

ORIGINAL RESEARCH

Open access

An Explainable Deep Survival Framework for Metastasis Risk Prediction in Prostate Cancer Using Serial PSA and Genomic Scores

Nikolai Ivanov^{1*}, Sergey Volkov¹, Elena Morozova²

Abstract

Prostate cancer metastasis to bone and lymph nodes marks a critical transition to incurable disease, with five-year survival dropping dramatically compared to localized disease. Early identification of patients at high risk of metastasis enables timely intensification of treatment, including androgen deprivation therapy, salvage radiation, or systemic therapies. Current deep survival models that integrate serial PSA measurements and genomic risk scores achieve high predictive accuracy for time-to-metastasis but operate as black boxes, providing no explanation for why a particular patient is predicted to have early or late metastasis. Clinicians cannot trust or act upon predictions without understanding which PSA features or genomic markers drive the risk assessment. We present an explainable deep survival framework that combines a deep survival model for time-to-metastasis prediction with Integrated Gradients attribution, a method that distributes the model's hazard prediction among input features. The framework produces patient-specific explanations showing how each serial PSA value and each genomic score component contributes to the predicted metastasis hazard. The framework consists of three core components: (1) a deep survival model (DeepSurv architecture) with a PSA time-series encoder and genomic risk encoder, (2) Integrated Gradients attribution computed over the hazard function, and (3) visualization tools for individual and population-level interpretations. Integrated Gradients attributes the predicted hazard to individual PSA measurements across time and specific genomic markers, enabling clinicians to distinguish between risk driven by rapid PSA kinetics versus high genomic risk scores. This interpretability transforms a black-box survival prediction into an actionable clinical decision support tool.

Keywords Explainable AI, Survival analysis, Prostate cancer, Integrated gradients, PSA kinetics, Genomic risk scores

*Correspondence:

Nikolai Ivanov
nikolai.ivanov@gmail.com

¹ Department of Healthcare Intelligence Analytics, Novosibirsk State University, Novosibirsk, Russia

² Department of Clinical AI Systems, Tomsk State University, Tomsk, Russia

Introduction

Prostate cancer metastasis represents a decisive clinical turning point, as patients with metastatic disease face substantially reduced survival and require intensive systemic therapies rather than localized curative interventions. Sundararajan *et al.* established the axiomatic foundations for attributing neural network predictions to input features, a principle that can be extended to survival models for cancer prognosis [1]. Clinical management decisions following primary treatment—such as whether to initiate salvage radiation, add androgen deprivation therapy, or pursue active surveillance—depend critically on accurate assessment of individual metastasis risk [2].

Multiple clinical biomarkers have demonstrated prognostic value for prostate cancer outcomes, yet they are rarely integrated into a unified predictive framework. Serial PSA measurements, including PSA velocity (ng/mL/year), PSA doubling time, and absolute PSA values, provide dynamic information about disease activity [3]. Concurrently, genomic risk scores such as Decipher (22-gene panel), Oncotype DX Genomic Prostate Score (17-gene panel), Prolaris (cell cycle progression), and Prolaris (homologous recombination deficiency) independently predict metastasis and prostate cancer-specific mortality [4]. However, clinical practice typically treats these as separate data streams rather than integrating them into a multimodal predictive model.

Deep survival models have demonstrated superior performance compared to traditional Cox proportional hazards models for time-to-event prediction with complex, high-dimensional inputs. Yao *et al.* developed DeepSurv, a deep neural network that extends the Cox model by learning nonlinear hazard functions directly from patient data [5]. Wang *et al.* introduced DeepHit, which handles competing risks without the proportional hazards assumption and learns the joint distribution of event times [6]. Despite their predictive power, these models are not interpretable by design, which limits their adoption in clinical settings where understanding the rationale behind a risk prediction is essential for treatment decisions [7, 8].

Background

Prostate cancer metastasis

Prostate cancer most commonly metastasizes to bone (spine, pelvis, ribs, long bones) and lymph nodes (pelvic, retroperitoneal), with visceral metastases to liver or lung occurring less frequently but indicating poor prognosis. Time to metastasis is a clinically meaningful endpoint that drives treatment intensification decisions, including the initiation of androgen deprivation therapy, docetaxel, or novel hormonal agents [8, 9]. Risk groups for metastasis are defined by clinical stage, Gleason score, and baseline PSA, but these categorical assignments fail to capture individual variability in disease trajectory.

Serial PSA measurements

Serial PSA measurements capture dynamic patterns of disease activity, with PSA velocity (rate of change in ng/mL per year) and PSA doubling time (time in months for PSA to double) showing strong associations with metastasis risk and prostate cancer-specific mortality. Baseline PSA prior to treatment, PSA nadir following definitive therapy, and post-treatment PSA kinetics each provide complementary prognostic information [10, 11]. The irregular timing and varying numbers of PSA measurements across patients pose challenges for standard time-series methods but can be addressed using sequence models with masking.

