine Learning Model for the Accurate Prediction of Prostate Cancer Risk Stratification. *Br. J. Radiol.* **2023**, *96* (1143). https://doi.org/10.1259/bjr.20220238.
47. Gupta, S.; Kumar, M. Prostate Cancer Prognosis Using Multi-Layer Perceptron and Class Balancing Techniques. In *2021 Thirteenth International Conference on Contemporary Computing (IC3-2021)*; ACM: New York, NY, USA, 2021; pp 1–6. https://doi.org/10.1145/3474124.3474125.
48. Yang, T.; Zhang, H.; Peng, H.; Niu, X.; Yang, F.; Zhang, J.; Wang, Q.; Fan, J.; Song, Y.; Tao, W. PSMA PET/MRI-based Swin Transformer Architecture for Gleason Score Prediction in Prostate Cancer. *Med. Phys.* **2026**, *53* (1). https://doi.org/10.1002/mp.70274.
49. Nguyen, C.; Hulsey, G.; James, K.; James, T.; Carlson, J. M. Deep Learning Classification of Prostate Cancer Using MRI Histopathologic Data. *Radiol. Imaging Cancer* **2025**, *7* (5). https://doi.org/10.1148/rycan.240381.
50. Li, D.; Han, X.; Gao, J.; Zhang, Q.; Yang, H.; Liao, S.; Guo, H.; Zhang, B. Deep Learning in Prostate Cancer Diagnosis Using Multiparametric Magnetic Resonance Imaging With Whole-Mount Histopathology Referenced Delineations. *Front. Med. (Lausanne).* **2022**, *8*. https://doi.org/10.3389/fmed.2021.810995.
51. He, M.; Cao, Y.; Chi, C.; Yang, X.; Ramin, R.; Wang, S.; Yang, G.; Mukhtorov, O.; Zhang, L.; Kazantsev, A.; Enikeev, M.; Hu, K. Research Progress on Deep Learning in Magnetic Resonance Imaging–Based Diagnosis and Treatment of Prostate Cancer: A Review on the Current Status and Perspectives. *Front. Oncol.* **2023**, *13*. https://doi.org/10.3389/fonc.2023.1189370.
52. Quarta, L.; Stabile, A.; Pellegrino, F.; Scilipoti, P.; Longoni, M.; Cannoletta, D.; Zaurito, P.; Santangelo, A.; Viti, A.; Barletta, F.; Scuderi, S.; Leni, R.; Pellegrino, A.; Mazzone, E.; Nocera, L.; Brembilla, G.; De Cobelli, F.; Karnes, R. J.; Rouprêt, M.; Montorsi, F.; Gandaglia, G.; Briganti, A. Tailored Use of PSA Density According to Multiparametric MRI Index Lesion Location: Results of a Large, Multi-Institutional Series. *Prostate Cancer Prostatic Dis.* **2025**. https://doi.org/10.1038/s41391-025-00987-4.
53. Ciobanu-Caraus, O.; Aicher, A.; Kernbach, J. M.; Regli, L.; Serra, C.; Staartjes, V. E. A Critical Moment in Machine Learning in Medicine: On Reproducible and Interpretable Learning. *Acta Neurochir. (Wien).* **2024**, *166* (1), 14. https://doi.org/10.1007/s00701-024-05892-8.
54. McDermott, M. B. A.; Wang, S.; Marinsek, N.; Ranganath, R.; Foschini, L.; Ghassemi, M. Reproducibility in Machine Learning for Health Research: Still a Ways to Go. *Sci. Transl. Med.* **2021**, *13* (586). https://doi.org/10.1126/scitranslmed.abb1655.
55. Guillen-Aguinaga, M.; Aguinaga-Ontoso, E.; Guillen-Aguinaga, L.; Guillen-Grima, F.; Aguinaga-Ontoso, I. Data Quality in the Age of AI: A Review of Governance, Ethics, and the FAIR Principles. *Data (Basel).* **2025**, *10* (12), 201. https://doi.org/10.3390/data10120201.
56. Clark, T.; Caufield, H.; Parker, J. A.; Al Manir, S.; Amorim, E.; Eddy, J.; Gim, N.; Gow, B.; Goar, W.; Haendel, M.; Hansen, J. N.; Harris, N.; Hermjakob, H.; Joachimiak, M.; Jordan, G.; Lee, I.-H.; K. McWeeney, S.; Nebeker, C.; Nikolov, M.; Shaffer, J.; Sheffield, N.; Sheynkman, G.; Stevenson, J.; Chen, J. Y.; Mungall, C.; Wagner, A.; Kong, S. W.; Ghosh, S. S.; Patel, B.; Williams, A.; Munoz-Torres, M. C. AI-Readiness for Biomedical Data: Bridge2AI Recommendations. October 25, 2024. https://doi.org/10.1101/2024.10.23.619844.
57. Aksenova, A.; Johny, A.; Adams, T.; Gribbon, P.; Jacobs, M.; Hofmann-Apitius, M. Current State of Data Stewardship Tools in Life Science. *Front. Big Data* **2024**, *7*. https://doi.org/10.3389/fdata.2024.1428568.
58. D'Amico, A. V. Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer. *JAMA* **1998**, *280* (11), 969. https://doi.org/10.1001/jama.280.11.969.
59. Diamand, R.; Oderda, M.; Albisinni, S.; Fourcade, A.; Fournier, G.; Benamran, D.; Iselin, C.; Fiard, G.; Descotes, J.-L.; Assenmacher, G.; Svistakov, I.; Peltier, A.; Simone, G.; Di Cosmo, G.; Roche, J.-B.; Bonnal,

