AI-Assisted Clinical Decision Support for Urology Oncology: Baseline Machine Learning Experiments on Noisy EHR Data

Arnas Bosas [1], Eglė Butkevičiūtė [1]

[1] *Kaunas University of Technology, K. Donelaičio St. 73., Kaunas, 44249, Lithuania*

Abstract

Clinical decisions in urologic oncology are often complex and require combining patient history, test results, and multiple treatment options. Many studies report that AI-based clinical decision support systems (CDSS) can help in prostate and urology cancer care, but they commonly rely on well-structured or heavily curated inputs and do not fully exploit free-text documentation in electronic health records (EHRs) [1,2,6,10].

In this paper, baseline experiments are presented using real-world, de-identified EHR data from 23,110 urology oncology patients. The data include structured fields and a large amount of Lithuanian clinical text. Across all experiments, the prediction target is a clinical procedure (target exam/procedure), while the dataset formulations differ in how patient information is represented. Variant A uses a patient-level multi-label formulation based on structured fields and free text. Variants B and C predict the next target procedure from sequential clinical history, with Variant C additionally including examination results. Text was used with basic preprocessing only, without advanced clinical NLP or text structuring.

Four classical models were compared: logistic regression, support vector machines (SVM), Random Forest, and XGBoost. Accuracy, macro-F1, and weighted-F1 are reported due to strong class imbalance. Random Forest achieved the best overall performance (up to 0.84 accuracy and 0.82 weighted-F1), while macro-F1 remained below 0.49, indicating weak performance on rare labels. Logistic regression performed consistently poorly, and XGBoost results were inconsistent across formulations.

Overall, these baselines highlight that modelling directly from raw Lithuanian free-text EHR data is limited by noise, sparsity, and long-tail imbalance. Future work will focus on lightweight clinical text structuring using an NLP pipeline with entity extraction and normalization, as well as imbalance-aware modelling to improve rare-label coverage and support more reliable decision-support use.

Keywords

Clinical decision support; Urology oncology; Prostate cancer; Machine learning; Electronic health records; Text mining; Sequence prediction

1. Introduction

Prostate cancer management involves complex clinical decision-making across diagnosis, risk stratification, treatment selection, and follow-up. In urologic oncology, artificial intelligence (AI)–based clinical decision support systems (CDSS) are increasingly explored as tools to support clinicians by combining different types of information and providing consistent recommendations. Recent reviews show that a considerable amount of recent work has focused on AI-CDSS in prostate cancer and clinical urology, while also pointing out recurring issues such as validation, workflow integration, and real-world applicability [1,2,8,10].

Published CDSS efforts range from classical expert systems to machine learning (ML) models trained on electronic health record (EHR) data. For example, clinical documentation and EHR-derived signals have been used to predict post-treatment outcomes such as urinary incontinence following prostatectomy [3], and decision-support platforms such as Watson for Oncology have been assessed in prostate cancer care [4]. At the same time, reviews of AI applications in urology indicate expanding use cases and increasing interest in translating these tools into routine clinical workflows [8].

*IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania

arnbos@ktu.lt (A. Bosas); egle.butkeviciute@ktu.lt (E. Butkeviciute)

0000-0002-6652-2536 (E. Butkeviciute)

Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-06

CEUR-WS.org/Vol-4317/paper12.pdf

Despite progress, one practical problem remains the nature of real-world EHR data. Clinically relevant information is often recorded in free text, which motivates natural language processing (NLP) approaches for structured extraction in prostate cancer care [6]. More recently, interest has grown in combining structured standards such as openEHR with large language models for prostate cancer decision support [5]. However, surveys of diagnostic decision support paradigms repeatedly emphasize limitations that hinder deployment, including heterogeneous and noisy inputs, temporal complexity, and the need for robust evaluation and reproducibility [10].

In this paper, baseline ML experiments are presented on de-identified urology oncology EHR data from 23,110 patients containing extensive Lithuanian clinical text. Importantly, the initial data export was not modelling-ready. It followed an administrative wide-table format with repeated blocks such as Exam1…ExamN and Treatment1…TreatmentN. This structure produced systematic missingness and made longitudinal reconstruction non-trivial. Therefore, a key prerequisite for modelling was a restructuring and preparation pipeline that preserves temporal order while removing technical artefacts of the export format.

