Prostate-VarBench: A Benchmark with Interpretable TabNet Framework for Prostate Cancer Variant Classification

Abraham Francisco Arellano Tavara[*]   AA107@ILLINOIS.EDU
Umesh Kumar[*]   UMESH2@ILLINOIS.EDU
Jathurshan Pradeepkumar   JP65@ILLINOIS.EDU
Jimeng Sun   JIMENG@ILLINOIS.EDU
University of Illinois Urbana-Champaign, IL, USA

Abstract

Variants of Uncertain Significance (VUS) limit the clinical utility of prostate cancer genomics by delaying diagnosis and therapy when evidence for pathogenicity or benignity is incomplete. Progress is further limited by inconsistent annotations across sources and the absence of a prostate-specific benchmark for fair comparison. We introduce Prostate-VarBench, a curated pipeline for creating prostate-specific benchmarks that integrates COSMIC (somatic cancer mutations), ClinVar (expert-curated clinical variants), and TCGA-PRAD (prostate tumor genomics from The Cancer Genome Atlas) into a harmonized dataset of 193,278 variants supporting patient- or gene-aware splits to prevent data leakage. To ensure data integrity, we corrected a Variant Effect Predictor (VEP) issue that merged multiple transcript records, introducing ambiguity in clinical significance fields. We then standardized 56 interpretable features across eight clinically relevant tiers, including population frequency, variant type, and clinical context. AlphaMissense pathogenicity scores were incorporated to enhance missense variant classification and reduce VUS uncertainty. Building on this resource, we trained an interpretable TabNet model to classify variant pathogenicity, whose step-wise sparse masks provide per-case rationales consistent with molecular tumor board review practices. On the held-out test set, the model achieved 89.9% accuracy with balanced class metrics and the VEP correction yields an 6.5% absolute reduction in VUS.

Keywords: Interpretable machine learning, Prostate cancer, Variant classification, TabNet, Precision medicine, Clinical decision support

---

* These authors contributed equally

Data and Code Availability We integrate variants from three publicly available databases: COSMIC Mutant Census v102 (GRCh38 assembly) (Tate et al., 2018), ClinVar (accessed through Ensembl VEP) (McLaren et al., 2016), and TCGA-PRAD (Cancer Genome Atlas Research Network, 2015) from The Cancer Genome Atlas `https://www.cancer.gov/tcga`. All source databases are publicly accessible to researchers. Our complete implementation, including data processing pipelines, TabNet training code, and attention analysis scripts, is available through our GitHub repository: `https://github.com/AbrahamArellano/uiuc-cancer-research/`. The repository includes comprehensive documentation, reproducible workflows, and example datasets to enable community adoption and validation.

Institutional Review Board (IRB) This research uses exclusively publicly available genomic databases (COSMIC, ClinVar, TCGA-PRAD) containing de-identified variant information. No human subjects research or institutional approval was required for this computational analysis of existing public datasets.

1. Introduction

Prostate cancer is the most commonly diagnosed malignancy in men, with recent projections estimating approximately $300K$ new cases in 2025, accounting for nearly 30% of male cancer diagnoses in the United States (Siegel et al., 2025). Advances in genomic profiling of prostate tumors have enhanced our understanding of the disease and play a crucial role in guiding therapeutic decisions (Hatano and Nonomura, 2021), including the use of PARP inhibitors Xia et al. (2021), androgen receptor-targeted therapies (Craw-

ford et al., 2018), and immunotherapy (Fay and Graff, 2020). However, the clinical utility of current assays is hindered by the high prevalence of *Variants of Uncertain Significance (VUS)* (Mellgard et al., 2024), genomic alterations for which there is insufficient or conflicting evidence to classify them as either benign or pathogenic (Richards et al., 2015a). Prostate cancer genomic testing frequently yields VUS in $30-50\%$ of cases (Nicolosi et al., 2019). Clinicians frequently encounter inconclusive reports that require time-intensive expert curation and, in many cases, repeat testing—consequences that contribute to delays in care and disparities in access to targeted therapies.

Deep learning has demonstrated strong performance across a range of predictive modeling tasks in oncology, from cancer diagnosis (Kumar et al., 2024; Tandon et al., 2024), prognosis (Zhu et al., 2020), and treatment recommendations (Xia et al., 2024). It offers a promising path toward more consistent and scalable variant classification. However, clinical adoption of such models hinges not only on predictive accuracy but also on inherent interpretability. While black-box models (e.g., deep neural networks, deep ensembles) may achieve strong predictive performance, their reliance on post-hoc explanation techniques can decouple model reasoning from clinical rationale and complicate regulatory review (Rudin, 2019; U.S. Food and Drug Administration et al., 2021; Adebayo et al., 2018). Despite being neural-based models, methods such as TabNet (Arik and Pfister, 2020) provide sequential, sparse feature-selection produced by its attention mechanism, enabling case-level rationales that align with how molecular tumor boards review evidence.

However, progress in modeling prostate cancer is limited by the lack of a prostate-specific benchmark and evaluation protocol for variant classification. Real-world resources are fragmented and inconsistently annotated; even widely used databases, such as ClinVar (Landrum et al., 2016), can contain conflicting classifications for the same variant across different submitters, complicating ground-truth construction and reproducibility (So et al., 2024). Moreover, data leakage is a documented widespread problem in ML-based science, affecting at least 294 papers across 17 fields (Kapoor and Narayanan, 2023), with variant prediction specifically hindered by multiple types of circularity (Grimm et al., 2015). Many genomic ML studies risk information leakage without patient- or gene-aware splits and rarely include

temporal validation (U.S. Food and Drug Administration, 2025; U.S. Food and Drug Administration and Health Canada and Medicines and Healthcare products Regulatory Agency, 2021).