Genomic risk scores

Genomic risk scores derived from tumor tissue provide molecular prognostic information independent of clinical variables. The Decipher score (22-gene panel) predicts metastasis and prostate cancer-specific mortality across multiple validation cohorts, with higher scores indicating greater genomic aggression [12, 13]. The Oncotype DX Genomic Prostate Score (17-gene panel) predicts adverse pathology and biochemical recurrence, while Prolaris (cell cycle progression score) and Prolaris (homologous recombination deficiency score) offer additional but partially overlapping risk information.

Deep survival models

DeepSurv implements a Cox proportional hazards model within a deep neural network, learning a nonlinear function

$f_{\theta}(x)$ such that the hazard function
$$h(t|x) = h_0(t) \times \exp(f_{\theta}(x))$$
, with parameters optimized via a partial likelihood loss that accounts

for censored observations [14, 15]. DeepHit extends this framework by using a discrete-time approach that directly learns the probability of an event occurring in each time interval, eliminating the proportional hazards assumption and naturally

handling competing risks such as death from other causes [16]. Both architectures can accommodate time-dependent covariates, including serial PSA measurements, through appropriate input encoding.

Framework Overview

High-level architecture

The framework accepts two input modalities: a sequence of serial PSA measurements with associated timestamps (irregularly sampled) and a genomic risk score vector (continuous values from one or more commercial assays such as Decipher or Oncotype DX). The deep survival model outputs a predicted hazard function over follow-up time, and Integrated Gradients computes per-input-feature attributions by integrating the gradient of the hazard output with respect to each input along a straight-line path from a baseline input to the actual input [17].

Figure 1 illustrates the proposed explainable deep survival framework linking multimodal prostate cancer inputs, DeepSurv-based hazard prediction, Integrated Gradients attribution, and clinically actionable metastasis-risk interpretation.

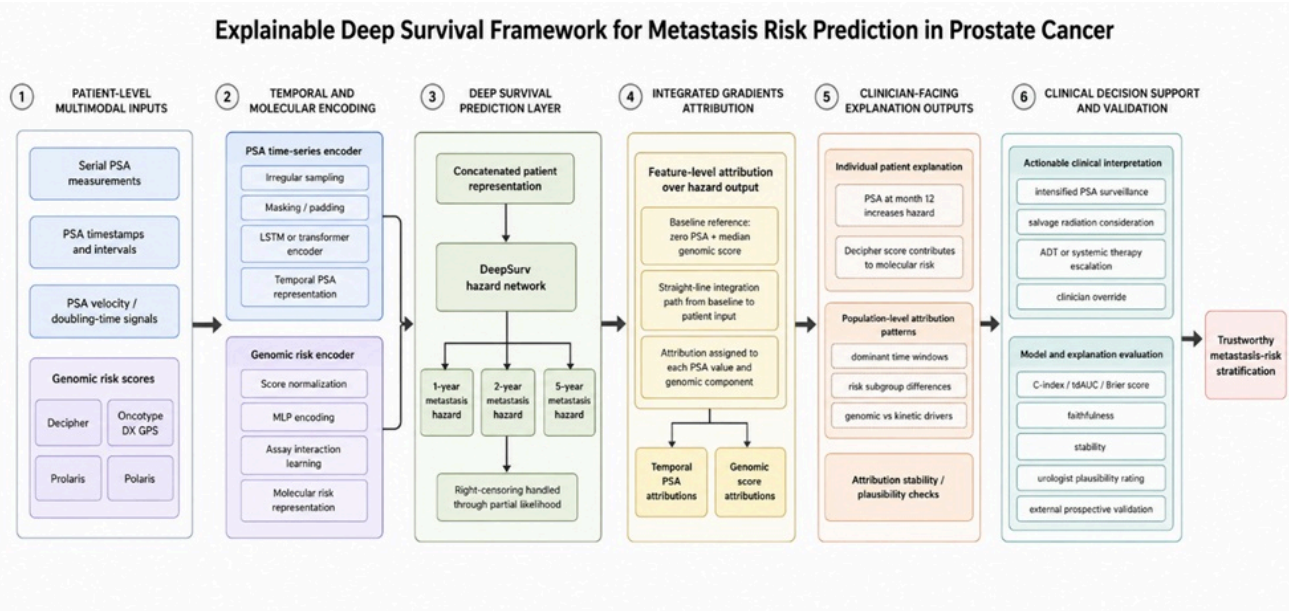


Figure 1. Explainable deep survival architecture for prostate cancer metastasis-risk prediction using serial PSA trajectories, genomic risk scores, and Integrated Gradients attribution.