Rather than proposing a final CDSS, classical baseline models are compared across three dataset formulations derived from the same underlying EHR source: a multi-label formulation and two sequential formulations that differ by whether examination results are included. Accuracy, macro-F1, and weighted-F1 are reported under strong class imbalance to provide a baseline for later improvements. Next steps include structured information extraction from free text through an NLP pipeline with entity extraction and normalization, as well as more advanced modelling approaches [6,10].

This study makes two main contributions. First, it presents a practical pipeline for restructuring a noisy wide-table EHR export into linked longitudinal tables suitable for modelling. Second, it evaluates simple baseline models across three dataset formulations derived from the same EHR source. Together, these experiments provide a clear baseline for future work on Lithuanian clinical NLP, richer temporal modelling, and better handling of class imbalance [6,10].

2. Methodology

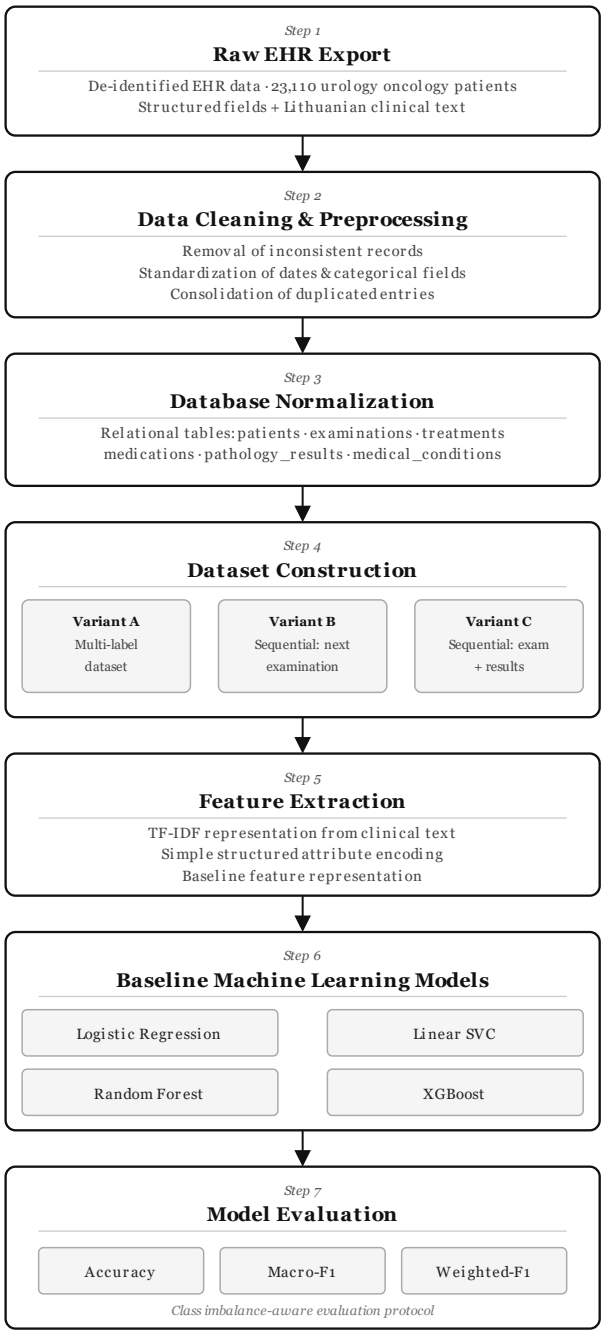


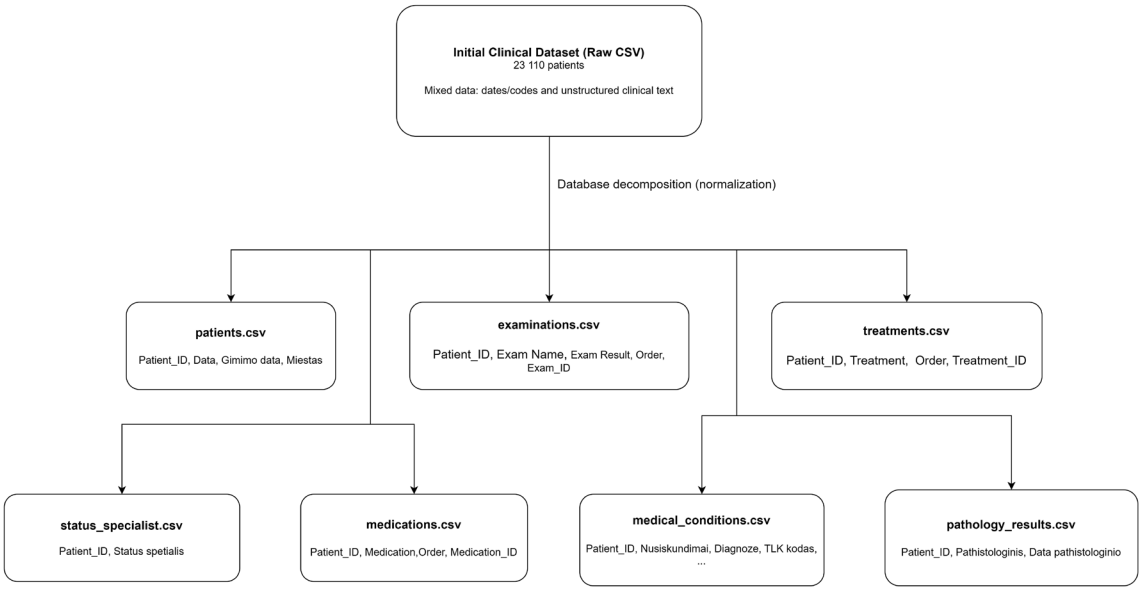
Figure 1. Workflow of the EHR data preparation and baseline machine learning pipeline used in this study.

2.1. Data source and restructuring

The study uses a de-identified clinical dataset provided by a urology oncology clinic. The initial flat CSV file contained 23,110 patients and included demographic and administrative information such as birth date, city, and registration dates, as well as free-text complaints and diagnoses, examination descriptions, and details on treatments, medications, and pathology results.

A key issue in the raw export was that it was not modelling-friendly. Many columns related to later treatment stages or specific procedures had more than 90–98% missing values, while other columns were almost fully populated. Simple solutions such as dropping columns or applying heavy imputation would either remove useful clinical information or introduce artefacts.

