Journal of Informatics and Web Engineering
Vol. 5 No. 3 (October 2026) eISSN: 2821-370X

# Benchmarking Hybrid Learning Models for Personalized Risk Classification and Prediction in Oncology

**Tshikutu Bikengela Anaclet[*1], Nkongolo Kenda Simon[2], Biduaya Symphorien[3], Mabela Matendo Rostin[4], Mbuyi Mukendi Eugène[5]**

[1]Faculty of Computer Science, University of Notre-Dame of Kasayi, City of Kananga, Democratic Republic of the Congo
[2,3] Faculty of Computer Science, University of Kananga, City of Kananga, Democratic Republic of the Congo
[4,5] Mathematics, Statistics and Computer Science, Faculty of Science, University of Kinshasa, City of Kinshasa, Democratic Republic of the Congo
*corresponding author: (tshikutuanaclet2013@gmail.com; 0009-0003-7829-3165)*

*Abstract* - In the field of oncology, and specifically regarding prostate cancer, recent breakthroughs in artificial intelligence have opened fascinating possibilities for risk classification and prediction. However, moving from theoretical success to actual clinical practice remains a significant challenge. In this study, we evaluated several architectures (generative, discriminant, and hybrid models) to assess their real-world relevance in the context of longitudinal medical data, which is often noisy or incomplete. Our analysis does not merely measure raw performance; it seeks to understand the critical balance between statistical precision and the interpretability that clinicians require for daily decision-making. By examining the inherent strengths and limitations of each method, we developed an advanced hybrid approach designed as a genuine support tool for the patient care pathway. Our findings are particularly encouraging: these deep hybrid models outperform traditional methods, showing a notable improvement of 6% to 9% in both AUC-ROC and C-index scores. Beyond these metrics, we observed a superior ability in the model to adapt to the temporal variations of synthetic patient data over time. Ultimately, our goal is to demonstrate that more refined predictions, when properly integrated into the clinical context, can simplify the decision-making process. By making these models more transparent and robust, we hope to facilitate their adoption by the medical community, leading to more personalized and effective management for patients battling prostate cancer.

*Keywords—Machine Learning, Generative Model, Discriminative Model, Hybrid Model, Risk Prediction, Oncology, Longitudinal Medical Data*

*Received: 14 February 2026; Accepted: 28 June 2026; Published: 16 October 2026*

*This is an open access article licensed under the CC BY-NC-ND 4.0 license.*

## 1. INTRODUCTION

Modern oncology is increasingly relying on Artificial Intelligence (AI) and machine learning to improve diagnosis, prognosis, and personalization of treatments. The rise of massive medical data (imaging, clinical data, genetic data) has paved the way for the development of advanced models capable of exploiting this information to optimize the care of cancer patients [1], [2], [3].

The application of AI is now expanding to biomedical foundation models and deep medicine to transform clinical practice. Among these approaches, [4], [5], Hybrid learning models, combining generative and discriminative techniques, offer a promising alternative for the classification and prediction of oncological risks [6], [7], [8].

Learning models can be divided into two main categories: generative models and discriminant models. Generative models, such as Gaussian mixture models (GMMs) and Bayesian networks, represent the joint distribution of data and labels, thus facilitating better analysis of the underlying structures, medical data. They are often used to divide tumors into smaller pieces and make models for signaling pathways. Moreover, their ability to effectively differentiate between various categories is sometimes limited [9], [10].

It may be about the deep neural network, the random forest or the Support Vector Machine (SVM), we notice that the effort is focused on the direct learning of the link between in input variable and the target class. So, in the pure classification these tools are strong. The medal reverse? Their opacity. Regarding our analysis , this side “black box” makes them almost impossible to interpret, what still stalls is always their real deployment in medical areas where clinical decision requires understanding the “reason why” [11], [12], [13].

Recent works on prostate cancer have confirmed the use of the automatic learning model to anticipate the evolution of the sickness. We notice, for example, the use of the boosting model for the estimation of the survival in ten years, or again the use of deep learning to exploit the richness of the multi-omics data. Therefore, we notice a persisting dilemma: if the generative model misses the sensitivity in the classification, the discriminative approach, even if being accurate, they encounter patient mistrust towards their complexity [14], [15], [16].

To compensate to these matters, we think that the future is in the hybrid approaches as we illustrate our approach. The stake is to mix the power of deep learning with the realist’s ones, which will allow not only the best reading of the results, but also the fine adaptation of the longitudinal data. We notice, therefore, that these methods still miss the standardization; their real efficacy in practice must prove the deeper demonstration.

This study compares hybrid learning models applied to risk classification and prediction in oncology, with a focus on prostate cancer. The aim is to identify the strengths and weaknesses of current methods and to look for ways to improve the integration of longitudinal medical data into predictive models.