Core assumptions

The framework assumes that each patient has at least three PSA measurements recorded following primary treatment (surgery or radiation) and that a genomic risk score is available from biopsy or surgical specimen, which may be obtained at diagnosis or post-prostatectomy. Metastasis status (event or censored) is assumed to be reliably determined through imaging (bone scan, CT, PSMA-PET) or clinical documentation, with the time from primary treatment to metastasis or last follow-up recorded as the survival time [18].

Design principles

Three design principles guide the framework: interpretability (every prediction must be accompanied by feature attributions that a clinician can understand), time-awareness (PSA values are attributed to specific time points, enabling

explanations such as "PSA at month 18 contributed most to increased risk"), and clinical actionability (explanations should suggest specific interventions, such as shortening surveillance intervals or initiating systemic therapy). The framework must also handle right-censored data without bias, as patients who have not metastasized by the end of follow-up provide partial information [19].

Table 1 maps each component of the proposed framework to its technical function, clinical interpretation, explanation output, and decision-support relevance.

Table 1. Conceptual Mapping between Model Components, Clinical Meaning, and Explanation Outputs

Framework component	Technical role in the survival model	Clinical meaning	Explanation output generated	Decision-support value
Serial PSA values	Provide longitudinal biomarker input to the time-series encoder	Captures post-treatment disease activity and biochemical recurrence patterns	Attribution assigned to each PSA measurement at each recorded time point	Identifies whether risk is driven by specific PSA elevations or sustained trajectory change
PSA timestamps and intervals	Encode irregular measurement spacing and temporal context	Reflects surveillance timing and rate of PSA evolution	Attribution linked to clinically meaningful months after treatment	Helps clinicians locate the temporal window in which risk becomes most informative
PSA velocity and doubling-time signals	Represent dynamic change in tumor activity over time	Indicates aggressive biochemical progression when PSA rises rapidly	Positive attribution to rapid PSA kinetics	Supports intensified monitoring, salvage therapy consideration, or systemic escalation
Genomic risk score vector	Provides molecular prognostic input independent of PSA kinetics	Reflects intrinsic tumor aggressiveness and metastatic potential	Attribution assigned to Decipher, Oncotype DX GPS, Prolaris, or Polaris components	Distinguishes biology-driven risk from trajectory-driven risk
PSA time-series encoder	Learns nonlinear temporal representation from irregular PSA sequences	Summarizes the patient's evolving post-treatment disease course	Time-specific PSA attribution pattern	Explains whether early, mid, or late PSA measurements dominate predicted hazard
Genomic MLP encoder	Learns interactions among genomic scores	Captures complementary molecular risk information across assays	Assay-specific attribution profile	Clarifies which molecular signal contributes most to predicted metastasis hazard
DeepSurv hazard network	Estimates individualized time-to-metastasis hazard	Converts multimodal patient information into survival-risk estimates	Horizon-specific hazard attribution at 1, 2, and 5 years	Aligns predictions with clinically relevant treatment windows
Integrated Gradients module	Decomposes hazard output into feature-level contributions	Makes model reasoning inspectable by clinicians	Signed positive or negative contribution for each PSA value and genomic feature	Enables verification, trust calibration, and clinician override

Population-level attribution aggregation	Summarizes explanation patterns across cohorts	Reveals common prognostic drivers across patient subgroups	Group-level ranking of PSA windows and genomic markers	Supports simplified risk-score design and subgroup-specific clinical interpretation
--	--	--	--	---

Deep Survival Model

PSA time-series encoder

The PSA time-series encoder processes irregularly sampled measurements using a Long Short-Term Memory (LSTM) or transformer architecture with time point encodings that capture both PSA values and the intervals between measurements. Missing data between recorded measurements is handled through masking or imputation, while variable-length sequences are padded to a maximum length with attention masks to exclude padded positions from computation [20]. For each time point, the encoder outputs a hidden state that summarizes PSA history up to that measurement, which can be pooled or passed to subsequent layers.

Genomic risk encoding

Genomic risk scores are encoded through a small multilayer perceptron (MLP) with one or two hidden layers that maps the input vector (e.g., a single Decipher score value between 0 and 100, or a panel of multiple scores) into a latent representation. When multiple genomic assays are available, the encoder learns interactions among them, allowing the model to weight complementary information from Decipher, Oncotype DX, and other assays relative to their empirical prognostic value [21]. The encoded genomic representation is concatenated with the pooled PSA sequence representation to form the full input feature vector for the hazard prediction layer.