To handle this, the dataset was restructured into several domain-specific tables, including patients, medical_conditions, status_specialist, examinations, treatments, medications, and pathology_results. All tables were linked through a common patient identifier, which allowed us to reconstruct each patient's longitudinal clinical history. This step removed artificially empty columns created by squeezing many event types into a single row and kept only events that actually occurred. An Order attribute was also stored to preserve the sequence of examinations and treatments over time.



Note: *Order* denotes the relative index of patient events, reconstructed from numbered source fields (e.g., *Examination1–5*, *Treatment1–6*), and is used in sequence modelling tasks.

Figure 2. Raw-to-relational restructuring of the initial EHR export.

The original wide CSV contained repeated stage-specific blocks such as Exam1…ExamN and Treatment1…TreatmentN, which led to systematic sparsity. The export was normalized into linked domain tables, and event order was preserved using the Order attribute, enabling reconstruction of longitudinal patient histories for downstream dataset generation. This restructuring did not solve all problems, but it made the data usable for modelling by representing patients as sequences of clinical events rather than a single flattened summary.

2.2. Experimental setup and dataset formulations

2.2.1. Dataset variants

The three dataset variants are based on the same clinical prediction setting, but they differ in how much prior patient information is given to the model. This makes it possible to compare whether richer longitudinal context improves baseline performance after the raw export has been restructured into linked tables.

Variant A – Patient-level multi-label formulation.

In Variant A, each patient is represented by a feature vector built from structured fields and free text. The target is represented as the set of target procedures associated with that patient, so the task is treated as a multi-label prediction problem. This variant tests whether baseline models can learn

common target-procedure combinations from a compressed patient-level representation under strong class imbalance.

Variant B – Sequential formulation using examination history.
In Variant B, an ordered patient history is constructed, and the next target procedure is predicted from the preceding examination history and related context. Compared with Variant A, this formulation is closer to a next-step clinical scenario in which temporal order matters.

