JMIR Bioinformatics and Biotechnology

Methods, devices, web-based platforms, open data and open software tools for big data analytics, understanding biological/medical data, and information retrieval in biology and medicine.
Volume 7 (2026) ISSN 2563-3570 Editor in Chief: Ece D. Uzun, MS, PhD, FAMIA

Contents

Editorial

Original Papers

Research Letter

A Strategic Partnership to Advance AI Applications in Genomics and Bioinformatics for Health Innovation

Aik Choon Tan[1,2*], PhD; Ece Dilber Gamsiz Uzun[3,4,5,6*], PhD, MS

[1]MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Salt Lake City, UT, United States

[2]Department of Oncological Sciences and Biomedical Informatics, Huntsman Cancer Institute, University of Utah, 1950 Circle of Hope Dr, Salt Lake City, UT, United States

[3]Department of Pathology and Laboratory Medicine, Alpert Medical School, Brown University, 593 Eddy Street, Providence, RI, United States

[4]Department of Pathology and Laboratory Medicine, Brown University Health, Providence, RI, United States

[5]Center for Clinical Cancer Informatics and Data Science (CCIDS), Brown University, Providence, RI, United States

[6]Center for Computational Molecular Biology (CCMB), Brown University, Providence, RI, United States

[*]all authors contributed equally

Corresponding Author:
Aik Choon Tan, PhD
MidSouth Computational Biology and Bioinformatics Society (MCBIOS), Salt Lake City, UT, United States

Abstract

JMIR Bioinformatics and Biotechnology announced a strategic partnership with the MidSouth Computational Biology and Bioinformatics Society (MCBIOS) in late 2025; this partnership establishes JMIR Bioinformatics and Biotechnology as the official journal of MCBIOS. This collaboration reflects a shared commitment to advancing computational biology, bioinformatics, and biotechnology through open science, interdisciplinary collaboration, and real-world data. By connecting MCBIOS' community-building and professional development initiatives with JMIR Publications' open-access publishing platform, this partnership aims to support emerging researchers, accelerate the dissemination of innovative computational methods and artificial intelligence applications, and strengthen data-driven research across medicine and biology.

(JMIR Bioinform Biotech 2026;7:e93272) doi:10.2196/93272

KEYWORDS

artificial intelligence; bioinformatics; genomics; AI; health innovation

In late 2025, *JMIR Bioinformatics and Biotechnology* announced a new strategic partnership with the MidSouth Computational Biology and Bioinformatics Society (MCBIOS), under which it will serve as the official journal of MCBIOS. This collaboration represents a shared commitment to advancing computational biology, bioinformatics, and biotechnology through open science, interdisciplinary collaboration, and real-world data.

MCBIOS is dedicated to advancing the fields of computational biology and bioinformatics by fostering collaboration, networking, and professional development among scientists, educators, and trainees. The society promotes interdisciplinary research, supports educational initiatives, and provides opportunities—particularly for trainees and early-career researchers to engage with peers and leaders in the field. Through annual conferences and community outreach, MCBIOS works to enhance scientific knowledge, encourage innovation, and contribute to solving complex biomedical questions [1].

The vision of *JMIR Bioinformatics and Biotechnology* is to promote high-quality, interdisciplinary research at the intersection of bioinformatics, computational biology, biotechnology and artificial intelligence (AI). The journal aims to provide an open-access platform for innovative computational methods, data-driven biological discoveries, and translational applications that bridge algorithm development with real-world biomedical questions. Emphasizing collaboration among bioinformaticians, biologists, clinicians, and data scientists, the journal seeks to foster rigorous, reproducible research that leverages emerging technologies such as AI, machine learning, and big data analytics to accelerate scientific progress and translational impact in the life sciences and medicine.

The partnership between MCBIOS and *JMIR Bioinformatics and Biotechnology* is well-timed within today's scientific and academic landscape for several key reasons. First, computational biology and bioinformatics have entered a phase defined by large-scale data generation, AI, and translational research. Second, aligning a scientific society that fosters community, mentorship, and scholarship with a global, open-access journal amplifies both visibility and impact. Furthermore, in an era where rapid dissemination and interdisciplinary collaboration have become key components in research, this partnership creates a streamlined pathway from conference presentation and networking to peer-reviewed publication in a journal that prioritizes accessibility and methodological innovation.

Together, these organizations will support emerging scholars and strengthen open, collaborative, and AI-enabled data science.

Strengthening a Community at the Intersection of Bioinformatics, Biotechnology, and AI

MCBIOS has long served as a vibrant forum for academic researchers, industry scientists and trainees working across computational biology, bioinformatics, systems biology, machine learning, medicine and data science. This partnership offers not only a conference or a journal to showcase research, but also an integrated professional ecosystem that amplifies visibility, supports career development, and fosters innovation—making participation both professionally valuable and intellectually impactful. As part of this partnership, *JMIR Bioinformatics and Biotechnology* will work closely with MCBIOS leadership to support the society's mission through:

- **Official society journal designation:** Being the official journal of MCBIOS reinforces *JMIR Bioinformatics and Biotechnology*'s credibility in the publishing landscape wherein researchers must identify trusted journals. This designation will provide quality, community oversight, and alignment with the interdisciplinary standards of computational biology and bioinformatics, giving authors confidence that their work will reach the right audience.
- **Annual conference-linked theme issues:** By publishing issues connected to MCBIOS annual meetings, *JMIR Bioinformatics and Biotechnology* will help capture emerging trends and high-impact research that might otherwise remain limited to conference presentations. This addresses the challenge of slow publication process and provides authors, especially early-career researchers, a pathway from presentation to peer-reviewed publication, increasing the visibility and impact of their work in a timely manner.
- **Publishing benefits for MCBIOS members:** Discounted article processing fees in reducing financial barriers to open access publishing, addressing a major concern for researchers.
- **Education and capacity building:** Through educational and career development webinars, and resources on peer review, publishing ethics, and research dissemination, *JMIR Bioinformatics and Biotechnology* and MCBIOS aim to address a gap in formal education on modern publishing practices.

These initiatives are designed to amplify the visibility and impact of work emerging from the MCBIOS community, while welcoming impactful contributions from the broader global research ecosystem to the *JMIR Bioinformatics and Biotechnology* .

A Shared Focus: From Computational Methods to Meaningful Clinical Impact

MCBIOS and *JMIR Bioinformatics and Biotechnology* share a vision of advancing interdisciplinary data-driven research with clinical impact through collaboration, transparency, and open dissemination. Both partners are committed to fostering innovation at the intersection of bioinformatics, computational biology, biotechnology and AI while supporting emerging investigators and strengthening research communities. Together, MCBIOS and *JMIR Bioinformatics and Biotechnology* aim to provide a scholarly home for research that not only advances methods and technologies but also demonstrates meaningful applications across biomedical, clinical, population health, and AI contexts. For example, machine learning tools that improve early disease detection or risk stratification, computational approaches that predict drug response, or real-world data analyses that evaluate the viability of computational tools in clinical settings.

We invite submissions that address, but are not limited to:

- Novel computational methodologies, algorithms, and statistical approaches
- AI and machine learning applications in genomics, bioinformatics, biotechnology, and health
- Data integration and multimodal analytics, including genomics, transcriptomics, proteomics, imaging, and clinical data
- Demonstrated real-world applications in translational bioinformatics, including clinical, public health, and biotechnological use cases
- Scalable technologies and platforms enabling reproducible and open science

By emphasizing both methodological rigor and impactful applications, *JMIR Bioinformatics and Biotechnology* seeks to serve as a bridge between innovation and implementation of computational methods, supporting data-driven research that advances science while contributing to improved health outcomes.

Looking Ahead

This partnership between *JMIR Bioinformatics and Biotechnology* and MCBIOS is not only a recognition of shared values but also an investment in the future of bioinformatics and biotechnology research. As the pace of innovation accelerates, there is a growing need for journals that can thoughtfully integrate advances in bioinformatics, data science, AI, and biotechnology within a health-focused, open access framework. We look forward to working with the MCBIOS community to shape this next chapter through collaborative theme issues, high-quality submissions, and ongoing dialogue about the evolving role of bioinformatics and biotechnology in transforming health.

We warmly invite MCBIOS members and the wider research community to submit their work to *JMIR Bioinformatics and Biotechnology* [2] and to join us in advancing research that bridges bioinformatics, data, AI, and innovation for real-world health impact.

Conflicts of Interest

EDGU serves as the editor in chief of *JMIR Bioinformatics and Biotechnology* at the time of this publication. AT serves as the president of MCBIOS and is an associate editor for *JMIR Bioinformatics and Biotechnology* at the time of this publication.

References

1. MCBIOS – midsouth computational biology and bioinformatics society. URL: https://mcbios.com [accessed 2026-02-26]
2. JMIR bioinformatics and biotechnology. Theme Issue 2026: Bridging Data, AI, and Innovation to Transform Health. URL: https://bioinform.jmir.org/themes/1678-theme-issue-2026-bridging-data-ai-and-innovation-to-transform-health [accessed 2026-02-26]

Abbreviations

AI: artificial intelligence
MCBIOS: MidSouth Computational Biology and Bioinformatics Society

Edited by T Leung; submitted 10.Feb.2026; this is a non–peer-reviewed article;accepted 10.Mar.2026; published 27.Mar.2026.

Please cite as:
Tan AC, Gamsiz Uzun ED
A Strategic Partnership to Advance AI Applications in Genomics and Bioinformatics for Health Innovation
JMIR Bioinform Biotech 2026;7:e93272
URL: https://bioinform.jmir.org/2026/1/e93272
doi:10.2196/93272

The AudioGene Translational Dashboard for Diagnosing Autosomal Dominant Nonsyndromic Hearing Loss: Phenotypic Data Visualization and Analysis Study

Benjamin DeSollar[1], BSE, MSE; Nathan Schaefer[1], BSE; Daniel Walls[2], MS; Amanda M Odell[2], MS; Kevin T A Booth[3,4], PhD; Hela Azaiez[2], PhD; Michael Schnieders[5], DSC; Richard J H Smith[2,6], PhD; Terry Braun[7], PhD; Thomas Casavant[1], PhD

[1]Department of Electrical and Computer Engineering, University of Iowa, 103 South Capitol Street, Room 5316, Iowa City, IA, United States

[2]Department of Otolaryngology, Head and Neck Surgery, University of Iowa, Iowa, IA, United States

[3]Department of Medical and Molecular Genetics, Indiana University, Indianapolis, IN, United States

[4]Department of Otolaryngology—Head and Neck Surgery, School of Medicine, Indiana University, Indianapolis, IN, United States

[5]Department of Biochemistry and Molecular Biology, University of Iowa, Iowa City, IA, United States

[6]Department of Molecular Physiology and Biophysics, University of Iowa, Iowa City, IA, United States

[7]Department of Biomedical Engineering, University of Iowa, Iowa City, IA, United States

Corresponding Author:
Thomas Casavant, PhD
Department of Electrical and Computer Engineering, University of Iowa, 103 South Capitol Street, Room 5316, Iowa City, IA, United States

Abstract

Background: Autosomal dominant nonsyndromic hearing loss (ADNSHL) is highly heterogeneous, with more than 64 genes implicated in its etiology. This complexity limits the diagnostic power of clinical examinations and audiometry alone, while existing computational approaches have achieved only moderate accuracy and often lack interpretability. As precision medicine increasingly emphasizes genotype-phenotype correlations, there is a recognized need for diagnostic tools that provide clinicians with transparent, interpretable outputs.

Objective: This study aimed to develop and evaluate the AudioGene Translational Dashboard, an interpretable clinical informatics tool that integrates machine learning models and interactive visualizations to enhance genotype-phenotype correlations and support diagnostic decision-making in ADNSHL.

Methods: We developed the AudioGene Translational Dashboard, integrating 2 machine learning models (AudioGene version 4 and AudioGene version 9.1) with 6 interactive visualization tools. AudioGene version 4 uses a multi-instance support vector machine classifier for patients with multiple audiograms, while AudioGene version 9.1 combines adaptive boosting, k-nearest neighbors, random forest models, and logistic regression for patients with a single audiogram. Visualizations include audiometric profile plots, audioprofile surfaces, clustering analyses, and data distribution charts designed to facilitate clinical interpretation.

Results: The AudioGene Translational Dashboard was developed to address the “70/30” phenomenon, indicating a 74% likelihood that the causative gene is among the top 3 predicted genes, thereby providing clinicians with a clear confidence indicator (“green flag”) or a caution alert (“red flag”) during diagnosis. While this level of performance is well suited for hypothesis generation, the remaining uncertainty underscores the need for interpretive context in clinical decision-making. Visualization tools enhanced clinicians’ ability to interpret and correlate phenotypic data with predicted genetic outcomes, improving diagnostic confidence and interpretability.

Conclusions: The AudioGene Translational Dashboard advances clinical informatics in genetic diagnosis of ADNSHL by integrating explainable artificial intelligence with interactive visualizations, enhancing clinical interpretability and diagnostic accuracy. This approach facilitates informed clinical decision-making, highlights the translational potential of genotype-phenotype computational models, and supports precision medicine in hearing loss diagnostics. Future enhancements will target improving class balance and incorporating additional user-customizable features to further optimize clinical applicability.

(JMIR Bioinform Biotech 2026;7:e85212) doi:10.2196/85212

KEYWORDS

autosomal dominant nonsyndromic hearing loss; machine learning; explainable artificial intelligence; clinical decision support systems; genotype-phenotype correlation; audiometry; genetic testing

Introduction

Background

Autosomal dominant nonsyndromic hearing loss (ADNSHL) presents a significant genetic diagnostic challenge due to its underlying heterogeneity—more than 64 genes are implicated in its etiology [1]. Because of this complex genetic landscape, computational tools designed to correlate audiogram profiles (commonly called audioprofiles) with specific genes have achieved only moderate success [2,3]. One such tool, which we developed approximately 15 years ago, is AudioGene. AudioGene uses numerous machine learning (ML) approaches to improve diagnostic precision [2,4,5]. These approaches include semisupervised support vector machines (SVMs), ensemble models, and hyper-tuning methods. However, challenges such as data imbalance and class sparsity continue to restrict the accuracy of these models.

Precision medicine harnesses information about the genome of an individual, environment, and lifestyle to guide medical care. With heterogeneous disorders such as ADNSHL, genetic variant interpretation can be challenging, complicating the diagnostic process and impacting patient care. Computational tools may improve the precision and reliability of genetic assessments by capitalizing on genotype-phenotype associations [6,7].

Current diagnostic methods for ADNSHL largely rely on clinical examination and audiometry, which do not provide sufficient resolution for the complex genetic landscape of ADNSHL [2,4]. However, with the availability of ML and artificial intelligence–driven approaches, there has been a shift toward integrating computational and visualization tools with genetic diagnostics to improve accuracy and predictive power [8,9].

To address these challenges, we have developed the AudioGene Translational Dashboard with the goal of enhancing both the accuracy and interpretability of genetic predictions. A feature of the AudioGene Translational Dashboard is the “70/30” phenomenon: by integrating the results from both models on a training dataset comprising 3189 audiograms from 1445 patients, we observed that the correct disease-causing gene was predicted within the top 3 predictions 74% of the time, with incorrect predictions accounting for the remaining 26%, hence “70/30.” This observation signals to health care providers when they can have confidence in the top predictions, serving as a “green flag” or “red flag” in the diagnostic process. Having a true positive rate of 70% is beneficial from a research perspective; however, for a diagnostic tool, the remaining 30% represents some risk that necessitates additional interpretative context. By providing this context, the AudioGene Translational Dashboard enables health care providers to weigh their confidence in the predictions, supporting more informed diagnostic decisions.

The AudioGene Translational Dashboard was introduced into the AudioGene toolset to increase transparency into the “black box” underlying the models by providing explainable artificial intelligence (XAI) to enhance model interpretability and utility in clinical settings, in line with trends in precision medicine that emphasize the importance of genotype-phenotype associations in improving diagnostic outcomes [7,10].

Related Works

Early attempts to map audiometric phenotypes to their underlying genotypes were spearheaded by AudioGene version 4 (AG4), a semisupervised multi-instance SVM that treats the collection of audiograms for a single patient as a “bag” and ranks loci according to pair-wise–coupled probability estimates [2,11]. Building on this foundation, AudioGene version 9.1 (AG9.1) introduced selective intraensemble data partitioning: training examples are first divided by gene-specific data volume, patient age, and audiogram shape, then modeled with a committee of k-nearest neighbor (KNN), adaptive boosting, and random forest subclassifiers, whose outputs are fused by logistic regression. AG9.1 offers a top-3 accuracy of 77.8%, with a precision of 0.51 and a recall of 0.56, at the cost of introducing a more complex model. We report top-3 accuracy rather than top-1 accuracy because, in the context of gene prioritization for validation sequencing, the cost of excluding the true causative gene is higher than the cost of evaluating a small number of candidate genes. In addition, the top-3 threshold represents a practical trade-off between high confidence in predictions and an acceptable loss of significance when selecting genes for sequence-based validation [4,5,12,13]. Both frameworks have improved locus-ranking accuracy for the 23 well-curated ADNSHL genes that account for roughly three-quarters of cases in populations of European ancestry [14]. Nonetheless, their predictions can still be difficult to interpret when class imbalance, sparse age coverage, or atypical audiogram morphologies are present.

Complementary to algorithmic advances, domain-specific visualization has been welcomed as a potentially beneficial tool for clinical use. Audioprofile surfaces (APS) plot 3D trajectories of frequency-specific threshold shift over time, revealing gene-characteristic progression patterns that are not obvious in 2D audiograms [15]. Circle-based genome views (eg, Circos [version 0.69-10; Krzywinski, Canada’s Michael Smith Genome Sciences Center] enable high-density comparison of structural variation or copy number events [16], while integrative genome browsers such as Integrative Genomics Viewer (version 2.19.7; UC San Diego and Broad Institute of MIT and Harvard) allow rapid inspection of read evidence at candidate loci [17]. More recent health care dashboards use fuzzy logic overlays and interactive filtering to expose outliers or low-confidence regions directly to end users [18]. Despite these advances, few systems combine genotype-prediction engines with audiogram-aware visual contexts; therefore, clinicians must cross-reference separate tools, a workflow that can erode trust in algorithmic suggestions and slow decision-making [19]. These limitations reflect shortcomings in how models communicate their reasoning and how results are presented to end users.

Accordingly, the literature reveals two unmet needs:

1. Model transparency—while ensemble and semisupervised approaches improve predictive accuracy, they do not inherently communicate *why* a particular gene is ranked highly, especially when training data are imbalanced or noisy.
2. Unified, clinician-friendly interfaces—existing genomic viewers excel at sequence-level detail but lack

phenotype-specific visualization; conversely, stand-alone audiogram tools rarely link observed hearing profiles back to the underlying variant evidence.

The AudioGene Translational Dashboard addresses these gaps by (1) merging the complementary strengths of AG4 and AG9.1 and (2) embedding 6 interactive visual modules—APS, 2D audioprofiles, uniform manifold approximation and projection (UMAP) cluster projections, gene count bar charts, region-of-origin pie charts, and age distribution plots—around the model output. This hybrid XAI-driven design supports clinicians in validating or questioning the algorithm's "70/30" confidence observation and thus advances the state of practice in ADNSHL diagnostics.

Methods

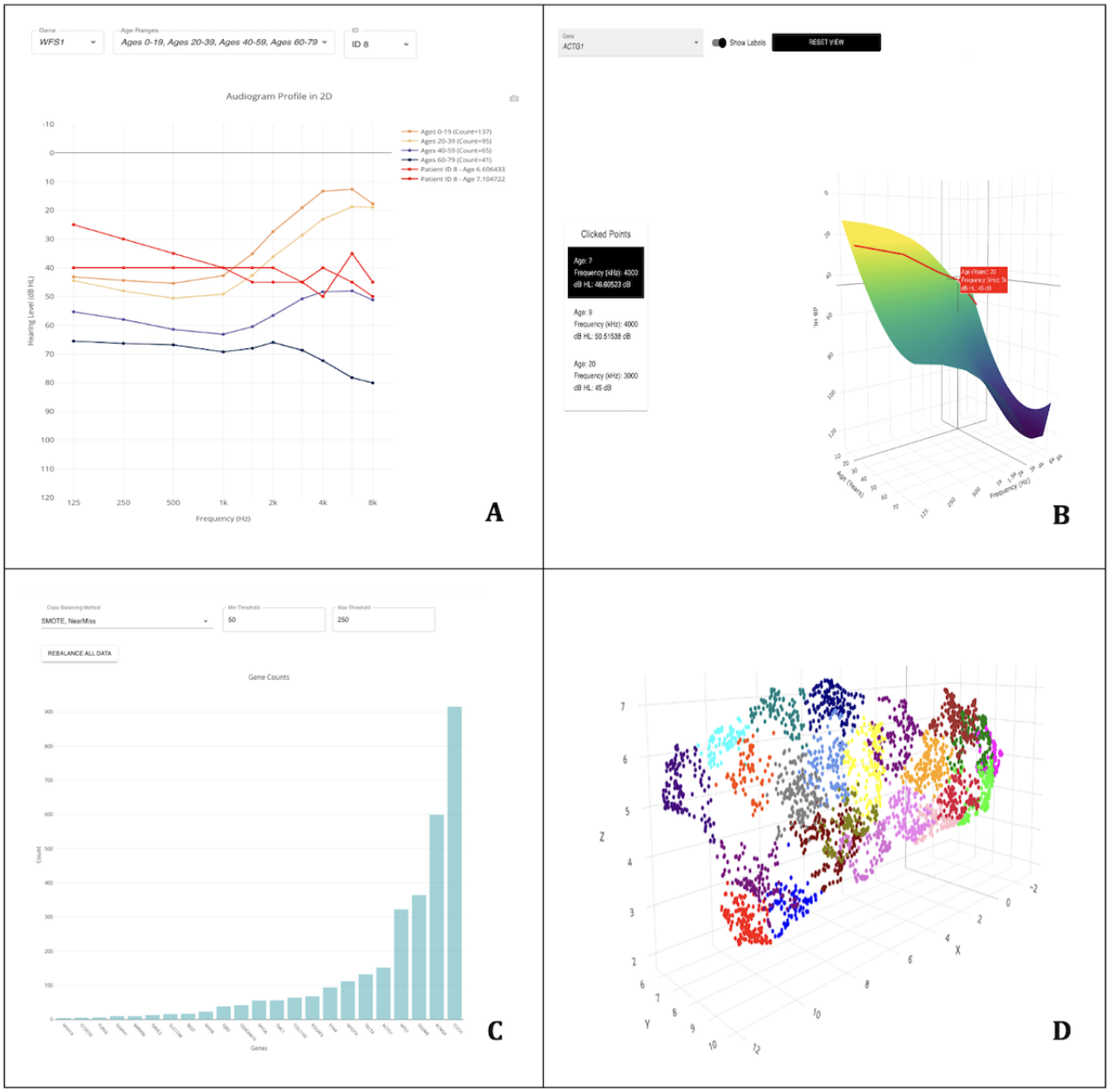
Tool Overview

The AudioGene Translational Dashboard integrates 2 ML models to increase diagnostic accuracy for ADNSHL:

- AG4—it is a multi-instance classifier designed for patients with multiple audiograms that uses a semisupervised SVM and ranks loci based on modified SVM probability outputs [4]. It was developed using the Waikato Environment for Knowledge Analysis (version 3.7.2; WekaIO Inc) platform [11].
- AG9.1—it is a single-instance classifier for patients with only 1 audiogram, developed using the scikit-learn library in Python. It comprises multiple submodels: 3 KNNs, 6 adaptive boosting models, 2 random forest models, and a logistic regression module for combining outputs [2,5,8].

The AudioGene Translational Dashboard interface provides six distinct visualization tools for health care providers and researchers to interactively assess genetic data and model predictions in real time: (1) audioprofile, a 2D plot displaying the average hearing loss (in dB) over 10 frequencies (125 Hz to 8000 Hz) for each age group, allowing comparison with a patient's hearing loss over time (Figure 1A); (2) APS, a 3D surface plot depicting gene-specific hearing loss progression over time across frequencies, illustrating age-related changes in decibel loss (Figure 1B) [20]; (3) a region-of-origin pie chart, which displays the geographic origin distribution (eg, Dutch, German, and Chinese) for audiograms associated with each gene; (4) a count bar chart, which illustrates the audiogram count for each gene, highlighting class imbalance challenges (Figure 1C); (5) spatial analysis and clustering, which shows the cluster position of each gene and a 3D plot of audiograms (clusters are created using the k-means algorithm to partition data into 23 gene-specific groups [15]; the 3D plot compresses 11 features [age plus 10 frequencies] into 3 dimensions using the UMAP method (Figure 1D) [21]); and (2) 6) an age distribution scatter plot, which shows the age distribution for each gene within the training dataset, providing context for the predictive model outputs.

Figure 1. Visualization components of the AudioGene Translational Dashboard. (A) Audioprofile for the selected gene (*WFS1*), displaying data from all age ranges along with patient data; (B) audioprofile surface view for the selected gene (*ACTG1*); (C) count bar chart showing the counts of each gene in the training data; and (D) 3D uniform manifold approximation and projection of genetic case data used in the AudioGene Translational Dashboard.(each point represents a classified genetic case, and the color coding corresponds to 1 of the 23 unique clusters identified through different genetic diagnoses).



The first 3 visualization tools were developed to compare patient-specific data to average thresholds for each of the 23 ADNSHL-associated genes. This comparison allows patient audiograms to be contextualized for each gene. The audioprofile visualization shows how a patient's audiogram compares to the expected audiograms associated with each gene.

The APS adds time as the third axis to provide a 3D rendering of gene-specific audiometric thresholds over time. Audiometric thresholds are represented as a 3D plane, depicting the expected dB loss over time (in years) at each frequency, thereby enabling comparisons between a patient's hearing thresholds and gene-specific expectations.

The spatial analysis and clustering tool uses a bar chart to show the distribution of each prediction among 23 different clusters. These clusters are created using k-means clustering [15]. Additionally, a 3D plot visualizes the audiograms within our data that have a confirmed genetic diagnosis, compressing 11 features into 3 dimensions using UMAP [21]. This feature allows users to interact with the bar chart, highlighting corresponding clusters in the 3D plot (Figure 1). Users can compare their patient's audiogram, represented by a large red dot, with others in the cluster, facilitating the identification of similar audiograms and associated genetic diagnoses.

By using these 3 visualizations, we aim to either enhance or reduce confidence in the predictions. For example, if the model ranks the *COCH* gene second among the top 3 predictions, health care providers can analyze the APS and spatial analysis tools to determine the degree of correlation with patterns

typically associated with the *COCH* gene, potentially increasing confidence in that diagnosis.

The last 3 visualization tools provide context for the data in the AudioGene dataset. The count bar chart highlights significant class imbalance, showcasing the challenge the model encounters in predicting smaller classes due to underrepresentation. The region-of-origin pie chart and age distribution scatter plot provide additional context about data distribution, allowing health care providers to understand model limitations and adjust diagnostic strategies accordingly.

The AudioGene Translational Dashboard integrates into the workflow of a clinician as a secondary validation layer. For example, within our clinical workflow, clinicians, genetic counselors, and bioinformaticians review the results of a clinical genetic test, including patient history, family structure, audiograms, and identified variants in hearing loss genes. This team can then inspect a patient's audiometric data relative to the landscape of audiometric data across all genes and patients, considering variance within a gene, rarity or abundance of cases within a cluster, and distance to genetically validated cases.

System Design

The system was designed using the SERN (SQL, Express.js, React.js, and Node.js) stack, which uses a client-server architecture where computationally intensive tasks are performed by the server, deployed in a Docker (version 28.5.1; Docker Inc) container [19,22]. The client is supported by React (version 18.2.0; React Foundation), a JavaScript library facilitating user interactions [23]. Data preprocessing used linear interpolation and extrapolation for missing values. The same methods for handling missing values were applied in the ML models [2,5,8].

The *Pandas* library in Python was used for data manipulation and analysis [8], and visualization libraries such as Plotly were used to create interactive graphs and plots [24].

For more information, please refer to the master's thesis by DeSollar [3] and the GitLab repository.

Ethical Considerations

This study was reviewed and approved by the University of Iowa Institutional Review Board (199701065). The institutional review board granted a waiver of informed consent under US federal regulation 45 CFR 46.116(f) (also known as the "Common Rule"), because, although audiograms were originally collected in clinical and research settings, the dataset used for this study was fully deidentified prior to analysis [25]. All procedures adhered to the ethical standards of the institutional and national research committees and to the 1964 Declaration of Helsinki and its later amendments. This paper does not contain any individual's data in any form, including individual details, images, or videos. No compensation was provided to participants, as this study involved secondary analysis of a fully deidentified existing dataset and no participants were directly recruited or enrolled.

Results

Introduction to AudioGene Translational Dashboard

The AudioGene Translational Dashboard combines advanced ML models with several visualization tools to create a platform that facilitates the prioritization of ADNSHL-associated genes in genetic testing results. Using patient data, the AudioGene Translational Dashboard generates gene rankings and enables auditory scientists and health care providers to explore these predictions interactively through various visualization tools. Gene ranking in phenotype-genotype associations can aid in the interpretation of complex genetic data, thereby providing greater context and confidence in diagnostic decisions [5,15].

The "70/30" phenomenon serves as an indicator for health care providers, providing them with the necessary context through visualizations to assess the reliability of the predictions. When the top 3 predictions include the correct gene, health care providers can have greater confidence in proceeding with targeted genetic testing.

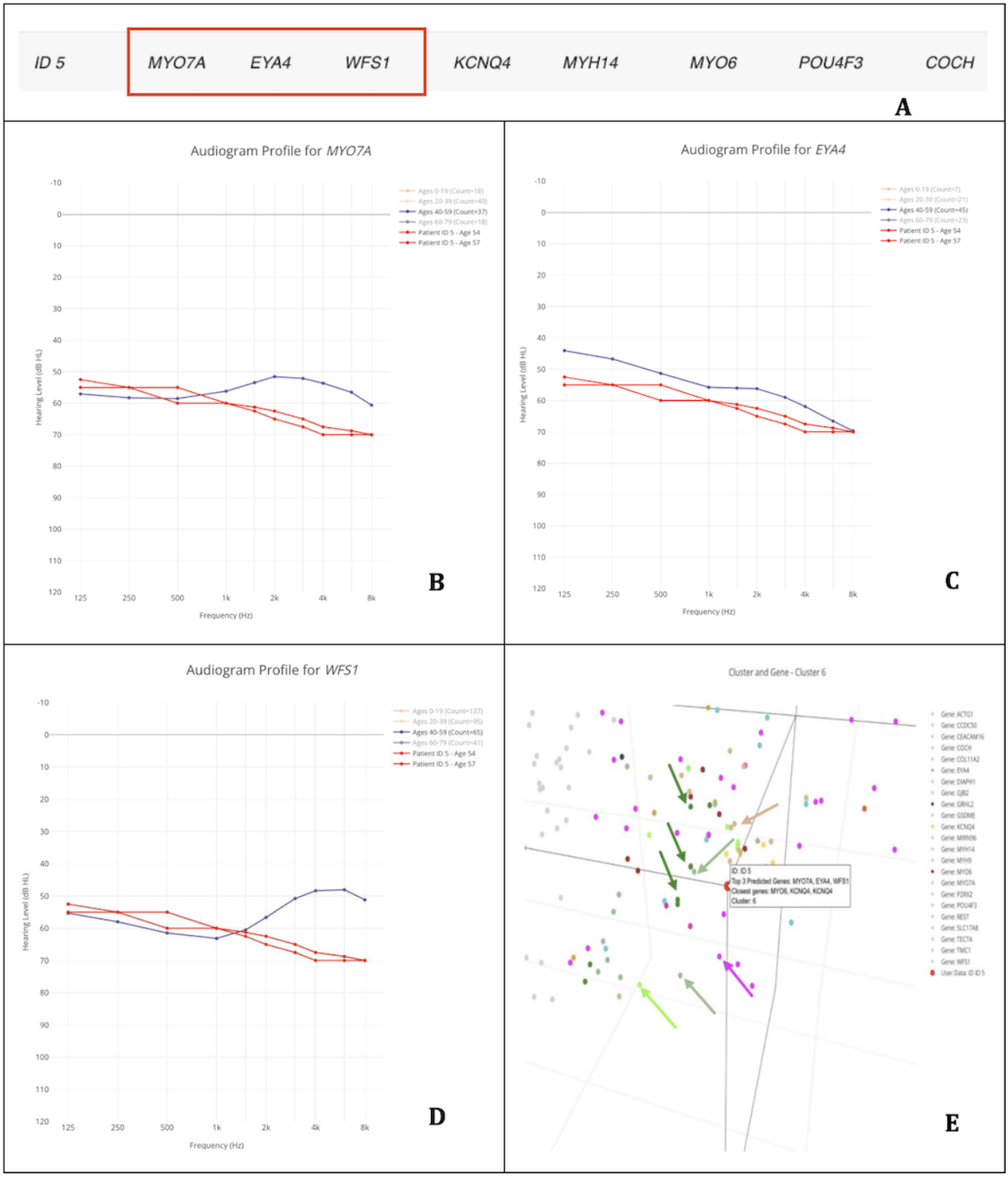
Case Studies and Clinical Implications

Several case studies have been carried out to demonstrate the effectiveness of the AudioGene Translational Dashboard in the diagnosis of specific genetic types of ADNSHL [3].