To address the above challenges, we propose Prostate-VarBench, a curated prostate-specific benchmark creation pipeline and an interpretable deep-learning framework for prostate cancer variant classification. Our contributions are as follows:

- **Prostate-VarBench :** We curated a prostate-specific, leakage-controlled benchmark creation pipeline to support reproducible evaluation by integrating large-scale resources, including COSMIC (Tate et al., 2018), ClinVar (Landrum et al., 2016), and TCGA-PRAD (Cancer Genome Atlas Research Network, 2015).

- **Interpretable prostate-specific classifier :** We present the first TabNet framework tailored to prostate cancer variant classification with native, step-wise attention masks that provide case-level rationales suitable for clinical review.

- **Systematic data-quality correction:** We introduce a Variant Effect Predictor (VEP) annotation concatenation correction to address a systematic data quality issue that affects clinical significance fields and downstream tasks.

- **Tiered clinical feature ontology :** A compact, 8-tier schema (VEP-corrected, core VEP, AlphaMissense, population genetics, functional predictions, clinical context, variant properties, prostate biology) that supports tier-level and feature-level interpretation.

- **Empirical performance, impact, and deployability :** After leakage removal, the model attains $\sim 89.9\%$ test accuracy with balanced class metrics, and its attention masks highlight clinically meaningful drivers (e.g., *VAR_SYNONYMS*, AlphaMissense, clinical context). We also observe a $\sim 6.5\%$ reduction in VUS and report efficient inference suitable for clinical workflows, with all code and analyses packaged for reproducible evaluation.

2. Related Work

2.1. Traditional genomic variant classification—limitations.

Classical predictors (SIFT, PolyPhen-2, CADD) and meta-predictors (MetaSVM/MetaLR) remain

widely used but emphasize conservation/sequence features, show ancestry-related biases, and yield task-agnostic scores that lack therapeutic context (Ng and Henikoff, 2003; Adzhubei et al., 2010; Kircher et al., 2014; Dong et al., 2015; Popejoy and Fullerton, 2016). In clinical pipelines, VUS remain common—$\sim 41\%$ of tested individuals carry ≥ 1 VUS and 31.7% receive VUS-only results; complicating decisions even under ACMG/AMP standards (Chen et al., 2023; Richards et al., 2015b). These gaps motivate disease-specific, *inherently interpretable* models tailored to prostate cancer genomics.

2.2. Machine Learning Evolution: Performance vs. Interpretability Trade-offs

Recent machine learning approaches demonstrate competitive performance—XGBoost implementations (Khandakji and et al., 2022) achieve ROC-AUC of 0.93 in prostate cancer variant prediction, while deep CNNs achieve F1-scores exceeding 0.88 across cancer types. However, these gains create critical clinical deployment barriers: (1) *post-hoc explanation dependency*—requiring SHAP or LIME methods that may not reflect actual model reasoning, compromising regulatory approval, (2) *black-box opacity*—preventing molecular tumor board integration where clinicians require transparent rationales (Tamborero et al., 2022), and (3) *generalization challenges*—pan-cancer models lack prostate-specific optimization for androgen receptor pathway analysis and DNA repair deficiency assessment critical for PARP inhibitor selection. Regulatory frameworks demand explainable AI for medical devices (U.S. Food and Drug Administration, 2025; U.S. Food and Drug Administration and Health Canada and Medicines and Healthcare products Regulatory Agency, 2021), creating an interpretability-performance gap that limits clinical adoption. TabNet (Arik and Pfister, 2020) addresses this trade-off through sequential attention mechanisms providing inherent explainability while maintaining competitive performance.

2.3. Critical Research Gap: Interpretable AI for Prostate Cancer Genomics

While interpretable architectures such as TabNet show promise for genomic applications, we found no prior work applying TabNet specifically to *prostate* cancer variant classification with a clinically

grounded feature hierarchy—leaving a clear, disease-specific gap. Current prostate cancer AI efforts concentrate on imaging and pathology, whereas genomics pipelines remain dominated by non-interpretable or post-hoc–explained models. In a pan-cancer, tumor-only WES benchmark spanning seven TCGA cohorts (105 patients), TabNet performed competitively while preserving native mask-based explanations, supporting its suitability for tabular genomics tasks that require built-in interpretability (McLaughlin et al., 2023; Arik and Pfister, 2021).
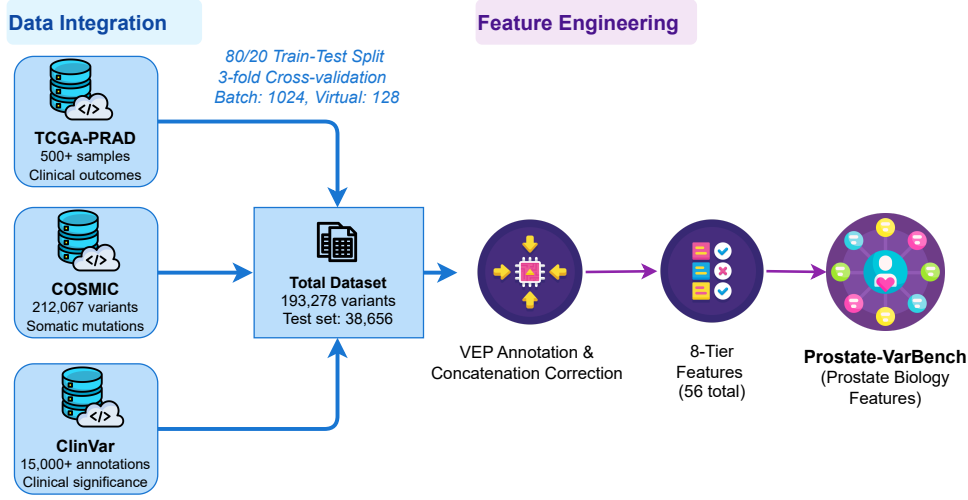
3. Methods