Variant C – Sequential formulation using examination history and examination results.
Variant C extends Variant B by adding examination results to the sequential representation. The target remains the next target procedure; the only difference is that the model receives richer clinical context than in Variant B.

Filtering note.
All variants are affected by strong class imbalance with a long tail of rare target procedures. To keep the baseline experiments tractable and the reported metrics interpretable, extremely infrequent target labels were excluded from the evaluated label space. The target concept itself was not changed by this step; only labels with too little support for stable training and evaluation were removed.

2.2.2. Features and baseline models

To keep the experiments simple and comparable across dataset variants, a baseline feature setup was used that mixes structured fields with free-text clinical content. The goal was not to optimize the feature representation, but to assess how far standard models can go on noisy real-world EHR exports before moving to more advanced NLP or temporal modelling.

Input features.
Three groups of inputs were used. First, basic structured patient information such as demographics and administrative context when available. Second, structured clinical fields including examinations, procedures, treatments and medications, and pathology-related attributes depending on the dataset variant. Third, Lithuanian free-text fields such as patient complaints, diagnoses, examination descriptions, and specialist notes.

Because most of the text is domain-specific and in Lithuanian, preprocessing was intentionally light. Basic cleanup of spacing and punctuation was applied, and obvious formatting artefacts were removed. Lemmatization, ICD/SNOMED mapping, and named entity recognition were not applied at this stage, since the aim was to establish a baseline before building a dedicated clinical NLP pipeline.

Vectorization / representation.
Free text was represented with TF-IDF. Structured categorical items were encoded as simple indicator features (one-hot / bag-of-codes), and numeric values were kept as numeric features when they were consistently available. This produces a high-dimensional sparse feature space, which is typical for EHR + text settings.

Baseline models.
Four baseline classifiers were compared: Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and XGBoost (XGB). The goal was to establish a clear baseline performance range across dataset formulations and see which failure modes appear first, especially under strong class imbalance.

3. Evaluation metrics

The prediction problems in this study involve many classes and a strongly imbalanced label distribution. A model that always predicts only the most frequent label can still get a surprisingly high overall accuracy, but it would be of limited use because it ignores less typical cases.

To obtain a more realistic view of model behavior, three standard metrics are reported:

- Accuracy – the proportion of correctly predicted labels across all samples. Easy to interpret, but it can hide poor performance on rare classes.
- Macro-F1 – the F1 score computed per class and then averaged. Each class has the same weight, so macro-F1 is sensitive to performance on rare labels.
- Weighted F1 – the F1 score computed per class and averaged using class support as weights. This is closer to performance on common classes but still reflects precision/recall trade-offs.

Looking at accuracy together with macro-F1 and weighted F1 helps separate models that mostly "memorize" frequent labels from models that cover more of the long tail. This combination of metrics is commonly used in medical decision support and EHR-based prediction settings [2,10].

4. Analysis and experiments

4.1. Dataset characteristics

The dataset represents real-world urology oncology EHR data and includes both structured fields and Lithuanian free-text notes. As expected in this setting, the data is not clean or uniform: it contains short and long narrative fragments, abbreviations, repeated phrases from copy–paste documentation, and varying levels of detail depending on the clinician and the stage of care. From a modelling perspective, three properties dominate.

Noise and heterogeneity. The same clinical idea can be written in multiple ways in free text, and structured fields are not always consistent across patients. Some patients have rich histories and detailed examinations, while others have sparse records. This makes the feature space messy and the signal uneven.

High dimensionality and sparsity. Once free text is represented with TF-IDF and combined with structured indicators, the resulting representation becomes high-dimensional and sparse. This often favours models that can handle sparse inputs, but it also increases the risk that models learn spurious correlations instead of stable clinical patterns.

Targets such as 'Target Exam' follow a long-tail distribution: a small set of common labels dominates the dataset, while many labels are rare. This means that overall accuracy can look acceptable even when a model performs poorly on minority classes. Because of this, macro-F1 and weighted-F1 are used in addition to accuracy to better reflect behavior across both frequent and rare labels [2,10].

4.2. Experimental setup

All experiments used a fixed 80/20 train–test split with a random seed of 42 to ensure reproducibility. Text features were represented using TF-IDF with a capped vocabulary, with max_features ranging from 500 to 1000 depending on the dataset variant.