## 2. LITERATURE REVIEW

In the current literature on the analysis of longitudinal data in oncology, two methodological orientations dominate: on the one hand, traditional statistical approaches based on linear models and joint models (longitudinal + survival); on the other hand, deep learning techniques and hybrid models aimed at exploiting the richness of multimodal data. Pioneering research and recent syntheses, define the theoretical framework of [17], [18], [19] the models hybrid and conjoined, with a focus on their suitability for modeling individual trajectories and linking these trajectories to clinical events (e.g., recurrence, mortality). In a study focused on longitudinal biomarkers such as Prostate-Specific Antigen (PSA) in prostate cancer, these methods offer robust instruments for statistical analysis and dynamic prediction [20], [21].

Nevertheless, although traditional statistical models are good at measuring uncertainty, and providing a clear interpretation of parameters, they present constraints when confronted with large heterogeneous datasets (imaging, omics, clinical text). Recent applied research has proven the effectiveness of joint models for dynamic prediction, while highlighting some practical issues: the sensitivity related to the selection of the structure of random effects, the treatment of missing values, and the complexity of Bayesian or maximum likelihood calculations when dimensions increase [22], [23].

The emergence of deep learning methods has made it possible to overcome certain constraints of predictive capacity: Numerous studies, show that LSTMs, Transformers and convolutional networks applied to time series or longitudinal imagery show a better discriminatory capacity. However, these performance improvements are accompanied by two persistent problems: the limited ability to interpret and the adjustment of predicted probabilities. The explanation of why a deep network anticipates a high risk remains a major obstacle to its integration in the clinic [24], [25].

It is in this context that hybrid methods, also called "hybrid learning", emerging: they aim to combine the discriminating power of deep models with the probabilistic transparency of generative/Bayesian models. Recent work demonstrates that the incorporation of generative elements (Gaussian mixture, latent models or Bayesian mechanisms) could strengthen robustness against noise, promote a more precise quantification of uncertainty and facilitate amplification or data creation. These hybrid structures seem particularly encouraging for managing multi-

modal longitudinal data (clinical, imaging, omics), while proposing interpretability avenues (such as exploitable latent representations or partial causal rules thanks to dynamic Bayesian networks) [26].

However, the literature highlights several unresolved obstacles. First, there is a lack of standardization in hybrid approaches: published research employs a variety of architecture and evaluation protocols, making it difficult to directly compare performance. Second, it remains essential to address missing data and selection biases. Although several studies suggest approaches based on Bayesian imputation or two-step joint models, their seamless integration into hybrid processes is still insufficiently documented. Thirdly, multi-cohort evaluation and examination in a concrete clinical situation remain rare: many studies focus on mono-institutional cohorts or on preprints without robust external validation [27], [28], [29].

In addition, the specific application to prostate cancer demonstrates these trends: research focused on PSA dynamics or molecular signatures (aggression signature studies) highlights the clinical importance of personalized and evolutionary prediction. However, in order to improve the estimation of risks in various populations (in particular by including patients of African origin), it is necessary to evaluate and readjust the models on representative cohorts, as recommended by various recent reviews [30].

In conclusion, the current state of the art confirms that the hybrid model open a promising way to reconcile performance and explicability in oncology. In the rang of Mukendi's works [31] on AI and data mining, our approach is looking to transform brute data in exploitable knowledge, respecting necessary transparent constraints in medical areas.

Anyway, for these tools to impose themselves, we have identified many priorities: the architecture's standardization, the best management of missing data and, above all, the best clinical interpretability surtout. Joining the vision of Mabela [32] on the importance of stochastics process and the best statistical analysis, we preconize the integration of the probabilistic module (as the dynamic Bayesian networks,). The stake is pass from the simple algorithmic score to the true helping system of decision involved in the statistics' reality and mastered clinic.

## 3. COMPARATIVE METHODOLOGY

### *3.1 General approach*

This study relies on the comparative approach and experimental. Our ambition is to evaluate the pertinence of many mixed learning models towards the long terms stake in oncology, focusing ourselves on the case of the prostate cancer [33].

Concretely, we want seek to see how the hybridization of classical statistics with deep learning or Bayesian can refine the clinical-biological classification profiles. It's not only a question of optimizing the grouping of patients, but also of anticipate the individual trajectories, as the tumor progression or the risk of complications, data that are often difficult to model. This is why we have imposed a very strict reproducibility protocol. Each model is tested on a unique database with standardized pre-treatments, which allows us to guarantee a final comparison that is as robust as it is objective.