Case 1: MYO7A Gene—Patient 1 (ID 5)

In this case study, we analyzed the results from a patient diagnosed with *MYO7A*-related hearing loss. Our dataset included 2 audiograms, which were predicted by AG4 to be associated with *MYO7A*-related, *EYA4*-related, or *WFS1*-related hearing loss (Figure 2A).

Figure 2. Application of the AudioGene Translational Dashboard for patient-level gene prediction and visualization. (A) Predictions for patient (ID 5), highlighting the top 3 genes associated with the audiological characteristics observed; (B) audioprofile of MYO7A with the patient's (ID 5) audiograms in red, taken at the ages of 54 and 57 years; (C) audioprofile of EYA4 with the patient's (ID 5) audiograms in red, taken at 54 and 57 years of age; (D) audioprofile of WFS1 with the patient's (ID 5) audiograms in red, taken at the ages of 54 and 57 years; and (E) 3D plot of audiograms in the training set reduced to 3 dimensions for visualization, with genes in cluster 6 colored (genes not in cluster are light gray; patient [ID 5] is the red dot hovered over by the displayed label; and the green arrows point to MYO7A [green dots], the pale green arrows point to EYA4 [pale green dots], the pale yellow arrow points to WFS1 [pale yellow dots], the pink arrow points to COCH [pink dots], and the light green arrow points to KCNQ4 [light green dots]).



Examining these predictions relative to the patient's audiograms, the following observations can be made: (1) *MYO7A's* audioprofile is similar in the low-to-mid frequencies but diverges in the high frequencies (Figure 2B); (2) *EYA4* displays a comparable shape, but the patient's thresholds are consistently lower than typical values (Figure 2C); and (3) *WFS1* shares some similarities in the low-to-mid frequencies but diverges in the high frequencies, as observed with *MYO7A* (Figure 2D).

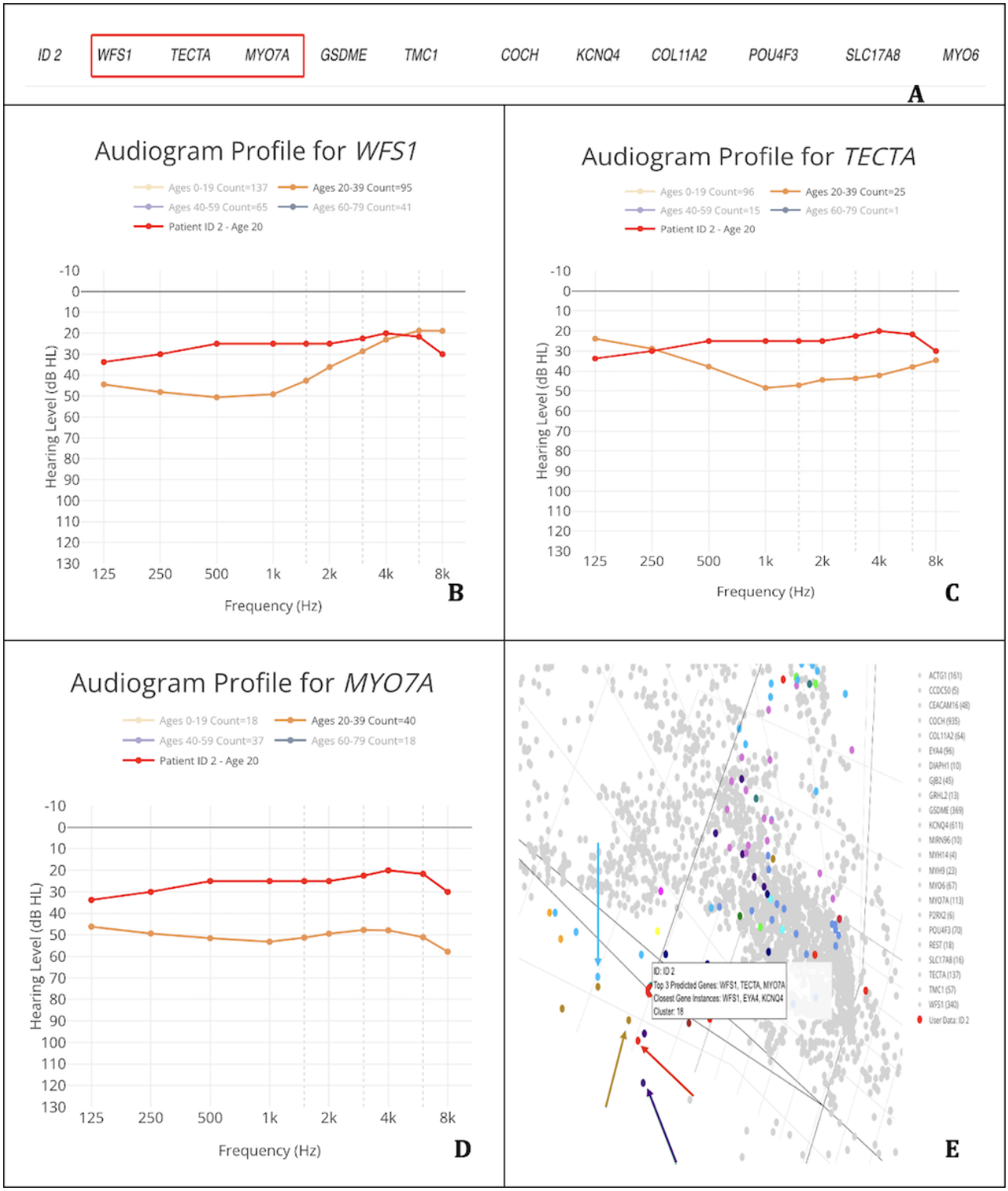
The audioprofiles for the 3 genes all show moderate to moderately severe hearing loss thresholds, with either close

similarity (<5 dB) at several frequencies or similar shapes. Therefore, we can conclude that the correct gene is likely captured within the top 3 predictions.

Case 2: MYO6 Gene—Patient 2 (ID 2)

In our second case, we explored the results from a patient (ID 2) previously diagnosed with *MYO6*-related hearing loss. Our dataset contains only 1 audiogram, with gene predictions shown in Figure 3A by AG9.1.

Figure 3. Application of the AudioGene Translational Dashboard for patient-level gene prediction, audioprofile comparison, and cluster-based visualization. (A) Predictions for the patient (ID 2), with the top 3 genes being WFS1, TECTA, and MYO7A; (B) audioprofile of WFS1 with the patient's (ID 2) audiograms in red, taken at 20 years of age; (C) audioprofile of TECTA with the patient's (ID 2) audiograms in red, taken at the age of 20 years; (D) audioprofile of MYO7A with the patient's (ID 2) audiograms in red, taken at the age of 20 years; and (E) 3D plot of audiograms in the training set reduced to 3 dimensions for visualization, with genes in cluster 18 colored (genes not in the cluster are light gray; patient [ID 2] is the red dot hovered over by the displayed label; and the light blue arrow points to KCNQ4 [light blue dots], the red arrow points to WFS1 [red dots], the purple arrow points to POU4F3 [purple dots], and the brown arrow points to GSDME [brown dots]).



None of the top 3 candidate genes (*WFS1*, *TECTA*, or *MYO7A*) display an audioprofile that closely aligns with the patient's thresholds (Figure 3B–D). This mismatch strongly suggests that the true causative gene (*MYO6*) does not appear among the model's top 3 predictions for this patient.

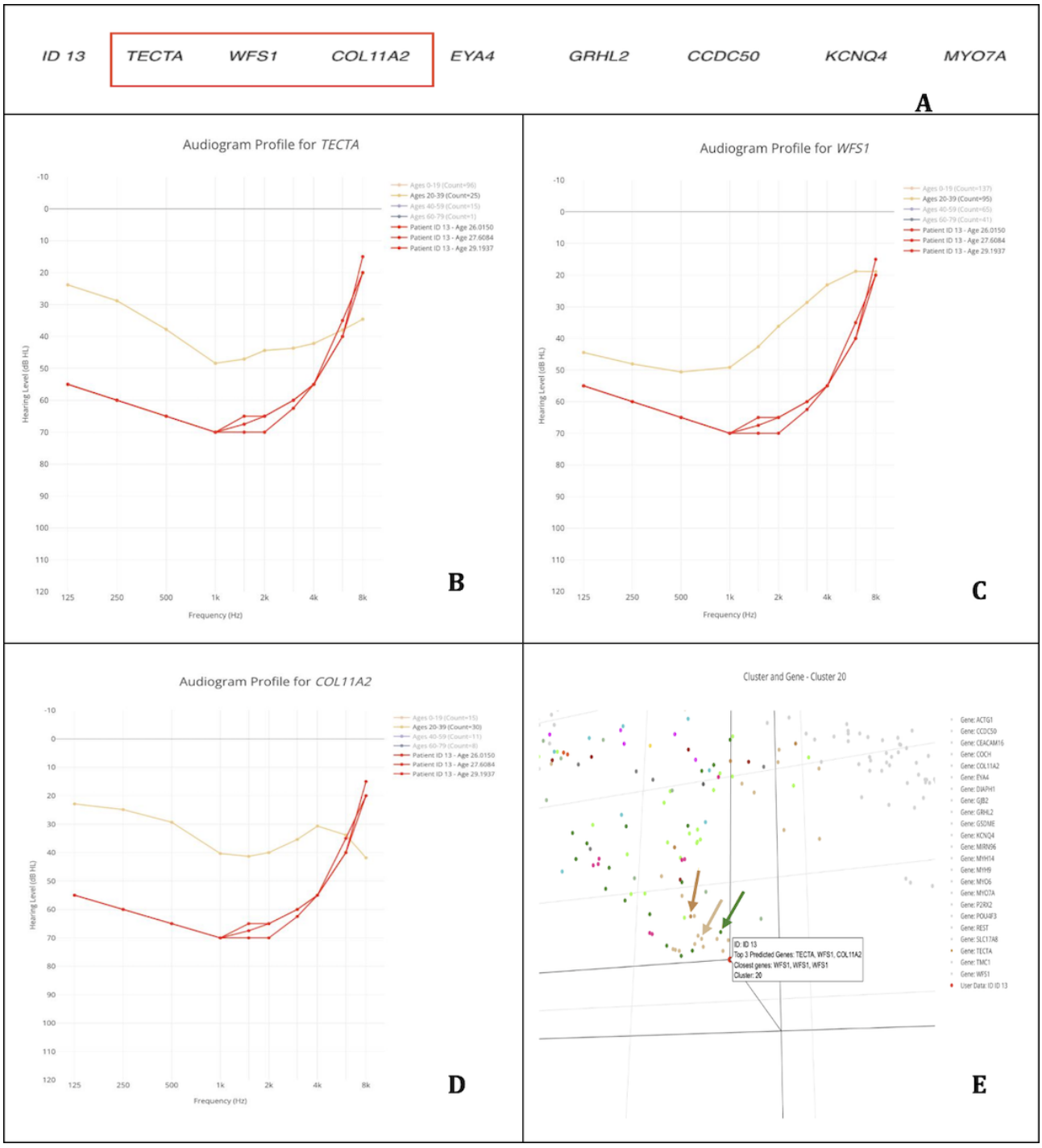
From the clustering interface, we observe that the genes closest to the patient's audiogram by the KNN metric (*WFS1*, *EYA4*, and *KCNQ4*) also fail to match the patient's observed audioprofile in any convincing way. Furthermore, *KCNQ4*, *GSDME*, and *POU4F3*, which are noted to have multiple data

points near the patient's cluster, likewise show audioprofiles inconsistent with the patient's hearing loss. These factors combine to produce inconclusive predictions in this case. When we then integrated these data with the genetic data—which identified no genetic variants in *WFS1*, *TECTA*, or *MYO7A* and confirmed a known variant in *MYO6*—we verified that the correct gene was not found among the model's top 3 predictions. This outcome highlights how conflicting audioprofiles and clustering results can indicate that a prediction should be viewed with caution.

Case 3: WFS1 Gene—Patient 3 (ID 13)

In our third case, the patient (ID 13) was diagnosed with *WFS1*-related hearing loss and, based on 3 audiograms, was predicted by AG4 to have *TECTA*-related, *WFS1*-related, or *COL11A2*-related hearing loss (Figure 4A).

Figure 4. Application of the AudioGene Translational Dashboard for patient-level gene prediction, audiometric profile comparison, and cluster-based visualization. (A) Predictions for the patient (ID 13), with the top 3 genes being TECTA, WFS1, and COL11A2; (B) audioprofile of TECTA with the patient's (ID 13) audiograms in red, taken at the ages of 26, 27, and 29 years; (C) audioprofile of WFS1 with the patient's (ID 13) audiograms in red, taken at the ages of 26, 27, and 29 years; (D) audioprofile of COL11A2 with the patient's (ID 13) audiograms in red, taken at the ages of 26, 27, and 29 years; and (E) 3D plot of audiograms in the training set converted into 3 dimensions, with genes in cluster 20 colored (genes not in cluster are light gray; patient [ID 13] is the red dot hovered over the displayed label; and the light brown arrow points to WFS1 [light brown dots], the green arrow points to MYO7A [green dots], and the brown arrow points to TECTA [brown dots]).



When examining these predictions in relation to the patient's audiograms, the following conclusions emerge regarding why *WFS1* is likely the correct gene and is captured within the top 3 predictions. First, the audioprofile for *TECTA* (Figure 4B) shows some similarities; however, it does not fully capture the nuanced relationship between low-frequency and high-frequency thresholds observed in the patient's data. Second, *COL11A2* (Figure 4D) also exhibits differences that diverge from the patient's pattern. Finally, *WFS1* (Figure 4C) demonstrates an especially close match to the patient's audiometric profile, particularly in the way it mirrors better hearing at the high frequencies relative to the low frequencies. Although one might contend that *TECTA* or *COL11A2* could also be considered candidates based on partial pattern matches, the overall evidence—supported by the 3D clustering in Figure 4E—reinforces that *WFS1* provides the best fit.

Thus, whether one emphasizes the possibility of *TECTA* or *COL11A2* as contenders, the integrated data confirm that the correct gene, *WFS1*, is indeed within the top 3 predictions. This close alignment between the patient's audiometric data and the *WFS1* reference profile, combined with supporting clustering analysis, enhances confidence in the diagnostic utility of the AudioGene Translational Dashboard.

Discussion

Principal Findings

These studies demonstrate how the model can raise or lower confidence in variant interpretation based on whether the correct genetic cause of ADNSHL appears among the top 3 predicted genes. Cases 1 and 3 illustrate scenarios in which the model successfully includes the causative gene in its top predictions and closely matches the patient's audiometric data, thereby justifying a higher level of trust in the result. In contrast, case 2 underscores how mismatched audioprofiles and inconclusive clustering can reveal when the actual gene of interest is likely missing from the top 3 predictions. The interactive visualizations of the AudioGene Translational Dashboard, such as the APS and spatial analysis tools, remain valuable in identifying gene-specific patterns that align with clinical observations [20].

However, there are important limitations of the AudioGene Translational Dashboard, especially concerning smaller gene classes. The sparsity of data and the lower accuracy of models in these categories can make the AudioGene Translational Dashboard and phenotypic predictions less reliable. However, by presenting visualizations of the data distribution and class imbalance, these limitations become more apparent, allowing data interpretation to be adjusted accordingly [2,4].

Conclusions

The AudioGene Translational Dashboard represents an advancement in the field of genetic diagnostics for ADNSHL. By integrating advanced ML algorithms with interactive visualization tools, the AudioGene Translational Dashboard enhances health care providers' ability to interpret genetic data and make more informed diagnostic decisions.

A central feature of the AudioGene Translational Dashboard is the "70/30" phenomenon, which provides health care providers with critical context for confidence in genetic predictions. When the top 3 predictions are likely to contain the correct gene, the tool serves as a "green flag" for health care providers, increasing diagnostic confidence. Conversely, it alerts health care providers when predictions may be less reliable, serving as a "red flag" and prompting further investigation.

The AudioGene Translational Dashboard is an example of XAI in clinical settings, offering a context-driven method with increased transparency for the diagnosis of ADNSHL. Future developments will focus on incorporating custom model building, enhancing class imbalance functionality, and implementing user suggestions. The AudioGene Translational Dashboard not only advances genetic diagnostics for hearing loss but also serves as an example of a hybrid ML system.

Funding

This research was supported by the National Institutes of Health and the National Institute on Deafness and Other Communication Disorders through grants DC002842, DC012049, and DC017955. These funding sources provided financial support for the development and testing of the AudioGene Translational Dashboard tool, contributing to advancements in machine learning and visualization methodologies for the diagnosis of autosomal dominant nonsyndromic hearing loss.

Data Availability

The datasets generated or analyzed during this study are not publicly available but are available from the corresponding author on reasonable request. The source code for the AudioGene Translational Dashboard is publicly available [26] under the GNU General Public License version 3.0 or later.

Authors' Contributions

BD led the design and development of the AudioGene Translational Dashboard, implemented the machine learning models, and drafted the manuscript. NS contributed to the integration of visualization tools and assisted with manuscript preparation. DW and AMO were responsible for data acquisition and preprocessing and contributed to tool validation. KTAB and HA provided expertise in genetic diagnostics and guided the tool's clinical relevance. MS supported statistical analysis and interpretation of the results. RJHS and TB provided project oversight, secured funding, and critically revised the manuscript. TC supervised the software engineering components and contributed to system architecture design. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Walls WD, Azaiez H, Smith RJ. Hereditary Hearing Loss Homepage. URL: https://hereditaryhearingloss.org [accessed 2026-03-28]
2. Taylor KR, Deluca AP, Shearer AE, et al. AudioGene: predicting hearing loss genotypes from phenotypes to guide genetic screening. Hum Mutat 2013 Apr;34(4):539-545. [doi: 10.1002/humu.22268] [Medline: 23280582]
3. DeSollar BR. AGTD - The AudioGene Translational Dashboard: a hybrid machine learning and visualization interface for genetic diagnosis of autosomal dominant non-syndromic hearing loss [Master's thesis]. : University of Iowa; 2024 URL: https://iro.uiowa.edu/esploro/outputs/graduate/9984647256502771 [accessed 2026-03-28]
4. Ryan S. Machine learning prediction of genetic hearing loss via selective intraensemble data partitioning [Master's thesis]. : University of Iowa; 2024 URL: https://iro.uiowa.edu/esploro/outputs/graduate/Machine-learning-prediction-of-genetic-hearing/9984647557802771 [accessed 2026-03-28]
5. Nwakama CC. AudioGene 9.0: novel ensemble machine learning classification of 23 classes of autosomal non-syndromic hearing loss (deafness) [Master's thesis]. : University of Iowa; 2021 URL: https://iro.uiowa.edu/view/pdfCoverPage?instCode=01IOWA_INST&filePid=13841170450002771&download=true [accessed 2026-03-28]
6. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-Explainable artificial intelligence. Sci Robot 2019 Dec 18;4(37):eaay7120. [doi: 10.1126/scirobotics.aay7120] [Medline: 33137719]
7. Smith RJH, Bale JF Jr, White KR. Sensorineural hearing loss in children. Lancet 2005 Mar;365(9462):879-890. [doi: 10.1016/S0140-6736(05)71047-3] [Medline: 15752533]
8. Venkatesh MD, Moorchung N, Puri B. Genetics of non syndromic hearing loss. Med J Armed Forces India 2015 Oct;71(4):363-368. [doi: 10.1016/j.mjafi.2015.07.003] [Medline: 26663965]
9. API reference—Pandas 1.5.3 documentation. Pandas. URL: https://pandas.pydata.org/pandas-docs/version/1.5/reference/index.html [accessed 2026-03-28]
10. Albarrak AM. Improving the trustworthiness of interactive visualization tools for healthcare data through a medical fuzzy expert system. Diagnostics (Basel) 2023 May 13;13(10):1733. [doi: 10.3390/diagnostics13101733] [Medline: 37238218]
11. Frank E, Hall MA, Witten IH. The WEKA workbench. In: Witten IH, Frank E, Hall MA, Pal CJ, editors. Data Mining: Practical Machine Learning Tools and Techniques, 4th edition: Morgan Kaufmann; 2016.
12. Deza E, Deza MM. Encyclopedia of Distances: Springer; 2009.
13. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009 Sep;19(9):1639-1645. [doi: 10.1101/gr.092759.109] [Medline: 19541911]
14. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory 1967;13(1):21-27. [doi: 10.1109/TIT.1967.1053964]
15. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Berkeley Symposium on Mathematical Statistics and Probability: University of California Press; 1967:281-297.
16. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013 Mar;14(2):178-192. [doi: 10.1093/bib/bbs017] [Medline: 22517427]
17. Wu TF, Lin CJ, Weng RC. Probability estimates for multi-class classification by pairwise coupling. J Mach Learn Res 2004;5:975-1005 [FREE Full text]
18. Weininger O, Warnecke A, Lesinski-Schiedat A, Lenarz T, Stolle S. Computational analysis based on audioprofiles: a new possibility for patient stratification in office-based otology. Audiol Res 2019 Sep 2;9(2):230. [doi: 10.4081/audiores.2019.230] [Medline: 31728177]
19. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J 2014;2014(239):2. [doi: 10.5555/2600239.2600241]
20. Taylor KR, Booth KT, Azaiez H, et al. Audioprofile surfaces: the 21st century audiogram. Ann Otol Rhinol Laryngol 2016 May;125(5):361-368. [doi: 10.1177/0003489415614863] [Medline: 26530094]
21. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw 2018;3(29):861. [doi: 10.21105/joss.00861]
22. Docker Compose. Docker Docs. URL: https://docs.docker.com/compose [accessed 2026-03-28]
23. React. URL: https://reactjs.org [accessed 2026-03-28]
24. Plotly JavaScript open source graphing library. Plotly. URL: https://plotly.com/javascript [accessed 2026-03-28]
25. 45 CFR §46.116 - General requirements for informed consent. Code of Federal Regulations. 2018. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.116 [accessed 2026-04-10]
26. Schaefer N. AudioGene. URL: https://research-git.uiowa.edu/morl/audiogene/website/AudioGene [accessed 2026-04-07]

Abbreviations

ADNSHL: autosomal dominant nonsyndromic hearing loss

AG4: AudioGene version 4
AG9.1: AudioGene version 9.1
APS: audioprofile surface
KNN: k-nearest neighbor
ML: machine learning
SERN: SQL, Express.js, React.js, and Node.js
SVM: support vector machine
UMAP: uniform manifold approximation and projection
XAI: explainable artificial intelligence

Edited by S Hacking; submitted 03.Oct.2025; peer-reviewed by H Wu, Sunny, CL Au, W Goar; revised version received 12.Mar.2026; accepted 12.Mar.2026; published 14.Apr.2026.

Please cite as:
DeSollar B, Schaefer N, Walls D, Odell AM, Booth KTA, Azaiez H, Schnieders M, Smith RJH, Braun T, Casavant T
The AudioGene Translational Dashboard for Diagnosing Autosomal Dominant Nonsyndromic Hearing Loss: Phenotypic Data Visualization and Analysis Study
JMIR Bioinform Biotech 2026;7:e85212
URL: https://bioinform.jmir.org/2026/1/e85212
doi:10.2196/85212

Temporal Reproducibility of a Genetic Algorithm–Derived Health Risk Score: Standardized Out-of-Fold Validation Framework (2021-2023)

Yoichiro Aoki[1], MD, PhD; Hiroki Takeda[2], MD; Kinichi Yokota[3], MD, PhD; Ryoko Yoshida[1], BA

[1]Yoshida Hospital-Keiyukai Medical Corporation, 1-2, Nishi 4-chome, 4-jyo, Asahikawa, Hokkaido, Japan

[2]Department of Cardiovascular Medicine, Yoshida Hospital- Keiyukai Medical Corporation, Asahikawa, Hokkaido, Japan

[3]Department of Gastroenterology, Yoshida Hospital-Keiyukai Medical Corporation, Asahikawa, Hokkaido, Japan

Corresponding Author:
Yoichiro Aoki, MD, PhD
Yoshida Hospital-Keiyukai Medical Corporation, 1-2, Nishi 4-chome, 4-jyo, Asahikawa, Hokkaido, Japan

Abstract

Background: Genetic algorithm (GA)–based scoring has been proposed as a data-driven approach for health risk stratification . However, performance estimates may be inflated when preprocessing, optimization, and evaluation are not strictly separated within a prespecified validation framework. Demonstrating temporal reproducibility under a standardized out-of-fold (OOF) evaluation framework with transparent uncertainty quantification is therefore essential for ensuring translational reliability in preventive health screening.

Objective: This study aimed to evaluate the temporal reproducibility of a GA-derived composite health risk score across three consecutive annual cohorts (2021 - 2023) under a standardized OOF validation pipeline and to assess robustness to policy-driven structural HbA_{1c} missingness through a prespecified ON/OFF sensitivity analysis.

Methods: Annual health examination datasets from 2021 (n=3744), 2022 (n=5153), and 2023 (n=5352) were analyzed using an identical preprocessing and modeling pipeline. Thirteen clinical indicators and eight lifestyle questionnaire variables were included as predictors. The outcome was based on an A–D grading framework and binarized using an OR rule across domains (grade ≥B in any domain). Continuous variables were median-imputed and standardized within each training fold to prevent information leakage. GA optimization was performed using fixed random seeds, and fitness estimation employed stratified K-fold cross-validation. Predicted probabilities were obtained by fitting logistic regression models to GA-derived composite scores within the OOF framework. Discrimination and overall predictive performance were quantified using the area under the receiver operating characteristic curve (AUC) and the Brier score calculated from OOF predicted probabilities. Uncertainty was estimated using 2,000-replicate percentile bootstrap resampling. A prespecified sensitivity analysis excluded HbA_{1c} while maintaining an identical evaluation framework.

Results: OOF AUC values were stable across cohorts (2021: 0.810; 2022: 0.814; 2023: 0.812), with overlapping 95% percentile bootstrap confidence intervals. Brier scores ranged from 0.172 to 0.176. Exclusion of HbA_{1c} resulted in small changes in discrimination (median ΔAUC was ≤0.007), consistent with the prespecified ON/OFF sensitivity analysis.

Conclusions: Under a harmonized OOF validation framework, the GA-derived composite risk score showed stable temporal discrimination and consistent overall predictive performance across three consecutive annual cohorts. These findings underscore the methodological importance of prespecified, standardized evaluation procedures and transparent uncertainty quantification when assessing reproducibility of risk stratification models in routine health screening data.

(JMIR Bioinform Biotech 2026;7:e85659) doi:10.2196/85659

KEYWORDS

genetic algorithm; health risk scoring; reproducibility; cross-validation; ROC; AUC; preventive medicine; area under the receiver operating characteristic curve

Introduction

In preventive health screening, risk classification commonly relies on threshold-based evaluation of individual clinical indicators (eg, blood pressure, lipids, or HbA_{1c}). While such approaches ensure procedural uniformity, they do not integrate multiple biological and lifestyle dimensions into a composite risk representation. Data-driven optimization approaches have therefore been proposed to enhance structural consistency and interpretability of health risk scoring.

Genetic algorithms (GAs), originally introduced by Holland [1] and further formalized by Goldberg [2], provide a flexible framework for feature weighting and optimization under cross-validated conditions. However, GA-based scoring models applied to real-world health checkup data require careful attention to reproducibility, methodological harmonization, and internal validation procedures. In particular, performance estimates may vary depending on whether preprocessing, optimization, and evaluation steps are strictly separated within a prespecified validation framework.

Bayesian estimation can be incorporated as an interpretive layer to express predicted risk probabilistically, aligning composite scores with calibrated predicted probabilities. Rather than emphasizing peak discrimination, evaluating temporal reproducibility under a standardized analytical framework is essential for ensuring methodological consistency.

The objective of this study was to evaluate the temporal reproducibility of a GA-derived composite health risk score across three consecutive annual cohorts (2021 - 2023) using a prespecified, standardized out-of-fold validation pipeline.

Methods

Data Source and Participants

We analyzed a deidentified dataset from annual health checkups conducted at the Preventive Medicine Center (Ningen Dock Division), Yoshida Hospital, Keiyukai Medical Corporation (Asahikawa, Hokkaido, Japan). The analytic cohorts comprised examinees from 2021 (n=3744), 2022 (n=5153), and 2023 (n=5352), each analyzed as an independent annual cohort.

Ethical Considerations

The study was approved by the Institutional Review Board of Yoshida Hospital, Keiyukai Medical Corporation (Approval No. 20251002001) and conducted in accordance with the Declaration of Helsinki. Data were deidentified prior to analysis. Written informed consent, including consent for secondary use of de-identified data, was obtained at the time of the health checkup. No additional interventions or participant contact occurred as part of this study.

Measures and Preprocessing

Thirteen routine clinical indicators were included: BMI, waist circumference, systolic blood pressure, diastolic blood pressure, fasting plasma glucose, hemoglobin A1c (HbA_{1c}), triglycerides, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, aspartate aminotransferase, alanine aminotransferase, γ-glutamyl transferase, and uric acid. Sex-specific thresholds were applied for waist circumference according to institutional criteria (see Table S2B in Multimedia Appendix 1).

Eight lifestyle questionnaire items were included (eg, smoking, alcohol consumption, breakfast habits, snacking, eating speed, mastication, physical activity/walking, and motivation for health improvement). Lifestyle questionnaire items were coded as binary variables according to the facility codebook. Variable definitions are provided in Table S2A in Multimedia Appendix 1.

Continuous variables were median-imputed and standardized within each training fold of each annual cohort. The imputation and standardization parameters were estimated within the training fold and applied to the corresponding held-out fold to prevent information leakage.

Indicator-level missingness rates are summarized in Table S1 in Multimedia Appendix 1. Missingness in indicators other than HbA_{1c} was low (<2% in each y).

Outcome Definition

The primary outcome was defined as a composite abnormality label derived from routine health-check classification rules used in the screening program. For each clinical domain (eg, glucose metabolism, blood pressure, lipids, liver enzymes, and anthropometric indices), examinees were categorized according to prespecified threshold-based grades (A–D).

In the institutional screening system, grade A indicates no abnormality; grade B indicates mild abnormality typically requiring lifestyle guidance; grade C indicates follow-up or re-evaluation; and grade D indicates recommendation for further diagnostic evaluation or treatment.

The composite outcome was binarized using an OR rule: participants were labeled outcome-positive if any domain met or exceeded the predefined abnormality threshold (grade B or higher); otherwise, they were labeled outcome-negative. This definition was applied consistently across all annual cohorts to evaluate structural reproducibility of the operational framework rather than severity-specific prognostic discrimination. Detailed domain-specific thresholds corresponding to grade B or higher are provided in Table S2B in Multimedia Appendix 1.

The grading thresholds (A–D) were defined according to the standardized health-check classification framework established by the Japanese Society of Ningen Dock and Preventive Medicine, which is based on national health screening standards and specialty society guidelines. These classifications are widely used in routine health-check programs across Japan to guide follow-up recommendations (eg, observation, repeat testing, referral, or treatment). This study adopted these externally defined operational criteria without modification and dichotomized grade B or higher to capture any clinically relevant abnormality that warrants structured follow-up under this program. Participants categorized as grade E (under active treatment) were excluded from the outcome classification process.

Handling of HbA1c Structural Missingness

HbA1c was structurally missing for a subset of participants due to the program's screening policy-based test selection. Indicator-level missingness rates are summarized in Table S1 in Multimedia Appendix 1.

Because the composite outcome was defined using an OR rule across multiple domains, outcome ascertainment did not depend solely on HbA_{1c}.

To evaluate robustness to structural HbA1c missingness, we conducted a prespecified sensitivity analysis excluding HbA1c

from the predictor set while maintaining an identical evaluation pipeline (HbA_{1c} included vs excluded).

Model Development and Evaluation

A composite score was generated from standardized features using a genetic algorithm (GA). All stochastic components were controlled by fixing the random seed (SEED=42) for both Python's random module and NumPy. Fitness estimation used stratified K-fold cross-validation with shuffling (random_state=42).

The GA-derived composite score was subsequently entered into a logistic regression model to generate calibrated predicted probabilities within a prespecified out-of-fold (OOF) validation framework (Platt scaling [3]).

Bayesian updating was applied post hoc for interpretability purposes to the calibrated predicted probabilities and did not influence GA optimization or probability estimation.

For each fold, the model was trained on the training subset and evaluated on the corresponding held-out subset. OOF predictions were aggregated across folds to obtain a single internally validated prediction for each participant within each annual cohort.

Discrimination and Calibration

Model discrimination was assessed using the area under the receiver operating characteristic curve (AUC) calculated from OOF predicted probabilities generated within the prespecified cross-validated pipeline.

Overall predictive performance was quantified using the OOF-based Brier score. Calibration was examined descriptively using calibration plots based on OOF predicted probabilities (Figure S1 in Multimedia Appendix 1). No additional recalibration or threshold optimization was performed beyond the prespecified validation framework.

Bootstrap Uncertainty Estimation

To improve statistical transparency, 95% percentile bootstrap confidence intervals for OOF AUC and Brier score were computed using 2000 participant-level resamples within each annual cohort and HbA_{1c} condition (ON/OFF). Performance metrics were recalculated from the previously generated OOF predicted probabilities without refitting the model, thereby preserving internal validation and avoiding information leakage while maintaining the integrity of the original cross-validated predictions.

Year-stratified OOF performance estimates and confidence intervals are reported in Table S4 in Multimedia Appendix 1 and prespecified ON–OFF differences are summarized in Table S5 in Multimedia Appendix 1.