Hyperparameter tuning was intentionally kept minimal to maintain the focus on baseline model comparison. The Random Forest classifier was configured with 100 trees and a maximum depth of 10. Linear models were trained using class_weight = "balanced" to mitigate the effects of class imbalance. Conservative iteration limits were applied (max_iter = 200 for Logistic Regression and max_iter = 10,000 for LinearSVC).

In selected experiments, SMOTE oversampling was applied only when feasible. Specifically, oversampling was used only when the smallest class in the training set contained at least two samples. In those cases, k_neighbors = 1 was used; otherwise, SMOTE was omitted.

4.3. Data cleaning and preprocessing

Data preparation was done to make the dataset usable for modelling while keeping it close to real clinical documentation. The main issue was not random missing values, but systematic sparsity caused by the original wide-table export design with repeated stage-specific column blocks. Missingness increases in later exams and treatments because many patients never reach those stages, even though the export still allocates columns for them.

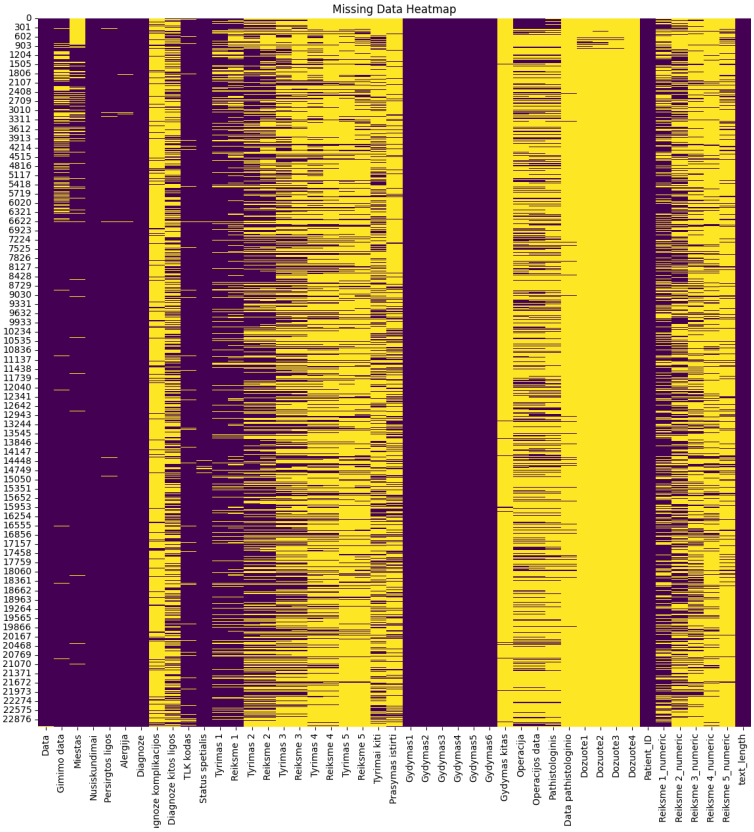


Figure 3. Missing data heatmap of the original wide-table export. Missingness is not random: repeated stage-specific column blocks become increasingly sparse, which motivates restructuring and careful preprocessing before modelling.