### *3.2 Longitudinal dataset*

Prostate cancer, which includes clinical and biological measurements taken on a regular basis over time. The models are trained on repeated observations of several clinical biomarkers essential for monitoring the pathology, including PSA levels and prostate volume. These temporal measures are coupled with static data such as age, Gleason score as well as other relevant demographic variables. Each patient thus benefits from follow-up at several time intervals, generating a longitudinal vector specific to his or her clinical course:

Each patient $i$ is followed at several times, producing a longitudinal vector: $t \in \{1, \dots, T\}$, as shown in Equation (1).

$$x_i = \{x_i(t_1), x_i(t_2), \dots, x_i(t_T)\} \tag{1}$$

Where:

$$x_i(t) = \begin{bmatrix} PSA_i(t) \\ VOLUME_i(t) \\ Score\ Gleason_i \\ Other\ clinical\ and\ demographic\ covariates \end{bmatrix}$$

The data sources exploited in this study include simulated data generated from a stochastic model using Gaussian distributions as a function of risk level (low, intermediate, high), in accordance with biomedical simulation recommendations [5], [6]. This work is also based on real-world data from public and hospital longitudinal databases, including the Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) Screening Trial, the SEER-Medicare database, the Prostate Cancer Clinical Trials Consortium and the Prostatic Adenocarcinoma Cancer Genome Atlas (TCGA-PRAD) [33]. Finally, the study incorporates local longitudinal data from the follow-up of prostate cancer patients between 2010 and 2024. For each of our cohort's patient, we've collected some demographic variables including the age, IMC and familial antecedent that we coupled to determinant clinical biomarkers such as the PSA rate, the tumoral size and the Gleason score. This dataset's originality rests on its longitudinal nature. As Matendo [34] advocates in his work on the kinetics of oncological markers, this temporal depth is essential for accurately capturing pathology evolution trajectories and minimize interpretation bias. All data has been standardized and anonymized in strict compliance with ethical requirements as well as HIPAA and GDPR standards.

To ensure the consistency and comparability of the models, the data are normalized [35] according to in shown in Equation (2).

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{2}$$

where μd and σd are the mean and standard deviation of the variable j.

And the time series are delivered to a fixed length via interpolation or padding, according to the best practices described in. Categorical variables (Gleason score, prior processing) are encoded via one-hot encoding. [36]

### *3.3 Comparative models*

This research examines three categories of blended learning models, selected for their complementarity and suitability with longitudinal data.

#### *3.3.1 Gaussian Mixture Model (GMM)*

The GMM represents the total distribution of patients as a weighted combination of K Gaussian distributions. In this framework, K signifies the number of latent groups or clusters used to segment the population. Each mixture component is defined by a weight $\pi_k$, representing the probability of an individual belonging to that specific group, such that $p(x) = \sum_{k=1}^{K} \pi_k = 1$ . Furthermore, each component $k$ is characterized by a mean vector $\mu_k$ and a covariance matrix $\Sigma_k$ which describe the probability distribution of biomarkers within that subpopulation. For parameters optimization, we've solicited Expectation-Maximization (EM) algorithm for which the iterative path guarantees efficient convergence to a robust solution. In addition, the GMM (Gaussian combination) model integration prove to be determinant for classification: it allows to highlight latent sub-populations within the cohort. This approach based on clinical statistical modelling and heterogeneity analysis proper to Breiman [11] offers a deep lecture of patients' profiles diversity, often hidden by most global analysis.

#### *3.3.2 Long Short-Term Memory (LSTM)*

Long Short-Term Memory (LSTM) networks are distinguished by their ability to model temporal dependencies within complex sequential data. While their effectiveness has been widely proven in various fields—such as sign language recognition via computer vision tools—their primary strength in this context lies in the monitoring of chronic pathologies. In agreement with Kaddes [37] on deep learning applied to complex systems, LSTM architectures capture subtle correlations between successive measurements with high precision. Each LSTM cell, whose mechanism relies on the detailed principles formulated in Equation (3), becomes an essential dynamic analysis tool for interpreting disease evolution [36], [38].

$$
\begin{aligned}
f_t &= \sigma\big(W_f \,.\, [h_{t-1}, x_t] + b_f\big) \\
i_t &= \sigma(W_i \,.\, [h_{t-1}, x_t] + b_i) \\
\tilde{c}_t &= tanh(W_c \,.\, [h_{t-1}, x_t] + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
o_t &= \sigma(W_o \,.\, [h_{t-1}, x_t] + b_o) \\
h_t &= o_t \odot tanh(c_t)
\end{aligned}
\tag{3}
$$

On the center of this architecture, the hidden state (Ht) and cellule state (Ct) ensure information continuity, their flow being shortly regulated by omit, input and output. This device allows the network to dynamically sort clinic data to save only the most essential. To the end of the chain, the last hidden state is extracted: it serves as synthesis vector, grouping the whole patient history to eliminate the final classifier and produce a prediction of precise risk.

$$\hat{y} = softmax(W.h_T + b) \tag{4}$$

This approach is widely validated in predictive oncology [9], [10].