 doi:10.20944/preprints202603.1207.v1

J.-L.; Van Damme, J.; Rossi, M.; Mandron, E.; Gontero, P.; Roumeguère, T. External Validation of the Briganti Nomogram Predicting Lymph Node Invasion in Patients with Intermediate and High-Risk Prostate Cancer Diagnosed with Magnetic Resonance Imaging-Targeted and Systematic Biopsies: A European Multicenter Study. *Urologic Oncology: Seminars and Original Investigations* **2020**, *38* (11), 847.e9-847.e16. https://doi.org/10.1016/j.urolonc.2020.04.011.

60. Hansen, J.; Rink, M.; Bianchi, M.; Kluth, L. A.; Tian, Z.; Ahyai, S. A.; Shariat, S. F.; Briganti, A.; Steuber, T.; Fisch, M.; Graefen, M.; Karakiewicz, P. I.; Chun, F. K. -H. External Validation of the Updated Briganti Nomogram to Predict Lymph Node Invasion in Prostate Cancer Patients Undergoing Extended Lymph Node Dissection. *Prostate* **2013**, *73* (2), 211–218. https://doi.org/10.1002/pros.22559.
61. Małkiewicz, B.; Ptaszkowski, K.; Knecht, K.; Gurwin, A.; Wilk, K.; Kiełb, P.; Dudek, K.; Zdrojowy, R. External Validation of the Briganti Nomogram to Predict Lymph Node Invasion in Prostate Cancer—Setting a New Threshold Value. . **2021**, *11* (6), 479. https://doi.org/10.3390/life11060479.
62. International Organization for Standardization. ISO/IEC 42001:2023 Information Technology — Artificial Intelligence — Management System; Geneva, 2023.
63. Wang, R. Y.; Strong, D. M. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* **1996**, *12* (4), 5–33. https://doi.org/10.1080/07421222.1996.11518099.
64. Faul, F.; Erdfelder, E.; Buchner, A.; Lang, A.-G. Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behav. Res. Methods* **2009**, *41* (4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.
65. Faul, F.; Erdfelder, E.; Lang, A.-G.; Buchner, A. G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behav. Res. Methods* **2007**, *39* (2), 175–191. https://doi.org/10.3758/BF03193146.
66. Soe MM. Dean AG, S. K. M. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Versión. 2013.
67. Cytel Inc. LogXact-11: Software for Exact Logistic Regression, A, 2015. Cytel Inc: Waltham, MA, US 2015.
68. Sur, P.; Candès, E. J. A Modern Maximum-Likelihood Theory for High-Dimensional Logistic Regression. *Proceedings of the National Academy of Sciences* **2019**, *116* (29), 14516–14525. https://doi.org/10.1073/pnas.1810420116.
69. Yee, T. W. On the Hauck-Donner Effect in Wald Tests: Detection, Tipping Points, and Parameter Space Characterization. **2022**. https://doi.org/10.1080/01621459.2021.1886936.
70. Hauck, W. W.; Donner, A. Wald's Test as Applied to Hypotheses in Logit Analysis. *J. Am. Stat. Assoc.* **1977**, *72* (360a), 851–853. https://doi.org/10.1080/01621459.1977.10479969.
71. Van Calster, B.; Collins, G. S.; Vickers, A. J.; Wynants, L.; Kerr, K. F.; Barreñada, L.; Varoquaux, G.; Singh, K.; Moons, K. G.; Hernandez-Boussard, T.; Timmerman, D.; McLernon, D. J.; van Smeden, M.; Steyerberg, E. W. Evaluation of Performance Measures in Predictive Artificial Intelligence Models to Support Medical Decisions: Overview and Guidance. *Lancet Digit. Health* **2025**, *7* (12), 100916. https://doi.org/10.1016/j.landig.2025.100916.
72. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A. I.; Etmann, C.; McCague, C.; Beer, L.; Weir-McCall, J. R.; Teng, Z.; Gkrania-Klotsas, E.; Ruggiero, A.; Korhonen, A.; Jefferson, E.; Ako, E.; Langs, G.; Gozaliasl, G.; Yang, G.; Prosch, H.; Preller, J.; Stanczuk, J.; Tang, J.; Hofmanninger, J.; Babar, J.; Sánchez, L. E.; Thillai, M.; Gonzalez, P. M.; Teare, P.; Zhu, X.; Patel, M.; Cafolla, C.; Azadbakht, H.; Jacob, J.; Lowe, J.; Zhang, K.; Bradley, K.; Wassin, M.; Holzer, M.; Ji, K.; Ortet, M. D.; Ai, T.; Walton, N.; Lio, P.; Stranks, S.; Shadbahr, T.; Lin, W.; Zha, Y.; Niu, Z.; Rudd, J. H. F.; Sala, E.; Schönlieb, C.-B. Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans. *Nat. Mach. Intell.* **2021**, *3* (3), 199–217. https://doi.org/10.1038/s42256-021-00307-0.
73. Varoquaux, G.; Cheplygina, V. Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future. *NPJ Digit. Med.* **2022**, *5* (1), 48. https://doi.org/10.1038/s41746-022-00592-y.
74. Yu, A. C.; Mohajer, B.; Eng, J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. *Radiol. Artif. Intell.* **2022**, *4* (3). https://doi.org/10.1148/ryai.210064.