Statistical Software

All analyses were performed in Python (version 3.13.5) using *scikit-learn* (version 1.6.1) [4] and DEAP (version 1.4) [5]. Analyses were conducted in a Jupyter-based environment (Anaconda distribution). Detailed GA configuration parameters, including population size, number of generations, weight initialization range, crossover and mutation settings, and random-seed control, are provided in Multimedia Appendix 1 ("Genetic Algorithm Implementation") to facilitate reproducibility.

Results

Discrimination and Overall Predictive Performance Under Harmonized Out-of-Fold (OOF) Validation

Across the three annual cohorts, sample sizes ranged from 3744 to 5352 individuals, with outcome prevalence between 36.7% and 37.9%, indicating comparable class balance across years.

Under the standardized OOF validation framework, discrimination remained stable across cohorts. OOF AUC values were 0.810 (2021), 0.814 (2022), and 0.812 (2023), with overlapping 95 % bootstrap confidence intervals (Table 1). Brier scores ranged from 0.172 to 0.176 across cohorts, indicating stable overall predictive performance across years (Table 1).

Table . Discrimination and overall predictive performance under harmonized out-of-fold (OOF) validation (2021 - 2023; primary model with HbA1c included). AUC and Brier score were calculated exclusively from OOF predicted probabilities generated within the prespecified cross-validation pipeline. Values in parentheses represent 95% percentile bootstrap confidence intervals based on 2000 resamples. Calibration plots based on OOF predicted probabilities are provided in Figure S1 in Multimedia Appendix 1. The prespecified HbA_{1c} ON/OFF sensitivity analysis and ON−OFF differences are summarized in Table S5 in Multimedia Appendix 1.

Year	Individuals, n	Outcome prevalence	OOF[a] AUC[b] (95% CI)	Brier score (95% CI)
2021	3744	0.375	0.810 (0.794 - 0.820)	0.176 (0.170 - 0.183)
2022	5153	0.379	0.814 (0.802 - 0.825)	0.173 (0.168 - 0.178)
2023	5352	0.367	0.812 (0.800 - 0.824)	0.172 (0.166 - 0.177)

[a]OOF: out-of-fold.

[b]AUC: area under the receiver operating characteristic curve.

Corresponding calibration plots are provided in Figure S1 in Multimedia Appendix 1.

In the prespecified HbA_{1c} ON/OFF sensitivity analysis, exclusion of HbA1c resulted in minimal changes in discrimination and Brier score (Table S5 in Multimedia Appendix 1), suggesting limited sensitivity of model performance to policy-driven structural HbA1c missingness under the harmonized OOF validation framework.

As the primary objective was to assess temporal reproducibility under standardized analytical procedures, formal hypothesis testing of between-year differences was not performed.

Discussion

Principal Findings

This study evaluated the temporal reproducibility of a genetic algorithm (GA)–derived composite health risk score across three consecutive annual health checkup cohorts under a prespecified OOF validation framework. Cross-validated OOF AUC values ranged from 0.810 to 0.814, with overlapping bootstrap confidence intervals, indicating stable discrimination under standardized analytical procedures. Earlier exploratory analyses yielded higher apparent AUC values; however, these were not derived exclusively from OOF predictions under the same pipeline and are therefore not presented as primary performance estimates. The primary contribution of this study is methodological: it demonstrates that a GA-derived score can achieve stable OOF discrimination across consecutive cohorts when preprocessing, optimization, and evaluation are uniformly applied, rather than emphasizing peak performance under heterogeneous analytical conditions.

Because performance estimates were based exclusively on OOF predicted probabilities, the evaluation preserved internal validation and minimized information leakage. Variability observed in earlier exploratory analyses likely reflected differences in analytical procedures rather than underlying cohort characteristics, highlighting the importance of consistent preprocessing, optimization, and evaluation when assessing artificial intelligence–based risk stratification models.

Importantly, the outcome definition reflects operational screening classification rather than confirmed clinical diagnoses. The composite label was constructed using threshold-based OR combinations across correlated clinical domains, and the dominant abnormality domain contributing to classification may vary across cohorts depending on distributional shifts. In this context, AUC values represent the model's ability to consistently reconstruct the structured screening framework under harmonized analytical conditions rather than discrimination of severity-specific disease states.

Methodologically, GA optimization produced a composite score from standardized predictors, which was subsequently mapped to calibrated predicted probabilities using logistic regression. Bayesian updating was applied post hoc as an interpretability add-on to the calibrated predicted probabilities and did not influence GA optimization or probability estimation. The explicit specification of evolutionary hyperparameters, weight initialization range, and random-seed control further strengthens the reproducibility of the optimization procedure and reduces the likelihood that the observed discrimination reflects stochastic artifacts.

The prespecified HbA1c ON/OFF sensitivity analysis showed minimal changes in AUC and Brier score, suggesting limited sensitivity to policy-driven structural HbA1c missingness under the standardized evaluation framework; however, re-optimizing GA weights under an altered feature set represents a distinct modeling exercise.

This study has limitations. It was conducted at a single center in Japan using an occupational health checkup population. The outcome was cross-sectional, and prospective or external validation was not performed. In addition, because grade B or higher includes heterogeneous categories with varying clinical implications (ranging from lifestyle guidance to referral for diagnostic evaluation), model discrimination reflects detection of operational abnormality rather than exclusively high-severity disease states. Furthermore, because the outcome was defined using threshold-based criteria that partially overlap with included predictors, discrimination should be interpreted as reconstruction of the operational classification within the same workflow rather than independent prognostic discrimination. This endpoint is an operational classification designed for workflow consistency, not a clinically adjudicated diagnosis. Accordingly, the present OOF-based results provide an internally validated assessment of temporal reproducibility within this screening system.

Conclusions

In conclusion, under a prespecified, harmonized OOF validation framework within a single institutional screening system, the GA-derived composite risk score demonstrated stable temporal discrimination and consistent overall predictive performance across three consecutive annual cohorts. These findings support methodological reproducibility and structural consistency under standardized analytical procedures. However, the present results do not establish external generalizability or clinical effectiveness, and independent external and prospective validation is required before broader clinical implementation can be inferred.

Acknowledgments

The authors thank the staff of the Preventive Medicine Center (Ningen Dock Division), Yoshida Hospital-Keiyukai Medical Corporation, including Junko Suzuki, Masami Takahashi, Eri Kagaya, Miki Sato, Mikiko Shibuya, Toshiharu Hazeyama, and Kouichi Kagi, for their assistance with data collection and management. The authors also thank the participants of the health checkup programs for their cooperation. The institution was certified as a government-recognized clinical research center (MEXT-approved, ID: 90106, August 2024).

Funding

This research received no external funding and was conducted as part of the institutional research activities of Keiyukai Medical Corporation.

Data Availability

The datasets generated and analyzed during the current study are not publicly available due to institutional policy and ethical restrictions. Deidentified data may be made available from the corresponding author upon reasonable request and with approval by the institutional review board.

Authors' Contributions

Conceptualization: YA
Data curation: RY
Formal analysis: YA
Methodology: YA
Validation: HT, KY
Writing – original draft: YA
Writing – review & editing: YA, HT, KY, RY

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials including Tables S1–S5, additional methodological details (GA implementation and Bayesian risk update), calibration metrics, and prespecified HbA_{1c} ON/OFF sensitivity analyses.
[DOCX File, 429 KB - bioinform_v7i1e85659_app1.docx]

References

1. Holland JH. Adaptation in Natural and Artificial Systems: University of Michigan Press; 1975.
2. Goldberg DE. Genetic Algorithms in Search, Optimization, and Machine Learning: Addison-Wesley; 1989.
3. Platt J. Probabilities for SV machines. In: Advances in Large Margin Classifiers: MIT Press; 1999:61-74. [doi: 10.7551/mitpress/1113.003.0008]
4. Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
5. Fortin FA. DEAP: evolutionary algorithms made easy. J Mach Learn Res 2012;13:2171-2175 [FREE Full text]

Abbreviations

AUC: area under the receiver operating characteristic curve
GA: genetic algorithm
HbA_{1c}: hemoglobin A1c
OOF: out-of-fold

Edited by Z Yue; submitted 14.Oct.2025; peer-reviewed by F Khamesipour, M Shannawaz; revised version received 28.Feb.2026; accepted 13.Mar.2026; published 21.Apr.2026.

Please cite as:
Aoki Y, Takeda H, Yokota K, Yoshida R
Temporal Reproducibility of a Genetic Algorithm–Derived Health Risk Score: Standardized Out-of-Fold Validation Framework (2021-2023)
JMIR Bioinform Biotech 2026;7:e85659
URL: https://bioinform.jmir.org/2026/1/e85659
doi:10.2196/85659

Systematic Mining of Bioactive Compounds for Wound Healing From Cayratia Japonica Exosome-Like Nanovesicles: A Workflow Combining LC-MS and DeepSeek Models

Qiang Fu[1,2], PhD; Wei Ji[3], MS; Yu-Ping Fan[4], MBBS; Jian Yao[5], PhD; Ming-Xia Song[2,6], PhD; Qiao-Jing Yan[2,6], PhD

[1]School of Basic Medical Sciences, Jinggangshan University, Ji'an, China

[2]Jiangxi Province Key Laboratory of Organ Development and Epigenetics, Clinical Medical Research Center, Affiliated Hospital of Jinggangshan University, College of Jinggangshan University, 28 Xueyuan Road, Qingyuan District, Ji'an, China

[3]University of Montpellier, Montpellier, France

[4]Department of Epidemiology & Biostatistics, School of Public Health, Southeast University, Nanjing, China

[5]Division of Molecular Signaling, Department of the Advanced Biomedical Research, Interdisciplinary Graduate School of Medicine, University of Yamanashi, Chuo, Japan

[6]College of Traditional Chinese Medicine and Pharmacy, Jinggangshan University, Ji'an, China

Corresponding Author:
Qiao-Jing Yan, PhD
Jiangxi Province Key Laboratory of Organ Development and Epigenetics, Clinical Medical Research Center, Affiliated Hospital of Jinggangshan University, College of Jinggangshan University, 28 Xueyuan Road, Qingyuan District, Ji'an, China

Abstract

Background: Plant-derived exosome-like nanovesicles (P-ELNs) effectively deliver bioactive compounds due to their high biocompatibility and low immunogenicity. While liquid chromatography-mass spectrometry (LC-MS) profiles compounds in complex samples, its analysis of large datasets remains limited by traditional methods. Recent advances in large language models (LLMs) and domain-specific systems have enhanced Chinese biomedical data processing and cross-modal pharmaceutical research.

Objective: This study aimed to create a multimodal framework of LC-MS combined with DeepSeek models for data mining of compounds with wound-healing properties from exosome-like nanovesicles derived from *Cayratia japonica* (CJ-ELNs).

Methods: LC-MS identified compounds enriched in CJ (n=3) and CJ-ELNs (n=3), and then compounds specifically enriched in CJ-ELNs were filtered via a four-step filtering workflow. The CJ-ELNs-specific compounds were processed by DeepSeek models for screening naturally active compounds with targeted functions of antioxidation, anti-inflammation, anticellular damage, antiapoptosis, wound healing and tissue regeneration, and cell proliferation.

Results: A multimodal framework of LC-MS combined with the DeepSeek-DF model was created. With the assistance of artificial intelligence (AI), a total of 46 naturally active compounds derived from CJ-ELNs with targeted functions were identified.

Conclusions: A self-designed multimodal framework of LC-MS, combined with DeepSeek models, rapidly and accurately identifies naturally active compounds from CJ-ELNs. This AI-powered system innovatively integrates the traditional analytical technique with modern LLMs, thus greatly favoring data mining of active ingredients in traditional Chinese medicine herbs.

(JMIR Bioinform Biotech 2026;7:e80539) doi:10.2196/80539

KEYWORDS

DeepSeek; liquid chromatography-mass spectrometry; LC-MS; Cayratia japonica exosome-like nanovesicles; CJ-ELNs; artificial intelligence; AI-powered multimodal framework; wound healing and tissue regeneration

Introduction

Plant-derived exosome-like nanovesicles (P-ELNs) contain abundant bioactive molecules, serving as novel carriers of natural products to mediate intercellular communication and mediate physiological processes [1,2]. P-ELNs are superior to conventional mammalian-derived exosomes, possessing unique advantages such as high biocompatibility, high skin permeability, low cytotoxicity and low immunogenicity [3,4]. Multiple in vitro and in vivo studies indicate that these P-ELNs possess intrinsic therapeutic activity, offering promise for disease treatment and enhancing human health [5,6]. *Cayratia japonica*, a traditional Chinese medicinal herb, is widely used for the treatment of traumatic injuries such as contusions and lacerations [7]. Recent clinical studies have confirmed that

topical application of CJ ointment effectively alleviates local inflammation and promotes the repair and regeneration of damaged tissue, demonstrating favorable therapeutic outcomes in the management of postoperative infectious wounds around the anus [8]. However, research and application of exosome-like nanovesicles (ELNs) derived from CJ remain incomplete. Our research team successfully extracted and characterized a novel type of P-ELNs from the traditional Chinese medicinal herb *Cayratia japonica*, namely *Cayratia japonica* exosome-like nanovesicles (CJ-ELNs). They possess efficient delivery of bioactive compounds to wound sites, thus favoring tissue regeneration from infectious wound-related disorders. Bioactive constituents encapsulated within CJ-ELNs are dominant in wound healing. Consequently, the identification and characterization of bioactive compounds responsible for wound healing are of paramount significance.

Great strides have been made in the screening of active ingredients from natural products via omics techniques [9]. Liquid chromatography–mass spectrometry (LC-MS) has emerged as a powerful tool for profiling trace-level compounds in complex samples, although its performance in processing massive data is limited by traditional manual or rule-based analytical approaches [10,11]. In recent years, large-scale pretrained language models (LLMs), such as ChatGPT, GPT-4, and domain-specific systems like DeepSeek, have significantly transformed the landscape of biomedical data analysis and knowledge discovery [11,12]. These models exhibit powerful capabilities in natural language understanding, semantic reasoning, and prompt-based knowledge retrieval [13-15]. They are promising tools to assist omics analysis. In particular, DeepSeek models have been widely adopted for optimizing Chinese-language biomedical contexts, and supporting cross-modal tasks in pharmaceutical research, such as entity recognition, document summarization, and semantic ranking [16,17].

In this study, we innovatively created a multimodal framework of LC-MS combined with DeepSeek models for data mining of compounds with wound-healing properties from CJ-ELNs. This work illustrates the potential of artificial intelligence (AI) as a computational engine in natural compound discovery and offers a scalable solution for mining multimodal biochemical data.

Methods

Preprocessing of LC-MS Data

Untargeted metabolomic profiling of CJ and CJ-ELNs was performed by LC-MS. A total of 6 samples (including 3 CJ samples and 3 CJ-ELNs samples) were analyzed using a ultra-high-performance liquid chromatography (UHPLC) system coupled to a Q Exactive HF-X mass spectrometer (Thermo Scientific). Chromatographic separation was performed on an HSS T3 column (maintained at 40°C) with a 12-minute linear gradient from 2% to 98% mobile phase B at a flow rate of 0.3 mL/min. Mass spectrometry (MS) data were acquired in both positive and negative electrospray ionization (ESI) mode (± ESI) using a data-dependent acquisition strategy (top 10 most intense ions). Raw data were first converted to the mzML format using ProteoWizard, followed by processing, using Compound Discoverer 3.3 (Thermo Fisher Scientific) for peak alignment (with maximum retention time shift of 0.5 min and mass tolerance of 10 ppm) and normalization (using the median of maximum peak areas). Compound identification was achieved by matching MS/MS spectra against the following databases: mzCloud, LipidMaps, KEGG, HMDB, and MassBank. The matching criteria were set to a mass tolerance of 10 ppm and a minimum match factor threshold of 10. A four-step filtering workflow was designed to quantitatively identify target compounds as follows (Figure 1).

1. Filtering of match confidence: compounds with spectral match scores ≥80 were retained [18];
2. Filtering of unique compounds of CJ-ELNs: compounds identified in CJ and CJ-ELNs were compared with isolated compounds unique to CJ-ELNs;
3. Filtering of biological relevance: candidate compounds were screened for associations with wound healing-related signaling pathways using the DeepSeek-Bio model;
4. Semantic recognition and prompt engineering: final candidate molecules were refined through semantic analysis and prompt-based selection.

Common and unique compounds derived from CJ and CJ-ELNs were visualized in a Venn diagram, and a word cloud analysis was conducted via Python. Functions and tools, and databases of key terms used in this study are listed in Table 1.

Figure 1. A four-step filtering workflow. CJ-ELNs: *Cayratia japonica* exosome-like nanovesicle; LC-MS: liquid chromatography-mass spectrometry.

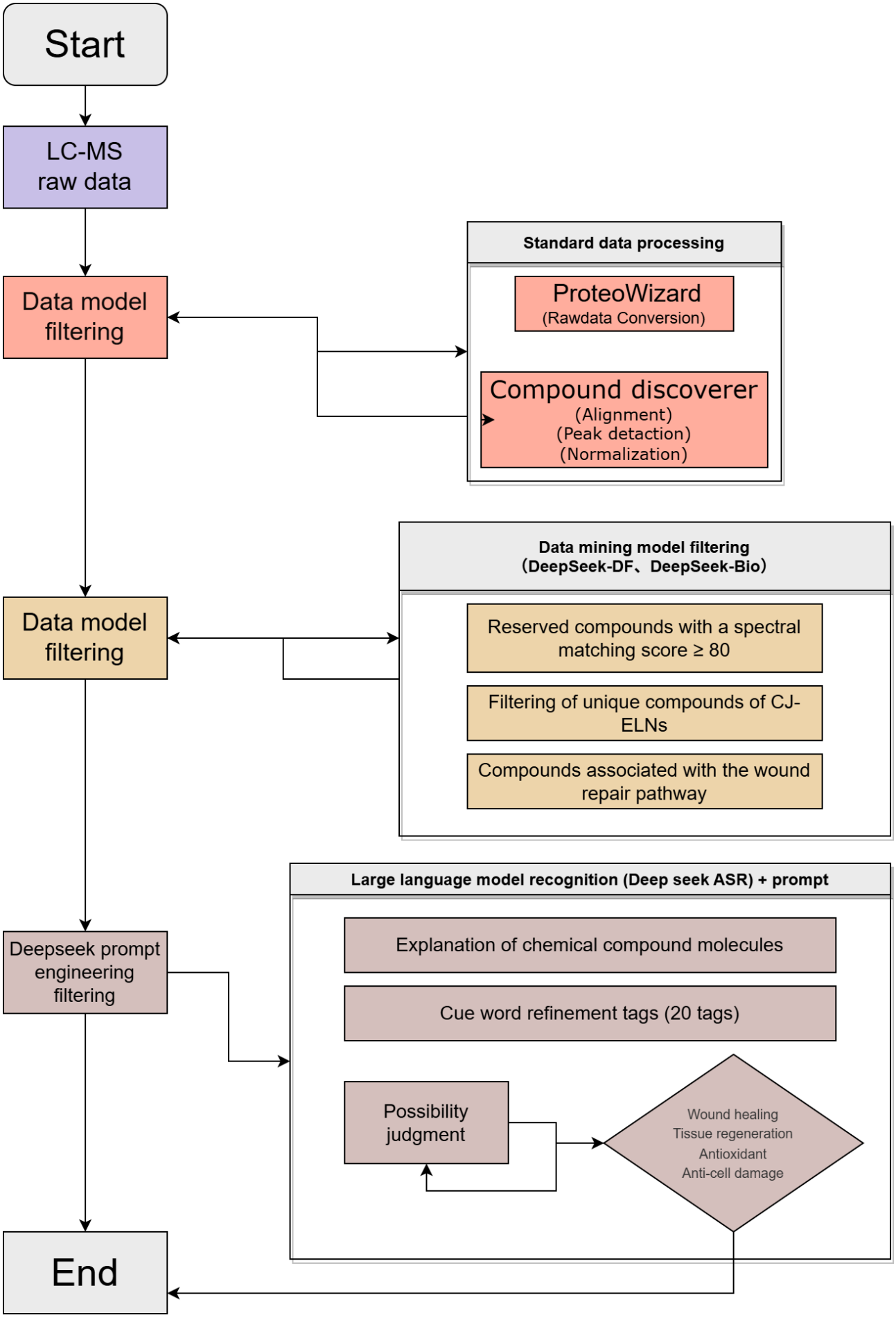


Table . Key terms, functions, tools and databases used in this study.

Key terms	Functions	Tools/databases
mzML	Standardized data storage	ProteoWizard
DeepSeek-Bio	Biological pathway association analysis	Deepseek 671B Model Network Edition KEGG database
Morgan	Digital characterization of molecular structures	Chemoinformatics software packages
PubMedBERT	Literature feature extraction	PubMed.pro
Grad-CAM	Visualization of model decisions	Deep learning frameworks (eg, PyTorch)
ASR	automatic semantic recognition	The Great Prophecy Model of Human-Computer Interaction

Construction of a Multimodal Framework of LC-MS Combined With DeepSeek Models

A multimodal framework of LC-MS combined with the DeepSeek-DF model was created, consisting of two major components of the input and output layers. The input layer integrated structural features of compounds (Morgan fingerprints), quantitative features (*z* score normalization), and literature-derived features (PubMedBERT embeddings). The core architecture was listed in Figure 2. Additionally, the output layer used multitask learning to simultaneously predict wound-healing activity via Sigmoid output and mechanism category via Softmax output.

Figure 2. The core architecture of the input layer.

```
class DualAttentionNN(nn.Module):
    def __init__(self):
        super().__init__()
        self.struct_net = GATv2Conv(
            in_channels=2048,
            hidden_channels=512
        )
        self.quant_net = TransformerEncoder(
            layers=4,
            d_model=256
        )
        self.fusion = DeepSeekCrossAttention(
            embed_dim=768
        )
```

Interpretability-Based Filtering

The Automated Semantic Recognition (ASR) module and prompt engineering techniques of DeepSeek-R1 32B, as well as web searching were used to interpret the potential biological functions of the screened candidate compound with an annotation of functional labels. A plausibility assessment was then performed based on predefined criteria, including antioxidation, anti-inflammation, anticellular damage, antiapoptosis, wound healing and tissue regeneration, and cell proliferation. Each compound was evaluated and categorized using the following scoring scheme: √ (confirmed), × (not supported), and ? (uncertain). Taking the metabolite (-)-Epicatechin 3-O-gallate as an example, its function, category and possibility in the involvement of wound healing, tissue regeneration, antioxidant, and anticellular damage were predicted via the multimodal framework (Table 2). Following this preliminary filtering, manual curation was conducted to eliminate compounds of nonplant origin and those with low abundances. Ultimately, a refined set of characteristic natural products from CJ-ELNs with potential wound-healing properties was selected.

Table . Functions, categories and possibility in the involvement of biological processes of representative metabolites.

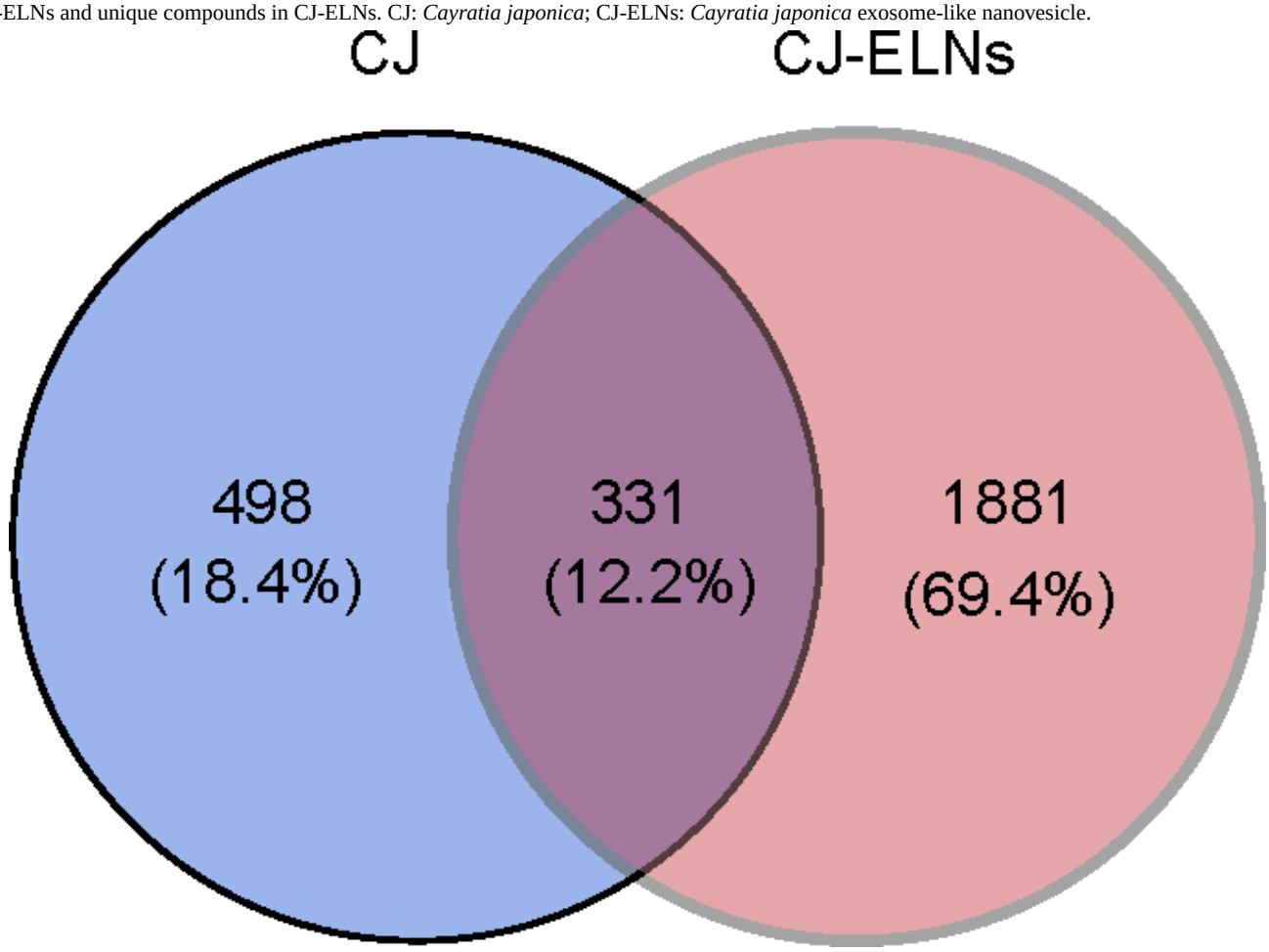
Compound	Functions	Categories	Possibility
(-)-Epicatechin 3-O-gallate	Antioxidant, anti-inflammatory, anti-cancer, cardiovascular protection, glucose and lipid metabolism regulation.	Organic compound, antioxidant factor, anti-inflammatory factor, energy metabolism, phenolic factor	Wound healing: ×, tissue regeneration: ×, antioxidant: √, anti-cellular damage: ?
Rutin	Antioxidant and anti-inflammatory, maintaining vascular resilience, reducing vascular permeability and fragility, exhibiting certain antiviral and anticancer effects.	Flavonoids, antioxidant, anti-inflammatory	Wound healing: ×, tissue regeneration: ×, antioxidant: √, anti-cellular damage: ?
Caffeine	Central nervous system stimulants, enhance mental alertness, alleviate fatigue.	Organic compounds, alkaloids, energy metabolism	Wound healing: ×, tissue regeneration: ×, antioxidant: ×, anti-cellular damage: ×

Results

Acidic Compounds Are Enriched in CJ-ELNs

After conversion and normalization of the raw LC-MS data, a total of 829 and 2212 compounds were identified from CJ and CJ-ELNs. A Venn diagram visualized 1881 specific compounds in CJ-ELNs (Figure 3). “Acid,” as the most frequent term across all entries of metabolite names, was detected by a word cloud analysis (Multimedia Appendix 1). It suggested that acidic compounds were highly enriched in CJ-ELNs.

Figure 3. Enrichment of acidic compounds in CJ-ELNs. (A) A Venn diagram visualizing an intersection of compounds identified from both CJ and CJ-ELNs and unique compounds in CJ-ELNs. CJ: *Cayratia japonica*; CJ-ELNs: *Cayratia japonica* exosome-like nanovesicle.

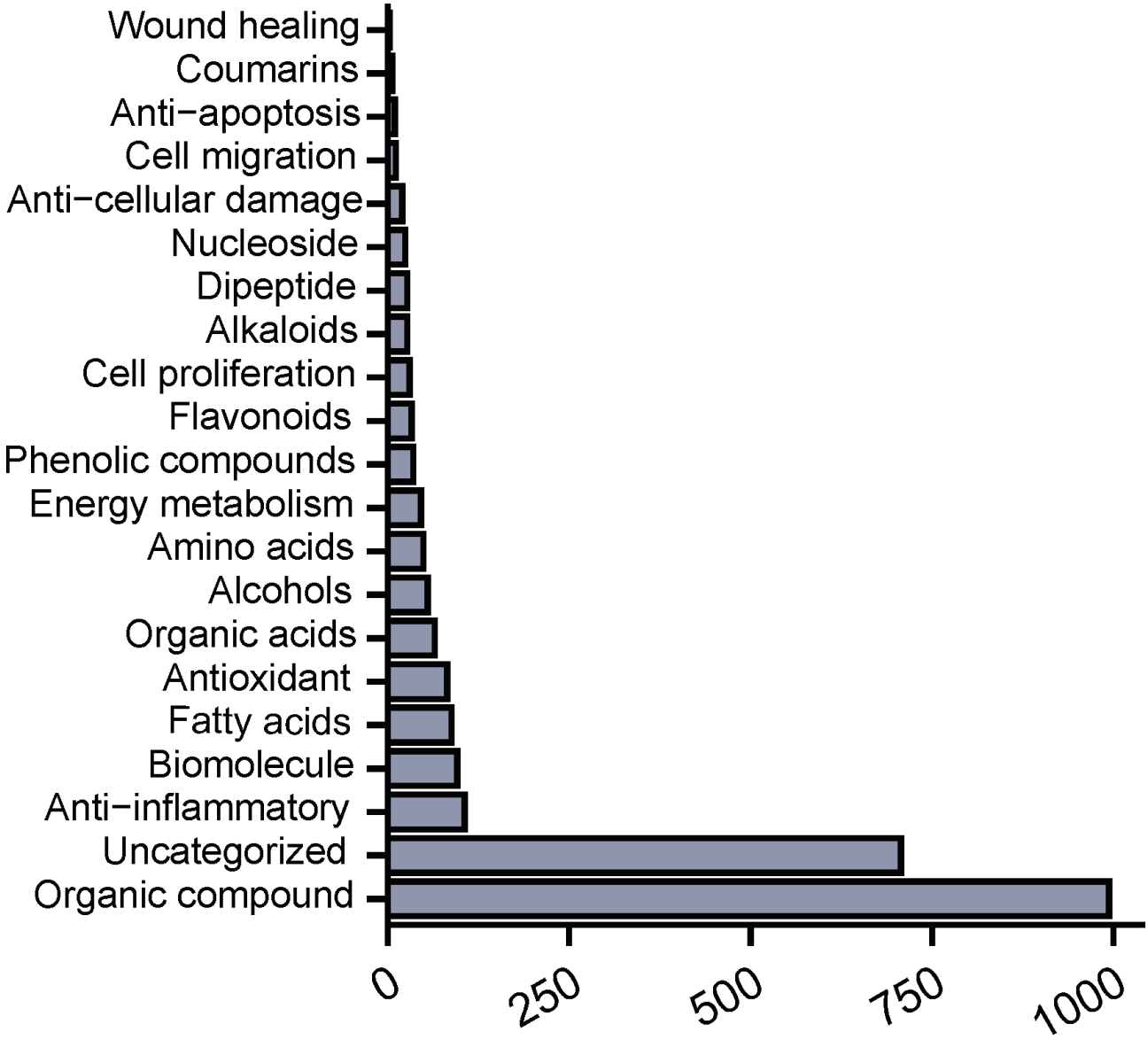


Rapid and Accurate Data Mining of Compounds in CJ-ELNs With Functional Properties

A total of 1881 candidate compounds enriched in CJ-ELNs were functionally annotated and classified using the self-designed multimodal framework of LC-MS combined with DeepSeek models. They were categorized into 20 distinct classes, including organic compounds, alkaloids, amino acids, biomolecules, organic acids, antioxidants, anti-inflammatory agents, energy metabolism-related molecules, phenolics, cytoprotective agents, alcohols, and others. Organic compounds were the leading category of compounds enriched in CJ-ELNs

(Figure 4, Multimedia Appendix 2). Functionally, 43.33% (n=39) of compounds enriched in CJ-ELNs possessed the antioxidant property. With the assistance of DeepSeek, we specifically screened compounds enriched in CJ-ELNs with targeted functions of antioxidation, anti-inflammation, anticellular damage, antiapoptosis, wound healing and tissue regeneration, and cell proliferation.

Figure 4. Rapid and accurate data mining of compounds in CJ-ELNs with functional properties. Top 20 classifications of compounds enriched in CJ-ELNs. CJ-ELN: *Cayratia japonica* exosome-like nanovesicle.