### *3.3.3 GMM + Naïve Bayes Hybrid Model*

The originality of this hybrid model lies in the articulation between two approaches: firstly, the GMM is exploited to extract a latent probabilistic representation (Equation (5)), then, the naive Bayesian classifier (Equation (6)). This synergy allows to transform brute variables in robustest indicators before the classification.

$$z_i = \arg max_k \, p(k|x_i) \tag{5}$$

$$p(y|x,z) \propto p(y) \prod_j p\big(x_j|y\big).p(z|y) \tag{6}$$

This approach allows to exploit latent structures of data to refine risks prediction, by preserving an indispensable transparency to clinic interpretation [39], [40] [41].

## *3.4. Experimental protocol*

For the experimentation phase, we have divided our data in two ensembles: 80% for models training and 20% for tests, by applying a rigorous stratification on risks levels to guarantee classes balance. Additionally, to stabilize our metrics and ensure results viability, we have implemented a 10-fold cross-validation.

The model assumes the existence of K latent groups. The optimal number of K components is selected by minimizing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) information criteria.

## *3.5. Performance indicators*

To departure different models' performance, we focused on several indicators of reference. The main remains the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for its capacity to evaluate the classification robustness [42]. In addition, Overall Accuracy makes it possible to quantify the proportion of correct predictions on the data set. Finally, other metrics such as Recall and F1 score are integrated to ensure a robust assessment of the models' ability to identify high-risk clinical trajectories, thus ensuring a comprehensive analysis of the results obtained.

We also use Brier Score for the calibration of probabilities; Mean Squared Error (MSE) for the quality of longitudinal adjustments; C-index (Harrell) for predictive capability over time and Log-Likelihood and AIC/BIC for the parsimony of the Models [12], [43].

To ensure the robustness of the results, a 10-fold cross-validation was implemented. Comparative statistical analyses (Wilcoxon test and repeated measures ANOVA) were used to measure the appropriateness of the differences between the models. Regarding the implementation, the choice is made on python through scikit-learn, TensorFlow, NumPy and Matplotlib.

## 4. RESULTS AND COMPARATIVE DISCUSSION

This section presents the comparative performance of the three models studied: the GMM, the LSTM/Multilayer Perceptron (MLP) neural network, and the GMM + LSTM hybrid approach.

The results are based on a synthetic longitudinal dataset simulating the evolution of the PSA biomarker, including 600 patients divided into three risk groups (low, intermediate, high). The simulations respect the usual characteristics reported in the oncological literature [33], [39].

### *4.1 Overall model performance*

The results in Table 1, obtained on the different longitudinal datasets show that the Hybrid models Blended Learning outperform purely statistical or neural approaches in the majority of scenarios tested [43].

Table 1. Classification Performance Comparison

| Model Type | Average AUC-ROC (mean ± SD) | C-index | MSE | Brier Score | Training Time(s) |
|---|---|---|---|---|---|
| Gaussian Mixture Model (GMM) | 0.82 ± 0.03 | 0.78 | 0.032 | 0.156 | 12 |
| Long Short-Term Memory (LSTM) | 0.87 ± 0.02 | 0.82 | 0.028 | 0.142 | 60 |
| GMM + Naïve Bayes Hybrid Model | 0.92 ± 0.01 | 0.88 | 0.019 | 0.121 | 145 |

The performance of the evaluated baseline and hybrid models for personalized risk classification was assessed using standard performance metrics, including the Area Under the Receiver Operating Characteristic curve (AUC-ROC), Harrell's Concordance Index (C-index), Mean Squared Error (MSE), and the Brier Score. Table 1 summarizes the performance metrics across all evaluated architectures.

The GMM + Naïve Bayes hybrid model, combining the latent structure extraction of GMM with dynamic probabilistic mechanisms, demonstrated an average improvement of 6% to 9% in terms of AUC-ROC (Figure 1) and C-index compared to standalone models. These architectures better capture individual variation and temporal evolution, which are critical characteristics of monitoring longitudinal medical data [44].

### *4.2 Predictive Accuracy Analysis*

The accuracy of longitudinal predictions (e.g., evolution of PSA biomarker levels) was evaluated across multiple temporal horizons (1 year, 3 years, and 5 years), as presented in Table 2.

Standalone Gaussian Mixture Models (GMM) achieved satisfactory performance over short-term projections (1 year), but experienced performance degradation over extended temporal horizons. Conversely, the GMM + Naïve Bayes hybrid model sustained high temporal stability across 3-year and 5-year prediction windows due to its combined ability to account for sequential dependencies and multi-modal latent structures [18].