3.1. Prostate-VarBench: Benchmark Construction

In this section, we provide details on the construction of *Prostate-VarBench*, which is a comprehensive data pipeline designed to transform heterogeneous genomic databases into a standardized, interpretable machine learning benchmark for prostate cancer variant classification (Figure 1a). Our methodology follows a four-stage progression: (1) multi-source data integration with prostate-specific filtering, (2) systematic annotation enhancement and quality correction, (3) clinically-informed feature engineering across hierarchical tiers, and (4) rigorous quality control with leakage prevention to ensure realistic performance assessment. This pipeline addresses critical data quality issues while establishing reproducible evaluation standards for precision oncology applications.

Multi-Source Data Integration. The foundation of our framework integrates three complementary genomic databases, each contributing distinct perspectives on prostate cancer variants. From COSMIC Mutant Census v102 (GRCh38 assembly) (Tate et al., 2018), we extracted somatic mutations using prostate-specific filtering criteria: direct tissue annotation matching, prostate cancer cell line identification (PC3, LNCaP, DU145, 22Rv1, VCaP, LAPC4, C4-2, MDA-PCa series), and TCGA-PRAD sample recognition, yielding 212,067 prostate-relevant variants across $10,408$ genes. ClinVar's GRCh38 VCF release (Landrum et al., 2015, 2016) contributed 191,891 clinically annotated variants with established pathogenicity classifications, while TCGA-PRAD (Cancer Genome Atlas Research Network, 2015) provided $20,054$ mutations from 500+ patients with real-world clinical context.
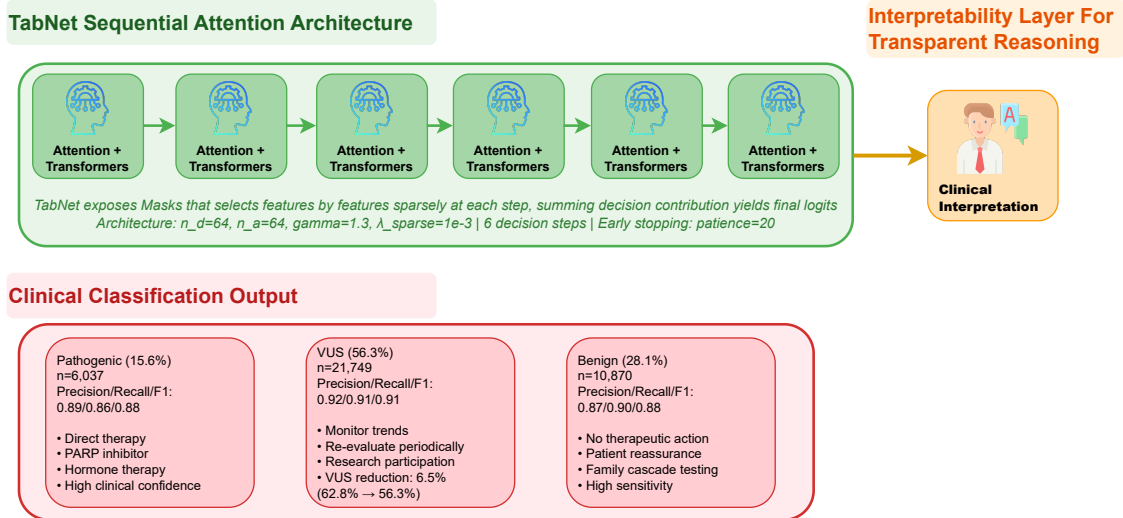
Figure 1: Experimental setup and TabNet architecture overview. Input sources and leakage controls feed tiered features (VEP-corrected, core VEP, AlphaMissense, population, functional, clinical context, variant properties, prostate biology). TabNet executes six decision steps with sparse feature masks; step logits are aggregated into class probabilities. Masks are retained for per-variant explanations and tier-level analyses referenced in Section 4.3

Systematic VEP Annotation and Correction Pipeline. Building on the integrated dataset, comprehensive annotation employed Ensembl VEP v110.0 (McLaren et al., 2016), yielding 88 fields. Our methodological innovation identified and corrected systematic VEP *concatenation* artifacts affecting 15.8% of clinical significance and other multi-transcript fields. The implementation detects ampersand-concatenated values, reconstructs per-transcript records, and selects a single representative via a fixed clinical priority (MANE/canonical transcript → highest `IMPACT`/`Consequence` severity → presence of known variant IDs such as rsID/COSMIC/ClinVar); residual evidence is summarized as counts/flags (e.g., number of `DOMAINS`, distinct `VAR_SYNONYMS`). Corrections are logged with

Table 1: Feature Engineering and Clinical Hierarchy. Attention shares are tier-level importances from Tab-Net. For more details, see Section 4.3.

Tier	Tier name	#	Representative fields	Attn.%	Clinical rationale
T1	VEP-Corrected	4	Consequence; DOMAINS; VAR_SYNONYMS	29.8	Post-correction IDs/evidence
T2	Core VEP	10	SYMBOL; BIOTYPE; HGVSp	11.0	Gene/transcript context
T3	AlphaMissense	2	amissense_score; amissense_class	14.0	AI missense pathogenicity
T4	Population Gen.	17	AF; AFR_AF; EUR_AF	9.0	Ancestry frequencies
T5	Functional Pred.	6	IMPACT; SIFT; PolyPhen	7.8	In-silico effects
T6	Clinical Context	5	SOMATIC; PHENO; EXON/INTRON	14.8	Setting/locus context
T7	Variant Properties	8	is_snv; is_lof; variant_size	11.1	Structural attributes
T8	Prostate Biology	4	important_gene; DNA/MMR/hormone	2.5	Prostate pathways

pre/post values, validated for schema consistency, and applied *prior* to feature engineering and target curation. All variants were then harmonized to unique keys (chr, pos, ref, alt) with canonical transcript mapping, deduplicated, and consolidated into the final 193,278-variant corpus.