Hazard prediction

Following the DeepSurv architecture, the model predicts the hazard function as
$$h(t|x) = h_0(t) \times \exp(f_\theta(x))$$
, where $f_\theta(x)$ is the output of a feedforward neural network taking the concatenated PSA and genomic representation as input, and $h_0(t)$ is the baseline hazard estimated nonparametrically from the training data [14]. Model parameters are optimized by minimizing the negative partial log-likelihood, treating censored observations as contributing only the survival probability up to their censoring time rather than a full event contribution. Time-dependent covariates are handled by constructing a separate input vector for each time interval using the PSA measurements available up to that interval, following the counting process formulation [22].

Integrated Gradients for Survival

Integrated gradients formulation

Integrated Gradients attributes the difference between the model's output at the actual input x and its output at a baseline input x' to each input feature i , computed as the integral of the gradient along the straight-line path from x' to x :
$$IG_i(x) = (x_i - x'_i) \int \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$
 where F is the hazard prediction function [1]. For survival applications, the baseline is defined as zero PSA values at all time points (representing a theoretical patient with no detectable PSA) and the median genomic risk score from the training population, which provides a neutral reference for attribution computation [23].

Attribution for time-to-event

Attributions are computed separately for predicted hazard at multiple clinically relevant time horizons—1 year, 2 years, and 5 years following primary treatment—enabling clinicians to understand how feature contributions evolve over time. For each time horizon, the attribution scores are aggregated across the hazard prediction, and the final interpretation presents the relative contribution of each PSA measurement (by time point) and each genomic score component. Standardization across patients allows comparison of attribution magnitudes, and the sign of each attribution indicates whether the feature increases (positive) or decreases (negative) the predicted hazard relative to baseline [24].

Feature Importance Interpretation

Individual patient explanations

For an individual patient, the framework produces a structured explanation stating how each PSA measurement and each genomic score component contributed to the predicted metastasis hazard at specified time horizons, such as "Your PSA value of 2.1 ng/mL at month 12 contributed +15% to your predicted 2-year metastasis risk, while your Decipher score of 0.65 contributed +30%." These attributions enable clinicians to identify whether hazard is driven by rapid PSA kinetics, high genomic risk, or unfavorable combinations thereof, and the baseline reference (zero PSA, median genomic score) makes the direction and magnitude of each contribution interpretable [25].

Population-level patterns

Across a patient cohort, the framework aggregates attribution scores to identify which PSA time points and which genomic markers are most informative for metastasis prediction on average, revealing that PSA velocity during months 6 to 18 post-treatment may be more prognostic than any single absolute PSA value. Population-level analysis can also compare attribution patterns between risk groups, showing that genomic scores dominate predictions for intermediate-risk patients while PSA kinetics dominate for high-risk patients, or vice versa [26]. These patterns can guide clinical understanding of how different prognostic factors operate across disease stages and inform the design of simplified risk scores for settings where full deep learning deployment is infeasible.

Clinical Utility

Actionable explanations

Actionable explanations directly inform treatment decisions by identifying modifiable or monitorable drivers of risk; for example, if the framework attributes high 2-year metastasis risk primarily to a PSA doubling time of less than three months, the clinician may escalate from semi-annual to quarterly PSA monitoring, initiate salvage radiation, or add a brief course of androgen deprivation therapy. Conversely, if high risk is attributed almost entirely to an elevated genomic score with stable PSA kinetics, the focus may shift toward systemic therapy rather than local salvage interventions, as the genomic risk reflects intrinsic tumor biology less responsive to local treatment [27]. The explanation format is designed to be readable at the point of care, requiring no specialized data science training.

Trust and adoption

Clinician trust in the framework depends fundamentally on the plausibility and consistency of its explanations: when the model attributes high 2-year metastasis risk to rapidly rising PSA (PSA doubling time less than three months) within the first year after radical prostatectomy, this aligns perfectly with established clinical knowledge that biochemical recurrence with short doubling time strongly predicts subsequent metastatic progression. Conversely, when the model attributes high risk to a single low PSA value while ignoring a clearly rising trend over multiple measurements, such contradictions would rightly prompt model skepticism and potential clinician override, creating a natural verification mechanism where clinical expertise serves as a check on model behavior rather than blind acceptance of its outputs [28].

Beyond pointwise plausibility, clinicians require consistency across similar patients and stability over time: if two patients with nearly identical PSA trajectories and Decipher scores receive substantially different attributions, or if the same patient evaluated twice with minimally updated PSA values receives dramatically different explanations, trust erodes rapidly, as clinicians perceive the model as capricious rather than reliable. The framework should therefore report attribution confidence intervals or stability metrics alongside point estimates, and prospective user studies should measure how explanation consistency affects clinician willingness to modify treatment plans based on model recommendations.