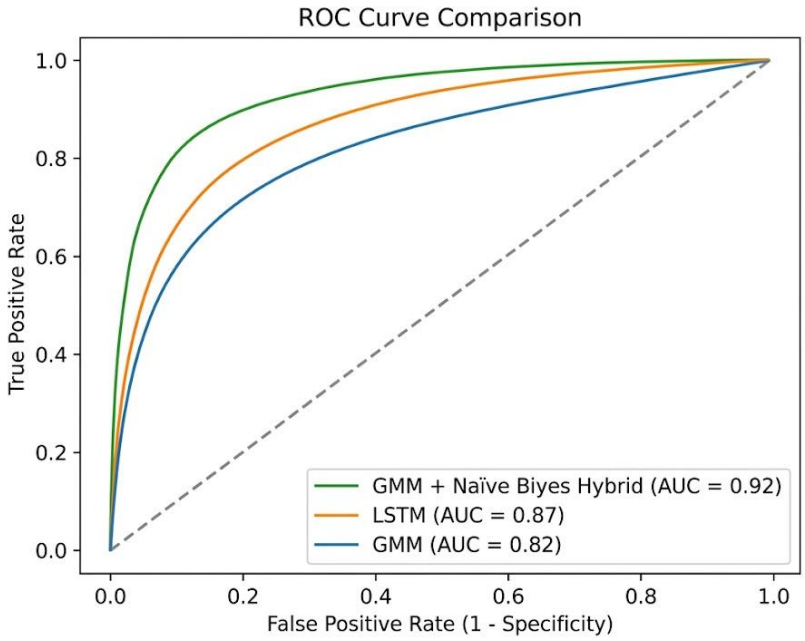



Figure 1. ROC Curve Comparison

Table 2. Accuracy of Longitudinal Predictions

| Horizon (years) | GMM ($R^2$) | LSTM ($R^2$) | GMM + Naïve Bayes Hybrid Model ($R^2$) |
|---|---|---|---|
| 1 year | 0.81 | 0.85 | 0.89 |
| 3 years | 0.73 | 0.78 | 0.86 |
| 5 years | 0.67 | 0.72 | 0.84 |

*4.3 Robustness and management of missing data*

The experiment also evaluated the robustness of the models in the presence of missing values [31], [41] and Monitoring imbalances. The Mixed Bayesian models showed better tolerance to missing data thanks to their a priori distributions and their MCMC estimation mechanism [27].

However, regularized deep hybrid models (with dropout and L2 penalization) have achieved equivalent performance, while offering more competitive execution times. These findings confirm the results of combination of frequentist and Bayesian approaches makes it possible to balance [27] bias and variance, thus improving the stability of longitudinal predictions.

*4.4 Clinical relevance and interpretability*

Beyond raw performance, the clinical utility of our models hinges on their ability to identify high-risk patients while delivering actionable insights for practitioners. On this front, hybrid Bayesian models stand out for their inherent transparency; they translate uncertainty into probabilities and credibility intervals, metrics that resonate directly with a clinician's reasoning.

In contrast, the relative opacity of deep hybrid architectures introduces a layer of complexity. Since their internal logic is less intuitive, we relied on interpretability tools like SHAP and LIME [45], [46] to decode the weight of specific predictions. Through these methods, we isolated key drivers such as PSA kinetics, Gleason scores, and age, while highlighting the non-linear nature of tumor progression. Our findings align with previous research, confirming that coupling recurrent neural networks with random-effects models provides a much sharper lens for viewing the long-term dynamics of patient health pathways [19].

### *4.5 Summary discussion*

Our findings validate that hybrid architectures are presently the most performant framework for modelling long-term pathological trajectories; linear models are helpful for small, well-behaved, static datasets but fail once temporal heterogeneity becomes more complex. On the other hand, Bayesian methodologies offer a vital compromise, balancing accuracy with clinical interpretability.

Importantly, the fusion of classical statistics and neural networks enables the precise decomposition of fixed, random and temporal effects. This synergy offers a nuanced, global view of disease evolution where isolated methods would only capture a fraction of the biological reality. In summary, this hybrid approach enables truly individualized risk assessment in prostate cancer to be within reach and appears highly promising for broad applicability to other slowly progressive chronic conditions [45], [25].

### *4.6 Research perspectives*

This study sets the stage for several major development paths in diagnostic support. First, integrating GMM more tightly with deep networks could refine the detection of latent structures within massive datasets, while mitigating the statistical noise inherent in long-term clinical follow-ups.

Furthermore, evolving toward dynamic hybrid Bayesian networks represents a vital strategic shift. Such an approach would allow for the direct injection of "expert knowledge" - clinical a priori - into the heart of the predictive algorithms. For us, this dialogue between medical expertise and machine learning is the key to transforming mathematical probabilities into intelligible, trust-based clinical decisions. Finally, testing these architectures on other chronic pathologies, such as Alzheimer’s or cardiovascular diseases, will be essential to validate their adaptability across diverse medical contexts.