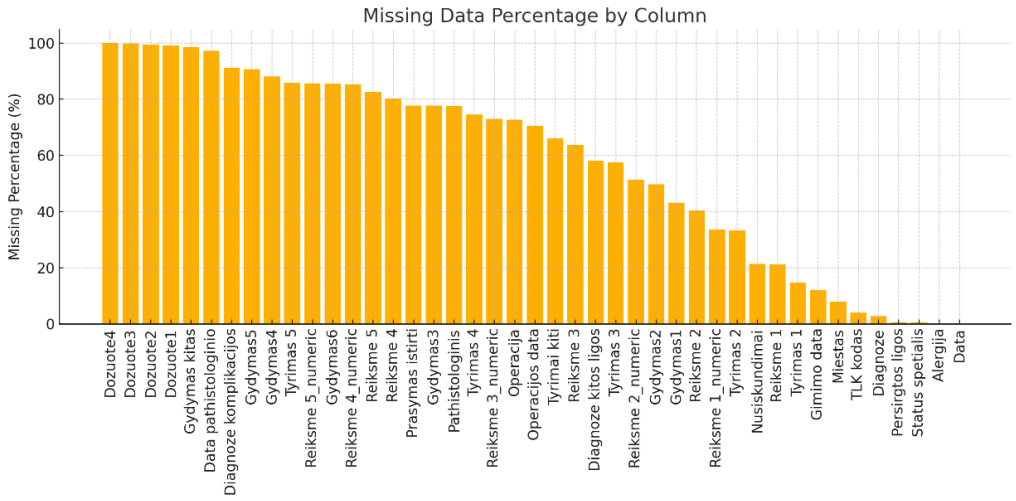


Figure 4. Missing value percentage per column in the original dataset. Many later-stage repeated fields exceed very high missingness levels, confirming that the raw export is not directly suitable for modelling without restructuring.

After restructuring the raw export into linked tables, preprocessing was applied at two levels:

1. Table-level cleaning. Records with missing essential identifiers were removed, formats for dates and categorical fields were standardized, and duplicated or partially repeated entries created by the export and merging process were consolidated. The restructuring step reduced the biggest technical problem of the flat file by storing only events that occurred for a patient, instead of keeping artificially empty repeated columns.
2. Text-level cleaning. Lithuanian free-text fields were kept with light normalization only, mainly whitespace and punctuation cleanup and removal of obvious formatting artefacts. Lemmatization, ICD/SNOMED mapping, and named entity recognition were not applied at this stage. The goal was to establish a reference point before introducing a dedicated clinical NLP pipeline.

5. Results

5.1. Quantitative results

Table 1 summarizes baseline performance across three dataset formulations and four model families. Accuracy, macro-F1, and weighted-F1 are reported to reflect both overall performance and behavior under strong class imbalance.

Table 1. Baseline results for the three dataset formulations. "—" indicates that the model was not evaluated in that specific setting.

Dataset variant	Model	Accuracy	Macro F1	Weighted F1
Variant A (multi-label)	Logistic Regression	0.1	0.16	0.08
	SVM (LinearSVC)	0.43	0.16	0.51
	Random Forest	0.84	0.49	0.82
	XGBoost	0.51	0.05	0.47
Variant B (seq-history)	Logistic Regression	0.12	0.17	0.08
	SVM (LinearSVC)	0.25	0.19	0.26
	Random Forest	0.69	0.4	0.69

	XGBoost	0.42	0.03	0.37
Variant C (seq-history+results)	Logistic Regression	0.12	0.17	0.08
	SVM (LinearSVC)	0.26	0.19	0.27
	Random Forest	—	—	—
	XGBoost	0.49	0.16	0.46

Random Forest provides the strongest baseline in the settings where it was successfully trained, achieving the best accuracy and weighted-F1 in both Variant A and Variant B, while also reaching the highest macro-F1 among the evaluated models in those settings. Logistic regression performs consistently poorly across all variants. Linear SVM improves weighted performance but retains low macro-F1, indicating continued bias toward common classes. XGBoost remains unstable in this setup, with very low macro-F1 in Variants A and B, although it improves when examination results are included in Variant C. Random Forest was not evaluated for Variant C because the combination of one-hot encoded structured features and sparse text features exceeded the available memory and made training impractically slow. Overall, the gap between macro-F1 and weighted-F1 confirms a strong long-tail imbalance problem, meaning that rare target procedures remain difficult under the current representation.

5.2. Discussion and limitations

The results show a clear pattern: performance on frequent labels looks reasonable for some baselines, particularly Random Forest, but rare labels remain difficult. This is visible in the gap between weighted-F1 and macro-F1 across variants. In practice, models tend to focus on the most common targets, which makes them less useful for covering the long tail of less frequent examinations.

One likely reason is the combination of high-dimensional sparse text features and noisy clinical documentation. In the richest sequential setting (Variant C), the feature space also became too large for Random Forest to train within the available memory and time limits. Lithuanian free text includes abbreviations, short fragments, and copy–paste patterns, so TF-IDF captures a lot of surface wording rather than the underlying clinical meaning. Without a dedicated clinical NLP layer, models can learn correlations that do not generalize well. A practical next step is to extract and normalize key entities from text and reduce synonym and formatting variation so that models learn more stable clinical signals.