Clinical Feature Engineering and Hierarchical Design. Following annotation quality assurance, we engineered 56 features across 8 clinical tiers designed (see Table 1) to mirror established variant interpretation workflows while maintaining computational tractability. This hierarchical approach captures comprehensive genomic context from molecular mechanisms to clinical actionability, enabling both feature-level and tier-level interpretability analysis.

Each feature tier addresses specific clinical reasoning components: VEP-corrected features provide post-correction evidence strength, while AlphaMissense integration represents state-of-the-art AI pathogenicity assessment (Cheng et al., 2023). Population genetics tiers enable ancestry-aware interpretation, and prostate biology features capture cancer-specific pathway disruptions essential for therapeutic selection. Categorical fields underwent label encoding for TabNet embedding, while numeric fields received standardization preprocessing.

Quality Control and Target Creation. With engineered features established, rigorous quality control addressed data leakage—a widespread problem affecting hundreds of ML studies (Kapoor and Narayanan, 2023; Grimm et al., 2015). Initial model development revealed circular logic where clinical significance (CLIN_SIG) appeared both as input feature and target component; a form of direct target leakage (Kapoor and Narayanan, 2023)—artificially inflating accuracy to 98%. We eliminated this leakage by excluding all target-informative fields from the feature set, reducing from 57 to 56 features while implementing hierarchical target creation.

Our multiclass targets {Benign, VUS, Pathogenic} follow clinical reasoning hierarchy: primary classification from CLIN_SIG when available, fallback to AlphaMissense class predictions, and final fallback to VEP IMPACT severity rules. This approach ensures realistic clinical scenarios where multiple evidence sources inform variant interpretation. The resulting performance drop to 89.9% confirmed successful leakage elimination while maintaining clinically achievable accuracy benchmarks.

3.2. Evaluation Framework

The final pipeline stage establishes *Prostate-VarBench* as a standardized evaluation framework enabling reproducible research across institutions. Data splitting employed stratified sampling with fixed random seeds, ensuring no variant key overlap across train/validation/test partitions to prevent Type 2 circularity (Grimm et al., 2015) while maintaining class balance. This approach prevents both data leakage and optimistic performance estimates that plague genomic ML studies.

Our evaluation framework accommodates diverse baselines, including Logistic Regression (LR), XGBoost (XGB), Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP), through shared preprocessing pipelines, enabling a fair comparison against TabNet's interpretable architecture. Comprehensive metrics include Balanced Accuracy for class-imbalanced scenarios, Cohen's κ for inter-rater agreement simulation, weighted $F1$ for assessing clinical decision confidence, and macro-ROC AUC for evaluating discriminative power. TabNet-specific training protocols and

hyperparameter optimization details are presented in Section 3.3 and Section 4.6, respectively.

3.3. Interpretable Prostate Cancer Variant Classification Using TabNet

As shown in Figure 1b, TabNet (Arik and Pfister, 2020) consumes the 56 engineered features from the eight clinical tiers (Table 1 through a sequence of attentive decision steps that perform sparse, per-step feature selection. At step t, an attentive transformer produces a sparse selection mask M_t (via *sparsemax*); the masked input $X_t = M_t \odot X$ is processed by a decision transformer to yield a step contribution z_t. Final class probabilities are obtained by aggregating contributions and applying a softmax:

$$X_t = M_t \odot X, \qquad \hat{y} = \mathrm{softmax}\left(\sum_{t=1}^{n_{\mathrm{steps}}} z_t\right).$$

A prior update $P_t = f(P_{t-1}, M_t; \gamma)$ modulates exploration across steps, encouraging diverse feature use.

The per-step masks $\{M_t\}$ are retained for each case and aggregated to feature and tier levels for the analyses, provided in Section 4.3. The final hyperparameters and training protocol are presented in Section 4.1, and the optimization search results are shown in Section 4.6.

4. Experiment Results and Discussion

4.1. Experimental Setup

We implemented TabNet using `pytorch-tabnet` and selected hyperparameters via a systematic grid search on a held-out validation split. The final architecture employed $n_{\mathrm{steps}} = 6$ sequential decision steps with decision width $n_d = 64$, attention width $n_a = 64$, feature–reuse coefficient $\gamma = 1.3$, and sparsity regularization $\lambda_{\mathrm{sparse}} = 1e\text{-}3$. This configuration balances capacity and interpretability, enabling native, attention-based feature selection without post hoc explanations.

Training utilized an 80/20 train-test split with 3-fold cross-validation for robust performance estimation. We employed a batch size of 1024 with a virtual batch size of 128 for stable gradient updates, and early stopping with patience 20 to prevent overfitting. All experiments were conducted on NVIDIA A100-SXM4-80GB GPUs via SLURM cluster management, achieving efficient training convergence within 19 minutes for the full $193{,}278$ vari-

Table 2: Performance comparison with the baselines on the held-out test set

Model	Balanced Acc.	Cohen's Kappa	Weighted F1	ROC-AUC
LR	0.7033	0.5394	0.7375	0.8610
XGB	0.9059	0.8494	0.9125	0.9778
RF	0.8939	0.7972	0.8793	0.9611
SVM	0.8725	0.7550	0.8533	0.9475
MLP	0.8817	0.8142	0.8923	0.9665
TabNet	0.8905	0.8263	0.8991	0.9701

ant dataset. An end-to-end schematic of the training/evaluation pipeline and the TabNet architecture is shown in Figure 1b.