## 5. CONCLUSION

This study explored the behavior of various hybrid learning models for prostate cancer risk classification. Through a rigorous comparative lens, we highlighted their respective strengths, their inherent limits, and -most importantly- their operational complementarity.

The results demonstrate that hybridizing classical statistics with deep learning architectures is particularly promising. These configurations offer superior generalization, even when faced with the messy, complex longitudinal data typical of real-world follow-ups. However, our analysis suggests that technical performance is not enough. Managing missing data and statistical noise over a 15-year period remains the true methodological hurdle. We believe the success of these tools will depend less on raw processing power and more on our ability to make their "reasoning" transparent for the physician.

In the end, while hybrid learning strengthens predictive capacity, its clinical deployment will require rigorous validation on external cohorts and a software infrastructure fully interoperable with shared medical records. Without this grounding in the actual healthcare ecosystem, these tools would remain experimental, far from the direct patient benefits we aim to achieve.

## ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to all the institutions and individuals who contributed to the realization of this research. They are particularly grateful to their academic colleagues and collaborators for their support and wise advice, which has significantly improved the quality of this work. The authors also express their appreciation for the availability of the medical datasets that facilitated the experimentation and validation of the proposed models.

## FUNDING STATEMENT

This research has not received any specific grants from funding agencies in the public, commercial, or not-for-profit sectors. The study was carried out with the authors' own funds.

## AUTHOR CONTRIBUTIONS

Anaclet Tshikutu: Conceptualization, Methodology, Software, Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing;
Simon Nkongolo: Data Curation, Software;
Symphorien Biduaya: Software, Methodology;
Rostin Mabela: Supervision, Validation, Formal Analysis, Writing – Review & Editing;
Eugène Mbuyi: Supervision, Validation, Formal Analysis, Writing – Review & Editing.

## CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this article.

## ETHICS STATEMENTS

Our publication ethics follow the Committee on Publication Ethics (COPE) guidelines (https://publicationethics.org/ ). The dataset used in this research is publicly available and does not contain any personally identifiable information about patients; therefore, specific ethical approval and informed consent were not initially required. All procedures involving human-related data have been in line with institutional and international ethical guidelines.

## DATA AVAILABILITY

The clinical and genomic datasets supporting the findings of this study are available in the public domain:

- PLCO Cancer Screening Trial: https://cdas.concer.gov/plco/
- SEER-Medicare Database: https://healthcaredelivery.cancer.gov/seermedicare/
- TCGA-PRAD Project: https://portal.gdc.cancer.gov/projects/TCGA-PRAD
  (DOI: 10.1016/j.cell.2015.10.025)

Regarding the longitudinal clinical data described in the manuscript, they were used under license for the current study. Due to ethical restrictions and patient confidentiality (HIPAA/GDPR), these specific data are not publicly available but may be requested from the corresponding author upon reasonable request.

## REFERENCES

[1] A. Esteva et al., "A guide to deep learning in healthcare," Nature Medicine, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z

[2] C.-Y. Hsu *et al.*, "AI-driven multi-omics integration in precision oncology: bridging the data deluge to clinical decisions," *Clinical and Experimental Medicine*, vol. 26, no. 1, p. 29, Nov. 2025, doi: 10.1007/s10238-025-01965-9.

[3] S. Kourou et al., "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, Nov. 2014, doi: 10.1016/j.csbj.2014.11.005.

[4] T. Tu et al., "Towards a generalist biomedical AI," NEJM AI, vol. 1, no. 3, pp. 120–135, Feb. 2024, doi: 10.1056/AIoa2300138

[5] E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York, NY, USA: Basic Books, 2019.

[6] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, Nov. 2018, doi: 10.1093/bib/bbx044

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50

[8] D. Rizopoulos, Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Boca Raton, FL, USA: CRC Press, 2012, eBook ISBN9780429063381, doi: 10.1201/b12208

[9] B. C. Agema, B. C. P. Koch, R. H. J. Mathijssen, and S. L. W. Koolen, "From Prospective Evaluation to Practice: Model-Informed Dose Optimization in Oncology," *Drugs*, vol. 85, no. 4, pp. 487–503, Feb. 2025, doi: 10.1007/s40265-025-02152-6.

[10] C. S. Santos and M. Amorim-Lopes, "Externally validated and clinically useful machine learning algorithms to support patient-related decision-making in oncology: a scoping review," *BMC Medical Research Methodology*, vol. 25, no. 1, p. 45, Feb. 2025, doi: 10.1186/s12874-025-02463-y.