Bioactive Compounds of CJ-ELNs Responsible for Wound Healing and Tissue Regeneration

We estimated the overall expression levels of compounds across the six target functions derived from the DeepSeek model within this multimodal framework, visualizing the results in radar chart format after log2-transformation. (Figure 5). Notably, compounds with the antioxidant function possessed the highest expression levels, proving the antioxidant mechanism of CJ-ELNs in wound repair. Finally, a secondary filtering of compounds with targeted functions was conducted. We manually excluded nonplant–derived compounds, including those of animal origin, synthetic chemicals, and other nonbotanical sources. In addition, compounds with low expression levels in CJ-ELNs were also removed. As a result, a total of 46 naturally active compounds derived from CJ-ELNs with targeted functions were identified (Figure 6 and Multimedia Appendix 3). Citric acid was the most abundant compound with the targeted functions, which was consistent with the finding from the word cloud analysis.

Figure 5. Radar plots visualizing bioactive compounds of *Cayratia japonica* exosome-like nanovesicles with targeted functions.

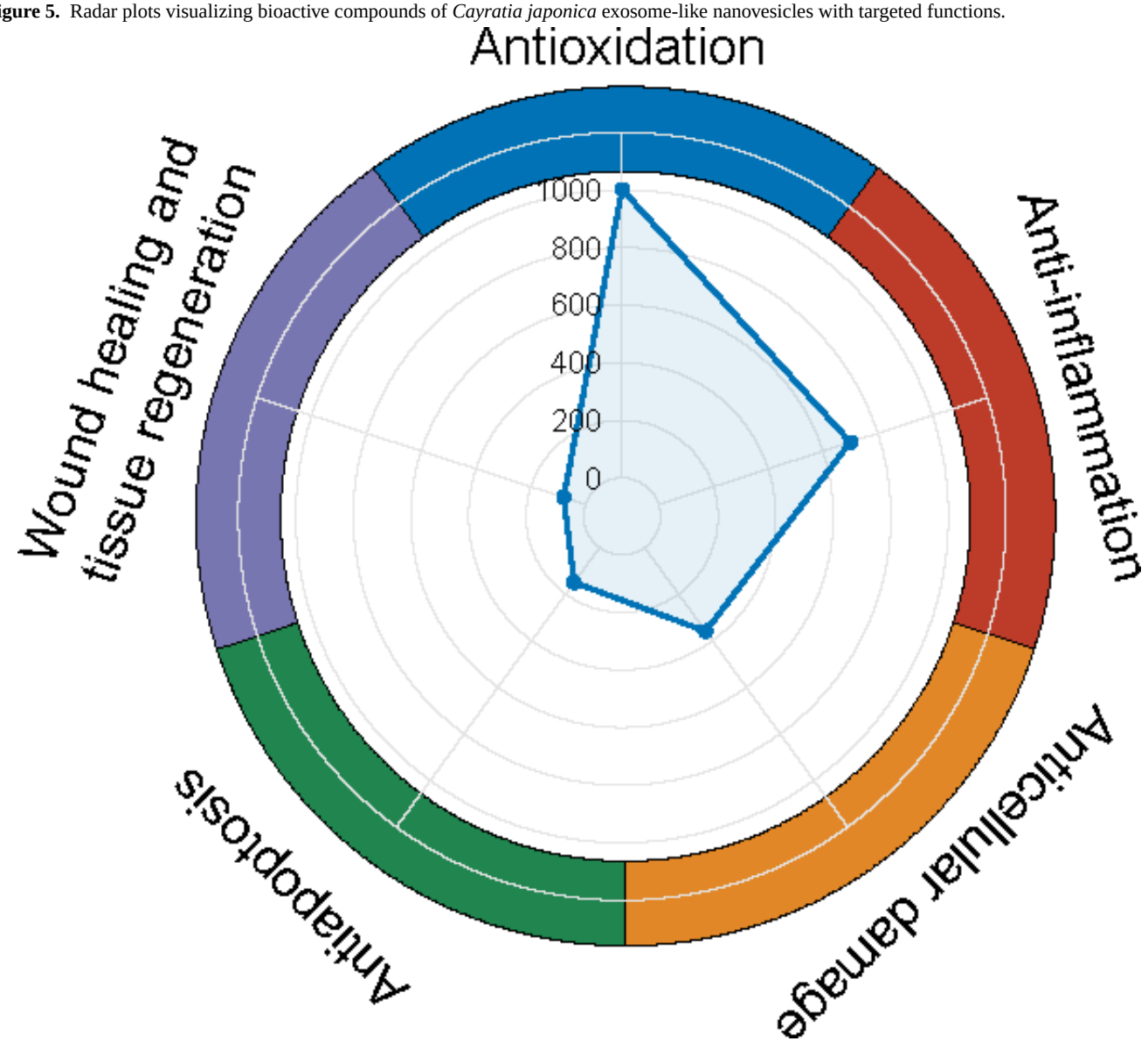
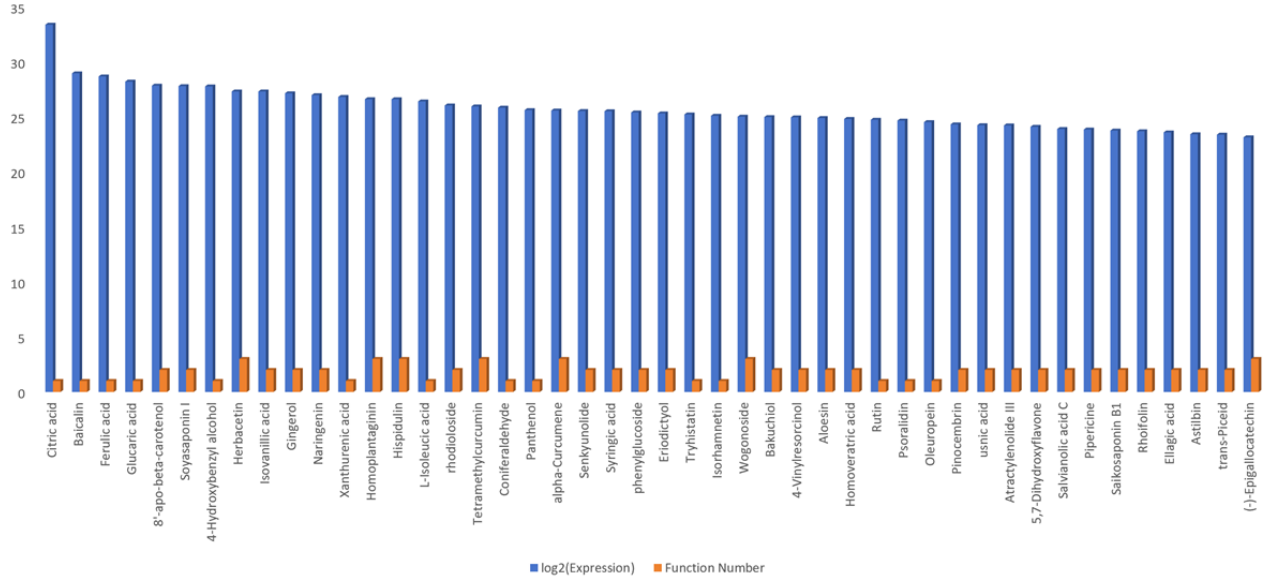


Figure 6. Expression levels (log_2-transformed) of naturally active compounds derived from *Cayratia japonica* exosome-like nanovesicle identified by an integration of liquid chromatography-mass spectrometry and DeepSeek models.



Discussion

Principal Findings

This study innovatively integrated DeepSeek models with LC-MS to successfully predict the major natural products of CJ-ELNs responsible for wound healing. DeepSeek's ASR semantic recognition and prompt engineering worked together to generate initial classification labels. Moreover, an automatic assessment effectively, rapidly, and accurately achieved the goal of data mining of specific compounds for targeted functions.

AI techniques, particularly LLMs, have become an unstoppable force for reshaping medical research [19,20]. Traditionally, LC-MS is a powerful analytical technique to identify and quantify active ingredients in traditional Chinese medicine (TCM) herbs. However, a rapid and accurate recognition of compounds with targeted functions, and a quantitative analysis of trace concentrations in complicated samples can be challenging [21]. We expected that an integration of LC-MS and LLMs would benefit TCM research, including the acceleration of active ingredient screen, precise targeting of interested compounds for certain diseases, and anchoring the promising candidates for developing new drugs. DeepSeek is an intelligent system based on a large-scale pre-trained language model, exhibiting strong capabilities in text understanding, knowledge reasoning, and cross-modal collaborative analysis, particularly excelling in processing information within Chinese-language contexts [22,23]. It enables rapid processing and analyzing massive volumes of both unstructured and structured data, thus digging biological insights out of complex omics datasets [24,25].

In the present study, we first created a four-step filtering workflow and quantitatively identified target compounds from CJ-ELNs by LC-MS. The cloud word analysis emphasized the term of acid among screened compounds enriched in CJ-ELNs. Acidic compounds derived from traditional Chinese herbals are established for the role of clearing heat and detoxifying [26]. Numerous studies have reported that acidic compounds in plants exert antioxidant, antibacterial, and anti-inflammatory effects through mechanisms such as scavenging free radicals, alleviating oxidative stress, modulating inflammatory factors, stimulating fibroblast proliferation, promoting collagen deposition, enhancing epithelialization, and inducing angiogenesis [27,28]. To achieve a precise data mining of compounds with relevant functions, DeepSeek models lent a hand that specifically screened compounds in CJ-ELNs with targeted functions of antioxidation, anti-inflammation, anticellular damage, antiapoptosis, wound healing and tissue regeneration, and cell proliferation. Finally, naturally active compounds in CJ-ELNs were resurfaced for their promising potentials in wound repair. For example, studies have shown that baicalin accelerates the wound healing process by downregulating the expression of pro-inflammatory cytokines (IL-6 and IL-1β) while upregulating the anti-inflammatory factor IL-10, and by promoting the secretion of various growth factors (VEGF, FGF-2, PDGF-β, and CTGF) [29]. The combination of LC-MS with DeepSeek paves the way to further analyses of therapeutic targets from traditional Chinese herbs for wound healing and tissue regeneration [30,31].

Limitations in this study should be noted. Firstly, bioactive compounds derived from CJ-ELNs were mined via LC-MS and a single LLM, namely, DeepSeek-R1. Other cutting-edge LLMs such as Claude, GPT-4 and Liama [32] can be further analyzed for the assistance of LC-MS in identifying interested compounds. Secondly, the 46 naturally active compounds derived from CJ-ELNs with targeted functions should be validated in in vivo and in vitro experiments. Lastly, the workflow we have established requires further validation on independent datasets. We shall address the aforementioned issues in subsequent work, including evaluating the efficacy of compounds through cell migration and transdermal tissue compatibility assays, verifying their efficacy via macroscopic imaging and H&E staining following animal wound modelling interventions, and validating potential pathways involved through Western blot and immunohistochemical analysis.

Conclusion

We innovatively designed a multimodal framework of LC-MS combined with DeepSeek models that rapidly and accurately identify naturally active compounds from CJ-ELNs. This AI-powered system innovatively integrates the traditional analytical technique with modern large language models, showing a huge potential in modern medicine and TCM research.

Acknowledgments

We would like to thank the National Center for Regional Technology Transfer and Commercialization in Biomedical Sciences for providing technical testing support and EVLIXIR for providing Cayratia japonica exosome-like nanovesicles (CJ-ELNs) samples. We are also truly grateful to Prof. Elvis Agbo for his comprehensive guidance and help in revising and polishing the manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (Grant No. 32460913), the Natural Science Foundation Project of Jiangxi Province (Grant No. 20232BAB205009), the Science and Technology Foundation of the Education Department of Jiangxi Province (Grant No. GJJ2201603), and the National Foreign Expert Projects (Y20240165).

Data Availability

The original data used for the current study are available upon reasonable request from the corresponding authors.

Authors' Contributions

Conceptualization: MXS, QF, QJY
Data curation: YPF, WJ
Formal analysis: YPF, WJ
Funding acquisition: MXS, QF, QJY
Investigation: WJ, YPF
Methodology: WJ, YPF
Project administration: MXS, QHY
Resources: MXS, QJY
Supervision: MXS, QF, QJY
Writing-original draft: WJ, YPF
Writing-review & editing: JX, JY, MXS, QF, QJY

Conflicts of Interest

None declared.

Multimedia Appendix 1
A word cloud of common compounds identified by liquid chromatography-mass spectrometry.
[PNG File, 283 KB - bioinform_v7i1e80539_app1.png]

Multimedia Appendix 2
Distribution of the classifications of compounds enriched in CJ-ELNs, distribution of functional compounds enriched in CJ-ELNs with targeted functions of wound healing and tissue regeneration, and distribution of compounds enriched in CJ-ELNs with all functional categories.
[TIF File, 1337 KB - bioinform_v7i1e80539_app2.tif]

Multimedia Appendix 3
Function of 46 compounds.
[XLSX File, 17 KB - bioinform_v7i1e80539_app3.xlsx]

References

1. Subha D, Harshnii K, Madhikiruba KG, Nandhini M, Tamilselvi KS. Plant derived exosome- like nanovesicles: an updated overview. Plant Nano Biology 2023 Feb;3:100022. [doi: 10.1016/j.plana.2022.100022]
2. Mu N, Li J, Zeng L, et al. Plant-derived exosome-like nanovesicles: current progress and prospects. Int J Nanomedicine 2023;18:4987-5009. [doi: 10.2147/IJN.S420748] [Medline: 37693885]
3. Dad HA, Gu TW, Zhu AQ, Huang LQ, Peng LH. Plant exosome-like nanovesicles: emerging therapeutics and drug delivery nanoplatforms. Mol Ther 2021 Jan;29(1):13-31. [doi: 10.1016/j.ymthe.2020.11.030]
4. Di Gioia S, Hossain MN, Conese M. Biological properties and therapeutic effects of plant-derived nanovesicles. Open Med 2020 Nov 21;15(1):1096-1122. [doi: 10.1515/med-2020-0160]
5. Lian MQ, Chng WH, Liang J, et al. Plant-derived extracellular vesicles: recent advancements and current challenges on their use for biomedical applications. J Extracell Vesicles 2022 Dec;11(12):e12283. [doi: 10.1002/jev2.12283] [Medline: 36519808]
6. Karamanidou T, Tsouknidas A. Plant-derived extracellular vesicles as therapeutic nanocarriers. Int J Mol Sci 2021 Dec 24;23(1):T-epublish. [doi: 10.3390/ijms23010191] [Medline: 35008617]
7. Sun J, Zhao P, Ding X, et al. Cayratia japonica prevents ulcerative colitis by promoting M2 macrophage polarization through blocking the TLR4/MAPK/NF-κB pathway. Mediators Inflamm 2022;2022:1108569. [doi: 10.1155/2022/1108569] [Medline: 36619207]
8. Zhao X, Dai R, Wang J, et al. Analysis of the permeable and retainable components of Cayratia japonica ointment through intact or broken skin after topical application by UPLC-Q-TOF-MS/MS combined with in vitro transdermal assay. J Pharm Biomed Anal 2024 Jan 20;238:115853. [doi: 10.1016/j.jpba.2023.115853] [Medline: 37976992]
9. Wolfender JL, Litaudon M, Touboul D, Queiroz EF. Innovative omics-based approaches for prioritisation and targeted isolation of natural products - new strategies for drug discovery. Nat Prod Rep 2019 Jun 19;36(6):855-868. [doi: 10.1039/c9np00004f] [Medline: 31073562]

10. Gros M, Petrović M, Barceló D. Development of a multi-residue analytical methodology based on liquid chromatography-tandem mass spectrometry (LC-MS/MS) for screening and trace level determination of pharmaceuticals in surface and wastewaters. Talanta 2006 Nov 15;70(4):678-690. [doi: 10.1016/j.talanta.2006.05.024] [Medline: 18970827]
11. Gika HG, Wilson ID, Theodoridis GA. LC–MS-based holistic metabolic profiling. Problems, limitations, advantages, and future perspectives. Journal of Chromatography B 2014 Sep;966:1-6. [doi: 10.1016/j.jchromb.2014.01.054]
12. Wang B, Xie Q, Pei J, et al. Pre-trained language models in biomedical domain: a systematic survey. ACM Comput Surv 2023 Oct 31;56:1-52. [doi: 10.1145/3611651]
13. Zhong W, Liu Y, Liu Y, et al. Performance of ChatGPT-4o and four open-source large language models in generating diagnoses based on China's rare disease catalog: comparative study. J Med Internet Res 2025;27:e69929-e69929. [doi: 10.2196/69929]
14. Liverpool S, Mota CP, Sales CMD, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317. [doi: 10.2196/16317] [Medline: 32442160]
15. Choudhury A, Shahsavar Y, Shamszare H. User intent to use Deepseek for health care purposes and their trust in the large language model: Multinational Survey Study. JMIR Hum Factors 2025 May 26;12:e72867. [doi: 10.2196/72867] [Medline: 40418796]
16. Tordjman M, Liu Z, Yuce M, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med 2025 Aug;31(8):2550-2555. [doi: 10.1038/s41591-025-03726-3] [Medline: 40267969]
17. W McGee R. Leveraging DeepSeek: an AI-powered exploration of traditional chinese medicine (Tai Chi and Qigong) for medical research. AJBSR 2025;25(5):645-654. [doi: 10.34297/AJBSR.2025.25.003362]
18. Alseekh S, Aharoni A, Brotman Y, et al. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. Nat Methods 2021 Jul;18(7):747-756. [doi: 10.1038/s41592-021-01197-1] [Medline: 34239102]
19. Haleem A, Javaid M, Khan IH. Current status and applications of artificial intelligence (AI) in medical field: an overview. Current Medicine Research and Practice 2019 Nov;9(6):231-237. [doi: 10.1016/j.cmrp.2019.11.005]
20. Tang X. The role of artificial intelligence in medical imaging research. BJR Open 2020;2(1):20190031. [doi: 10.1259/bjro.20190031] [Medline: 33178962]
21. Pang B, Zhu Y, Lu L, Gu F, Chen H. The applications and features of liquid chromatography‐mass spectrometry in the analysis of Traditional Chinese Medicine. Evid Based Complement Alternat Med 2016 Jan;2016(1). [doi: 10.1155/2016/3837270]
22. Du K, Li A, Zuo QH, et al. Comparing artificial intelligence-generated and clinician-created personalized self-management guidance for patients with knee osteoarthritis: blinded observational study. J Med Internet Res 2025 May 7;27:e67830. [doi: 10.2196/67830] [Medline: 40332991]
23. Huang T, et al. TCM-3ceval: a triaxial benchmark for assessing responses from large language models in traditional chinese medicine. arXiv. Preprint posted online on Mar 10, 2025. [doi: 10.48550/arXiv.2503.07041]
24. Li F, Chen J, Luo W, et al. DeepPGDB: a novel paradigm for AI-guided interactive plant genomic database. Bioinformatics. Preprint posted online on 2025. [doi: 10.1101/2025.06.01.657209]
25. Luo E, et al. Benchmarking AI scientists in Omics data-driven biological research. arXiv. Preprint posted online on May 13, 2025. [doi: 10.48550/arXiv.2505.08341]
26. Muluye RA, Bian Y, Alemu PN. Anti-inflammatory and antimicrobial effects of heat-clearing chinese herbs: a current review. J Tradit Complement Med 2014 Apr;4(2):93-98. [doi: 10.4103/2225-4110.126635]
27. Guan S, Ge D, Liu TQ, Ma XH, Cui ZF. Protocatechuic acid promotes cell proliferation and reduces basal apoptosis in cultured neural stem cells. Toxicol In Vitro 2009 Mar;23(2):201-208. [doi: 10.1016/j.tiv.2008.11.008] [Medline: 19095056]
28. Yang D, Moh S, Son D, et al. Gallic acid promotes wound healing in normal and hyperglucidic conditions. Molecules 2016;21(7):899. [doi: 10.3390/molecules21070899]
29. Kim E, Ham S, Jung BK, Park JW, Kim J, Lee JH. Effect of baicalin on wound healing in a mouse model of pressure ulcers. IJMS ;24(1):329. [doi: 10.3390/ijms24010329]
30. Zhao F, Li Q, Wang M, Xiong X. An AI agent-based system for retrieving compound information in Traditional Chinese Medicine. Information 2025;16(7):543. [doi: 10.3390/info16070543]
31. He J, et al. OpenTCM: a graphrag-empowered LLM-based system for traditional chinese medicine knowledge retrieval and diagnosis. arXiv. Preprint posted online on Apr 28, 2025. [doi: 10.48550/arXiv.2504.20118]
32. Jaleel A, Aziz U, Farid G, et al. Evaluating the potential and accuracy of ChatGPT-3.5 and 4.0 in Medical Licensing and In-Training Examinations: systematic review and meta-analysis. JMIR Med Educ 2025;11:e68070. [doi: 10.2196/68070]

Abbreviations

AI: artificial intelligence
ASR: Automated Semantic Recognition
CJ-ELN: *Cayratia japonica* exosome-like nanovesicle
ELN: exosome-like nanovesicles

ESI: electrospray ionization
LC-MS: liquid chromatography-mass spectrometry
LLM: large language model
MS: mass spectrometry
P-ELN: Plant-derived exosome-like nanovesicle
TCM: traditional Chinese medicine
UHPLC: ultra-high-performance liquid chromatography

Edited by Z Yue; submitted 12.Jul.2025; peer-reviewed by V Alevizos, X Kang; revised version received 14.Oct.2025; accepted 19.Oct.2025; published 08.Jan.2026.

Please cite as:
Fu Q, Ji W, Fan YP, Yao J, Song MX, Yan QJ
Systematic Mining of Bioactive Compounds for Wound Healing From Cayratia Japonica Exosome-Like Nanovesicles: A Workflow Combining LC-MS and DeepSeek Models
JMIR Bioinform Biotech 2026;7:e80539
URL: https://bioinform.jmir.org/2026/1/e80539
doi:10.2196/80539

Prevalence and Associated Risk Factors of Bovine Fasciolosis in Bahir Dar, Ethiopia: Cross-Sectional Study

Tesfaye Mesfin[1]; Theobesta Solomon[2]; Abraham Belete Temesgen[2]

[1]Department of Animal Sciences, College of Agriculture and Natural Resource, Debre Markos University, Debre Markos, Ethiopia

[2]Department of Veterinary Pathobiology, College of Veterinary Medicine and Animals Sciences, University of Gondar, P.O. Box 196, Central Gondar, Ethiopia

Corresponding Author:
Abraham Belete Temesgen
Department of Veterinary Pathobiology, College of Veterinary Medicine and Animals Sciences, University of Gondar, P.O. Box 196, Central Gondar, Ethiopia

Abstract

Background: Cattle are among the most important livestock resources in Ethiopia, contributing significantly to the agricultural economy and rural livelihoods. They provide meat, milk, hides, draft power for crop production, and serve as a major source of income for farmers. Despite their vital role, cattle productivity is often constrained by various diseases, particularly parasitic diseases. One of the most significant of these is bovine fasciolosis, a condition caused by ingestion of metacercariae of liver flukes belonging to the genus *Fasciola*.

Objective: This study aimed to assess the prevalence and associated risk factors of bovine fasciolosis in Bahir Dar, Ethiopia.

Methods: A cross-sectional study was conducted from November 2021 to April 2022. A total of 384 cattle were randomly selected from different locations within the study area. Animals of all age groups and both sexes were included. Fecal samples were collected directly from the rectum of each animal using clean, labeled containers. The samples were examined using standard coprological techniques, specifically the sedimentation method, to detect liver fluke eggs. All findings were recorded, and the data were analyzed using descriptive statistical methods.

Results: The overall prevalence of fasciolosis was 49.21% (n=189). Based on origin, Sebatamit had the most incidence at 61.84% (n=47), followed by Kebele 11 at 59.37% (n=57), Tikurit at 50% (n=59), and Latammba at 27.65% (n=26). Statistical analysis revealed significant disparities in occurrence among areas. Cattle in poor condition had the largest prevalence (n=80, 64%), followed by medium condition (n=85, 50%) and fat cattle (n=24, 26.96%). This variation was statistically significant. Age-group analysis revealed comparable prevalence rates, with young cattle at 50.38% (n=65), adults at 47.33% (n=71), and elderly cattle at 50.47% (n=53), with no significant differences found. There were no significant sex-related variations in prevalence, with males exhibiting a prevalence of 49.73% (n=93) and females 48.73% (n=96). Local cattle had a slightly higher prevalence (n=111, 51.62%) than crossbreeds (n=78, 46.15%), although the difference was not statistically significant (*P*=.29).

Conclusions: These findings underscore the need for targeted, location-specific control strategies and highlight the importance of improved nutritional and health management practices to reduce the burden of fasciolosis in cattle populations.

(JMIR Bioinform Biotech 2026;7:e81219) doi:10.2196/81219

KEYWORDS

Bahir Dar; bovine fasciolosis; fecal examination; prevalence; risk factors

Introduction

Ethiopia hosts one of the largest livestock populations in Africa, with an estimated 55.03 million cattle, 27.32 million sheep, and 28.16 million goats as of 2019. Cattle, in particular, play a central role in the country's agricultural economy, with the dairy sector contributing over 81% of total milk production. Despite this abundance, livestock productivity remains low due to constraints such as poor nutrition, inadequate management, and widespread infectious diseases. Among these, fasciolosis is one of the most impactful parasitic diseases, causing substantial economic losses through reduced growth, impaired fertility, decreased milk yield, and increased mortality [1].

Fasciolosis is caused by trematode parasites of the genus *Fasciola*, commonly known as liver flukes. Infection occurs when animals ingest metacercariae, the infective stage of the parasite, from contaminated pasture, water, or feed. Two species are primarily responsible: *Fasciola hepatica*, which predominates in temperate regions, and *Fasciola gigantica*, more common in tropical climates, including much of Africa and Ethiopia [2-4]. Transmission relies on the presence of aquatic snails, such as *Lymnaea truncatula* and *Lymnaea*

natalensis, which serve as intermediate hosts under suitable conditions like stagnant water and moist environments [5].

Infected cattle experience liver tissue damage due to migrating immature flukes, resulting in inflammation, bile duct obstruction, hepatocellular necrosis, and fibrosis. Clinical signs often include weight loss, jaundice, poor body condition, and reduced productivity. Severe infections compromise liver function, predispose animals to secondary infections, and may lead to significant morbidity and mortality [6]. Economically, fasciolosis causes extensive liver condemnation at slaughterhouses, higher veterinary costs, and financial losses for farmers and the meat industry [7].

Epidemiology of fasciolosis is influenced by host and environmental factors, including age, sex, breed, management practices, and ecological conditions. Older animals are often more affected due to cumulative exposure, while some breeds show variable susceptibility. Differences in grazing behavior and reproductive cycles may contribute to higher prevalence in females in some cases [8-10].

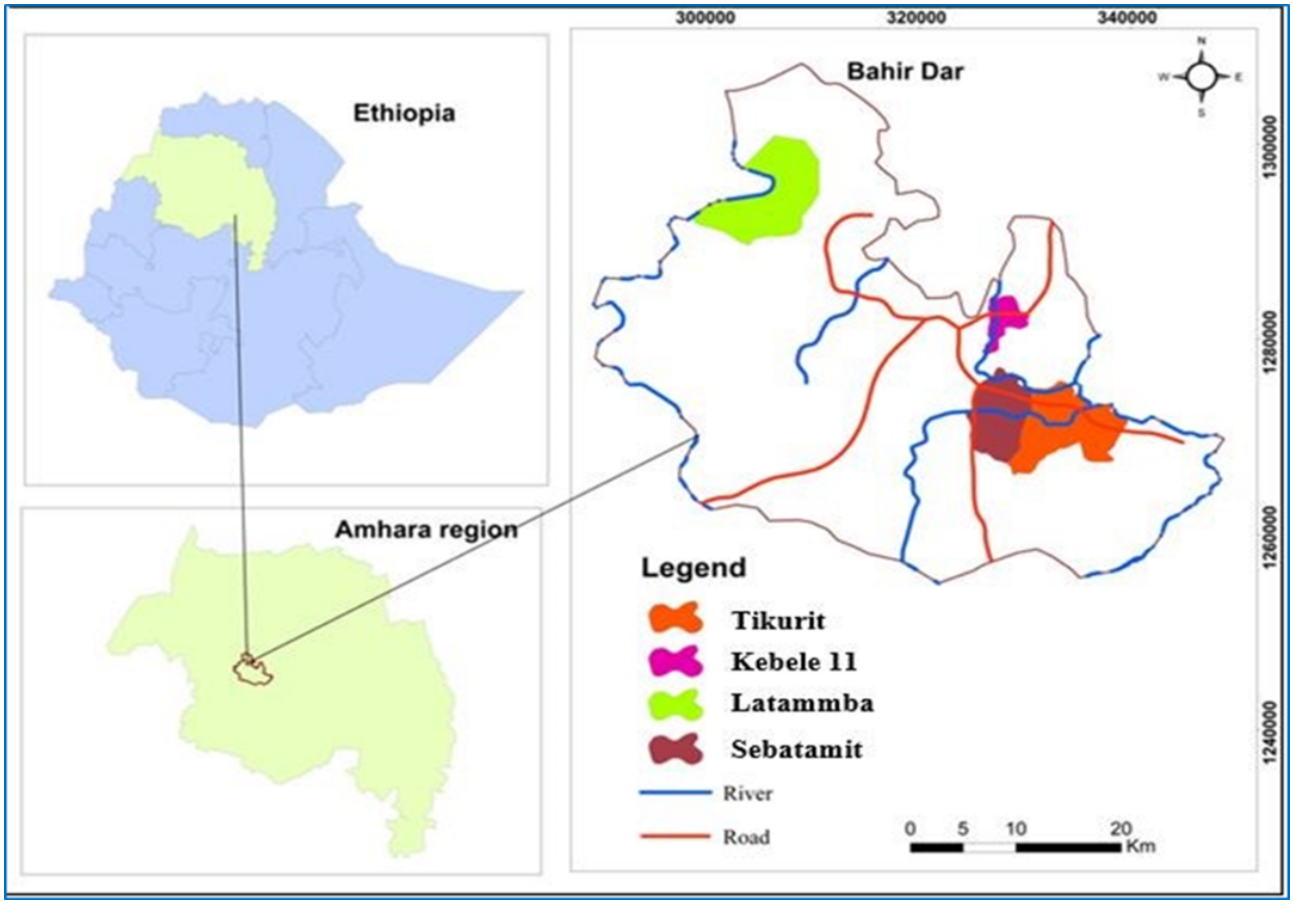
Additionally, pasture type, access to contaminated water, and use of anthelmintics play critical roles in transmission. Although previous studies have provided insights into fasciolosis in Ethiopia, many were localized, leaving gaps in regional prevalence and risk factors. Given the ecological and management diversity across the country, region-specific studies are essential to inform targeted control strategies. In particular, the central and northern highlands, where cattle production is economically significant, may present unique environmental conditions that affect parasite dynamics [11-14]. Therefore, this study aimed to determine the prevalence and associated risk factors of bovine fasciolosis in Bahir Dar, Ethiopia.

Methods

Study Area

The study was conducted in Bahir Dar, Ethiopia, from November 2021 to April 2022 (Figure 1). Bahir Dar is located approximately 575 km northwest of Addis Ababa, at an elevation of 1500 - 2600 meters above sea level, with geographic coordinates of 12°29’ N latitude and 37°29’ E longitude. The area receives an average annual rainfall of 1200 - 1600 mm, and temperatures range from 8 °C to 31 °C. The landscape is predominantly plain plateaus, covering roughly 70% of the region, and the vegetation includes shrub formations, low woodlands, evergreen areas, and semi-humid highland vegetation. Agriculture is a key livelihood, with major crops including teff (*Eragrostis tef*), wheat (*Triticum aestivum*), maize (*Zea mays*), and various pulses [1].

Figure 1. Map of the study area



Study Animal and Sampling Method

The study animals consisted of cattle from four selected sites in Bahir Dar: Kebele 11, Sebatamit, Tikurit, and Latammba. Both indigenous and crossbred Holstein Friesian cattle reared under local management conditions were included. A total of 384 cattle were randomly selected, representing both sexes and multiple age groups. Animal age was assessed based on dentition

and categorized as young or adult [15]. Body condition score was also evaluated for each animal following established guidelines to estimate nutritional and health status [16].

Study Design and Sample Size

A cross-sectional study was carried out in Bahir Dar, Ethiopia, from November 2021 to April 2022 to determine the prevalence and associated risk factors of bovine fasciolosis in Bahir Dar, Ethiopia. The risk factors considered included origin (location), breed, age, sex, and body condition of the cattle. The required sample size for fecal sample collection was calculated using a 95% confidence level, 5% absolute precision, and an expected prevalence of 50% in the absence of prior data for the study area [17].

$$n=Z2\times Pexp\times(1-Pexp)d2$$

where: n=required sample size; *P*exp=expected prevalence (0.5); *d*=desired absolute precision (0.05); and *Z*=Z-value for confidence level (1.96). Based on this formula, a total of 384 cattle were included in the study.

Fecal Examination and Identification of Fasciola Eggs

Fresh fecal samples were collected directly from the rectum of each animal using disposable gloves and placed into universal bottles containing 10% formalin as a preservative. Samples were transported under controlled conditions to the Bahir Dar Regional Veterinary Laboratory for parasitological examination. The sedimentation technique was used to detect Fasciola eggs, following standard procedures [18,19]. To differentiate Fasciola eggs from those of other trematodes, such as Paramphistomum species, the sediment was stained with methylene blue. *Fasciola* eggs are typically yellowish, large, and operculated, whereas Paramphistomum eggs stain blue [20,21].