Over time, consistent alignment between attributions and clinical expectations builds the trust necessary for routine adoption, while persistent contradictions signal the need for model retraining, architecture revision, or recalibration of the attribution baseline. A phased adoption strategy is recommended: initially, the framework serves as a silent decision-support tool whose explanations are reviewed retrospectively in tumor boards; after achieving high plausibility scores (e.g., >85% of explanations rated as clinically sensible by urologists), the framework can transition to prospective use with clinician override authority; and only after demonstrating improved clinical outcomes (e.g., reduced time to metastasis detection or more appropriate treatment intensification) should autonomous recommendation be considered [28]. This cautious pathway respects the primacy of clinical judgment while systematically building the evidence base for XAI in prostate cancer survival prediction.

Table 2 provides an evaluation matrix linking survival-model performance, explanation reliability, subgroup robustness, and clinical adoption readiness.

Table 2. Evaluation Matrix for Predictive Performance, Explanation Quality, and Clinical Adoption Readiness

Evaluation domain	Metric or assessment	What it tests	Minimum desirable interpretation	Why it matters for clinical deployment
Survival discrimination	Concordance index	Whether higher-risk patients are correctly ranked before lower-risk patients	C-index above 0.7 acceptable; above 0.8 strong, depending on censoring and cohort characteristics	Determines whether the model meaningfully stratifies metastasis risk
Time-specific discrimination	Time-dependent AUC at 1, 2, and 5 years	Whether the model separates patients who metastasize within clinically relevant windows	Strong performance should be consistent across short- and medium-term horizons	Supports treatment decisions tied to actionable follow-up periods
Calibration	Brier score and calibration plots	Whether predicted survival probabilities match observed metastasis rates	Predicted risk should correspond closely to observed event frequency	Prevents overconfident or systematically biased treatment recommendations
Explanation faithfulness	Feature perturbation or masking tests	Whether high-attribution features truly influence the model output	Removing highly attributed features should substantially change hazard estimates	Ensures explanations reflect model behavior rather than post-hoc artifacts
Explanation stability	Small input-perturbation sensitivity analysis	Whether minor PSA measurement variation causes large attribution shifts	Clinically negligible PSA variation should not radically change explanations	Protects against unreliable explanations caused by measurement noise

Clinical plausibility	Blinded urologist or radiation oncologist rating	Whether attributions align with expert clinical reasoning	Most explanations should receive high plausibility ratings across subgroups	Builds clinician trust and identifies implausible reasoning patterns
Subgroup reliability	Stratified evaluation by risk group, age, race, treatment type, and genomic-score availability	Whether predictions and explanations remain valid across patient populations	No subgroup should show substantial degradation in accuracy or plausibility	Reduces risk of inequitable decision support
Baseline comparison	Cox model, standard DeepSurv, random survival forest, and SHAP comparison	Whether the proposed framework improves prediction or explanation utility over alternatives	Should retain predictive performance while adding patient-specific interpretability	Justifies the added complexity of explainable deep survival modeling
Deployment readiness	Prospective external validation and silent-mode tumor-board review	Whether the framework performs reliably outside retrospective development data	Strong external performance and high clinician plausibility before active use	Prevents premature clinical deployment of an unvalidated model
Adoption safety	Clinician override and explanation review workflow	Whether clinicians can challenge or reject implausible recommendations	Override should remain explicit during early deployment	Maintains clinical accountability and prevents blind automation dependence

Survival metrics

Concordance index (C-index) measures the model's ability to order patients by risk, with values above 0.7 generally considered acceptable and above 0.8 indicating strong discriminative performance for time-to-metastasis prediction, though the baseline prevalence and censoring rate in the validation cohort must be reported to enable meaningful comparisons across studies. Time-dependent area under the ROC curve (tdAUC) evaluates discrimination at specific clinically relevant time horizons (1, 2, and 5 years), capturing whether the model distinguishes patients who metastasize within each window from those who do not, which is more directly interpretable for treatment decisions than a single global concordance measure. The Brier score assesses calibration by measuring the squared difference between predicted survival probabilities and observed outcomes, with lower values indicating better-calibrated predictions; a well-calibrated model's predicted 20% 5-year metastasis risk should correspond to approximately one in five patients actually metastasizing within five years [29]. These metrics collectively characterize predictive accuracy across different aspects of survival performance and should be reported alongside explanation quality metrics, as a model with excellent discrimination but poor calibration may produce systematically biased risk estimates, while a well-calibrated model with modest discrimination may still support clinical decision-making when combined with trustworthy explanations.