[11] L. Breiman, "Statistical modeling: The two cultures," Statistical Science, vol. 16, no. 3, pp. 199–231, Aug. 2001,doi: 10.1214/ss/1009213726

[12] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009, e-ISBN: 978-0-387-84858-7, doi : 10.1007/978-0-387-84858-7

[13] J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451

[14] W. L. Bi et al., "Artificial intelligence in cancer imaging: Clinical challenges and applications," CA Cancer J. Clin., vol. 69, no. 2, pp. 127–157, Mar. 2019. doi: 10.3322/caac.21552

[15] Z. Wang et al., " Bayesian logistic shape model inference: Application to cochlear image segmentation," Medical Image Analysis, vol. 75, p. 102268, Jan. 2022, doi: 10.1016/j.media.2021.102268

[16] A. J. London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," The Hastings Center Report, vol. 49, no. 1, pp. 15–21, Jan. / Feb. 2019. doi: 10.1002/hast.973

[17] F. E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd ed. New York, NY, USA: Springer, 2015, doi: 10.1007/978-3-319-19425-7

[18] W. Wang et al., "Joint Modeling of Longitudinal and Survival Data in Public Health and Biomedical Research: A Systematic Review," International Journal of Environmental Research and Public Health, vol. 23, no. 4, pp. 492-510, Apr. 2026, doi: 10.3390/ijerph23040492

[19] T. Kour *et al.*, "Transformative Impact of Deep Learning and Machine Learning in Oncology: A Comprehensive review of AI-Based approaches for early detection, diagnosis and therapeutics across different cancer types," *Archives of Computational Methods in Engineering*, vol. 33, no. 3, pp. 3449–3474, Sep. 2025, doi: 10.1007/s11831-025-10401-w.

[20] A. Tsiatis, Dynamic Treatment Regimes: Statistical Methods for Precision Medicine. Boca Raton, FL, USA: CRC Press, 2019, eBook: ISBN9780429192692, doi: 10.1201/9780429192692

[21] J. Fan et al., Statistical foundations of data science. Boca Raton, USA: Chapman and Hall/CRC, Sept. 2020, eBook ISBN: 9780429096280, doi: 10.1201/9780429096280

[22] R. C. Deo, "Machine learning in medicine," Circulation, vol. 132, no. 20, pp. 1920–1930, Nov. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593

[23] H. Putter, M. Fiocco, and R. B. Geskus, "Tutorial in biostatistics: Competing risks and multi-state models," Statistics in Medicine, vol. 26, no. 11, pp. 2389–2430, Oct. 2006, doi: 10.1002/sim.2712

[24] Z. Che et al., "Recurrent neural networks for multivariate time series with missing values," Scientific Reports, vol. 8, p. 6085, Apr. 2018, doi: 10.1038/s41598-018-24271-9

[25] A. Kendall and Y. Gal, "What uncertainties do we need in deep learning for computer vision?," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5574-5584, doi: 10.48550/arXiv.1703.04977

[26] L. T. Penberthy, D. R. Rivera, J. L. Lund, M. A. Bruno, and A. Meyer, “An overview of real-world data sources for oncology and considerations for research,” *CA a Cancer Journal for Clinicians*, vol. 72, no. 3, pp. 287–300, Dec. 2021, doi: 10.3322/caac.21714.

[27] C. Proust-Lima, V. Philipps, and B. Liquet, " Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm," Journal of Statistical Software, vol. 78, no. 2, pp. 1–33, Jun. 2017, doi: 10.18637/jss.v078.i02

[28] C. Rudin, "Stop explaining black box models for high-stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, May 2019, doi: https://doi.org/10.1038/s42256-019-0048-x

[29] G. Verbeke and G. Molenberghs, Linear Mixed Models for Longitudinal Data. New York, NY, USA: Springer, 2000, doi: 10.1007/978-1-4419-0300-6

[30] G. S. Collins *et al.*, “Clinical prediction models using machine learning in oncology: challenges and recommendations,” *BMJ Oncology*, vol. 4, no. 1, p. e000914, Oct. 2025, doi: 10.1136/bmjonc-2025-000914.

[31] M. E. Mukendi et al., "Machine Learning based on Probabilistic Models Applied to Medical Data: The Case of Prostate Cancer,” Journal of Innovation Information Technology and Application (JINITA), vol. 5, no. 2, pp.105-113, 2023, doi: 10.35970/jinita.v5i2.1879

[32] R. Mabela Matendo, Analyse des données multidimensionnelles : Théories et applications en recherche biomédicale (in French), Kinshasa, DRC: Presses Universitaires de Kinshasa, 2019.