Data Management and Analysis

The raw data collected from the study were organized and entered into a Microsoft Excel spreadsheet for initial management. Subsequently, the data were exported to STATA (version 16.0; StataCorp) for statistical analysis. A χ^2 test was used to evaluate the correlation between infection rates and risk factors such as age, sex, breed, and location. The test evaluated infection rates based on these parameters, with a significance level of $P<.05$.

Ethical Considerations

Ethical clearance was obtained from the Ethics Research Review Committee of the University of Gondar, College of Veterinary Medicine and Animal Sciences (Ref. No. CVMASc/UoG/RERC/10/11/2021; November 7, 2021). Cattle were handled according to animal welfare guidelines, and owner consent was obtained prior to data collection.

Results

Prevalence of Bovine Fasciolosis

A total of 384 cattle were examined in this study, and the overall prevalence of bovine fasciolosis in the study area was 49.21% (n=189). Analysis by origin showed the highest prevalence in Sebatamit (n=47, 61.84%), followed by Kebele 11 (n=57, 59.37%), Tikurit (50%), and Latammba (n=26, 27.65%), with a statistically significant difference among sites (χ^2=26.31, $P<.001$). Prevalence also varied according to body condition, with poor-conditioned cattle exhibiting the highest prevalence (n=80, 64%), followed by medium condition (n=85, 50%) and fat cattle (n=24, 26.96%), showing a significant difference (χ^2=28.6, $P<.001$). When grouped by age, the prevalence was 50.38% (n=65) in young cattle, 47.33% (n=71) in adults, and 50.47% (n=53) in older animals; however, differences were not statistically significant (χ^2=0.84, P=.35). Similarly, sex did not significantly influence prevalence: males had 49.73% (n=93), and females had 48.73% (n=96, χ^2=0.844; P=.35). Finally, breed showed no significant effect on infection rates, although local cattle had a slightly higher prevalence (n=111, 51.62%) compared to crossbreeds (n=78, 46.15%) (χ^2=0.287, P=.29) (Table 1).

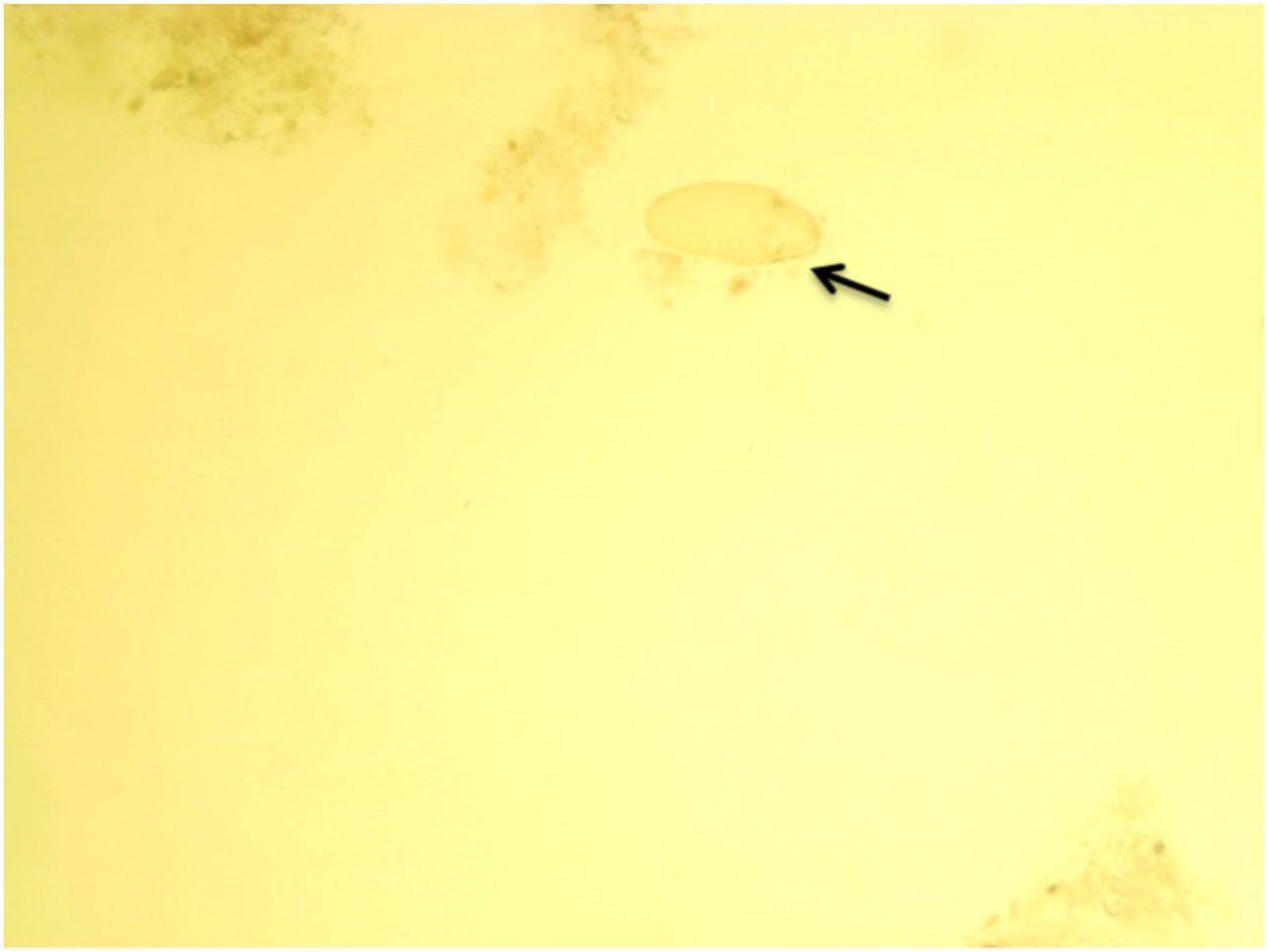
Table . Prevalence of bovine fasciolosis based on risk factors.

Risk factor and category	Examined (n=384)	Positive (n=189), %	Chi-square (*df*)	P value
Origin			26.31 (3)	<.001
Tikurit	118	59 (50)		
Sebatamit	76	47 (61.84)		
Latammba	94	26 (27.65)		
Kebele 11	96	57 (59.37)		
Body condition			28.60 (2)	<.001
Poor	125	80 (64)		
Medium	170	85 (50)		
Fat	89	24 (26.96)		
Age			0.839 (2)	.35
Young	129	65 (50.38)		
Adult	150	71 (47.33)		
Old	105	53 (50.47)		
Sex			0.844 (1)	.35
Male	187	93 (49.73)		
Female	197	96 (48.73)		
Breed			1.134 (1)	.29
Local	215	111 (51.62)		
Crossbreed	169	78 (46.15)		

Identification of Fasciola Eggs

The Fasciola eggs detected in cattle fecal samples are shown in Figure 2. Identification was based on morphological characteristics [20,21]. The eggs are ovoid, possess a thick yellowish-brown shell, and typically have an operculum at one end.

Figure 2. Egg of Fasciola species: yellowish egg (arrow).



Discussion

Principal Findings

Bovine fasciolosis remains one of the most significant parasitic diseases affecting cattle in Ethiopia, causing substantial economic losses through reduced growth, poor milk production, liver condemnation, and increased susceptibility to secondary infections. The present study recorded an overall prevalence of bovine fasciolosis of 49.21% in the study area, which aligns with previous reports from different regions of Ethiopia. For instance, prevalence rates of 41.41% and 54.5% have been reported in Woreta [22] and Jimma [23], respectively. Similar rates were observed in North-East Amhara (47.10%) [13], and Eastern Shoa, Kuyu District (54.2%) [19] likely associated with common agro-ecological factors such as the presence of Lymnaea snails and practices like communal water use and irrigation that facilitate pasture contamination [24]. Conversely, lower prevalence rates have been reported in some regions. Investigations in Soddo (4.9%) [25], Nekemte (15.9%) [7], and Southern Ethiopia (15.9%) [8] demonstrated lower frequencies, while Zenzelma, Bahir Dar (26%) [1], and Bahir Dar (32.3%) [26] reported slightly lower prevalence. These variations are likely influenced by differences in altitude, geography, climate, snail host abundance, management practices, and anthelmintic use [27].

Geographical and Management Factors

Prevalence varied significantly among study sites in the current work, with Sebatamit recording the highest prevalence (61.84%), followed by Kebele 11 (59.37%), Tikurit (50%), and Latammba (27.65%). This observation is consistent with previous studies indicating substantial geographical variation in fasciolosis prevalence [28-31], which may be influenced by climatic conditions, animal management practices, and access to veterinary services [32].

Influence of Animal Factors

Cattle in poor body condition exhibited the highest prevalence (64%), compared to medium (50%) and fat cattle (26.96%). This pattern corresponds with several Ethiopian reports that found significantly higher infection rates in animals with poor body condition, likely reflecting increased susceptibility due to malnutrition, compromised immunity, and concurrent infections that weaken host defenses [33-36]. However, some studies reported no significant differences in prevalence among body condition categories, indicating that body condition may not always predict *Fasciola* infection [23,37].

In this study, no significant differences were observed across age groups, with young cattle (50.38%), adults (47.33%), and older cattle (50.47%) showing similar prevalence. Several studies in Ethiopia have likewise reported no significant association between age and *Fasciola* infection [38,39], although

other reports documented age-related variation, likely reflecting differences in exposure or management practices [40,41].

Similarly, sex did not significantly influence prevalence, with males at 49.73% and females at 48.73%. This finding aligns with previous studies reporting no significant differences between male and female cattle [35], although one study did observe sex-related variation, possibly due to differences in management practices or environmental exposure [25].

Finally, breed had no significant effect on infection rates, with local cattle and crossbreeds showing prevalence of 51.62% and 46.15%, respectively. This finding agrees with earlier research [24,34], suggesting similar susceptibility between breeds, although a few studies reported breed-related differences, potentially attributable to genetic factors or breed-specific traits.

Limitations

The cross-sectional design and use of a single diagnostic method provide an accurate snapshot of prevalence and associated risk factors within the study population. While longitudinal or multimethod studies may provide additional detail, the current methodology is sufficient to address the study objectives and offers valuable insights for local disease control strategies.

Conclusion

This study revealed a high overall prevalence of bovine fasciolosis (49.21%) in the Bahir Dar area, confirming its significance as a major parasitic disease affecting cattle in the region. Prevalence varied notably across localities, with Sebatamit exhibiting the highest rate (61.84%) and Latammba the lowest (27.65%). Analysis of risk factors indicated a significant association between body condition and fasciolosis, with poorly conditioned cattle being more susceptible to infection. In contrast, no statistically significant differences were observed based on age, sex, or breed, suggesting that cattle across all demographic groups are at risk. These findings underscore the need for targeted, location-specific control strategies and highlight the importance of improved nutritional and health management practices to reduce the burden of fasciolosis in cattle populations.

Acknowledgments

The authors extend their sincere gratitude to all the dedicated staff members of the Bahir Dar Regional Veterinary Laboratory.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Funding

The authors declared no financial support was received for this work.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

TM: writing – review & editing, writing – original draft, resources, formal analysis, investigation, conceptualization, methodology, validation. TS: writing – review & editing, resources, validation, methodology, visualization. ABT: writing the original draft, review editing, funding acquisition, resources, methodology, conceptualization, formal analysis, supervision, and data curation

Conflicts of Interest

The authors declare that they have no competing interests.

References

1. Legesse S, Tsegaye S, Lamesgen S, Wolelaw Y, Garikipati D, Wondimagegn W. Coprological prevalence and associated risk factors of bovine fasciolosis in and around Zenzelma, Bahir Dar, Ethiopia. Eur J Exp Bio 2017;07(5):34. [doi: 10.21767/2248-9215.100034]
2. Ardo MB, Aliyara YH, Lawal H. Prevalence of bovine fasciolosis in major abattoirs of Adamawa State, Nigeria. Bayero Journal of Pure and Applied Sciences 2014;6:12-16. [doi: 10.4314/bajopas.v6i1.3]
3. Figtree M, Beaman MH, Lee R, et al. Fascioliasis in Australian travellers to Bali. Med J Aust 2015 Aug 17;203(4):186-188. [doi: 10.5694/mja15.00010] [Medline: 26268290]
4. Pfukenyi DM, Mukaratirwa S, Willingham AL, Monrad J. Epidemiological studies of Fasciola gigantica infections in cattle in the highveld and lowveld communal grazing areas of Zimbabwe. Onderstepoort J Vet Res 2006 Mar;73(1):37-51. [Medline: 16715877]
5. Mungube EO, Bauni SM, Tenhagen BA, Wamae LW, Nginyi JM, Mugambi JM. The prevalence and economic significance of Fasciola gigantica and Stilesia hepatica in slaughtered animals in the semi-arid coastal Kenya. Trop Anim Health Prod 2006;38(6):475-483. [doi: 10.1007/s11250-006-4394-4] [Medline: 17243475]

6. Mas-Coma S, Valero MA, Bargues MD. Chapter 2. Fasciola, lymnaeids and human fascioliasis, with a global overview on disease transmission, epidemiology, evolutionary genetics, molecular epidemiology and control. Adv Parasitol 2009;69:41-146. [doi: 10.1016/S0065-308X(09)69002-3] [Medline: 19622408]
7. Petros A, Kebede A, Wolde A. Prevalence and economic significance of bovine fasciolosis in cattle slaughtered at Nekemte Municipal Abattoir, Western Ethiopia. Journal of Veterinary Medicine and Animal Health 2013;5:202-205 [FREE Full text]
8. Asrese, NM, Ali MG. Bovine fasciolosis: prevalence and economic significance in Southern Ethiopia. Acta Parasitologica Globalis 2014;5:76-82 [FREE Full text]
9. Guteta M, Batu D. Prevalence of bovine fasciolosis and its financial loss at Gulliso Slaughter House, West Wallaga Zone Western Ethiopia. IJAAP 2022;2(21):15-24. [doi: 10.55529/ijaap.21.15.24]
10. Tulu D, Gebeyehu S. Study of prevalence and associated risk factors of bovine fasciolosis in Jimma Horro District of Kellem Wollega Zone, Western Ethiopia. Arch Vet Sci Med 2018;1:9-18. [doi: 10.26502/avsm.002]
11. Jilo SA, Abadura SZ, Adem ME, et al. Epidemiology of Fasciolosis in cattle in selected districts of bale zone, south eastern Ethiopia. Int J Vet Sci Anim Husbandry 2021 Jan 1;6(6):54-58 [FREE Full text] [doi: 10.22271/veterinary.2021.v6.i6a.471]
12. Bekele M, Tesfay H, Getachew Y. Bovine fasciolosis: prevalence and its economic loss due to liver condemnation at Adwa Municipal Abattoir, North Ethiopia. EJAST 2010;1:39-47 [FREE Full text]
13. Ayelign M, Alemneh T. Study on prevalence and economic importance of bovine fasciolosis in three districts of North-East Amhara Region, Ethiopia. Journal of Infectious & Non-Infectious Diseases 2017:1-5 [FREE Full text]
14. Tibbo M, Aragaw K, Teferi M, Haile A. Effect of strategic helminthosis control on mortality of communally grazed Menz lambs of smallholders in the cool central Ethiopian highlands. Small Rumin Res 2010 May;90(1-3):58-63. [doi: 10.1016/j.smallrumres.2010.01.002]
15. Mequaninit G, Mengesha A. Prevalence of bovine fasciolosis and associated economic loss in cattle slaughtered at Kombolcha Industrial Abbatior. J Vet Med Animal Sci 2021;4:1-8 [FREE Full text]
16. Yeneneh A, Kebede H, Fentahun T, Chanie M. Prevalence of cattle flukes infection at Andassa Livestock Research Center in north-west of Ethiopia. Vet Res Forum 2012;3(2):85-89 [FREE Full text] [Medline: 25653752]
17. Thrusfield M, Christley R. Veterinary Epidemiology: John Wiley and Sons; 2018. URL: https://www.wiley-vch.de/en/areas-interest/medicine-health-care/veterinary-epidemiology-978-1-118-28028-7 [accessed 2026-03-12]
18. Garg R, Yadav CL, Kumar RR, Banerjee PS, Vatsya S, Godara R. The epidemiology of fasciolosis in ruminants in different geo-climatic regions of north India. Trop Anim Health Prod 2009 Dec;41(8):1695-1700. [doi: 10.1007/s11250-009-9367-y] [Medline: 19455400]
19. Ayele Y, wondmnew F, Tarekegn Y. The prevalence of bovine and ovine fasciolosis and the associated economic loss due to liver condemnation in and around Debire Birhan, Ethiopia. SOJI 2018;6(3):1-11. [doi: 10.15226/2372-0948/6/3/00177]
20. Khan MK, Sajid MS, Khan MN, Iqbal Z, Iqbal MU. Bovine fasciolosis: prevalence, effects of treatment on productivity and cost benefit analysis in five districts of Punjab, Pakistan. Res Vet Sci 2009 Aug;87(1):70-75. [doi: 10.1016/j.rvsc.2008.12.013] [Medline: 19181352]
21. Amer S, Dar Y, Ichikawa M, et al. Identification of Fasciola species isolated from Egypt based on sequence analysis of genomic (ITS1 and ITS2) and mitochondrial (NDI and COI) gene markers. Parasitol Int 2011 Jan;60:5-12. [doi: 10.1016/j.parint.2010.09.003] [Medline: 20888427]
22. Tsegaye B, Abebaw H, Girma S. Study on coprological prevalence of bovine fasciolosis in and around Woreta, Northwestern Ethiopia. J Vet Med Anim Health 2012:89-92 [FREE Full text]
23. Tolosa T, Tigre W. The prevalence and economic significance of bovine fasciolosis at Jimma abattoir, Ethiopia. Internet J Vet Med 2007;3:1-7 [FREE Full text]
24. Yokananth S, Ghosh S, Gupta SC, Suresh MG, Saravanan D. Characterization of specific and cross-reacting antigens of Fasciola gigantica by immunoblotting. Parasitol Res 2005;97:41-48. [doi: 10.1007/s00436-005-1371-1] [Medline: 15952043]
25. Abunna F, Asfaw L, Megersa B, Regassa A. Bovine fasciolosis: coprological, abattoir survey and its economic impact due to liver condemnation at Soddo municipal abattoir, Southern Ethiopia. Trop Anim Health Prod 2010;42:289-292. [doi: 10.1007/s11250-009-9419-3] [Medline: 19680772]
26. Gebrie Y, Gebreyohannes M, Tesfaye A. Prevalence of bovine fasciolosis in and around Bahir Dar, Northwest Ethiopia: J Parasitol Vector Biol; 2015:74-79 URL: https://www.researchgate.net/publication/313652158 [accessed 2022-03-15]
27. Mulugeta S, Begna F, Tsegaye E. Prevalence of bovine fasciolosis and its economic significance in and around Assela, Ethiopia. Global Journal of Medical Research 2011;11:1-9 [FREE Full text]
28. Yilma JM, Mesfin A. Dry season bovine fasciolosis in northwestern part of Ethiopia. Rev Med Vet 2000;151:493-500 [FREE Full text]
29. Ibrahim N, Wasihun P, Tolosa T. Prevalence of bovine fasciolosis and its economic importance due to liver condemnation at Kombolcha Industrial Abattoir, Ethiopia. Internet J Vet Med 2010;8:1-7 [FREE Full text]
30. Mensur S, Ansuar I, Tesfaye A, Abdulkaf K, Ahmed Y. Small ruminant fasciolosis and its economic impact in an export abattoir of Ethiopia. Livest Res Rural Dev 2016 [FREE Full text]
31. Birhan M, Demewez G, Tewodros F, Tadegenge M. Prevalence and economic significance of bovine fasciolosis in cattle slaughtered at Debre. Online Journal of Animal and Feed Research 2019;9:256-259. [doi: 10.36380/scil.2019.ojafr35]

32. Nicholson MJ, Butterworth MH. A Guide to Condition Scoring of Zebu Cattle: Int Livest Cent Afr; 1986. URL: https://books.google.com/books?id=8rhIDD8XRBkC
33. Coles GC. The future of veterinary parasitology. Vet Parasitol (Amst) 2001:31-39. [doi: 10.1016/S0304-4017(01)00421-6]
34. Keyyu JD, Monrad J, Kyvsgaard NC, Kassuku AA. Epidemiology of Fasciola gigantica and amphistomes in cattle on traditional, small-scale dairy and large-scale dairy farms in the southern highlands of Tanzania. Trop Anim Health Prod 2005;37:303-314. [doi: 10.1007/s11250-005-5688-7] [Medline: 15934638]
35. Zeryehun T. Helminthosis of sheep and goats in and around Haramaya, Southeastern Ethiopia. J Vet Med Anim Health 2012;4:48-55 [FREE Full text]
36. Aregay F, Bekele J, Ferede Y, Hailemelekot M. Study on the prevalence of bovine fasciolosis in and around Bahir Dar, Ethiopia. Ethiop Vet J 2013;17:1. [doi: 10.4314/evj.v17i1]
37. Getahun A, Aynalem Y, Haile A. Prevalence of bovine fasciolosis infection in Hossana municipal abattoir, Southern Ethiopia. J Nat Sci Res 2017;7:65-70 [FREE Full text]
38. Maingi N, Otieno RO, Weda EH, Gichohi VM. Effects of three anthelmintic treatment regimes against Fasciola and nematodes on the performance of ewes and lambs on pasture in the highlands of Kenya. Vet Res Commun 2002;26:543-552. [doi: 10.1023/a:1020291531858] [Medline: 12416869]
39. Ahmad-Najib M, Wan-Nor-Amilah WAW, Kin WW, Arizam MF, Noor-Izani NJ. Prevalence and risk factors of bovine fascioliasis in Kelantan, Malaysia: a cross-sectional study. Trop Life Sci Res 2021;32:1-14. [doi: 10.21315/tlsr2021.32.2.1] [Medline: 34367511]
40. Kifle D, Hiko A. Abattoir survey on the prevalence and monetary loss associated with fasciolosis in sheep and goats. International Journal of Livestock Production 2011;2:138-141 [FREE Full text]
41. Opio LG, Abdelfattah EM, Terry J, Odongo S, Okello E. Prevalence of fascioliasis and associated economic losses in cattle slaughtered at Lira Municipality Abattoir in Northern Uganda. Animals (Basel) 2021;11:681. [doi: 10.3390/ani11030681] [Medline: 33806313]

Edited by A Uzun; submitted 24.Jul.2025; peer-reviewed by DM Okello, D Tamir; accepted 25.Nov.2025; published 17.Mar.2026.

Please cite as:
Mesfin T, Solomon T, Temesgen AB
Prevalence and Associated Risk Factors of Bovine Fasciolosis in Bahir Dar, Ethiopia: Cross-Sectional Study
JMIR Bioinform Biotech 2026;7:e81219
URL: https://bioinform.jmir.org/2026/1/e81219
doi:10.2196/81219

Development and Validation of a Generative Artificial Intelligence-Based Pipeline for Automated Clinical Data Extraction From Electronic Health Records: Technical Implementation Study

Marvin N Carlisle[1], BS; William A Pace[1], BS; Andrew W Liu[1,2], BA; Robert Krumm[1], BA; Janet E Cowan[1], MA; Peter R Carroll[1,3], MD, MPH; Matthew R Cooperberg[1,3,4], MD, MPH; Anobel Y Odisho[1,3,4,5], MD, MPH

[1]Department of Urology, University of California, San Francisco, 550 16th Street, Box 1695, San Francisco, CA, United States

[2]Chan Medical School, University of Massachusetts, Worcester, MA, United States

[3]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, United States

[4]Department of Epidemiology and Biostatistics, School of Medicine, University of California, San Francisco, San Francisco, CA, United States

[5]Department of Medicine, Division of Clinical Informatics and Transformation, School of Medicine, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Anobel Y Odisho, MD, MPH
Department of Urology, University of California, San Francisco, 550 16th Street, Box 1695, San Francisco, CA, United States

Abstract

Background: The manual abstraction of unstructured clinical data is often necessary for granular clinical outcomes research but is time consuming and can be of variable quality. Large language models (LLMs) show promise in medical data extraction yet integrating them into research workflows remains challenging and poorly described.

Objective: This study aimed to develop and integrate an LLM-based system for automated data extraction from unstructured electronic health record (EHR) text reports within an established clinical outcomes database.

Methods: We implemented a generative artificial intelligence pipeline (UODBLLM) utilizing a flexible language model interface that supports various LLM implementations, including Health Insurance Portability and Accountability Act-compliant cloud services and local open-source models. We used extensible markup language (XML)-structured prompts and integrated using an open database connectivity interface to generate structured data from clinical documentation in the EHR. We evaluated the UODBLLM's performance on the completion rate, processing time, and extraction capabilities across multiple clinical data elements, including quantitative measurements, categorical assessments, and anatomical descriptions, using sample magnetic resonance imaging (MRI) reports as test cases. System reliability was tested across multiple batches to assess scalability and consistency.

Results: Piloted against MRI reports, UODBLLM processed 1800 clinical documents with a 100% completion rate and an average processing time of 8.90 seconds per report. The token utilization averaged 2692 tokens per report, with an input-to-output ratio of approximately 13:2, resulting in a processing cost of US $0.009 per report. UODBLLM had consistent performance across 18 batches of 100 reports each and completed all processing in 4.45 hours. From each report, UODBLLM extracted 16 structured clinical elements, including prostate volume, prostate-specific antigen values, Prostate Imaging Reporting and Data System scores, clinical staging, and anatomical assessments. All extracted data were automatically validated against predefined schemas and stored in standardized JSON format.

Conclusions: We demonstrated the successful integration of an LLM-based extraction system within an existing clinical outcomes database, achieving rapid, comprehensive data extraction at minimal cost. UODBLLM provides a scalable, efficient solution for automating clinical data extraction while maintaining protected health information security. This approach could significantly accelerate research timelines and expand feasible clinical studies, particularly for large-scale database projects.

(JMIR Bioinform Biotech 2026;7:e70708) doi:10.2196/70708

KEYWORDS

generative artificial intelligence; artificial intelligence large language model; GPT-4; chatbot; pattern analysis; prostate cancer; kidney cancer

Introduction

Background

Electronic health record (EHR) systems contain extensive health data, but much of it is in unstructured notes such as radiology and pathology reports, making it hard to access for large-scale research. Granular clinical outcomes research often requires laborious manual chart review. The automation of this process requires significant investment, and algorithm performance varies with report parameters and automation type [1,2]. Previous attempts to automate this process have tried natural language processing on prostate cancer pathology reports, reporting a weighted F_1 score and accuracy as high as 0.97% and 93%, respectively [3].

Large language models (LLMs) represent a new opportunity for addressing this problem. LLMs are generative artificial intelligence programs capable of drafting human-like responses to specific queries. In oncological contexts, LLM applications can create medical notes, aggregate imaging findings, extract operative note data, and identify presenting symptoms [4-7]. Previous studies analyzing the overall data extraction capabilities have found accuracies ranging from 63.9% to 100% in retrieving data elements [5,8-13]. Specifically, several LLM models have also been developed to extract medical information from text, including early-stage LLM trained on medical encyclopedias and radiology datasets to read annotated radiology reports (71.6% accuracy) and inferring cancer disease response based on computed tomography reports (89% accuracy) [14,15]. Some of these groups also implemented or hypothesized implementing their systems into medical research pipelines for expediting data extraction [3,8]. Another group applied a customized, open-source LLM trained on medical data to read magnetic resonance imaging (MRI) reports with a sensitivity of 96% and specificity of 99%. In terms of data extraction, generative pre-trained transformer (GPT)-4 has been shown to extract hepatocellular carcinoma data from MRI reports with an overall accuracy of 93.4% [16]. LLMs have also proven to be flexible and frequently outperform traditional automated models, suggesting that powerful LLMs might be ready to support research endeavors via the extraction of unstructured data [5,8,17]. Implementing LLMs into practical, applicable tools remains challenging, and some private organizations have attempted to improve clinical data extraction through EHR integration [18]. Despite this, most efforts, such as the American Urological Association Quality Registry, remain dependent on manual data management, partially due to difficulty integrating new tools into existing workflows. While some larger institutions have begun implementing automated data extraction pipelines, traditional methods of data extraction require considerable technical expertise and resources to initiate, making these methods inaccessible for most institutions.

The University of California, San Francisco (UCSF) Department of Urology maintains the Urologic Outcomes Database (UODB) for prostate, bladder, and renal cancers [19]. The UODB is an SQL-based clinical data research database that holds structured manually abstracted clinical data for patients treated at the UCSF, including 7000 patients with prostate cancer over 20 years. Due to limited manual abstraction capacity and increasing patient volume, clinical events and data entry often lag. Previous in-house attempts to automate this process using traditional natural language processing solutions proved to be time-consuming to develop and maintain [1-3,20]. The aim of this study was to demonstrate a practical use of LLMs in academic clinical research by describing the successful implementation of a secure, baseline, institutional version of GPT-4 within the UODB to quickly and easily extract unstructured data and effectively reduce manual labor in gathering data from medical reports.

Related Work

Previous studies by our group have utilized UCSF's Versa, an internal, secure, Health Insurance Portability and Accountability Act (HIPAA)-compliant deployment of OpenAI's GPT models (OpenAI Inc.) that includes an application programming interface (API) for query automation [17,21]. We demonstrated that systems based on the Versa GPT-4 API can accurately extract structured data from real-world clinical reports. In one study involving 424 prostate MRI reports, our pipeline, using zero-shot prompting, achieved an overall median field-level accuracy of 98.1% (IQR 96.3% - 99.2%), with key elements such as prostate-specific antigen density (98.3%), extracapsular extension (97.4%), and TNM staging (98.1%) [21]. In a separate effort with 228 prostate MRI reports, the approach achieved similarly high concordance (over 95%) when compared with manual abstraction [17].

These validation efforts serve to confirm the accuracy of the underlying extraction prompts and Versa GPT-4 API performance. The focus of the current work, therefore, is not on additional accuracy testing; rather, we build upon this foundation to present a modular, scalable implementation pipeline that operationalizes LLM-driven extraction at scale, within a secure, clinical-grade environment.

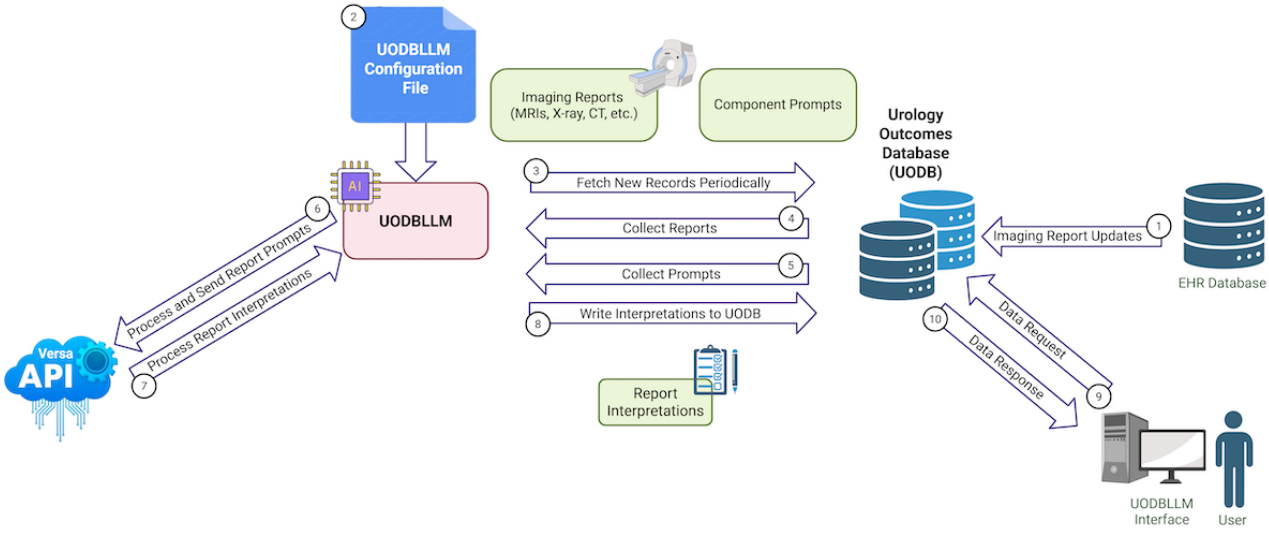
Methods

Overall Design

This study presents the implementation and performance evaluation of UODBLLM, a modular LLM-based pipeline designed for structured data extraction from a wide range of unstructured clinical reports. For this technical implementation, the system was evaluated using free-text prostate MRI radiology reports as the primary use case (Figure 1). The system was deployed within a secure, HIPAA-compliant clinical environment using the internal UCSF Versa GPT-4 API, ensuring that protected health information (PHI) remained confined to institutional systems. UODBLLM was designed with a flexible architecture to support multiple language models and API endpoints, enabling adaptability across varied clinical settings.

Prompts are stored as configurable components in dedicated database tables, allowing users to dynamically pair extraction templates with report sets without modifying the underlying code. This design supports rapid iteration, version control, and seamless adaptation to evolving information extraction needs.