Explanation metrics

Faithfulness measures whether features identified as important by Integrated Gradients actually affect the model's prediction when perturbed: if removing or masking a high-attribution PSA value (e.g., setting it to baseline zero) substantially changes the predicted hazard while perturbing a low-attribution value produces minimal change, the explanations are faithful to the model's actual decision boundary. Conversely, if a feature receives high attribution but the predicted hazard remains unchanged after masking, the explanation is misleading and could cause clinicians to focus on irrelevant variables while overlooking true drivers of risk. Stability requires that small, clinically negligible perturbations to the input (e.g., adding ± 0.1 ng/mL measurement error to PSA values) produce similar attribution scores; highly unstable

explanations undermine clinical confidence because clinicians cannot rely on attributions that change dramatically with routine measurement variability or with minor differences in how PSA values are recorded across laboratories.

Plausibility is assessed by having board-certified urologists or radiation oncologists review a blinded set of patient explanations (inputs, predictions, and attributions) and rate whether the attributed features align with clinical reasoning on a Likert scale from 1 (completely implausible) to 5 (highly plausible). High plausibility scores indicate that explanations are clinically sensible even if the underlying model remains a black box, and these scores should be disaggregated by patient subgroup (e.g., low-risk vs. high-risk, African American vs. white, younger vs. older) to identify populations where explanations systematically fail to align with clinical expectations. A minimum acceptable threshold might be that at least 80% of explanations receive plausibility scores of 4 or 5, with no subgroup falling below 70%, and qualitative feedback should be collected to identify recurring patterns of implausibility that could be addressed through baseline recalibration or model retraining.

Baseline comparisons

The explainable framework should be compared against several baselines: a standard DeepSurv model without attribution capabilities (to assess whether adding Integrated Gradients degrades predictive performance, as gradient computation and the path integral approximation might introduce numerical instability or optimization challenges), a traditional Cox proportional hazards model with the same inputs (to evaluate whether deep learning provides meaningful improvement over a well-established, interpretable baseline that already offers coefficient-based explanations). A random survival forest (an ensemble of survival trees) serves as a non-neural, non-proportional-hazards alternative that can handle nonlinearities and interactions while providing variable importance measures, though tree-based importance scores are global rather than patient-specific and cannot attribute risk to temporal sequences of PSA measurements in the same way Integrated Gradients can.

For explanation quality, Integrated Gradients should be compared with SHAP (SHapley Additive exPlanations) applied to the same deep survival model, evaluating both computational efficiency (integrated gradients requires one gradient evaluation per integration step, typically 50-100 steps, while SHAP requires an exponential number of coalition evaluations unless approximation methods like KernelSHAP are used, which may still be slower for high-dimensional time-series inputs) and clinically assessed plausibility using the urologist review protocol described above. However, SHAP's exponential computational cost for many features—particularly when the PSA time-series encoder produces dozens or hundreds of temporal features—may disadvantage it in time-sensitive clinical applications where explanations must be generated within seconds of a patient visit, while Integrated Gradients scales linearly with input dimensionality and can be further accelerated through gradient checkpointing or parallel integration step computation.

Limitations

Technical limitations

Integrated Gradients requires an integral approximation along the path from baseline to input, which is computationally expensive when repeated for each patient and each time horizon; practical deployment may require reducing the number of integration steps or precomputing gradients for common input patterns. Baseline selection significantly affects attribution magnitudes: an alternative baseline, such as the population median PSA trajectory rather than zero PSA, would produce different numerical attributions, though the relative ordering of feature importance is often preserved [1]. Correlated features—such as closely spaced PSA values that rise together—pose a challenge because Integrated Gradients distributes credit among correlated inputs arbitrarily, potentially attributing risk to only one of several similarly informative measurements despite their joint contribution [23].

Clinical limitations

The framework provides correlational attributions describing which input features influenced the model's hazard prediction, not causal explanations of why metastasis will or will not occur; a high attribution to PSA velocity does not prove that PSA velocity causes metastasis—only that the model learned to associate it with metastasis risk. Prospective validation in external cohorts is essential before clinical deployment, as retrospective datasets may contain unmeasured confounding or selection bias that affects both the predictive model and its explanations [24]. Genomic score availability varies substantially across clinical settings, with community practices less likely to order Decipher or Oncotype DX testing than academic centers, potentially limiting the framework's applicability to patients without molecular profiling.

Conclusion

This manuscript has presented an explainable deep survival framework that integrates serial PSA measurements and genomic risk scores to predict time-to-metastasis in prostate cancer while providing feature-level attributions using Integrated Gradients. The framework transforms black-box hazard predictions into interpretable explanations that identify which specific PSA values at which time points and which genomic markers drive an individual patient's predicted metastasis risk, addressing a critical barrier to clinical adoption of deep learning for survival analysis.