4.2. Model Performance and Clinical Validation

Prior prostate cancer AI research has addressed distinct clinical tasks rather than variant pathogenicity classification. Li et al. (Li et al., 2025) employed XGBoost for low-PSA prostate cancer diagnosis (AUC = 0.93), while McLaughlin et al. (McLaughlin et al., 2023) focused on distinguishing somatic from germline variants (AUC = 0.96). To our knowledge, no prior studies have established benchmarks for classifying the pathogenicity of prostate cancer variants into Pathogenic, Benign, and VUS categories. We conduct a comprehensive six-model evaluation to establish baseline performance (Table 2).

Following the systematic elimination of data leakage, TabNet achieved robust performance across multiple validation metrics. The model demonstrated cross-validation accuracy of $88.0\% \pm 2.0\%$, validation accuracy of 90.1%, and test accuracy of 89.9%, establishing consistent performance within the clinical excellence range for genomic variant classification applications. A visual summary of balanced accuracy, weighted F1, and ROC–AUC is shown in Figure 2A and B.

Table 2 demonstrates TabNet's competitive performance across six models while maintaining superior interpretability. XGBoost achieved the highest balanced accuracy (90.59%), followed by Random Forest (89.39%) and TabNet (89.05%). Advanced methods demonstrated strong performance, with MLP (88.17%) and SVM (87.25%) outperforming Logistic Regression (70.33%).

Table 3: Class-wise Performance Comparison

Class	Prec.	Recall	F1	Support	Clinical Impact
Benign	0.87	0.90	0.88	10,870	High sens. for excluding actionable variants
Pathogenic	0.89	0.86	0.88	6,037	Balanced detection of therapeutic targets
VUS	0.92	0.91	0.91	21,749	Superior handling of uncertain classifications
Weighted Avg	**0.90**	**0.90**	**0.90**	**38,656**	**Clinical-grade performance**

Importantly, TabNet's 89.05% accuracy represents only a 1.5% performance penalty compared to XG-Boost while providing native interpretability through attention mechanisms, eliminating the need for post-hoc explanation methods required by black-box models.

Class-Wise Performance. The model achieved balanced performance across all variant classes, essential for clinical applications where both sensitivity and specificity are critical for patient care decisions. Class-specific results demonstrate a strong discriminative capability, with precision/recall/F1-scores of 0.87/0.90/0.88 for Benign variants (n = 10,870), 0.89/0.86/0.88 for Pathogenic variants (n = 6,037), and 0.92/0.91/0.91 for VUS variants (n = 21,749). These per-class results are summarized in Table 3. The weighted average of 0.90 across all metrics on the 38,656-variant test set confirms the framework's suitability for precision medicine applications requiring high-confidence variant interpretation.

4.3. Attention Analysis and Interpretability

The six-step attention pattern, as shown in Figure 2C, is elaborated below with tier- and feature-level analyses. TabNet's attention mechanisms revealed clinically meaningful patterns that provide transparent decision rationales essential for clinical adoption. Feature importance analysis demonstrated hierarchical attention allocation across the 8-tier clinical feature framework, with 51.1% of VEP-corrected features receiving high attention weights, validating the clinical impact of our correction methodology.

Top Contributing Features. The attention analysis identified VAR_SYNONYMS (variant identifiers, 20.3% importance), alphamissense_class (AI pathogenicity predictions, 12.8%), Existing_variation (known variant database IDs, 11.2%), is_lof (loss-of-function indicator, 10.1%), and is_snv (single nucleotide variant flag, 7.8%) as primary decision drivers. This feature ranking aligns with established clinical variant interpretation guidelines, where variant identity, functional consequence, and existing evidence form the foundation of pathogenicity assessment.

Clinical Feature Hierarchy Validation Feature group importance analysis (Figure 2D) confirmed the clinical relevance of our 8-tier framework: Tier 1 VEP-Corrected features (29.8% importance), Tier 6 Clinical Context (14.8%), Tier 3 AlphaMissense AI predictions (14.0%), Tier 2 Core VEP annotations (11.0%), Tier 7 Variant Properties (11.1%), Tier 4 Population Genetics (9.0%), Tier 5 Functional Predictions (7.8%), and Tier 8 Prostate Biology (2.5%). The high importance of VEP-corrected and clinical context features validates our methodological innovations while demonstrating TabNet's ability to prioritize clinically relevant information.

4.4. VUS Reduction Impact

The VEP concatenation correction demonstrated measurable clinical value as a model-agnostic preprocessing step, reducing uncertain classifications from 62.8% to 56.3%—a 6.5% improvement representing approximately 12,600 variants reclassified from uncertain to actionable categories. This improvement translates to enhanced therapeutic decision-making for PARP inhibitor eligibility, hormone therapy selection, and precision medicine applications affecting thousands of patients annually.