Figure 1. System design and data flow of the UODBLLM application**.** The process begins with an initial connection between the electronic health record (EHR) and the Urologic Outcomes Database (UODB) for imaging report updates (1). The UODBLLM application is governed by a configuration file defining its core parameters (2). The application periodically fetches new records from the UODB (3), collects the relevant reports (4) and component prompts (5), and sends these to the Versa application programming interface (API) for processing (6). The API returns structured interpretations of the reports (7), which are then written back into the UODB (8). A user, via the UODBLLM interface, can send a data request to the UODB (9) and receive a data response for review and analysis (10).



Study Population

The study dataset comprised 1800 prostate MRI radiology reports retrieved from the institutional EHR system. Reports were selected based on procedural coding and metadata filters to ensure relevance to downstream urologic data extraction.

Intervention

UODBLLM is a Python-based (version 3.9.6, Python Software Foundation, worldwide) application designed to extract structured information from clinical reports using a modular, API-driven architecture. Source text is retrieved from the UODB using a parameterized SQL query passed via a secure Open Database Connectivity connection. Text blocks are staged and dispatched in configurable batches, controlled by a modifiable parameter specified in a configuration file or modifiable via command-line flag.

The pipeline retrieves a version-controlled extensible markup language (XML)-based prompt template at runtime using a parameterized SQL query from the UODB. This template specifies the role, task, JSON response schema, and a structured sub-prompt with 16 XML elements that each represent a clinical field of interest (eg, prostate volume, prostate-specific antigen density, and overall Prostate Imaging Reporting and Data System score), each with plain-language extraction instructions (Figure S1 in Multimedia Appendix 1). For every report, the program inserts the full free-text report into the template's designated placeholder, producing a complete prompt that is then submitted to the Versa GPT-4 model. Embedding the report within a constant, schema-constrained envelope ensures that returned JSON follows a predictable structure, enabling reliable downstream parsing and storage.

Each batch is passed to a thin wrapper around the Versa GPT-4 API. Requests are streamed to the API endpoint; results are captured, parsed, and validated against the predefined JSON schema. Error handling includes up to 5 retry attempts per request with exponential back-off (2^n seconds, capped at 30 seconds). Failed requests are logged, and the affected reports are re-queued for later processing. Element-level completeness is defined as the proportion of reports for which the pipeline returned a non-null value.

Extracted fields are transmitted back to the database using a set of parameterized SQL UPDATE statements mapped to internal column identifiers. A custom statistics tracking module records token usage, response latency, and processing cost per report by counting model-specific numerical tokens generated from text via Byte Pair Encoding. System-wide throughput and error frequency are also recorded. The pipeline was executed on a 2019 MacBook Pro (Intel Core i9, 2.4 GHz, 64 GB RAM, macOS Ventura 13.2.1). The system's computational workload is lightweight and not hardware dependent, making it executable on a standard consumer laptop. The source code will be made available to investigators for non-commercial purposed upon request.

Ethical Considerations

The study was approved by the University of California, San Francisco Institutional Review Board (IRB #11-05329), and the requirement for informed consent was waived. The system was deployed within a secure, HIPAA-compliant clinical environment using the internal UCSF Versa GPT-4 API, ensuring that PHI remained confined to institutional systems. All reports were de-identified prior to processing.

Results

Processing Performance and Resource Utilization

The analysis of system logs demonstrated consistent performance metrics, with an average processing speed of 8.90 seconds per report across 1800 reports. UODBLLM maintained 100% completion rates across all test runs, with batch sizes of 100 reports. Token utilization, representing the count of model-specific numerical tokens generated from the input and output text via Byte Pair Encoding (calculated using the tiktoken library), averaged 2692 tokens per report. Given the model's context window capacity relative to typical report lengths, specific token optimization techniques like input text chunking were not required for this implementation. This resulted in an input-to-output ratio of approximately 13:2 (4,196,697 input tokens, 648,723 output tokens), resulting in an average processing cost of US $0.009 per report. The total processing run successfully analyzed all 1800 test reports in 4.45 hours, showing sustained performance at scale.

Prior Validation

Although the present study did not re-evaluate extraction accuracy on this corpus, the underlying extraction logic and prompt structure have been previously validated in two independent studies by our group. In one effort involving 424 prostate MRI reports, the system achieved a median field-level accuracy of 98.1% (IQR 96.3% - 99.2%) for key clinical variables [21]. A subsequent study with 228 MRI reports demonstrated similarly high extraction fidelity, with all structured elements exceeding 95% accuracy [17]. These findings confirm the robustness of the prompt design and model configuration across settings, supporting their reliability in the context of the current implementation.

Experience

Researchers interact with UODBLLM by selecting the clinical report category (eg, MRI reports or pathology reports) through a secure web-based application that integrates with the UODB and is accessible only through local institutional network connections. UODBLLM displays quantitative processing metrics for the selected report type, including extraction completion timestamps, LLM prompts, and performance statistics from previous analyses. This longitudinal view enables investigators to evaluate existing structured data's temporal relevance and completeness before proceeding with additional processing.

Researchers can use previously extracted structured data or initiate a new extraction cycle with refined extraction parameters. When opting for new extraction, investigators can specify temporal bounds for report inclusion and modify extraction prompts stored in the database tables. This parameterization enables the analysis of specific clinical cohorts while ensuring consistent extraction methodology across research protocols.

Upon initiating the UODBLLM process, the system executes batch processing of identified reports, with real-time logging providing visibility into extraction progress. Researchers can monitor the system performance through logs that track processing times, success rates, and any encountered exceptions. The structured JSON output is automatically integrated into the UODB, enabling immediate access for researchers.

Quality assurance is implemented through a review interface where researchers can perform comparative analysis of extracted data elements against source reports and any pre-existing manually abstracted data with the opportunity to iteratively refine prompts. Successfully processed reports are flagged in the database, preventing duplicate processing while maintaining a comprehensive audit trail of all data extraction operations.

Discussion

Principal Findings and Comparison With Previous Works

In this study, we developed and validated an automated LLM-based integration for UODB management that achieved a 100% completion rate across 1800 clinical documents, with an average processing time of 8.90 seconds per report. The UODBLLM demonstrates an implementation of a PHI-secure, LLM-agnostic system for automated data extraction from urological outcomes documentation. By leveraging institutional cloud infrastructure and established database architecture, we created a scalable solution that significantly reduces the manual effort traditionally required for data extraction while maintaining high accuracy rates [19]. This advancement represents a crucial step toward efficient, accurate, and comprehensive research database management [18].

The integration of generative artificial intelligence in clinical data management has seen rapid evolution, with several institutions developing specialized approaches for extracting structured data from clinical documentation [1,2]. While the validation of a local GPT model showed promising accuracy in the low 90th percentile for biomedical data collection, their focus on chromatin expression in cell lines addressed a more constrained data domain [20]. UODBLLM demonstrates comparable accuracy rates with the ability for researcher customization. Recent oncology initiatives using LLMs for clinical note evaluation have shown potential, but our approach differs by providing a complete pipeline that not only extracts data but also integrates directly with existing database infrastructure [5,6]. The problem of integration from clinical care to research database is common in clinical trials, clinical record management, and safety reports, encouraging other groups to design automated data capture and transfer pipelines. These pipelines have historically been evaluated as successful by the variables they extract, efficiency gained, and interoperability they provide, aligning with our key performance indicators [22,23]. The pipeline here described and designed has been estimated to improve data extraction manual time efficiency by as much as 90% if pulling multiple variables from hundreds of reports, although this enhancement varies based on report type, variable, and iterations of prompt refinement.

The technical robustness of our approach is supported by key design decisions and validated through comprehensive testing. Our choice to leverage a PHI-secure institutional version of GPT-4 addresses performance and privacy requirements, crucial

considerations for clinical data management [5]. The system's integration within the UODB piggybacks off a validated foundation for data structure and management [19]. Our validation protocol included processing reports across various batch sizes, achieving consistent performance and reliable operation at scale. The ability of the UODBLLM to efficiently process clinical documentation while maintaining high accuracy suggests the potential for significant resource optimization in research operations [6]. These efficiency gains could dramatically accelerate research timelines and expand the scope of feasible clinical studies.

Although this study did not re-assess extraction accuracy, this was a deliberate design choice. The extraction framework employed here has already undergone validation in prior work, with element-level accuracies exceeding 95% across multiple prostate MRI cohorts [17,21]. In contrast, our current objective was to evaluate the system-level performance of a scalable, generalizable implementation pipeline deployed within a secure clinical environment. Notably, the architecture is model-agnostic and allows for future integration of various LLMs or prompt schemas. This decoupling of model validation from pipeline implementation facilitates adaptability while building on established, validated components.

The limitations of our approach warrant careful consideration. While UODBLLM performs robustly for current use cases, the accuracy of LLM-based data extraction still requires human validation for critical data points, a challenge noted across multiple studies [4,5,8]. The evolving nature of clinical research means that prompt engineering must continually adapt to new data types and research questions. Additionally, while our pipeline is LLM-agnostic, our specific performance results were achieved using a PHI-secure version of GPT-4, and performance may vary with different models or implementations. While this implementation focused on prostate MRI reports, the UODBLLM pipeline was designed for broad applicability across diverse clinical documents. This generalizability is enabled by its modular, model-agnostic architecture and a flexible prompting system where extraction templates are stored as configurable components in the database. The design allows the pipeline to be readily adapted for other unstructured texts, such as pathology results or operative notes, which aligns with plans to expand its use to other urologic cancers.

Conclusions

Our study demonstrates the feasibility and effectiveness of integrating LLM-based automation into UODB management. Our system's perfect completion rate, rapid processing speed, and cost-effective operation provides a robust framework for modernizing clinical research data management. Looking ahead, we aim to develop protocols for using LLMs to validate existing data entries and expanding to renal and bladder cancer radiology and pathology texts. The potential benefits of increased research efficiency and data quality suggest that LLM-based approaches will play an increasingly important role in clinical research infrastructure [4]. These advances may ultimately accelerate the pace of discovery in clinical oncology and serve as a model for other medical specialties.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of the UODBLLM Data Extraction Workflow. (A) The original unstructured text from a sample magnetic resonance imaging report. (B) The corresponding extensible markup language-structured prompt containing instructions and specific data extraction queries sent to the large language model (LLM). (C) The structured JSON data returned by the LLM based on the prompt and report.
[DOCX File, 9 KB - bioinform_v7i1e70708_app1.docx]

References

1. Park B, Altieri N, DeNero J, Odisho AY, Yu B. Improving natural language information extraction from cancer pathology reports using transfer learning and zero-shot string similarity. JAMIA Open 2021 Jul;4(3):ooab085. [doi: 10.1093/jamiaopen/ooab085] [Medline: 34604711]
2. Odisho AY, Bridge M, Webb M, et al. Automating the capture of structured pathology data for prostate cancer clinical care and research. JCO Clin Cancer Inform 2019 Jul;3(3):1-8. [doi: 10.1200/CCI.18.00084] [Medline: 31314550]
3. Odisho AY, Park B, Altieri N, et al. Natural language processing systems for pathology parsing in limited data environments with uncertainty estimation. JAMIA Open 2020 Oct;3(3):431-438. [doi: 10.1093/jamiaopen/ooaa029] [Medline: 33381748]
4. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. J Multidiscip Healthc 2023;16:1513-1520. [doi: 10.2147/JMDH.S413470] [Medline: 37274428]
5. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med 2024 May 1;7(1):106. [doi: 10.1038/s41746-024-01079-8] [Medline: 38693429]
6. Hsueh JY, Nethala D, Singh S, et al. Exploring the feasibility of GPT-4 as a data extraction tool for renal surgery operative notes. Urol Pract 2024 Sep;11(5):782-789. [doi: 10.1097/UPJ.0000000000000599] [Medline: 38913566]
7. Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 multimodal performance in radiological image analysis. Eur Radiol 2025 Apr;35(4):1959-1965. [doi: 10.1007/s00330-024-11035-5] [Medline: 39214893]

8. Truhn D, Loeffler CM, Müller-Franzes G, et al. Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4). J Pathol 2024 Mar;262(3):310-319. [doi: 10.1002/path.6232] [Medline: 38098169]
9. Lehnen NC, Dorn F, Wiest IC, et al. Data extraction from free-text reports on mechanical thrombectomy in acute ischemic stroke using ChatGPT: a retrospective analysis. Radiology 2024 Apr;311(1):e232741. [doi: 10.1148/radiol.232741] [Medline: 38625006]
10. Siepmann RM, Baldini G, Schmidt CS, et al. An automated information extraction model for unstructured discharge letters using large language models and GPT-4. Healthcare Analytics 2025 Jun;7:100378. [doi: 10.1016/j.health.2024.100378]
11. Verma A, Verma P, editors. Research Advances in Intelligent Computing: Volume 2: CRC Press; 2024. [doi: 10.1201/9781003433941]
12. Shah-Mohammadi F, Finkelstein J. Extraction of substance use information from clinical notes: generative pretrained transformer-based investigation. JMIR Med Inform 2024 Aug 19;12:e56243. [doi: 10.2196/56243] [Medline: 39037700]
13. Chiang CC, Luo M, Dumkrieger G, et al. A large language model-based generative natural language processing framework finetuned on clinical notes accurately extracts headache frequency from electronic health records. Neurology. . [doi: 10.1101/2023.10.02.23296403]
14. Tan R, Lin Q, Low GH, et al. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. J Am Med Inform Assoc 2023 Sep 25;30(10):1657-1664. [doi: 10.1093/jamia/ocad133] [Medline: 37451682]
15. Le Guellec B, Lefèvre A, Geay C, et al. Performance of an open-source large language model in extracting information from free-text radiology reports. Radiol Artif Intell 2024 Jul;6(4):e230364. [doi: 10.1148/ryai.230364] [Medline: 38717292]
16. Ge J, Li M, Delk MB, Lai JC. A comparison of a large language model vs manual chart review for the extraction of data elements from the electronic health record. Gastroenterology 2024 Apr;166(4):707-709. [doi: 10.1053/j.gastro.2023.12.019] [Medline: 38151192]
17. Odisho AY, Liu AW, Pace WA, et al. MP07-14 development of a generative artificial intelligence data pipeline to automate the capture of unstructured MRI data for prostate cancer care. Journal of Urology 2024 May;211(5S). [doi: 10.1097/01.JU.0001008728.41882.d7.14]
18. Flatiron health. Clinical Research Solutions. URL: https://flatiron.com/clinical-research-solutions [accessed 2024-12-11]
19. UCSF department of urology. Urologic Outcomes Database (UODB). URL: https://urology.ucsf.edu/research/cancer/urologic-oncology-database-uodb [accessed 2024-11-23]
20. Altieri N, Park B, Olson M, DeNero J, Odisho AY, Yu B. Supervised line attention for tumor attribute classification from pathology reports: Higher performance with less data. J Biomed Inform 2021 Oct;122:103872. [doi: 10.1016/j.jbi.2021.103872] [Medline: 34411709]
21. Pace W, Liu A, Carlisle M, et al. 23 Generative artificial intelligence for automated unstructured MRI data extraction in prostate cancer care. J Clin Trans Sci 2025 Apr;9(s1):8-8. [doi: 10.1017/cts.2024.714]
22. Mueller C, Herrmann P, Cichos S, et al. Automated electronic health record to electronic data capture transfer in clinical studies in the German health care system: feasibility study and gap analysis. J Med Internet Res 2023 Aug 4;25(1):e47958. [doi: 10.2196/47958] [Medline: 37540555]
23. Ebbers T, Takes RP, Smeele LE, Kool RB, van den Broek GB, Dirven R. The implementation of a multidisciplinary, electronic health record embedded care pathway to improve structured data recording and decrease electronic health record burden. Int J Med Inform 2024 Apr;184:105344. [doi: 10.1016/j.ijmedinf.2024.105344] [Medline: 38310755]

Abbreviations

API: application programming interface
EHR: electronic health record
GPT: generative pre-trained transformer
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
MRI: magnetic resonance imaging
UCSF: University of California, San Francisco
UODB: Urologic Outcomes Database
XML: extensible markup language

Edited by E Uzun; submitted 30.Dec.2024; peer-reviewed by A Abe, JGD Ochoa, S Mohanadas, T Ekundayo; revised version received 29.Aug.2025; accepted 21.Oct.2025; published 06.Jan.2026.

Please cite as:
Carlisle MN, Pace WA, Liu AW, Krumm R, Cowan JE, Carroll PR, Cooperberg MR, Odisho AY
Development and Validation of a Generative Artificial Intelligence-Based Pipeline for Automated Clinical Data Extraction From Electronic Health Records: Technical Implementation Study
JMIR Bioinform Biotech 2026;7:e70708
URL: https://bioinform.jmir.org/2026/1/e70708
doi:10.2196/70708

Unpacking Genomic Biomarkers for Programmed Cell Death Receptor-1 Immunotherapy Success in Non–Small Cell Lung Cancer Using Deep Neural Networks: Quantitative Study

Rayan Mubarak[1]; Fahim Islam Anik[2], MS; Jean T Rodriguez[3], MS; Nazmus Sakib[4], PhD; Mohammad A Rahman[3]

[1]Cypress Bay High School, Weston, FL, United States

[2]Department of Mechanical Engineering, Khulna University of Engineering and Technology, Khulna, Bangladesh

[3]School of Computing and Information Sciences, Florida International University, Miami, FL, United States

[4]Department of Information Technology, Kennesaw State University, Atrium Building J3218, 1100 South Marietta Pkwy SE, Marietta, GA, United States

Corresponding Author:
Nazmus Sakib, PhD
Department of Information Technology, Kennesaw State University, Atrium Building J3218, 1100 South Marietta Pkwy SE, Marietta, GA, United States

Abstract

Background: Non–small cell lung cancer (NSCLC) is one of the leading causes of cancer-related mortality. Programmed cell death receptor-1 (PD-1) immunotherapy has shown results in the treatment of NSCLC; however, not all patients respond effectively to it. Identifying predictive biomarkers for PD-1 therapy response is critical to improving patient outcomes and treatment strategies. Traditional methods of biomarker discovery often fall short in terms of accuracy and comprehensiveness. Recent advancements in deep learning provide a powerful approach to analyze complex genomic data to resolve this issue.

Objective: This study aims to leverage deep neural networks (DNNs) to identify genomic biomarkers predictive of patient responses to PD-1 immunotherapy in NSCLC. DeepImmunoGene is a model designed using a reduced feature set to identify the most critical biomarkers. We use feature selection to reduce the space and apply deep learning to identify the highly predictive gene subset.

Methods: Differentially expressed genes were identified in RNA-seq data from 355 patients with NSCLC using the LIMMA package in R, followed by preprocessing with log2 transformation, removing outliers, and detecting easily identified genes. Machine learning models, including support vector machines, extreme gradient boosting (XGBoost), and DNNs, were applied to gene expression data to predict patient responses to immunotherapy. Key predictive genes were identified through model interpretation techniques, and differences in model performance were assessed for statistical significance. Primarily, the metric used identifies which genes serve as key biomarkers in regard to immunotherapy detection.

Results: Initially, we identified 1093 differentially expressed genes from RNA-seq data of 355 patients. We then trained models using SVM, XGBoost, and DNN to predict immunotherapy response. The DNN model outperformed both SVM and XGBoost with an accuracy of 82%, an area under the curve of 90%, and recall of 85%. To identify key biomarkers, we performed a permutation importance analysis, narrowing down the gene set to 98 genes. DeepImmunoGene, trained on these 98 genes, showed superior results, with an accuracy of 87% and an area under the curve of 95%. The top 36 upregulated genes in responders and 62 upregulated genes in nonresponders were identified, which could serve as potential biomarkers for predicting response to PD-1 inhibitors. These findings suggest that DeepImmunoGene can reliably forecast immunotherapy outcomes and aid in biomarker discovery, supporting the development of more personalized treatment strategies in NSCLC.

Conclusions: The DeepImmunoGene predictive model identified 36 upregulated genes that may represent candidate genomic biomarkers associated with response to PD-1 immunotherapy in patients with NSCLC. Notably, the 10 most significant genes offer valuable insights into the underlying mechanisms of treatment responses. These biomarkers may not only aid in predicting which patients are more likely to respond to PD-1 immunotherapy but also offer insights into the molecular differences associated with nonresponse.

(JMIR Bioinform Biotech 2026;7:e70553) doi:10.2196/70553

KEYWORDS

lung cancer; machine learning; deep neural network; DeepImmunoGene; biomarkers; RNA-seq analysis; differential gene expression; programmed cell death receptor-1; immunotherapy

Introduction

Lung cancer is a leading cause of cancer-related deaths globally, with approximately 238,340 new cases and 127,070 deaths annually in the United States [1,2] and 2.5 million new cases and 1.8 million deaths worldwide [3]. Smoking accounts for approximately 90% of lung cancer cases [4], whereas the remaining cases in nonsmokers are due to other factors, including environmental exposure to asbestos, arsenic, nickel, pesticides, other toxic chemicals, and air pollution [5,6]. Lung cancer is classified into 2 main groups: small cell lung cancer (SCLC) and non–small cell lung cancer (NSCLC) [4]. SCLC is a rare, fast-growing form of lung cancer that primarily develops in individuals with a long history of tobacco smoking, whereas NSCLC is more common, accounting for 85% of lung cancer cases compared to 15% for SCLC [5]. Although tobacco smoking is a major risk factor for NSCLC, it can also develop in nonsmokers. NSCLC is divided into 3 main types: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma [5,6]. Among these, adenocarcinoma is the most prevalent type, typically developing in the outer parts of the lung and being more common in individuals aged <45 years [5,6]. In contrast, squamous cell carcinoma originates from the epithelial cells of the central airways and is strongly associated with smoking [7,8].

Over the last 10 years, lung cancer treatment has undergone significant changes, with advancements in understanding its biology leading to the development of immunotherapy, which has emerged as a promising therapeutic option [9,10]. Immunotherapy works by enhancing the immune system through the use of drugs that block inhibitory signaling pathways, allowing it to better recognize and eliminate cancer cells [9,10]. Cancer can evade immunosurveillance by expressing ligands for inhibitory checkpoint molecules, such as programmed cell death receptor-1 (PD-1) and cytotoxic T-lymphocyte–associated protein-4, which prevent T cells from recognizing and destroying cancer cells [11]. Thus, immune checkpoint inhibitors (ICIs) have become an effective cancer therapy [12]. In recent years, ICIs have been used as the first line of treatment for metastatic NSCLC as well as consolidation therapy after surgical removal and chemotherapy [10]. PD-1 is a surface receptor found on T cells in lung cancer that acts as a negative regulator of immune responses [13-15]. Recent studies have shown that inhibiting PD-1 or programmed cell death-ligand 1 (PD-L1) restores T cell function, enabling the immune system to recognize and destroy cancer cells, suggesting their potential as promising therapeutic targets for NSCLC treatment [15-17]. However, only a fraction of patients respond to this immunotherapy. Therefore, we aimed to investigate genomic features that may help distinguish responders from nonresponders to PD-1 inhibitors and to gain insight into potential underlying biological differences. Furthermore, researchers have increasingly turned to bioinformatics and machine learning (ML) techniques to discover more precise biomarkers by analyzing large-scale genomic and molecular data. Among ML techniques, deep neural networks (DNNs) are particularly well suited for these tasks due to their ability to process and analyze vast, high-dimensional datasets. The use of ML in this research is indispensable for tackling the complexity of RNA-seq data and addressing the limitations of traditional analytical methods. Traditional statistical methods, such as ANOVA and *t* tests, rely on assumptions such as a normal distribution of the data, which is generally violated in gene expression data. Furthermore, as sample sizes and feature dimensions expand, these approaches also face computational constraints. In contrast, deep learning (DL) methods are particularly well suited to capturing the complex patterns present in genomic data [18]. Such models enable the identification of high-impact biomarkers, uncover nonlinear relationships in gene expression, and generate robust predictions for patient responses to PD-1 immunotherapy.

Several DL approaches have previously been proposed to predict immunotherapy outcomes, including survival-focused models such as DeepSurv and attention-based architectures designed to capture complex transcriptomic interactions [19-23]. These models demonstrate the growing interest in applying advanced DL to immunogenomics. We build upon this foundation by integrating interpretability into our approach. Furthermore, other existing approaches typically rely heavily on imaging-based methods, which can suffer from scanner or protocol heterogeneity and spurious correlation, among others. This study highlights the potential of ML techniques, particularly DNNs, in advancing precision medicine for patients with NSCLC undergoing PD-1 immunotherapy. We applied permutation importance in conjunction with DeepImmunoGene, which identified 98 important genes from a large RNA-seq dataset of 19,911 genes in the Gene Expression Omnibus (GEO) Repository [24]. We trained the DeepImmunoGene model on these genes, which outperformed linear models, achieving an accuracy of 87% and an area under the receiver operating characteristic curve (AUC) of 95%. This model identified a set of 36 upregulated genes in patients with NSCLC who are responders, which may serve as potential biomarkers for predicting responses to PD-1 immunotherapy for this group. Additionally, it identified another set of 62 upregulated genes in patients with NSCLC who are nonresponders, which could act as potential biomarkers for developing ICI therapy for this subgroup. These findings not only offer a foundation for improving patient stratification but also provide insights for tailoring therapeutic strategies. Despite significant advancements in treatment over the past decade, including the development of immunotherapy as a promising strategy for NSCLC, the prognosis for many patients remains poor [25,26]. Although ICIs targeting PD-1 and PD-L1 have shown potential as immunotherapy for patients with NSCLC, only a small fraction of patients respond to PD-1 inhibitors [24].

This underscores the need for more reliable biomarkers to accurately identify patients who will benefit from PD-1 inhibitors. The core work tries to answer 2 research questions (RQs) as follows:

- RQ1: How do ML models perform in predicting patient response to PD-1 immunotherapy based on differentially expressed genes (DEGs)?
- RQ2: What are the key biomarkers identified through feature selection and DL that predict patient response to

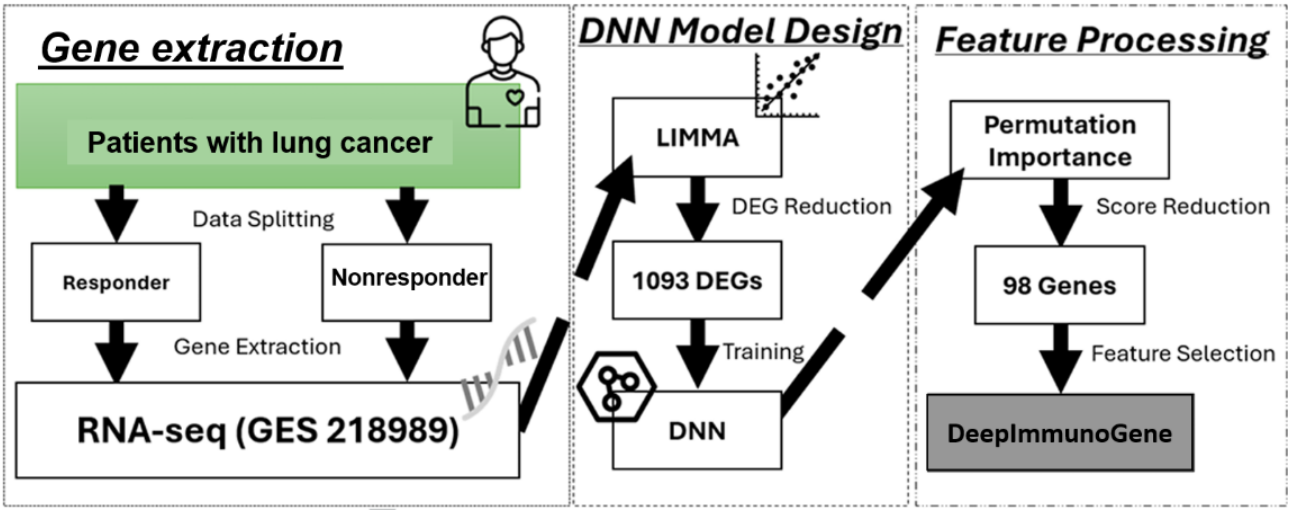
PD-1 immunotherapy, and how do they contribute to model performance?

Methods

Overview

The study was carried out according to the workflow presented in Figure 1. This workflow delineates the steps, beginning with the identification of significant DEGs from RNA-seq data [27] using the LIMMA package and culminating in the application of the DeepImmunoGene framework to identify and validate key genes associated with the response to PD-1 immunotherapy in patients with NSCLC.

Figure 1. Workflow for identifying biomarkers and predicting programmed cell death receptor-1 immunotherapy response in non–small cell lung cancer. DEG: differentially expressed gene; DNN: deep neural network; SVM: support vector machine; XGBoost: extreme gradient boosting.



Data Acquisition and Preprocessing

We used one RNA-Seq dataset (GSE218989) from the GEO public database GEO Repository [24]. This dataset included gene expression data for 19,911 genes across 355 patients with lung cancer who were treated with either PD-1 or PD-L1 inhibitors. It consisted of 187 nonresponders and 168 responders. Responsiveness was determined by Kang et al [24] using Response Evaluation Criteria in Solid Tumors (RECIST; version 1.1) [28]. Progression-free survival [29] was measured from the start of PD-1/PD-L1 inhibitor therapy to either documented disease progression or death from any cause. Overall survival was measured from the start of PD-1/PD-L1 inhibitor therapy to death from any cause [24]. A responder is therefore classified as a patient who showed improvement under the RECIST criteria or, in other words, a patient who experienced improvements after the PD-1 immunotherapy was administered. At the same time, a nonresponder is a patient who did not meet the criteria showcased by a worsening or stable disease.

The raw gene expression count data were already normalized in the transcripts per million (TPM) value for the 19,911 protein-coding genes. We first identified the DEGs between the responders and nonresponders using the LIMMA package [30] in R (version 4.4.1; Bioconductor, USA). LIMMA was used to create a linear function to model the entire dataset and to develop correlations with response status as the main variable in the design matrix. Empirical Bayes moderation was performed to model and stabilize the gene-wise variances using a prior marginal distribution of the data [30]. Genes with a LIMMA-calculated *P* value less than .05 were considered significantly differentially expressed and were selected for all subsequent analyses and modeling. For model training and testing, the data were further processed by performing a log2 (TPM+1) transformation on each gene expression value to stabilize the variance in gene expression.

ML Models

Overview

The application of ML is vital in this research due to the complexity, scale, and dimensionality of RNA-seq data, as well as the intricate, nonlinear biological mechanisms underlying immunotherapy response in patients with NSCLC [31]. Traditional statistical methods struggle with high-dimensional datasets, such as the 19,911-gene RNA-seq data used here, often succumbing to the “curse of dimensionality” and failing to capture subtle gene interactions. ML models such as support vector machines (SVMs) [32], extreme gradient boosting (XGBoost) [33], and DNN [34] overcome these challenges by effectively handling high-dimensional inputs, modeling complex nonlinear relationships, and identifying important gene features through built-in feature selection techniques. This enables the discovery of meaningful gene patterns that differentiate responders from nonresponders while enhancing predictive power and model generalizability.

Moreover, ML methods excel in managing noise and variability inherent in biological data, offering robust performance through techniques such as regularization and early stopping [35,36]. Their scalability and automation allow for efficient analysis of massive RNA-seq datasets, ensuring accuracy and rapid processing, essential for clinical translation. By integrating advanced techniques for hyperparameter tuning, ML provides a unified, systematic workflow that optimizes predictive performance [37]. These capabilities facilitate the identification

of potential predictive biomarkers from gene expression data, which may serve as a foundation for future precision medicine efforts aimed at tailoring immunotherapy strategies in patients with NSCLC. This study used several ML models, including SVM, XGBoost, and DNN [11]. Their predictive performance was evaluated to identify the model that worked best. We built the SVM model using the Python package Scikit-learn (sklearn); for XGBoost, we used the XGBoost Python package [38]; and for the DNN, we used the Keras and TensorFlow Python packages [11]. The details about each ML approach are further described below.

Support Vector Machine

SVM is a kernel-based binary classifier that separates key data features linearly into 2 groups in a high-dimensional space called the feature space [38,39]. It searches for the optimal decision boundary (hyperplane) to separate the features by maximizing the margin between the hyperplane and the nearest training data. SVM effectively extracts key but subtle patterns in a complex dataset, allowing for low-error, high-precision sample classification [40]. The model architecture's hyperparameter settings are given in Table 1.

Table . Summary of model architectures' hyperparameter settings.

Model	Key hyperparameters tuned	Final settings	Optimization approach
SVM[a]	C, kernel, gamma	C=0.1, kernel=linear, gamma=0.1	GridSearchCV (5-fold CV[b])
XGBoost[c]	n_estimators, max_depth, learning_rate, sampling	n_estimators=300, max_depth=100, learning_rate=0.1, sampling=uniform	GridSearchCV (5-fold CV)
DNN[d]	batch_size, epochs, initializer, optimizer, activation, dropout, layers, nodes	Input=256; hidden layers=[128, 100, 100]; activation=ELU[e]; optimizer=Adam; dropout=0; epochs=100; batch size=100	Multistage GridSearchCV

[a]SVM: support vector machine.

[b]CV: cross-validation.

[c]XGBoost: extreme gradient boosting.

[d]DNN: deep neural network.

[e]ELU: exponential linear unit.

XGBoost

XGBoost is an ensemble learning algorithm that builds gradient-boosted decision trees one by one and passes the residuals of the previous tree to train the following model. It uses the second partial derivative of the loss function and adds an L1 and L2 regularization term to reduce overfitting [41]. Similar to SVM, we optimized the hyperparameters using GridSearchCV to evaluate a combination of parameters. The hyperparameter settings are given in Table 1.