The key advantages of this approach include the ability to attribute risk to clinically meaningful temporal patterns (PSA velocity, doubling time, absolute values) and molecular features (Decipher, Oncotype DX, Prolaris, Polaris scores) within a unified model, enabling clinicians to distinguish between risk driven by dynamic disease activity versus intrinsic tumor biology. Actionable explanations directly inform treatment decisions, from intensified surveillance to salvage therapy to systemic treatment, while providing a natural mechanism for clinician verification and override when explanations contradict established knowledge.

Several limitations must be acknowledged: Integrated Gradients imposes computational costs that may challenge real-time clinical use, the attribution baseline requires careful selection, correlated PSA measurements can produce unstable credit assignment, and explanations are correlational rather than causal. Prospective validation in external, multi-institutional cohorts is required before any clinical deployment, and the current framework does not address competing risks such as non-prostate-cancer mortality in older patients.

We call for implementation of this explainable survival framework on large-scale prostate cancer cohorts, including Surveillance, Epidemiology, and End Results (SEER) data linked to longitudinal PSA, Veterans Affairs (VA) corporate data warehouse with comprehensive genomic and outcomes data, CaPSURE (Cancer of the Prostate Strategic Urologic Research Endeavor) registry, and multi-institutional radical prostatectomy registries with complete follow-up. Such implementations will enable rigorous evaluation of explanation quality, clinical utility, and generalizability across diverse patient populations and healthcare settings.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 23 Nov 2024

Revised: 20 Jan 2025

Accepted: 15 Feb 2025

Published online: 20 July 2025

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *Proc Mach Learn Res (PMLR)*. 2017;70:3319–28.
- Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18(1):24.
- Lee C, Zame W, Yoon J, Van der Schaar M. DeepHit: a deep learning approach to survival analysis with competing risks. *Proc AAAI Conf Artif Intell*. 2018;32(1).
- Nagpal C, Li X, Dubrawski A. Deep survival machines: fully parametric survival regression and representation learning for censored data with competing risks. *IEEE J Biomed Health Inform*. 2021;25(8):3163–75.
- Yao J, Zhu X, Jonnagaddala J, Hawkins N, Huang J. Whole slide images-based cancer survival prediction using attention-guided deep multiple instance learning networks. *Med Image Anal*. 2020;65:101789.
- Wang Z, Gao Q, Yi X, Zhang X, Zhang Y, et al. Survformer: an interpretable pattern-perceptive survival transformer for cancer survival prediction from histopathology whole slide images. *Comput Methods Programs Biomed*. 2023;241:107733.
- Hao J, Kim Y, Mallavarapu T, Oh JH, Kang M. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Med Genomics*. 2019;12(Suppl 10):189.
- Langbein SH, Krzyżiński M, Spytek M, Baniecki H, Biecek P, et al. Interpretable machine learning for survival analysis. *Biom J*. 2025;67(6):e70089.
- Mohamed YA, Khoo BE, Asaari MS, Aziz ME, Ghazali FR. Decoding the black box: explainable AI (XAI) for cancer diagnosis, prognosis, and treatment planning—a state-of-the-art systematic review. *Int J Med Inform*. 2025;193:105689.
- Núñez JJ, Leung B, Ho C, Bates AT, Ng RT. Predicting the survival of patients with cancer from their initial oncology consultation document using natural language processing. *JAMA Netw Open*. 2023;6(2):e230813.
- Dai X, Park JH, Yoo S, D'Imperio N, McMahon BH, Rentsch CT, et al. Survival analysis of localized prostate cancer with deep learning. *Sci Rep*. 2022;12(1):17821.
<https://doi.org/10.1038/s41598-022-22118-y>.

Lee C, Light A, Saveliev ES, Van der Schaar M, Gnanapragasam VJ. Developing machine learning algorithms for dynamic estimation of progression during active surveillance for prostate cancer. *NPJ Digit Med.* 2022;5:110.

Lee HW, Kim E, Na I, Kim CK, Seo SI, Park H. Novel multiparametric MRI-based deep learning and clinical parameter integration for prediction of long-term biochemical recurrence-free survival in prostate cancer after radical prostatectomy. *Cancers.* 2023;15(13):3416.

Hu C, Qiao X, Huang R, Hu C, Bao J, Wang X. Development and validation of a multimodality model based on whole-slide imaging and biparametric MRI for predicting postoperative biochemical recurrence in prostate cancer. *Radiol Imaging Cancer.* 2024;6(3):e230143.