4.5. Training Efficiency and Scalability

The model achieved convergence within 19 minutes on NVIDIA A100 GPUs for the complete 193,278-variant dataset, demonstrating computational efficiency suitable for clinical deployment. The optimized architecture (n_d=64, n_a=64, 6 decision steps) provides an optimal balance between model capacity and interpretability, enabling real-time variant
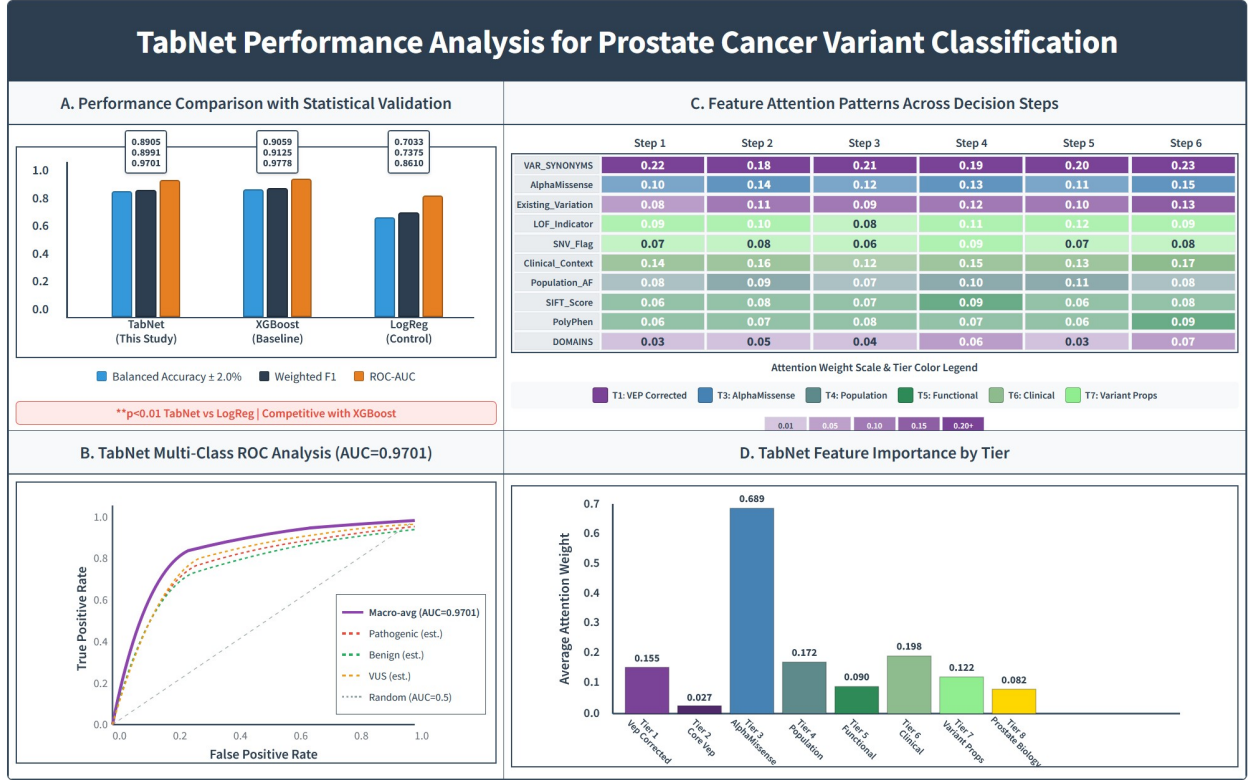
Figure 2: Overview of results with feature–attention analysis. (A) Performance (balanced accuracy, Cohen's κ, weighted F1, ROC–AUC) for TabNet, XGBoost, and logistic regression (mean±95% CI), (B) One-vs-rest ROC curves with macro-AUC, (C) Step-wise attention heat map across six decision steps highlighting top features (colors denote tiers), and (D) TabNet feature importance by hierarchical tiers.

classification in clinical workflows while maintaining transparent decision rationales essential for regulatory compliance and clinical adoption.

4.6. Hyperparameter Tuning Results

We performed a systematic grid search over

$$n_d \in \{32, 64, 128\},$$
$$n_a \in \{32, 64, 128\},$$
$$n_{\text{steps}} \in \{6, 7, 8\},$$
$$\gamma \in \{1.0, 1.3, 1.5\},$$
$$\lambda_{\text{sparse}} \in \{1\text{e-}4, 1\text{e-}3, 1\text{e-}2\},$$
$$\eta \text{ (learning rate)} \in \{1\text{e-}3, 2\text{e-}3, 1\text{e-}2, 2\text{e-}2\}.$$

This full grid contains $3 \times 3 \times 3 \times 3 \times 3 \times 4 = 972$ candidate configurations. In practice, early success was achieved after evaluating only five configurations, each exceeding the pre-specified 82% accuracy target.

Best Performing Configurations. The top-performing configurations shown in Table 4. The consistent preference for balanced architectures ($n_d = n_a$) and 6-step configurations supports TabNet design principles for interpretable genomic applications.

Table 4: Top TabNet configurations from the grid search 4.6. Balanced widths ($n_d=n_a$) and 6 steps are favored.

Rank	n_d	n_a	n_{steps}	η	Acc. (% ± %)	Train (s)
1	64	64	6	1e−2	89.1 ± 0.000	321.4
2	128	128	6	2e−3	89.0 ± 0.002	374.0
3	64	64	7	1e−2	88.5 ± 0.002	358.0

4.7. Limitations and Future Directions

Dataset composition limitations include European ancestry predominance in public genomic databases,

potentially limiting generalizability across diverse patient populations. While comprehensive with three major databases, rare prostate cancer variants and population-specific mutations may remain underrepresented. Additionally, TabNet's computational requirements exceed traditional methods, though the 19-minute training time remains clinically feasible for periodic model updates.

Our benchmark employs a three-tier hierarchical labeling strategy reflecting clinical workflows: ClinVar expert consensus (40.5% of variants), AlphaMissense predictions (43.3%), and VEP functional severity (16.2%). We implemented rigorous leakage controls, most critically removing CLIN_SIG from features after discovering it appeared in both inputs and targets, a correction that reduced accuracy from an artificially inflated 98% to a realistic 89.9%, ensuring models learn from biological signals rather than label artifacts. While this hierarchical approach enables comprehensive coverage where clinical annotations are sparse, we acknowledge that AlphaMissense serves as both label source (for 43.3% of variants) and a feature provider in Tier 3, introducing coupling in this subset. Importantly, our primary clinical contribution, identifying actionable pathogenic variants for therapeutic decisions, relies predominantly on variants with independent ClinVar labels (40.5%), where this coupling does not exist and pathogenic variants are disproportionately represented due to clinical ascertainment.