Another limitation is the export structure and missingness pattern. Even after restructuring, the available information is uneven across patients and care stages. Variant C adds examination results, but the improvement is not consistent across models. This suggests that results are missing in many cases or are not represented in a way that baseline models can use effectively.

Finally, the baseline setup is intentionally kept simple by using a fixed train/test split and minimal tuning. This makes comparisons transparent, but it also means the reported numbers are not an upper bound. Better imbalance handling, through cost-sensitive learning, calibrated decision thresholds, and class-aware sampling, together with richer temporal representations and structured text extraction with NER, should improve rare-label performance and make sequential formulations more informative.

6. Conclusion and future work

This paper presented baseline machine-learning experiments on de-identified urology oncology EHR data from 23,110 patients, including extensive Lithuanian clinical text. A major part of the work was making the raw export usable for modelling. The original wide-table format created systematic sparsity, so the data was restructured into linked tables that preserve event ordering and support longitudinal representations.

Across the evaluated dataset formulations, Random Forest provided the strongest baseline when training was computationally feasible, while logistic regression and linear SVM struggled to generalize beyond frequent target procedures. At the same time, the gap between weighted-F1 and macro-F1 shows that rare target procedures remain difficult across all baseline settings.

Future work will focus on improving how clinical text and temporal context are represented. The most promising direction is to add a lightweight clinical NLP layer for Lithuanian text with entity extraction and normalization, so models learn more stable clinical signals instead of surface-level wording. Imbalance-aware training strategies and richer sequential modelling are also planned for future work to better handle the long-tail distribution and to make better use of examination results when they are available.

7. Declaration on Generative AI

The authors have not employed any Generative AI tools.

8. References

[1] Naderian S, Soleimanzadeh F, Nikniaz L. *A systematic review of artificial intelligence–based clinical decision support systems in prostate cancer management.* Healthcare Technology Letters. 2025. **doi:10.1049/htl2.70026**.

[2] Salem H, Soria D, Lund JN, Awwad A. *A systematic review of the applications of Expert Systems (ES) and machine learning (ML) in clinical urology.* BMC Medical Informatics and Decision Making. 2021;21(1):268. **doi:10.1186/s12911-021-01585-9**.

[3] Li K, Banerjee I, Magnani CJ, Blayney DW, et al. *Clinical documentation to predict factors associated with urinary incontinence following prostatectomy for prostate cancer.* Research and Reports in Urology. 2020;12:355–366. **doi:10.2147/RRU.S234178**.

[4] Yu SH, Kim MS, Chung HS, Hwang EC, Jung SI, et al. *Early experience with Watson for Oncology: a clinical decision-support system for prostate cancer treatment recommendations.* World Journal of Urology. 2021;39:3169–3177. **doi:10.1007/s00345-020-03214-y**.

[5] Kaiser P, Yang S, Bach M, Breit C, Mertz K, et al. *The interaction of structured data using openEHR and large language models for clinical decision support in prostate cancer.* World Journal of Urology. 2025. **doi:10.1007/s00345-024-05423-1**.

[6] Bhatia A, Titus R, Porto JG, Katz J, et al. *Application of natural language processing in electronic health record data extraction for navigating prostate cancer care: a narrative review.* Journal of Endourology. 2024. **doi:10.1089/end.2023.0690**.

[7] Koutsojannis C, Nabil E, Tsimara M. *Using machine learning techniques to improve the behaviour of a medical decision support system for prostate diseases.* In: ISDA 2009. **doi:10.1109/ISDA.2009.110**.

[8] Chen J, Remulla D, Nguyen JH, Dua A, Liu Y. *Current status of artificial intelligence applications in urology and their potential to influence clinical practice.* BJU International. 2019;124(4):567–577. **doi:10.1111/bju.14852**.

[9] Talyshinskii A, Naik N, Hameed BMZ. *Potential of AI-driven chatbots in urology: revolutionizing patient care through artificial intelligence.* Current Urology Reports. 2024;25:395–405. **doi:10.1007/s11934-023-01184-3**.

[10] Wagholikar KB, Sundararajan V. *Modeling paradigms for medical diagnostic decision support: a survey and future directions.* Journal of Medical Systems. 2012;36:3029–3049. **doi:10.1007/s10916-011-9780-4**.