Ferguson T, Ravani P, Sood MM, Clarke A, Komenda P, Rigatto C, et al. Development and External Validation of a Machine Learning Model for Progression of CKD. *Kidney Int Rep.* 2022;7(8):1772-81.
<https://doi.org/10.1016/j.ekir.2022.05.004>.

Al Hussein Al Awamlh B, Wallis CJ, Penson DF, Huang LC, Zhao Z, et al. Functional outcomes after localized prostate cancer treatment. *JAMA.* 2024;331(4):302–17.

Pan H, Wang J, Shi W, Xu Z, Zhu E. Quantified treatment effect at the individual level is more indicative for personalized radical prostatectomy recommendation: implications for prostate cancer treatment using deep learning. *J Cancer Res Clin Oncol.* 2024;150(2):67.

Wong NC, Lam C, Patterson L, Shayegan B. Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy. *BJU Int.* 2019;123(1):51–7.

Lee SJ, Yu SH, Kim Y, Kim JK, Hong JH, Kim CH, et al. Prediction system for prostate cancer recurrence using machine learning. *Appl Sci.* 2020;10(4):1333.

Park J, Rho MJ, Moon HW, Kim J, Lee C, Kim D, et al. Dr. Answer AI for Prostate Cancer: Predicting Biochemical Recurrence Following Radical Prostatectomy. *Technol Cancer Res Treat.* 2021;20:15330338211024660.
<https://doi.org/10.1177/15330338211024660>.

Huang W, Randhawa R, Jain P, Hubbard S, Eickhoff J, Kummar S, et al. A Novel Artificial Intelligence-Powered Method for Prediction of Early Recurrence of Prostate Cancer After Prostatectomy and Cancer Drivers. *JCO Clin Cancer Inform.* 2022;6:e2100131.
<https://doi.org/10.1200/CCI.21.00131>.

Ekşi M, Evren İ, Akkaş F, Arıkan Y, Özdemir O, Özlü DN, et al, Taşçı AI. Machine learning algorithms can more efficiently predict biochemical recurrence after robot-assisted radical prostatectomy. *Prostate.* 2021;81(12):913-20.
<https://doi.org/10.1002/pros.24188>.

Bergero MA, Martínez P, Modina P, Hosman R, Villamil W, Gudiño R, et al. Artificial intelligence model for predicting early biochemical recurrence of prostate cancer after robotic-assisted radical prostatectomy. *Sci Rep.* 2025;15(1):30822.
<https://doi.org/10.1038/s41598-025-16362-1>.

Parker CC, Clarke NW, Cook AD, Kynaston HG, Petersen PM, Catton C, et al. Timing of radiotherapy after radical prostatectomy (RADICALS-RT): a randomised, controlled phase 3 trial. *Lancet.* 2020;396(10260):1413-21.
[https://doi.org/10.1016/S0140-6736\(20\)31553-1](https://doi.org/10.1016/S0140-6736(20)31553-1).

Van Den Eeden SK, Lu R, Zhang N, Quesenberry CP Jr, Shan J, et al. A biopsy-based 17-gene genomic prostate score as a predictor of metastases and prostate cancer death in surgically treated men with clinically localized disease. *Eur Urol.* 2018;73(1):129–38.

Brooks MA, Thomas L, Magi-Galluzzi C, Li J, Crager MR, Lu R, et al. GPS Assay Association With Long-Term Cancer Outcomes: Twenty-Year Risk of Distant Metastasis and Prostate Cancer-Specific Mortality. *JCO Precis Oncol.* 2021;5:PO.20.00325.

<https://doi.org/10.1200/PO.20.00325>.

Nguyen PL, Haddad Z, Ross AE, Martin NE, Deheshi S, Lam LLC, et al. Ability of a Genomic Classifier to Predict Metastasis and Prostate Cancer-specific Mortality after Radiation or Surgery based on Needle Biopsy Specimens. *Eur Urol.* 2017;72(5):845-52.

<https://doi.org/10.1016/j.eururo.2017.05.009>.

Van den Broeck T, Moris L, Gevaert T, Tosco L, Smeets E, Fishbane N, et al. Validation of the Decipher Test for Predicting Distant Metastatic Recurrence in Men with High-risk Nonmetastatic Prostate Cancer 10 Years After Surgery. *Eur Urol Oncol.* 2019;2(5):589-96.

<https://doi.org/10.1016/j.euo.2018.12.007>.

Jairath NK, Dal Pra A, Vince R Jr, Dess RT, Jackson WC, Tosoian JJ, et al. A Systematic Review of the Evidence for the Decipher Genomic Classifier in Prostate Cancer. *Eur Urol.* 2021;79(3):374-83.

<https://doi.org/10.1016/j.eururo.2020.11.021>.