Deep Neural Network

DNN is a nonlinear model that combines neurons that simulate the human brain to make predictions [41,42]. It consists of 3 layers: the input layer, hidden layers, and output layer, which are linked by weights to allow the model to understand complex patterns in the data. We used a DNN because they have been previously applied for genomic-based predictions for diseases [43]. Similar to the previous 2 models, we started with hyperparameter optimization using GridSearchCV. As the DNN has more parameters to tune, we split the Grid Search into 3 stages: (1) batch size and epoch; (2) weight initializer, optimizer, and activation function; and (3) hidden layers, nodes per hidden layer, and dropout optimization. The resulting network consisted of an input layer with 256 nodes, 3 hidden layers with 128 nodes, 100 nodes, and 100 nodes, respectively, an exponential linear unit activation function, Adam optimizer, zero dropout, and normal initializer. The details are summarized in Table 1. We applied the binary cross-entropy loss function as shown in Equation 1 so that the model minimizes to learn the optimal weights for each gene to classify responder and nonresponder patients.

$$(1) LBCE = -1N\sum i=1Nyi \times \log(p(yi)) + (1-yi) \times \log(1-p(yi))$$

The model was trained for 100 epochs with a batch size of 100 based on the GridSearchCV results. After identifying these optimal hyperparameters for the DNN, we used it to construct the architecture for the DeepImmunoGene network.

Permutation Importance

To develop the DeepImmunoGene framework, we used the permutation importance method from scikit-learn to identify the subset of genes that most significantly contributed to the DNN's prediction of patient outcomes to PD-1 immunotherapy [11]. Basically, this technique improves model accuracy by removing the "noisy" genes. First, we used the original DNN trained on the 1093 gene expression data to establish a baseline performance using the accuracy score. Then, we randomly shuffled each gene's expression values across the 71 testing patients one at a time to disrupt any existing association between that gene and the response classification. After shuffling a gene, the DNN was run again to recalculate the accuracy. If the accuracy decreased after shuffling, that gene was important for predicting the response. Conversely, if the accuracy increased or did not change after shuffling, that gene showed little to no correlation with response prediction. Given the nonlinearity of PD-1 immunotherapy genetics, a standard linear model, such as least absolute shrinkage and selection operator or stepwise regression, is unable to capture the noise in the genes. Feature permutation ignores this weakness by using a direct DNN

architecture to quantify the decrease in performance due to a change in the feature. By exploring the performance of the model directly, we remove the uncertainty of a linear model and guarantee the importance of the features in the deployed solution. To evaluate the stability of the features identified, we ran the analysis 3 additional times, each with 50 iterations. We then compared the resulting gene sets to quantify their overlap. We also trained and evaluated the model using each gene set to determine the superior cohort for all subsequent analyses. Equation 2 was used to calculate the importance score assigned to each gene.

(2)Importance score=accuracybaseline−accuracypermutation

Training and Testing

We executed our code for the ML models in Google Colab notebooks [44] using an NVIDIA T4 GPU [45] operating with 15 GB of RAM. For all models, 284 patients were used for training, and 71 patients were used for testing. This provided an 80/20 percentage split of the data. For the DNN, an additional validation split of 10% was applied to the training data to monitor model performance during training. This validation set was extracted from the training data, leaving the test set of 71 patients unchanged. During the training of the DNN, an early stopping method was used to monitor the validation loss after each epoch to stop training if the model's performance diminished. The state of the model was saved after each epoch so that it could revert to the optimal state for testing. This was done to mitigate any overfitting that might occur during training. All ML models were executed 15 times.

Evaluation Metrics

To evaluate the models' performance, we used accuracy, AUC score, recall, specificity, precision, and F_1-scores [46], which are standard metrics used to assess classification performance. These metrics can be found using the confusion matrix, a 2×2 matrix with the number of true positives, true negatives, false positives, and false negatives that the model predicts, with the equations listed below to calculate each metric.

(3)Accuracy=TP+TNTP+TN+FP+FN × ∈[0,1]

(4)Recall=TPTP+FN×∈0,1

(5)Specificity=TNTN+FP×∈0,1

(6)Precision=TPTP+FP×∈0,1

(7)F1=2 ×Precision ×Recall Precision+Recall × ∈[0,1]

Accuracy (Equation 3) measures the overall correct predictions out of all predictions made. Recall evaluates the model's ability to correctly identify PD-1 responders as positive out of all PD-1 responders, as shown in Equation 4. Specificity (Equation 5) is the opposite; it measures the model's ability to correctly identify PD-1 nonresponders out of all nonresponders. Precision (Equation 6) is the ratio of all correctly identified positive PD-1 respondents to all the patients the model assigns as positive, and the F_1-score (Equation 7) is a harmonic mean of precision and recall that penalizes extreme values [47]. AUC measures the trade-off between specificity and recall [38,48].

Bioinformatics and Statistical Analysis

All computations and analyses in this study were performed in Google Colab notebooks using Python (version 3.10) and R (version 4.4.1). Differentially expressed genes were analyzed with LIMMA in R [30]. Upregulated genes were classified for responders and nonresponders by calculating log fold changes (LogFC). Accuracy, AUC, recall, specificity, precision, F_1-score, true positives, true negatives, false positives, and false negatives were calculated using sklearn Metrics. Statistical analyses were conducted using GraphPad Prism (version 5.01; GraphPad Software). The Kruskal-Wallis nonparametric test, followed by the Dunn post hoc multiple comparison test, was used to compare predictive performance between the models. A P value less than .05 was considered statistically significant.

The next section delves into the detailed analysis of the genes identified through the DeepImmunoGene framework and their relevance in predicting immunotherapy response. It outlines how the permutation importance method was used to isolate key genes associated with positive or negative treatment outcomes and discusses the biological significance of these genes in the context of immune response modulation in NSCLC. Additionally, the section provides an in-depth comparison of the ML models' performance, highlighting the strengths and limitations of each approach, and evaluates their potential applications in clinical settings for improving patient stratification and personalized treatment strategies. By integrating these findings, the study aims to contribute to our understanding of molecular biomarkers that may inform future efforts to optimize the use of PD-1 inhibitors in cancer therapy.

External Validation

To externally validate the biomarkers identified by DeepImmunoGene, we obtained a bulk RNA-seq dataset (GSE207422) from the GEO public database. This dataset included gene expression data for 58,387 genes across 24 patients with NSCLC who were treated with PD-1 inhibitors combined with chemotherapy [49]. Patient responsiveness was determined using RECIST, where complete response and partial response were considered responders, whereas stable disease was considered a nonresponder. The cohort comprised 17 responders and 7 nonresponders. This external dataset was processed using the aforementioned workflow applied to the training dataset. The Mann-Whitney U test was used to determine whether the difference in gene expression between responders and nonresponders was statistically significant. We generated violin plots of the top-ranked responder and nonresponder biomarkers identified by DeepImmunoGene to assess whether their expression patterns in the test set were consistent with the model's predictions using the ggplot2 package [50].

Ethical Considerations

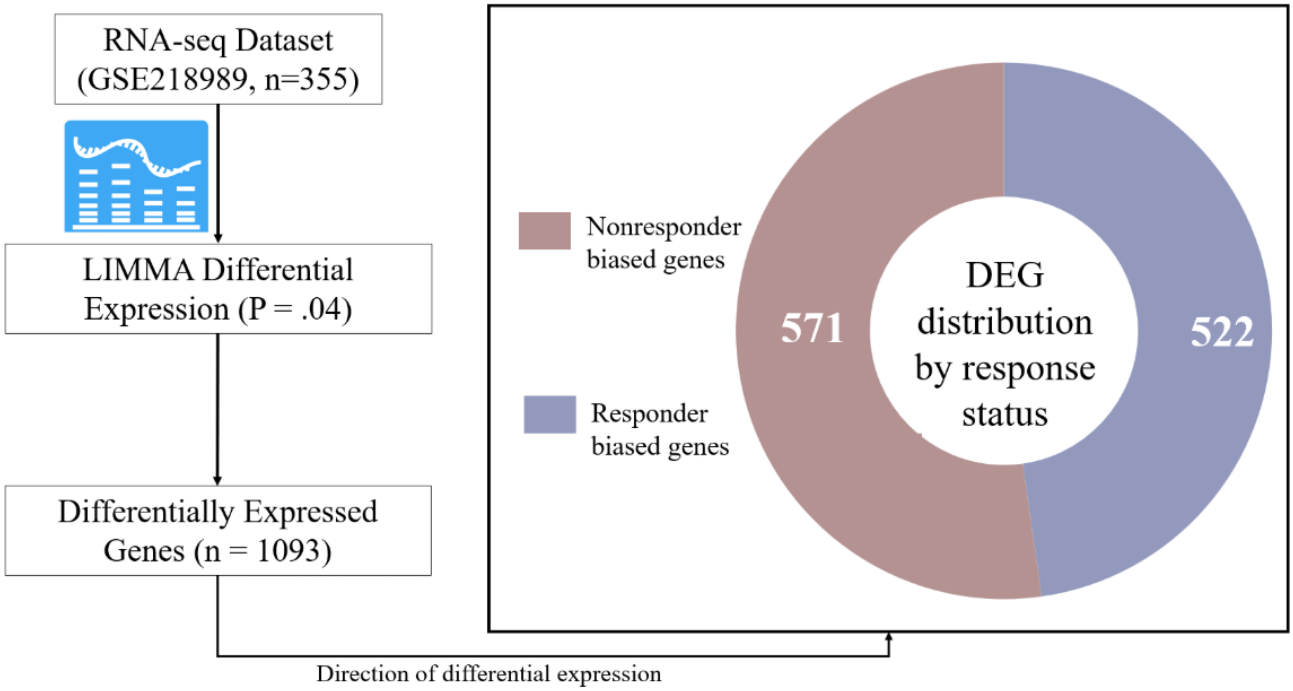
This study used only publicly available or fully deidentified secondary data; therefore, institutional review board approval and informed consent were not required. No personal identifiers were accessed, and privacy and confidentiality were strictly maintained.

Results

ML Predicts Response to PD-1 Immunotherapy (RQ1)

DEGs were identified using LIMMA power analysis of bulk RNA-seq data (GSE218989) from the GEO public database GEO Repository. LIMMA identified 1093 important DEGs from a total of 19,911 genes in patients with lung cancer, where 522 genes were upregulated in responders, and 571 genes were upregulated in nonresponders (P=.04), as shown in Figure 2.

Figure 2. Identification and stratification of differentially expressed genes associated with programmed cell death receptor-1 immunotherapy response in non–small cell lung cancer. Bulk RNA-seq data from 355 patients (GSE218989) were analyzed using LIMMA differential expression analysis (P=.04), identifying 1093 differentially expressed genes. These genes were stratified by direction of differential expression into responder-upregulated (n=522) and nonresponder-upregulated (n=571) gene sets, forming the initial feature space for downstream machine learning analyses. DEG: differentially expressed gene.



Here, we trained SVM and XGBoost models using the 1093 identified DEGs to predict patient response to PD-1 immunotherapy. The performance of the models was evaluated using several metrics, including accuracy, AUC, recall, specificity, precision, and F_1-score [46]. First, we applied SVM, and our data showed that it achieved an accuracy of 68% and an AUC score of 76% with recall, specificity, precision, and F_1-score values of 0.70, 0.65, 0.77, and 0.71, respectively (Figure 3A, 3B and Table 1). Next, we used XGBoost to see if its ensemble learning method could yield higher accuracy and AUC scores. Our data showed that XGBoost performed slightly better than SVM, with an accuracy of 72%, an AUC score of 77%, a recall of 0.73, a specificity of 0.71, a precision of 0.76, and an F_1-score of 0.74 (Figure 3A, 3B and Table 2). The suboptimal performance of these 2 models may be due to the large dataset, suggesting that a more complex and nonlinear approach, such as a DNN, is necessary for accurately predicting patient responses. We used SVM and XGBoost as baseline classifiers commonly applied in gene expression studies to provide context for the performance of our DNN. While these models differ in complexity from DNNs, the comparison helps demonstrate the added value of capturing nonlinear interactions in gene expression data.

Figure 3. Predictive performance comparison of support vector machine (SVM), extreme gradient boosting (XGBoost), and deep neural network (DNN) models. (A) Accuracy scores and (B) receiver operating characteristic (ROC) curve analysis demonstrate that the DNN model outperformed both SVM and XGBoost. The DNN achieved an accuracy of 82% and an area under the curve (AUC) of 90%, compared to 68% and 76% for SVM and 72% and 77% for XGBoost. These results highlight the advantage of deep learning for modeling complex, high-dimensional gene expression data.

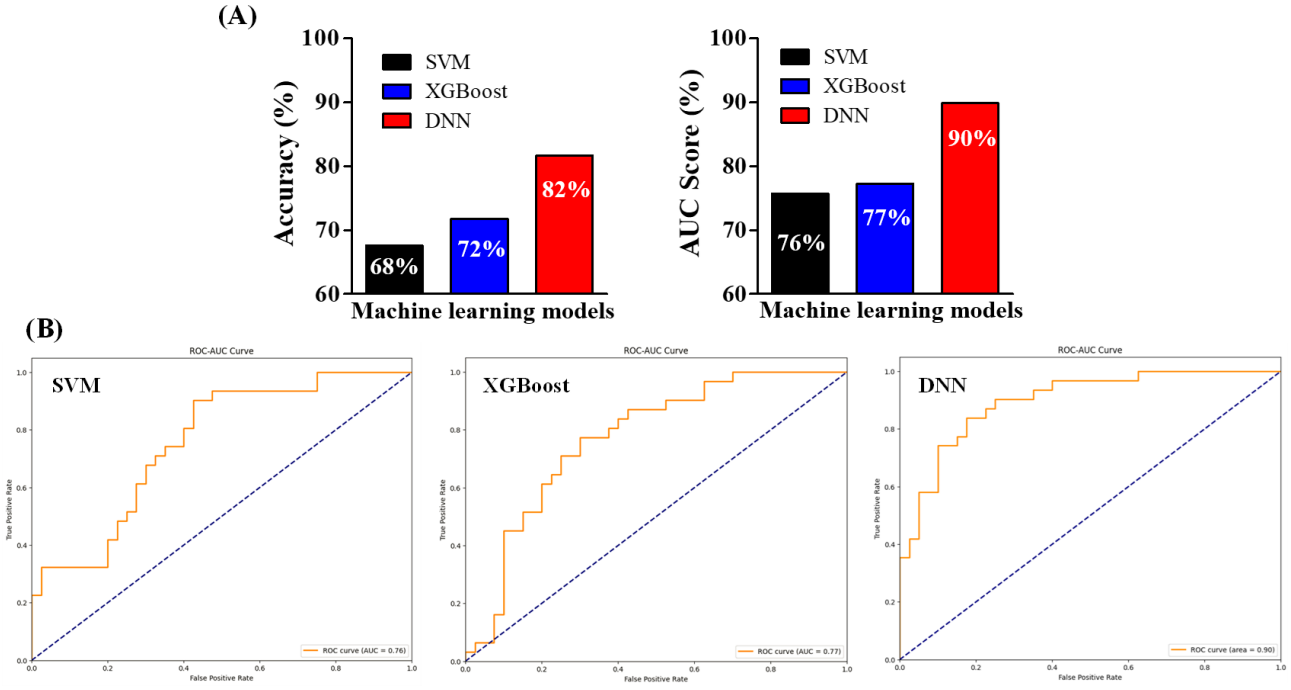


Table . Performance comparison of machine learning models for predicting response to programmed cell death receptor-1 immunotherapy.

Models	Accuracy	AUC[a]	Recall	Specificity	Precision	F_1-score
SVM[b] (1093 genes)	0.68	0.76	0.70	0.65	0.77	0.71
XGBoost[c] (1093 genes)	0.72	0.77	0.73	0.71	0.76	0.74
DNN[d] (1093 genes)	0.82[e]	0.90[e]	0.85[e]	0.78[e]	0.81	0.84[e]
SVM (98 genes)	0.65	0.75	0.65	0.65	0.70	0.68
XGBoost (98 genes)	0.77	0.81	0.80	0.74	0.80	0.80
DeepImmunoGene (98 genes)	0.87[e]	0.95[e]	0.87[e]	0.89[e]	0.93[e]	0.89[e]

[a]AUC: area under the receiver operating characteristic curve.

[b]SVM: support vector machine.

[c]XGBoost: extreme gradient boosting.

[d]DNN: deep neural network.

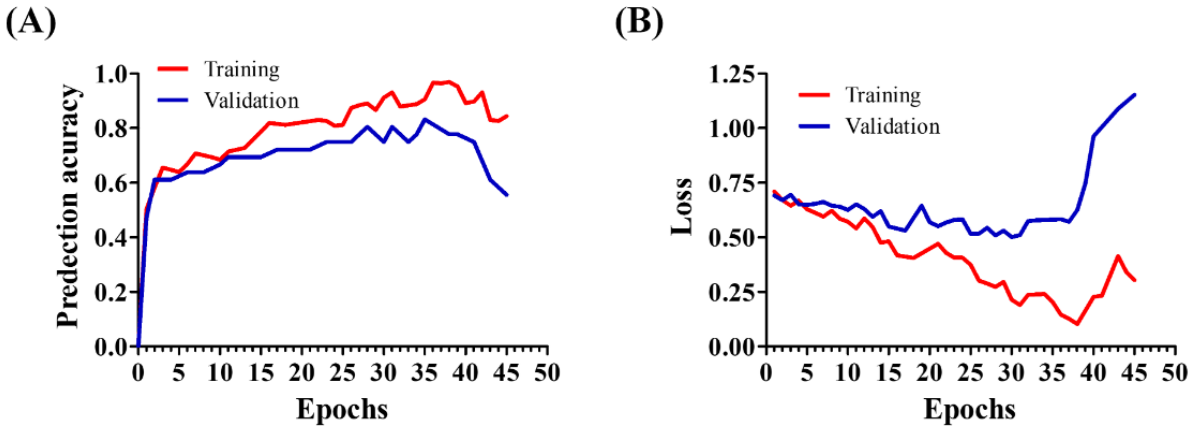
[e]A statistically significant difference from DeepImmunoGene when compared to SVM or XGBoost.

DNN Predicts Response to PD-1 Immunotherapy With Higher Accuracy

Given that the RNA-seq data includes the expression of more than 1000 genes, we implemented a DNN to enhance predictive accuracy. First, we set the DNN training for 100 epochs, but it stopped at 45 epochs due to early stopping, and the model was then reverted to the optimal state reached at 35 epochs (Figure 4). During the training process, both training and validation accuracy and loss were monitored. We found that the accuracy increased until it exhibited an asymptotic behavior (Figure 4A). Conversely, the training loss decreased steadily, while the validation loss showed some fluctuations (Figure 4B). These findings suggest that training the model for additional epochs would not further improve its performance. Next, we tested the predictive performance. Our data revealed that the DNN achieved excellent predictive performance compared to both SVM and XGBoost, achieving an accuracy of 82%, an AUC score of 90%, a recall of 0.85, a specificity of 0.78, a precision of 0.81, and an F_1-score of 0.84 (Figure 3A, 3B and Table 2). Given the nature of the data, DNN can analyze multidimensional genetic information more accurately than existing linear models.

This is showcased with a 21% accuracy improvement over more linear models, such as SVM, and a 14% improvement over XGBoost in our experiments. As a result, we can showcase that to capture the intricacies of the data, it is important to use a model capable of supporting complex multidimensional relationships such as a DNN architecture.

Figure 4. Deep neural network training and validation performance. (A) Training and validation accuracy over epochs shows a steady increase until convergence, with early stopping triggered at epoch 45 and the model reverting to optimal weights from epoch 35. (B) Training loss decreased continuously, whereas validation loss fluctuated slightly before stabilizing, indicating that further training would not significantly improve model performance.



Key Biomarker Identification (RQ2)

We applied DeepImmunoGene with scikit-learn permutation importance to a set of 1093 genes. To mitigate variability in feature importance estimates and to ensure the identification of robust features, this procedure was repeated 3 additional times with 50 iterations each. We then compared the gene sets identified across all 4 total runs and observed a high degree of overlap, with an average of 85.5% consistency among them. The resulting analysis (Figure 5) identified a final set of 98 genes with nonzero importance scores and ranked them according to their level of importance (Figure 6). Although individual gene importance scores below 0.0025 may appear low, the combined contribution of these genes accounts for approximately 18% of the total model importance, indicating they meaningfully improve the model's predictive performance. These 98 genes were subsequently used to train DeepImmunoGene. Testing this model revealed an accuracy of 0.87 and an AUC of 0.95, a recall of 0.87, a specificity of 0.89, a precision of 0.93, and an F_1-score of 0.89, demonstrating superior performance across all metrics. To validate the necessity of a DL approach for our feature selection and to better contextualize the significant performance improvement of DeepImmunoGene, we conducted a comparative analysis with the traditional ML models. We trained and tested both SVM and XGBoost on the same 98 genes identified via permutation importance. The 98-gene SVM model attained an accuracy of 65%, an AUC of 75%, a recall and specificity of 0.65, a precision of 0.70, and an F_1-score of 0.68. The 98-gene XGBoost model achieved an accuracy of 77%, an AUC of 81%, a recall of 0.80, a specificity of 0.74, a precision of 0.80, and an F_1-score of 0.80 (Table 2). This indicates that DeepImmunoGene outperformed all other models in every metric (Table 2). Genes with a LogFC greater than 0 were considered upregulated in responders, whereas genes with a LogFC less than 0 were considered upregulated in nonresponders. We discovered that 36 genes were upregulated in patients with NSCLC who responded to PD-1 immunotherapy, with the top 10 most significant being GSTT2B, HMGA2, AC135050.2, ANKRD33B, MMP13, PLA2G2D, RASGEF1A, BIRC7, DCAF4L2, and CHMP7 (Figure 7). These genes may serve as potential biomarkers for predicting response to PD-1 immunotherapy. Additionally, we identified 62 upregulated genes in nonresponder patients with NSCLC, with the top 10 most important being SPINK1, FEZF1, THBS4, BEST3, TESC, C6orf226, TSSK2, SFRP2, C1GALT1C1L, and RARRES1 (Figure 7).

The top 10 most significant upregulated genes were identified for both responder and nonresponder patients with NSCLC based on the DeepImmunoGene model. In responders, genes such as GSTT2B, HMGA2, and MMP13 were prominent, whereas SPINK1, FEZF1, and THBS4 were among the top in nonresponders. These genes may serve as potential predictive biomarkers for PD-1 treatment outcomes.

Figure 5. Workflow for identifying predictive biomarkers using DeepImmunoGene. Schematic of the DeepImmunoGene model pipeline. The 1093 differentially expressed genes were subjected to permutation importance analysis to extract the 98 most informative features, which were then used to train the final model. This approach enabled identification of key genes associated with programmed cell death receptor-1 (PD-1) immunotherapy response.

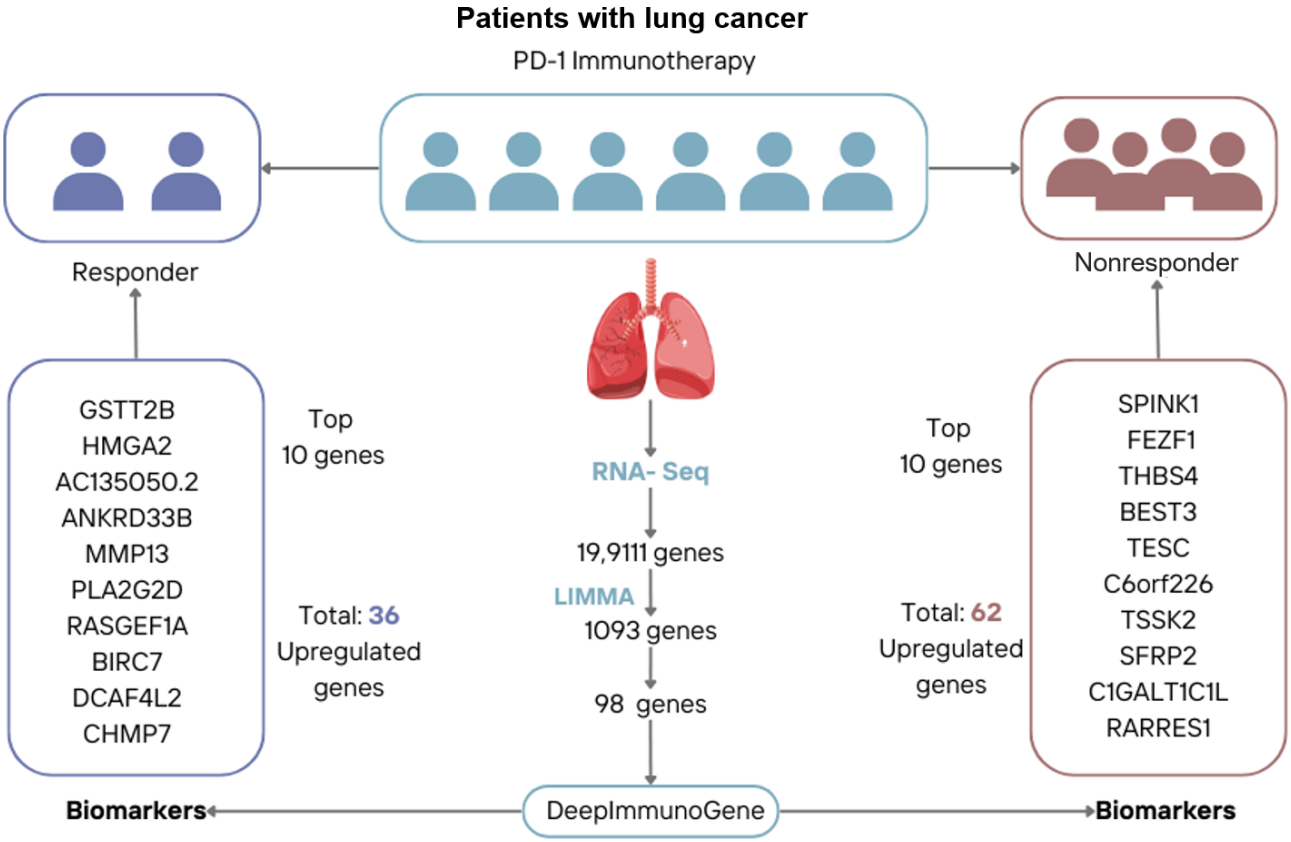


Figure 6. Gene importance ranking using permutation analysis. Permutation importance applied to the 1093 differentially expressed genes using the DeepImmunoGene model identified 98 genes with nonzero importance scores. These genes were ranked based on their contribution to model prediction performance, highlighting their potential as key features for programmed cell death receptor-1 response classification in patients with non–small cell lung cancer.

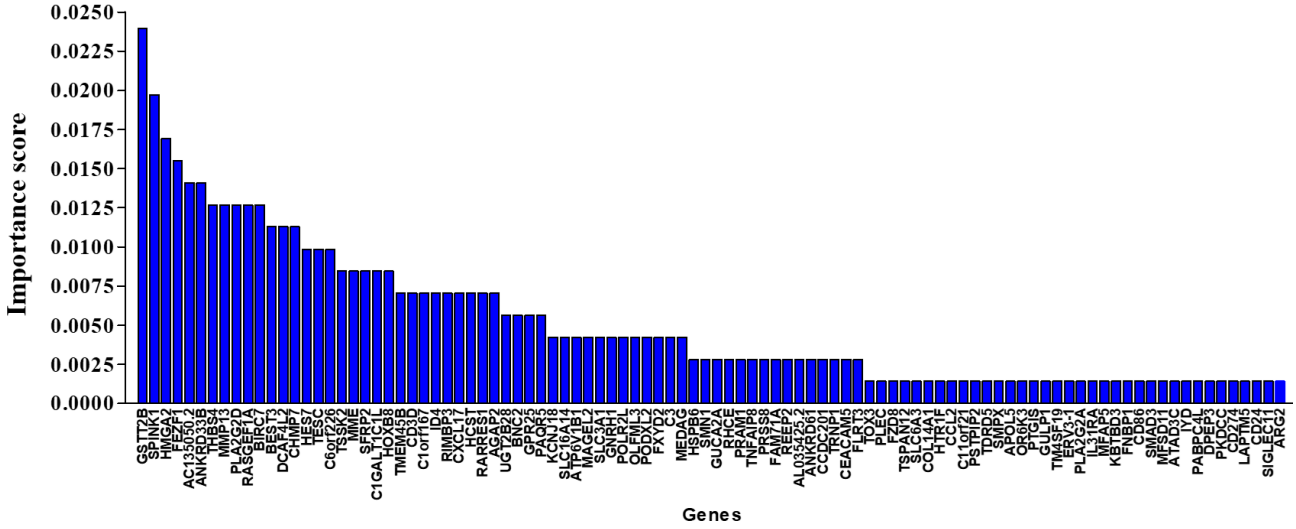
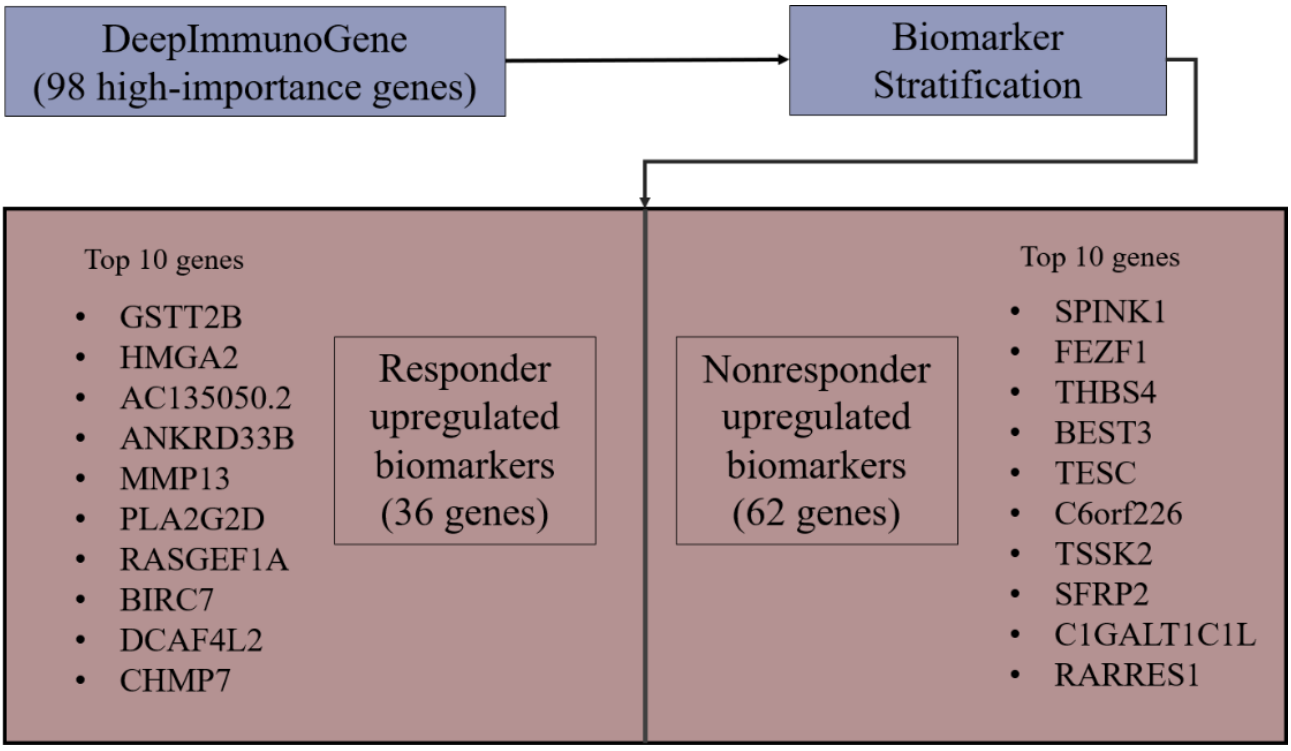


Figure 7. DeepImmunoGene-based stratification of predictive biomarkers associated with programmed cell death receptor-1 (PD-1) immunotherapy response. Using permutation importance and deep neural network modeling, 98 high-importance genes were identified and stratified based on direction of differential expression. Thirty-six genes were upregulated in responders and 62 in nonresponders. The top 10 genes in each group are shown as candidate biomarkers for predicting PD-1 treatment outcomes in non–small cell lung cancer.