 doi:10.20944/preprints202603.1207.v1

75. Arun, S.; Grosheva, M.; Kosenko, M.; Robertus, J. L.; Blyuss, O.; Gabe, R.; Munblit, D.; Offman, J. Systematic Scoping Review of External Validation Studies of AI Pathology Models for Lung Cancer Diagnosis. *NPJ Precis. Oncol.* **2025**, *9* (1), 166. https://doi.org/10.1038/s41698-025-00940-7.
76. Tseng, A. S.; Shelly-Cohen, M.; Attia, I. Z.; Noseworthy, P. A.; Friedman, P. A.; Oh, J. K.; Lopez-Jimenez, F. Spectrum Bias in Algorithms Derived by Artificial Intelligence: A Case Study in Detecting Aortic Stenosis Using Electrocardiograms. *European Heart Journal - Digital Health* **2021**, *2* (4), 561–567. https://doi.org/10.1093/ehjdh/ztab061.
77. Yu, L.; Gao, X.-S.; Zhang, L.; Miao, Y. Generalizability of Memorization Neural Networks. **2024**.
78. Zhang, C.; Bengio, S.; Hardt, M.; Mozer, M. C.; Singer, Y. Identity Crisis: Memorization and Generalization under Extreme Overparameterization. **2020**.
79. Belkin, M.; Hsu, D.; Ma, S.; Mandal, S. Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-Off. *Proceedings of the National Academy of Sciences* **2019**, *116* (32), 15849–15854. https://doi.org/10.1073/pnas.1903070116.
80. Hastie, T.; Tibshiran, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction., 2nd ed.; Springer: New York, 2017.
81. Crespo Márquez, A. The Curse of Dimensionality. In *Digital Maintenance Management*; Springer Science and Business Media Deutschland GmbH: New York, 2022; pp 67–86. https://doi.org/10.1007/978-3-030-97660-6_7.
82. Tishby, N.; Zaslavsky, N. Deep Learning and the Information Bottleneck Principle. In *2015 IEEE Information Theory Workshop (ITW)*; IEEE, 2015; pp 1–5. https://doi.org/10.1109/ITW.2015.7133169.
83. Koch, R. de M.; Ghosh, A. Two-Phase Perspective on Deep Learning Dynamics. *Phys. Rev. E* **2025**, *112* (2), 025307. https://doi.org/10.1103/3l3p-vkf2.
84. Chapman, B. P.; Weiss, A.; Duberstein, P. R. Statistical Learning Theory for High Dimensional Prediction: Application to Criterion-Keyed Scale Development. *Psychol. Methods* **2016**, *21* (4), 603–620. https://doi.org/10.1037/met0000088.
85. Yao, Y.; Rosasco, L.; Caponnetto, A. On Early Stopping in Gradient Descent Learning. *Constr. Approx.* **2007**, *26* (2), 289–315. https://doi.org/10.1007/s00365-006-0663-2.
86. Lee, F. What is Bias-Variance Tradeoff? https://www.ibm.com/think/topics/bias-variance-tradeoff (accessed 2026 -02 -15).
87. Hu, Y.; Zhang, X.; Slavin, V.; Belsti, Y.; Tiruneh, S. A.; Callander, E.; Enticott, J. Beyond Comparing Machine Learning and Logistic Regression in Clinical Prediction Modelling: Shifting from Model Debate to Data Quality. *J. Med. Internet Res.* **2025**, *27*, e77721. https://doi.org/10.2196/77721.