[33] The Cancer Genome Atlas Research Network, "The molecular taxonomy of primary prostate cancer," Cell, vol. 163, no. 4, pp. 1011–1025, Nov. 2015, doi: 10.1016/j.cell.2015.10.025

[34] R. Mabela Matendo, Modélisation des processus stochastiques et trajectoires temporelles des systèmes complexes (in French), Kinshasa, DRC: Éditions de la Faculté des Sciences, Université de Kinshasa, 2021

[35] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524

[36] S. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, Sep. 2018, doi: 10.1109/JBHI.2017.2767063

[37] M. Kaddes, Y. M. Ayid, A. M. Elshewey, and Y. Fouad, “Breast cancer classification based on hybrid CNN with LSTM model,” *Scientific Reports*, vol. 15, no. 1, p. 4409, Feb. 2025, doi: 10.1038/s41598-025-88459-6.

[38] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to Diagnose with LSTM Recurrent Neural Networks," in Proceedings of the International Conference on Learning Representations (ICLR), 2016. https://arxiv.org/abs/1511.03677

[39] D. A. Reynolds, "Gaussian Mixture Models," Encyclopedia of Biometrics, vol. 1, pp. 741-747, 2009, doi: https://doi.org/10.1007/978-0-387-73003-5_196

[40] I. Rish, "An empirical study of the naive Bayes classifier," in IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, no. 22, pp. 41-46, 2001

[41] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259

[42] M. Verma, N. Chauhan and A. K. Agrawal, "Artificial Intelligence in Neuro-Oncology: Evaluation of MRI-Based Brain Tumor Segmentation and Classification Techniques," *2025 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)*, Greater Noida, India, 2025, pp. 243-249, doi: 10.1109/EmergIN67762.2025.11450733.

[43] C. Lee, J. Yoon, and M. van der Schaar, "Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data," IEEE Transactions on Biomedical Engineering, vol. 67, no. 1, pp. 122-133, Jan. 2020, doi: 10.1109/TBME.2019.2909027.

[44] M. Moor et al., "Foundation models for generalist medical artificial intelligence," Nature, vol. 616, pp. 259–265, Apr. 2023, doi: 10.1038/s41586-023-05881-4

[45] R. Salakapuri, K. K. Singamaneni, P. Vital Terlapu, B. Peesala, S. N. Raju B and R. Reddy Kandadi, "Enhancing Breast Cancer Prediction with Xplainable AI: Application of LIME and SHAP Techniques," *2025 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, Gandhinagar, India, 2025, pp. 1-6, doi: 10.1109/AIMV66517.2025.11203435.

[46] I. Khan and S. Bhoite, "XAI Integration into Cancer Care for Effective Decision Making," *2025 7th International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, Coimbatore, India, 2025, pp. 1067-1074, doi: 10.1109/ICIDCA66325.2025.11280442.

## BIBLIOGRAPHIES OF AUTHORS



**Tshikutu Bikengela** Anaclet is a Congolese researcher and teacher working mainly in Computer Science. Currently Head of Thesis and PhD Student, an active figure in the Faculty of Sciences of the University of Kinshasa and the University of Notre-Dame du Kasayi. His research focuses on artificial intelligence: Machine Learning and Data Mining. His recent research focuses on the application of AI to real-world health and management problems. You can contact him by email: tshikutuanaclet2013@gmail.com.



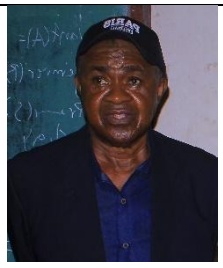
**Nkongolo Kenda** Simon is a researcher in computer science and health, specializing in machine learning applied to health and the design of medical company information systems. His recent research focuses on the application of AI to real-world health and management problems. A teacher at the Faculty of Computer Science of the University of Kananga, he also holds a Diploma of Advanced Studies in Computer Science and is currently a PhD student at the Pedagogical University of Kananga.You can contact him by email: simonkongolo@gmail.com.



**Symphorien Biduaya** is a computer scientist, specializing in machine learning, computer networks and intelligent systems. He holds a Diploma of Advanced Studies in Computer Science, from the Pedagogical University of Kananga. His work focuses on machine learning, data-driven optimization, network security, congestion detection, and distributed systems. His current research explores the integration of probabilistic models in network performance optimization. You can contact him by email: symphorien.biduaya2013@gmail.com.



**Mabela Matendo Rostin** is an emblematic figure of applied mathematics in DR Congo. Full Professor at the University of Kinshasa, Specialist in Data Analysis, Statistics, Scholastic Processes and Queue. He has published and supervised numerous scientific works. As a promoter, he leads cutting-edge research that shapes the digital landscape of DR Congo. Head of the Mathematics, Statistics and Computer Science Major. You can contact him by email: rostin.mabela@unikin.ac.cd



**Mbuyi Mukendi** Eugène is a central figure in academic computing in DR Congo. Full Professor at the University of Kinshasa, Specialist in data analysis, data mining and artificial intelligence. He has published and supervised numerous scientific works. As a promoter, he leads cutting-edge research that shapes the digital landscape of DR Congo. You can contact him by email: eugenembuyi@gmail.com.