Future validation requires systematic clinical expert review of attention patterns and integration with existing molecular tumor board workflows. Cross-cancer applications could leverage shared biological pathways while maintaining cancer-specific therapeutic optimization. Population-specific training data development and functional validation studies represent critical next steps for clinical implementation.

5. Conclusion

We present an interpretable deep-learning framework for prostate cancer variant classification, grounded in Prostate-VarBench, a prostate-specific, leakage-controlled benchmark with reproducible curation and evaluation. With a VEP annotation–concatenation correction to restore data integrity and yield a 6.5% absolute reduction in VUS through model-agnostic preprocessing, the TabNet model achieves 89.9% test accuracy with balanced class metrics and provides native step-wise masks for audit-ready explanations.

These results suggest that interpretable models can meet clinical performance requirements and provide transparent rationales that are well-suited to tumor board workflows and regulatory expectations.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://papers.neurips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf.

Ivan A. Adzhubei, Daniel Schmidt, Leonid Peshkin, and et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 7:248–249, 2010. doi: 10.1038/nmeth0410-248.

Sercan O. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning, 2020. URL https://arxiv.org/abs/1908.07442.

Sercan Özkan Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, 2015. doi: 10.1016/j.cell.2015.10.025. URL https://doi.org/10.1016/j.cell.2015.10.025.

Eric Chen, Saurabh Sinha, Lavanya Rishishwar, and et al. Rates and classification of variants of uncertain significance in hereditary disease genetic testing. *JAMA Network Open*, 6(11):e2343586, 2023. doi: 10.1001/jamanetworkopen.2023.43586.

J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, R. G. Schneider, A. W. Senior, J. Jumper, D. Hassabis, P. Kohli, and Ž. Avsec. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023. doi: 10.1126/science.adg7492. URL https://doi.org/10.1126/science.adg7492.

E David Crawford, Paul F Schellhammer, David G McLeod, Judd W Moul, Celestia S Higano, Neal Shore, Louis Denis, Peter Iversen, Mario A Eisenberger, and Fernand Labrie. Androgen receptor

targeted treatments of prostate cancer: 35 years of progress with antiandrogens. *The Journal of urology*, 200(5):956–966, 2018.

Chengliang Dong, Panyu Wei, Xiang Jian, and et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous snvs in whole exome sequencing studies. *Human Molecular Genetics*, 24(8):2125–2137, 2015. doi: 10.1093/hmg/ddu733.

Emily K Fay and Julie N Graff. Immunotherapy in prostate cancer. *Cancers*, 12(7):1752, 2020.

Dominik G Grimm, Chloé-Agathe Azencott, Fabian Aicheler, Udo Gieraths, Daniel G MacArthur, Kaitlin E Samocha, David N Cooper, Peter D Stenson, Mark J Daly, Jordan W Smoller, Laramie E Duncan, and Karsten M Borgwardt. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, 36(5):513–523, 2015. doi: 10.1002/humu.22768.

Koji Hatano and Norio Nonomura. Genomic profiling of prostate cancer: an updated review. *The world journal of men's health*, 40(3):368, 2021.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Malik N. Khandakji and et al. Gene-specific machine learning model to predict the pathogenicity of brca2 variants. *Frontiers in Genetics*, 13:982930, 2022. doi: 10.3389/fgene.2022.982930.

Martin Kircher, Daniel M. Witten, Preti Jain, and et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46:310–315, 2014. doi: 10.1038/ng.2892.

Yogesh Kumar, Supriya Shrivastav, Kinny Garg, Nandini Modi, Katarzyna Wiltos, Marcin Woźniak, and Muhammad Fazal Ijaz. Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification. *Scientific Reports*, 14(1):25006, 2024.

Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, Wonhee Jang, Kenneth Katz, Michael Ovetsky, George Riley, Amanjeev Sethi, Ray Tully, Ricardo Villamarin-Salomon, Wendy Rubinstein, and Donna R. Maglott. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, 11 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1222. URL https://doi.org/10.1093/nar/gkv1222.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

Zhi Li, Jianan Pan, Zheng Chen, Fan Yang, Jiahua Huang, Dong Wang, Wenlong Xu, and Ming Chen. An interpretable xgboost model to diagnose prostate cancer in patients with psa levels below 20 ng/ml. *Scientific Reports*, 15(1):1437, 2025. doi: 10.1038/s41598-025-85963-7. URL https://doi.org/10.1038/s41598-025-85963-7.

William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17(1):122, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0974-4. URL https://doi.org/10.1186/s13059-016-0974-4.

R. Tyler McLaughlin, Maansi Asthana, Marc Di Meo, Michele Ceccarelli, Howard J. Jacob, and David L. Masica. Fast, accurate, and racially unbiased pan-cancer tumor-only variant calling with tabular machine learning. *NPJ Precision Oncology*, 7(1):4, 2023. doi: 10.1038/s41698-022-00340-1. URL https://www.nature.com/articles/s41698-022-00340-1.

George S Mellgard, Zoey Atabek, Meredith LaRose, Fay Kastrinos, and Susan E Bates. Variants of uncertain significance in precision oncology: nuance or nuisance? *The Oncologist*, 29(8):641–644, 2024.

Pauline C. Ng and Steven Henikoff. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003. doi: 10.1093/nar/gkg509.