88. Condon, D. Performance of Artificial Neural Networks on Small Structured Datasets. ; 2019.
89. Bailly, A.; Blanc, C.; Francis, É.; Guillotin, T.; Jamal, F.; Wakim, B.; Roy, P. Effects of Dataset Size and Interactions on the Prediction Performance of Logistic Regression and Deep Learning Models. *Comput. Methods Programs Biomed.* **2022**, *213*, 106504. https://doi.org/10.1016/j.cmpb.2021.106504.
90. Issitt, R. W.; Cortina-Borja, M.; Bryant, W.; Bowyer, S.; Taylor, A. M.; Sebire, N. Classification Performance of Neural Networks Versus Logistic Regression Models: Evidence From Healthcare Practice. *Cureus* **2022**. https://doi.org/10.7759/cureus.22443.
91. Nguyen, T.; Nguyen, H.-T.; Nguyen-Hoang, T.-A. Data Quality Management in Big Data: Strategies, Tools, and Educational Implications. *J. Parallel Distrib. Comput.* **2025**, *200*, 105067. https://doi.org/10.1016/j.jpdc.2025.105067.
92. Liu, J.; Zhao, J.; Zhang, M.; Chen, N.; Sun, G.; Yang, Y.; Zhang, X.; Chen, J.; Shen, P.; Shi, M.; Zeng, H. The Validation of the 2014 International Society of Urological Pathology (ISUP) Grading System for Patients with High-Risk Prostate Cancer: A Single-Center Retrospective Study</P>. *Cancer Manag. Res.* **2019**, *Volume 11*, 6521–6529. https://doi.org/10.2147/CMAR.S196286.
93. Offermann, A.; Hupe, M. C.; Sailer, V.; Merseburger, A. S.; Perner, S. The New ISUP 2014/WHO 2016 Prostate Cancer Grade Group System: First Résumé 5 Years after Introduction and Systemic Review of the Literature. *World J. Urol.* **2020**, *38* (3), 657–662. https://doi.org/10.1007/s00345-019-02744-4.

 doi:10.20944/preprints202603.1207.v1

94. Epstein, J. I.; Egevad, L.; Amin, M. B.; Delahunt, B.; Srigley, J. R.; Humphrey, P. A. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *American Journal of Surgical Pathology* **2016**, *40* (2), 244–252. https://doi.org/10.1097/PAS.0000000000000530.
95. Barberis, A.; Aerts, H. J. W. L.; Buffa, F. M. Robustness and Reproducibility for AI Learning in Biomedical Sciences: RENOIR. *Sci. Rep.* **2024**, *14* (1), 1933. https://doi.org/10.1038/s41598-024-51381-4.
96. Colliot, O.; Thibeau-Sutre, E.; Burgos, N. Reproducibility in Machine Learning for Medical Imaging. In *Machine Learning for Brain Disorders*; Colliot, O., Ed.; Humana, 2023; Vol. 197, pp 631–653. https://doi.org/10.1007/978-1-0716-3195-9_21.
97. Beam, A. L.; Manrai, A. K.; Ghassemi, M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA* **2020**, *323* (4), 305. https://doi.org/10.1001/jama.2019.20866.