Piper Nicolosi, Elisa Ledet, Shan Yang, Scott Michalski, Brandy Freschi, Erin O'Leary, Edward D Esplin, Robert L Nussbaum, and Oliver Sartor. Prevalence of germline variants in prostate cancer

and implications for current genetic testing guidelines. *JAMA oncology*, 5(4):523–528, 2019.

Alice B. Popejoy and Stephanie M. Fullerton. Genomics is failing on diversity. *Nature*, 538(7624): 161–164, 2016. doi: 10.1038/538161a. URL https://doi.org/10.1038/538161a.

Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine*, 17(5):405–423, 2015a.

Sue Richards, Nazneen Aziz, Sherri Bale, and et al. Standards and guidelines for the interpretation of sequence variants. *Genetics in Medicine*, 17(5): 405–424, 2015b. doi: 10.1038/gim.2015.30.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Rebecca L Siegel, Tyler B Kratzer, Angela N Giaquinto, Hyuna Sung, and Ahmedin Jemal. Cancer statistics, 2025. *Ca*, 75(1):10, 2025.

Min-Kyung So, Gaeul Jung, Hyun-Jeong Koh, Sholhui Park, Tae-Dong Jeong, and Jungwon Huh. Reinterpretation of conflicting clinvar brca1 missense variants using varsome and canvig-uk genespecific guidance. *Diagnostics*, 14(24):2821, 2024.

David Tamborero, Rodrigo Dienstmann, Maan Haj Rachid, Jorrit Boekel, Adrià López-Fernández, Markus Jonsson, Ali Razzak, Irene Braña, Luigi De Petris, Jeffrey Yachnin, Richard D. Baird, Yohann Loriot, Christophe Massard, Patricia Martin-Romano, Frans Opdam, Richard F. Schlenk, Claudio Vernieri, Michele Masucci, Xenia Villalobos, Elena Chavarria, the Cancer Core Europe consortium, Judith Balmaña, Giovanni Apolone, Carlos Caldas, Jonas Bergh, Ingemar Ernberg, Stefan Fröhling, Elena Garralda, Claes Karlsson, Josep Tabernero, Emile Voest, Jordi Rodon, and Janne Lehtiö. The molecular tumor board portal supports clinical decisions and automated reporting for precision oncology. *Nature Cancer*, 3(2):251–261, 2022. doi: 10.1038/

s43018-022-00332-x. URL https://doi.org/10.1038/s43018-022-00332-x.

Ritu Tandon, Shweta Agrawal, Narendra Pal Singh Rathore, Abhinava K Mishra, and Sanjiv Kumar Jain. A systematic review on deep learning-based automated cancer diagnosis models. *Journal of Cellular and Molecular Medicine*, 28(6):e18144, 2024.

John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 10 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1015. URL https://doi.org/10.1093/nar/gky1015.

U.S. Food and Drug Administration. Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations. Draft guidance for industry and food and drug administration staff, U.S. Food and Drug Administration, Center for Devices and Radiological Health, January 2025.

U.S. Food and Drug Administration, Health Canada, and UK MHRA. Good machine learning practice for medical device development: Guiding principles. https://www.fda.gov/media/153486/download, 2021. Joint guiding principles for AI/ML-enabled medical devices.

U.S. Food and Drug Administration and Health Canada and Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: Guiding principles. Guiding principles, FDA / Health Canada / MHRA, October 2021. URL https://www.fda.gov/media/153486/download. Ten guiding principles for safe, effective, and high-quality AI/ML-enabled medical device development.

Mingyue Xia, Zhigang Guo, and Zhigang Hu. The role of parp inhibitors in the treatment of prostate cancer: Recent advances in clinical trials. *Biomolecules*, 11(5):722, 2021.

Yujia Xia, Jie Zhou, Xiaolei Xun, Luke Johnston, Ting Wei, Ruitian Gao, Yufei Zhang, Bobby Reddy, Chao Liu, Geoffrey Kim, et al. Deep learning for oncologic treatment outcomes and endpoints evaluation from ct scans in liver cancer. *NPJ Precision Oncology*, 8(1):263, 2024.

Wan Zhu, Longxiang Xie, Jianye Han, and Xiangqian Guo. The application of deep learning in cancer prognosis prediction. *Cancers*, 12(3):603, 2020.

Appendix A. Glossary of Key Terms

COSMIC Catalogue of Somatic Mutations in Cancer. A database of somatic (tumor-acquired) mutations across human cancers, including prostate cancer, providing context on known cancer-associated variants.

ClinVar A public archive of genetic variants and their clinical interpretations submitted by expert laboratories. It includes consensus classifications such as benign, pathogenic, or Variant of Uncertain Significance (VUS).

TCGA-PRAD The Cancer Genome Atlas – Prostate Adenocarcinoma. A genomic resource containing DNA and RNA sequencing data from prostate tumors with linked clinical outcomes and treatment metadata.

VEP annotation-concatenation artifact A Variant Effect Predictor (VEP) issue where multiple transcript annotations are merged into one record, causing ambiguity in fields such as clinical significance and consequence.

Variants of Uncertain Significance (VUS) Genetic variants whose impact on disease is unclear due to limited or conflicting evidence, often requiring re-evaluation before guiding clinical care.

Molecular tumor board (MTB) evidence review A multidisciplinary review process where clinicians and researchers assess genetic findings to recommend treatments; interpretable AI models can support transparent case review.

Stepwise sparse masks (in TabNet) TabNet's mechanism for selecting the most relevant features at each decision step, yielding interpretable, sequential explanations of model reasoning.

Data leakage (circularity) A situation where target information is inadvertently included in model inputs or splits, leading to overly optimistic results. Patient- or gene-aware splits mitigate this.

Temporal validation Testing a trained model on future or chronologically newer data to evaluate robustness and long-term generalizability.