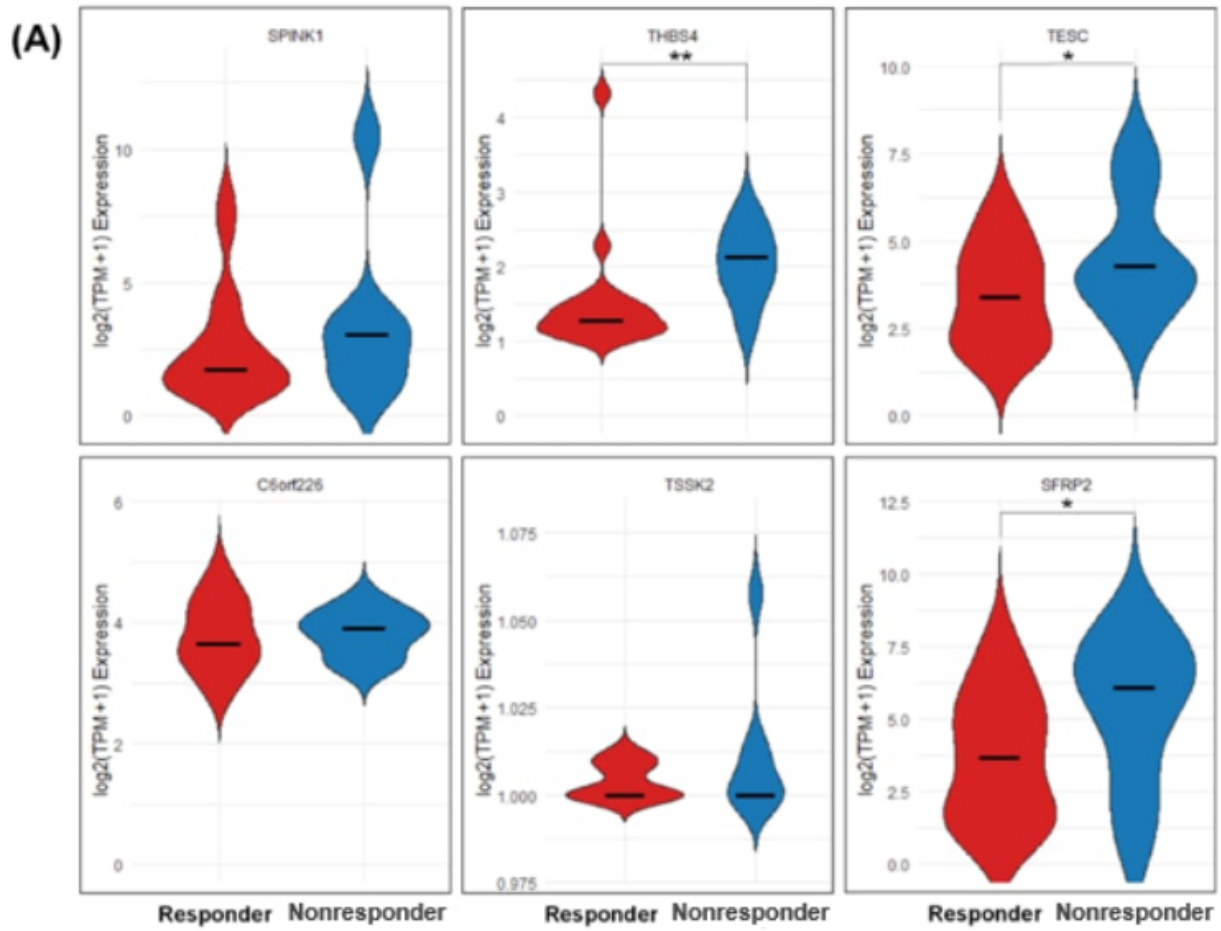


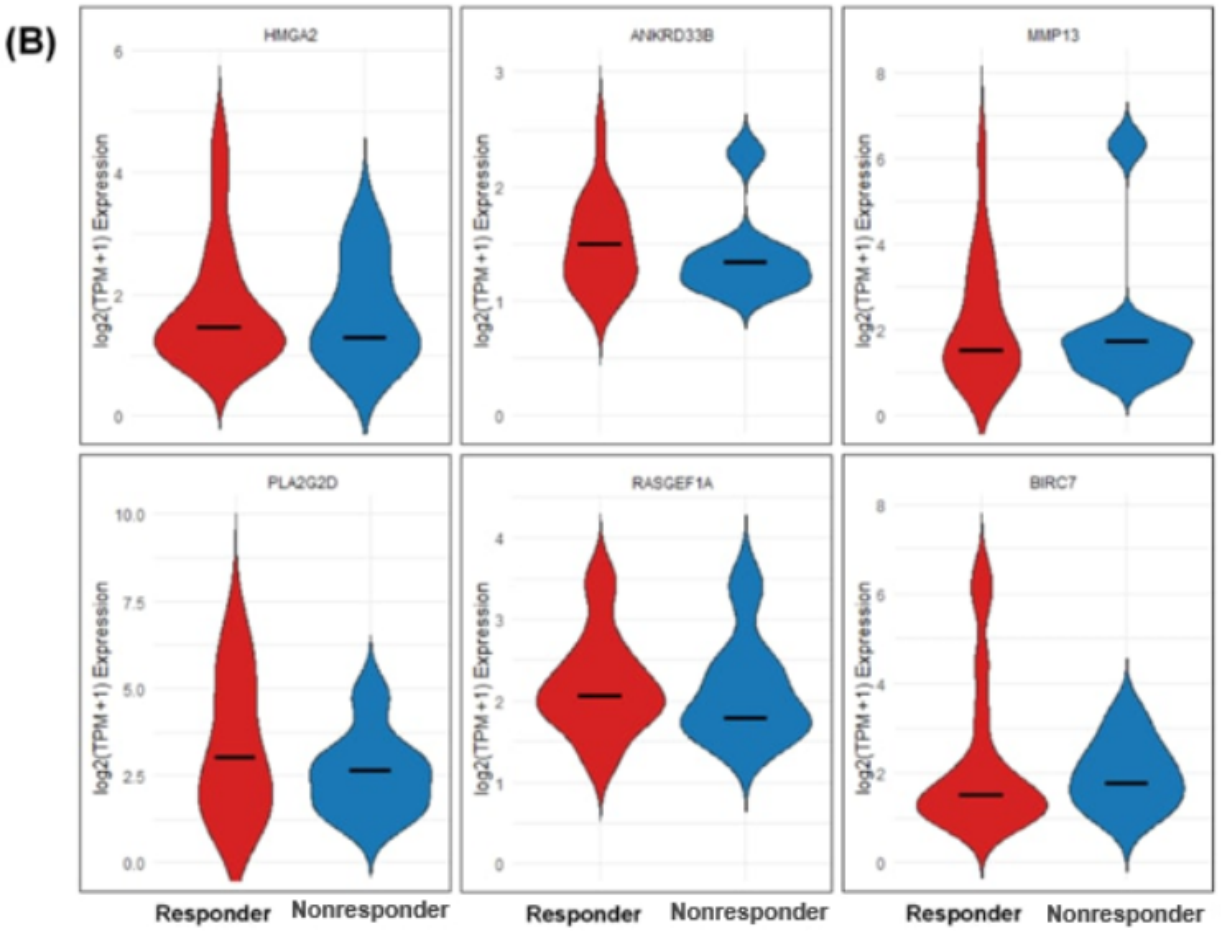
External Validation of Biomarkers Identified by DeepImmunoGene

Here, we sought to determine whether DeepImmunoGene's predicted biomarkers showed consistent expression patterns in an independent dataset. We generated violin plots comparing log2 (TPM +1) gene expression between responders and nonresponders. Of the top 10 nonresponder-upregulated biomarkers identified by DeepImmunoGene, 6 genes were present in the independent dataset and analyzed. We found that 4 of these 6 genes (SPINK1, THBS4, TESC, and SFRP2) showed a consistent trend of higher median expression in nonresponders (Figure 8A). Of these, 3 genes (THBS4, TESC, and SFRP2) demonstrated statistically significantly higher expression (P=.04) in nonresponders.

Of the top 10 responder-upregulated biomarkers identified, 6 genes were present in the independent dataset and analyzed. We found that 4 of these 6 genes (HMGA2, ANKRD33B, PLA2G2D, and RASGEF1A) showed higher median expression in responders (Figure 8B). BIRC7 and MMP13 had similar median expression in both groups; however, their violin plots displayed extended upper tails, indicating that some patients exhibited markedly higher expression levels. While these patterns suggest differences in expression between responders and nonresponders, statistical significance was not reached in this analysis.

Figure 8. Validation of biomarkers identified by DeepImmunoGene. Violin plots showing differences in the expression of (A) 6 nonresponder-upregulated biomarkers and (B) 6 responder-upregulated biomarkers. *P* values determined by Mann-Whitney *U* test. **P*=.05, ***P*<.01.





Discussion

Principal Findings

We aimed to identify DEGs associated with response to PD-1 immunotherapy in patients with lung cancer using a DNN model to explore the biological mechanisms underlying immunotherapy response. Here, we developed DeepImmunoGene, a computational framework that uses an advanced neural network with an integrated approach to predict patient response to PD-1 immunotherapy with high accuracy. Our model identified 36 upregulated genes, including the top 10 (GSTT2B, HMGA2, AC135050.2, ANKRD33B, MMP13, PLA2G2D, RASGEF1A, BIRC7, DCAF4L2, and CHMP7), which were associated with positive responses to PD-1 immunotherapy in patients with NSCLC. However, apart from the 10 described, our model was able to find approximately 96 total critical genes. If we were to leverage only differential gene expression rather than DeepImmunoGene, more than 1000 genes would be present, many of which are not significant biomarkers for identifying responders. As a result, we deployed a permutation importance feature selector to identify from the potential 1000 expressive genes the ones that are critical in the identification of the patient, reducing the quantity of noisy biomarkers in the dataset. These findings suggest that these genes could serve as the candidate biomarkers for predicting patients who respond to PD-1 inhibitors. Some of these genes (HMGA2, MMP13, BIRC7, and PLA2G2D) have been reported to be overexpressed in various cancers, including lung adenocarcinoma, and are associated with tumor progression and metastasis [51-54], supporting their potential as biomarkers for PD-1 immunotherapy. We can identify these genes by ranking based on feature importance. We identify the most important genes, given the decrease in performance once permutated. The 10 most critical genes show the greatest decline in model accuracy once they are shifted. Furthermore, existing literature has shown many of these genes to be capable identifiers of immunotherapy. Genes such as HMGA2 and MMP13 are currently in the literature to identify a high likelihood of therapy success [55,56]. Our primary contribution lies not in introducing a novel DL architecture, but in developing DeepImmunoGene, a framework that complements prior frameworks, integrating interpretability and ML with the novelty to identify key genomic markers for PD-1 immunotherapy response.

In addition to their differential expression patterns, several of the top-ranked genes identified in our model have established roles in cancer-related biological processes. HMGA2 is a well-characterized architectural transcription factor associated with epithelial-mesenchymal transition and metastatic progression [57]. MMP13 contributes to extracellular matrix degradation and tumor invasion [55]. BIRC7 (also known as Livin) has been implicated in the inhibition of apoptosis and immune evasion mechanisms in solid tumors [58]. PLA2G2D is known for its involvement in inflammatory signaling and has been shown to modulate dendritic cell function and T-cell recruitment in the tumor microenvironment [59]. These functional insights, drawn from existing literature, suggest that many of the identified genes may influence immunotherapy response through diverse oncogenic and immune-related pathways. Although a formal pathway enrichment analysis was not performed, the biological relevance of these genes supports their potential as markers of therapeutic response.

Our analysis began with the application of the LIMMA method to bulk RNA-seq data, which identified 1093 DEGs from a total of 19,911 genes in patients with lung cancer [24]. LIMMA is a widely used tool for differential gene expression analysis, facilitating the identification of genes linked to disease pathogenesis, particularly in RNA-seq and microarray data [30]. We evaluated these 1093 genes using 3 different ML models, including SVM, XGBoost, and DNN, to assess their predictive performance. The SVM showed moderate performance in classifying patient response with an accuracy of 0.68 and an AUC of 0.76, suggesting that it was unable to effectively capture the underlying correlations between gene expression and patient response. This may be due to the nonlinear nature of gene expression data, which likely hindered the SVM model's ability to generalize its predictions across patients [11,60]. While XGBoost outperformed SVM by a slight margin (0.04 for accuracy and 0.01 for AUC), there is no significant difference between these models, indicating that neither model could provide sufficiently robust predictions. These findings suggest that the high dimensionality, small sample size, and categorical imbalance of RNA-seq data pose significant challenges for traditional ML approaches [61].

To address the limitations of traditional ML models, we applied a DNN, a nonlinear model capable of capturing complex relationships within large gene expression datasets by mimicking the information-processing patterns of the human brain to generate predictions [11,40,60]. Unlike traditional models such as SVM and XGBoost, the DNN consists of multiple layers of neurons connected by weighted links, which allow the model to learn intricate patterns within the data. DNNs have shown strong performance in genomic predictions for various diseases [43]. The DNN model using the 1093 DEGs significantly outperformed both SVM and XGBoost. It exceeded SVM by 14% in both accuracy and AUC and outperformed XGBoost by 10% in accuracy and 13% in AUC. This improved performance of the DNN is attributed to its ability to capture and learn from the high-dimensional, nonlinear interactions inherent in gene expression data, which are challenging for traditional linear models to predict accurately [61]. This capability allows the DNN to generalize more effectively across diverse patient data, leading to more accurate and robust predictions than those made by more basic, linear computational models.

To reduce the number of genes and enhance the reliability of our model, we performed a permutation importance analysis using the scikit-learn framework. This analysis was repeated 4 times, each with 50 iterations to ensure the identification of a robust gene set to build DeepImmunoGene on. This subsequently reduced the set of 1093 genes to 98 genes based on nonzero importance scores, which were correlated with the response to PD-1 inhibitors and ranked according to their importance [62]. The DeepImmunoGene model was then trained using this refined set of 98 genes. Compared to our previous models, DeepImmunoGene demonstrated superior performance and robustness across all metrics (Table 2), indicating that the

application of permutation importance effectively eliminated irrelevant, noisy genes, allowing the model to focus exclusively on the most relevant genes without interference during training, such as overfitting. However, we also observed that specificity was consistently slightly lower than recall across all models, indicating that the models had more difficulty discerning nonresponders. This suggests that nonresponders may not have responded to immunotherapy due to external factors, such as the tumor microenvironment, age, or gender [24]. The comparative analysis with traditional ML models using the 98-gene subset found through permutation importance validates the core framework of DeepImmunoGene. The results highlight a specific synergistic effect between our feature selection method and the DNN, which is critical for achieving superior predictive performance. Although reducing the feature set to 98 genes improved computation efficiency no less, the fact that SVM and XGBoost trained on this same reduced feature set still failed to achieve comparable performance suggests that the DNN is better suited to capture the complex, nonlinear relationships and subtle gene-gene interactions underlying the RNA-seq data. Ultimately, the strength of DeepImmunoGene lies in this integrative approach of first identifying the most influential genes for accurate prediction and then leveraging a sophisticated DL model to interpret their combined predictive signal.

Further analysis revealed that 36 genes were upregulated (LogFC>0) in patients who responded to PD-1 immunotherapy, whereas 62 genes were upregulated (LogFC<0) in nonresponders [63]. These results suggest that DeepImmunoGene could serve as a robust ML-based tool for predicting immunotherapy outcomes in patients with lung cancer. The identification of these genes linked to responders and nonresponders not only offers potential biomarkers for predicting immunotherapy success but also enhances our understanding of the molecular mechanisms underlying the immune response in cancer. This could help guide more personalized treatment strategies, ultimately reducing unnecessary side effects and financial burdens for patients and health care systems, as immunotherapy is currently administered without prior knowledge of its effectiveness or safety for each patient [24,26]. Recent studies showed that only approximately 25% of patients show a positive response to immunotherapy, as PD-1/PD-L1 expression is not a sufficient biomarker to select patients who are likely to benefit [25,26]. Therefore, in addition to PD-1/PD-L1 expressions, these genes could be used as clinically actionable biomarkers for predicting response to ICIs with high accuracy.

Finally, we externally validated the predictive biomarkers identified by DeepImmunoGene using an independent bulk RNA-seq dataset of patients with NSCLC treated with PD-1 inhibitors (GSE207422) [49]. Given the small size of the external validation cohort (n=24) and the notable class imbalance (17 responders vs 7 nonresponders), we anticipated limited statistical power to detect meaningful differences (67). Additionally, the dataset itself includes patients receiving PD-1 inhibitors in combination with chemotherapy, which introduces treatment heterogeneity that may cause much of the variations observed in the expression patterns. Despite these limitations inherent to the available data, our analysis found that 4 of 6 nonresponder-upregulated genes showed higher median expression in nonresponders, with 3 achieving statistically significant differences in the predicted direction ($P<.05$). Similarly, 4 of 6 responder-upregulated genes demonstrated higher median expression in responders, although none reached statistical significance. This partial agreement offers encouraging evidence that the model-identified biomarkers capture biologically meaningful expression trends even in an independent, clinically realistic cohort. While these results should be interpreted cautiously, given the small sample size, class imbalance, and treatment variability, they support the potential utility of these gene markers for predicting immunotherapy response. Future validation in larger, well-annotated cohorts with consistent PD-1 treatment protocols is warranted to confirm their clinical relevance, fully validate the model's predictive classification performance, and further refine the list of biomarkers.

To contextualize DeepImmunoGene among existing approaches, we compared our method to previously published biomarker studies in NSCLC using PD-1 datasets. For example, Hwang et al [64] developed immune gene signatures derived from small patient cohorts with a limited number of features, which can restrict the model's ability to generalize to diverse patient populations or capture variability in gene expression. In contrast, Ravi et al [65] applied regression-based linear models that assume compounding, independent effects of genes on treatment response, which may fail to capture complex, nonlinear gene-gene interactions. By leveraging a DNN architecture, DeepImmunoGene is designed to learn these nonlinear dependencies across large-scale gene expression data, enabling more comprehensive and potentially generalizable biomarker discovery for predicting immunotherapy response. Other approaches, such as Lee et al [66], propose an ensemble method incorporating different models for the classification from gene expression profiles and additional information. This adds informative features, which may not always be available; in contrast, DeepImmunoGene reduces the feature space of RNA sequencing, helping isolate and detect features that are more likely to carry correct information.

Conclusions

Our DeepImmunoGene predictive model identified 36 upregulated genes in patients with NSCLC who responded to PD-1 immunotherapy. Among these, the 10 most significant genes (GSTT2B, HMGA2, AC135050.2, ANKRD33B, MMP13, PLA2G2D, RASGEF1A, BIRC7, DCAF4L2, and CHMP7) may serve as potential genomic biomarkers for predicting which patients with NSCLC are most likely to respond to PD-1 immunotherapy. Our external validation on an independent cohort supported several of the model-identified biomarkers, demonstrating partial agreement with DeepImmunoGene's predicted expression patterns despite the small sample size and class imbalance. These findings offer a promising foundation for future research aiming to improve patient stratification for PD-1 immunotherapy. Further validation in larger, well-annotated datasets and biological systems is needed to confirm their correlation with PD-1 inhibitors, which could lead to the development of more personalized and effective immunotherapies for lung cancer. Although the

DeepImmunoGene model demonstrated promising predictive performance, this study has several limitations. First, the analysis was conducted on a relatively small cohort of 355 patients with lung cancer. Second, we relied on a single publicly available RNA-seq dataset, which limited our ability to perform external validation. Third, key demographic and clinical variables, such as cancer stage, NSCLC subtype, age, and sex, were not available in the dataset. These factors are known to influence both immune response and gene expression, and their absence restricts the model's robustness assessment across patient subgroups. As a result, we were unable to evaluate the potential influence of demographic biases on model predictions. Future work with more comprehensive and diverse datasets is essential to validate the model's generalizability and to assess its consistency across clinically relevant subpopulations. We plan to conduct a follow-up study using external datasets when available and collaborate with clinics to validate our findings and further refine the list of biomarkers.

We also acknowledge that more advanced DL models exist for this task. Future work will involve evaluating DeepImmunoGene against state-of-the-art architectures, incorporating multimodal data, and validating performance on larger and more diverse cohorts. In this study, while DeepImmunoGene demonstrated strong performance metrics, future research should focus on improving the model's robustness through external validation across diverse datasets, including those from different geographical regions, patient demographics, and cancer stages. This would help assess how well the model generalizes beyond the current cohort of 355 patients. Moreover, the bias-variance tradeoff is crucial in this context. Our current model, which is highly sophisticated (DNN), likely strikes a balance between bias and variance, but there may still be room for improvement. High bias could occur if the model is overly simplified, missing important patterns in the data, whereas high variance could result from overfitting the model to the training data, leading to poor performance on new, unseen data.

Funding

This research received no external funding. JMIR Publications provided APF support for the publication of this article.

Data Availability

The patient data used can be found from the Gene Expression Omnibus public database GEO Repository (accessed on August 26, 2024).

Conflicts of Interest

None declared.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021 Jan;71(1):7-33. [doi: 10.3322/caac.21654] [Medline: 33433946]
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023 Jan;73(1):17-48. [doi: 10.3322/caac.21763] [Medline: 36633525]
3. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263. [doi: 10.3322/caac.21834] [Medline: 38572751]
4. Schabath MB, Cote ML. Cancer progress and priorities: lung cancer. Cancer Epidemiol Biomarkers Prev 2019 Oct;28(10):1563-1579. [doi: 10.1158/1055-9965.EPI-19-0221] [Medline: 31575553]
5. Zappa C, Mousa SA. Non-small cell lung cancer: current treatment and future advances. Transl Lung Cancer Res 2016 Jun;5(3):288-300. [doi: 10.21037/tlcr.2016.06.07] [Medline: 27413711]
6. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc 2008 May;83(5):584-594. [doi: 10.1016/S0025-6196(11)60735-0]
7. Wen J, Fu JH, Zhang W, Guo M. Lung carcinoma signaling pathways activated by smoking. Chin J Cancer 2011 Aug;30(8):551-558. [doi: 10.5732/cjc.011.10059] [Medline: 21801603]
8. Anusewicz D, Orzechowska M, Bednarek AK. Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of Notch, Hedgehog, Wnt, and ErbB signalling. Sci Rep 2020 Dec 3;10(1):21128. [doi: 10.1038/s41598-020-77284-8] [Medline: 33273537]
9. Lahiri A, Maji A, Potdar PD, et al. Lung cancer immunotherapy: progress, pitfalls, and promises. Mol Cancer 2023 Feb 21;22(1):40. [doi: 10.1186/s12943-023-01740-y] [Medline: 36810079]
10. Mamdani H, Matosevic S, Khalid AB, Durm G, Jalal SI. Immunotherapy in lung cancer: current landscape and future directions. Front Immunol 2022;13:823618. [doi: 10.3389/fimmu.2022.823618] [Medline: 35222404]
11. Kang Y, Vijay S, Gujral TS. Deep neural network modeling identifies biomarkers of response to immune-checkpoint therapy. iScience 2022 May 20;25(5):104228. [doi: 10.1016/j.isci.2022.104228] [Medline: 35494249]
12. Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. Nat Rev Immunol 2020 Nov;20(11):651-668. [doi: 10.1038/s41577-020-0306-5] [Medline: 32433532]

13. Ishida Y, Agata Y, Shibahara K, Honjo T. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. EMBO J 1992 Nov;11(11):3887-3895. [doi: 10.1002/j.1460-2075.1992.tb05481.x] [Medline: 1396582]
14. Nishimura H, Nose M, Hiai H, Minato N, Honjo T. Development of lupus-like autoimmune diseases by disruption of the PD-1 gene encoding an ITIM motif-carrying immunoreceptor. Immunity 1999 Aug;11(2):141-151. [doi: 10.1016/s1074-7613(00)80089-8] [Medline: 10485649]
15. Zhang Y, Zhang Z. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. Cell Mol Immunol 2020 Aug;17(8):807-821. [doi: 10.1038/s41423-020-0488-6] [Medline: 32612154]
16. Pitter MR, Zou W. Uncovering the immunoregulatory function and therapeutic potential of the PD-1/PD-L1 axis in cancer. Cancer Res 2021 Oct 15;81(20):5141-5143. [doi: 10.1158/0008-5472.CAN-21-2926] [Medline: 34654698]
17. Iwai Y, Terawaki S, Honjo T. PD-1 blockade inhibits hematogenous spread of poorly immunogenic tumor cells by enhanced recruitment of effector T cells. Int Immunol 2005 Feb;17(2):133-144. [doi: 10.1093/intimm/dxh194] [Medline: 15611321]
18. Ng S, Masarone S, Watson D, Barnes MR. The benefits and pitfalls of machine learning for biomarker discovery. Cell Tissue Res 2023 Oct;394(1):17-31. [doi: 10.1007/s00441-023-03816-z] [Medline: 37498390]
19. Yang B, Liu C, Wu R, et al. Development and validation of a DeepSurv nomogram to predict survival outcomes and guide personalized adjuvant chemotherapy in non-small cell lung cancer. Front Oncol 2022;12:895014. [doi: 10.3389/fonc.2022.895014] [Medline: 35814402]
20. Lei J, Xu X, Xu J, et al. The predictive value of modified-DeepSurv in overall survivals of patients with lung cancer. iScience 2023 Nov 17;26(11):108200. [doi: 10.1016/j.isci.2023.108200] [Medline: 38033628]
21. Supriya K, Anitha A. Survival analysis of superficial bladder cancer patients using DeepSurv and Cox models. Presented at: 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE); Feb 22-23, 2024. [doi: 10.1109/ic-ETITE58242.2024.10493319]
22. Vanitha K, Manimaran A, Chokkanathan K, et al. Attention-based feature fusion with external attention transformers for breast cancer histopathology analysis. IEEE Access 2024;12:126296-126312. [doi: 10.1109/ACCESS.2024.3443126]
23. Souza MD, Ananth Prabhu G, Kumara V. Advanced breast cancer detection using Spatial Attention and Neural Architecture Search (SANAS-Net). SN Comput Sci 2025;6(1):1-12. [doi: 10.1007/s42979-024-03568-9] [Medline: 40092049]
24. Kang J, Lee JH, Cha H, et al. Systematic dissection of tumor-normal single-cell ecosystems across a thousand tumors of 30 cancer types. Nat Commun 2024 May 14;15(1):4067. [doi: 10.1038/s41467-024-48310-4] [Medline: 38744958]
25. Rossi G, Russo A, Tagliamento M, et al. Precision medicine for NSCLC in the era of immunotherapy: new biomarkers to select the most suitable treatment or the most suitable patient. Cancers (Basel) 2020 Apr 30;12(5):1125. [doi: 10.3390/cancers12051125] [Medline: 32365882]
26. Cho JH. Immunotherapy for non-small-cell lung cancer: current status and future obstacles. Immune Netw 2017 Dec;17(6):378-391. [doi: 10.4110/in.2017.17.6.378] [Medline: 29302251]
27. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. J Vis Exp 2021 Sep 18(175):e62528. [doi: 10.3791/62528] [Medline: 34605806]
28. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009 Jan;45(2):228-247. [doi: 10.1016/j.ejca.2008.10.026] [Medline: 19097774]
29. Progression-free survival. National Cancer Institute. 2024. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/progression-free-survival [accessed 2024-12-20]
30. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015 Apr 20;43(7):e47-e47. [doi: 10.1093/nar/gkv007] [Medline: 25605792]
31. Restrepo JC, Dueñas D, Corredor Z, Liscano Y. Advances in genomic data and biomarkers: revolutionizing NSCLC diagnosis and treatment. Cancers (Basel) 2023 Jul 3;15(13):3474. [doi: 10.3390/cancers15133474] [Medline: 37444584]
32. Simes RJ. Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. J Chronic Dis 1985;38(2):171-186. [doi: 10.1016/0021-9681(85)90090-6] [Medline: 3882734]
33. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
34. Yi H, Shiyu S, Xiusheng D, et al. A study on deep neural networks framework. Presented at: 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC); Oct 3-5, 2016; Xi'an, China p. 1519-1522. [doi: 10.1109/IMCEC.2016.7867471]
35. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform 2018 Mar 1;19(2):325-340. [doi: 10.1093/bib/bbw113] [Medline: 28011753]
36. Teli TA, Masoodi FS. Application of ML and DL on biological data. In: Applications of Machine Learning and Deep Learning on Biological Data: Taylor Francis; 2023:159-180. [doi: 10.1201/9781003328780-10]
37. Manakitsa N, Maraslidis GS, Moysis L, Fragulis GF. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. Technologies (Basel) 2024;12(2):15. [doi: 10.3390/technologies12020015]

38. Chen J, Hao L, Qian X, Lin L, Pan Y, Han X. Machine learning models based on immunological genes to predict the response to neoadjuvant therapy in breast cancer patients. Front Immunol 2022;13:948601. [doi: 10.3389/fimmu.2022.948601] [Medline: 35935976]
39. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing 2020 Sep;408:189-215. [doi: 10.1016/j.neucom.2019.10.118]
40. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics 2018;15(1):41-51. [doi: 10.21873/cgp.20063] [Medline: 29275361]
41. Mesut B, Başkor A, Buket Aksu N. Role of artificial intelligence in quality profiling and optimization of drug products. In: A Handbook of Artificial Intelligence in Drug Delivery: Elsevier; 2023:35-54. [doi: 10.1016/B978-0-323-89925-3.00003-4]
42. Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digit Signal Process 2018 Feb;73:1-15. [doi: 10.1016/j.dsp.2017.10.011]
43. Ye J, Wang S, Yang X, Tang X. Gene prediction of aging-related diseases based on DNN and Mashup. BMC Bioinformatics 2021 Dec 17;22(1):597. [doi: 10.1186/s12859-021-04518-5] [Medline: 34920719]
44. Sukhdeve SR, Sukhdeve SS. Google Colaboratory. In: Google Cloud Platform for Data Science: Springer; 2023:11-34. [doi: 10.1007/978-1-4842-9688-2_2]
45. Mei X, Brei N, Lawrence D. Towards high-performance AI4NP applications on modern GPU platforms. EPJ Web of Conf 2024;295:11023. [doi: 10.1051/epjconf/202429511023]
46. Ayalew AM, Salau AO, Tamyalew Y, Abeje BT, Woreta N. X-Ray image-based COVID-19 detection using deep learning. Multimed Tools Appl 2023 Apr 26;82:1-19. [doi: 10.1007/s11042-023-15389-8] [Medline: 37362655]
47. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep 2022 Apr 8;12(1):5979. [doi: 10.1038/s41598-022-09954-8] [Medline: 35395867]
48. Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr 2011 Apr;48(4):277-287. [doi: 10.1007/s13312-011-0055-4] [Medline: 21532099]
49. Hu J, Zhang L, Xia H, et al. Tumor microenvironment remodeling after neoadjuvant immunotherapy in non-small cell lung cancer revealed by single-cell RNA sequencing. Genome Med 2023 Mar 3;15(1):14. [doi: 10.1186/s13073-023-01164-9] [Medline: 36869384]
50. Wickham H. Data analysis. In: ggplot2: Elegant Graphics for Data Analysis: Springer; 2016:189-211. [doi: 10.1007/978-3-319-24277-4_9]
51. Liu K, Yu Q, Li H, et al. BIRC7 promotes epithelial-mesenchymal transition and metastasis in papillary thyroid carcinoma through restraining autophagy. Am J Cancer Res 2020;10(1):78-94. [Medline: 32064154]
52. Wang H, Jiang Z, Chen H, Wu X, Xiang J, Peng J. MicroRNA-495 inhibits gastric cancer cell migration and invasion possibly via targeting High Mobility Group AT-Hook 2 (HMGA2). Med Sci Monit 2017 Feb 4;23:640-648. [doi: 10.12659/msm.898740] [Medline: 28159956]
53. Salucci S, Aramini B, Bartoletti-Stella A, et al. Phospholipase family enzymes in lung cancer: looking for novel therapeutic approaches. Cancers (Basel) 2023 Jun 19;15(12):3245. [doi: 10.3390/cancers15123245] [Medline: 37370855]
54. Hsu CP, Shen GH, Ko JL. Matrix metalloproteinase-13 expression is associated with bone marrow microinvolvement and prognosis in non-small cell lung cancer. Lung Cancer (Auckl) 2006 Jun;52(3):349-357. [doi: 10.1016/j.lungcan.2006.01.011] [Medline: 16569461]
55. Li S, Pritchard DM, Yu LG. Regulation and function of matrix metalloproteinase-13 in cancer progression and metastasis. Cancers (Basel) 2022 Jul 3;14(13):3263. [doi: 10.3390/cancers14133263] [Medline: 35805035]
56. Wang X, Wang J, Zhao J, Wang H, Chen J, Wu J. HMGA2 facilitates colorectal cancer progression via STAT3-mediated tumor-associated macrophage recruitment. Theranostics 2022;12(2):963-975. [doi: 10.7150/thno.65411] [Medline: 34976223]
57. Ma Q, Ye S, Liu H, Zhao Y, Mao Y, Zhang W. HMGA2 promotes cancer metastasis by regulating epithelial-mesenchymal transition. Front Oncol 2024;14:1320887. [doi: 10.3389/fonc.2024.1320887] [Medline: 38361784]
58. Altieri B, Sbiera S, Della Casa S, et al. Livin/BIRC7 expression as malignancy marker in adrenocortical tumors. Oncotarget 2017 Feb 7;8(6):9323-9338. [doi: 10.18632/oncotarget.14067] [Medline: 28030838]
59. Liu H, Xu R, Gao C, et al. Metabolic molecule PLA2G2D is a potential prognostic biomarker correlating with immune cell infiltration and the expression of immune checkpoint genes in cervical squamous cell carcinoma. Front Oncol 2021;11:755668. [doi: 10.3389/fonc.2021.755668] [Medline: 34733790]
60. Zeng Z, Mao C, Vo A, et al. Deep learning for cancer type classification and driver gene identification. BMC Bioinformatics 2021 Oct 25;22(Suppl 4):491. [doi: 10.1186/s12859-021-04400-4] [Medline: 34689757]
61. Li Q, Yang H, Wang P, Liu X, Lv K, Ye M. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. J Transl Med 2022 Apr 18;20(1):177. [doi: 10.1186/s12967-022-03369-9] [Medline: 35436939]
62. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010 May 15;26(10):1340-1347. [doi: 10.1093/bioinformatics/btq134] [Medline: 20385727]

63. Yu K, Zhang D, Yao Q, et al. Identification of functional genes regulating gastric cancer progression using integrated bioinformatics analysis. World J Clin Cases 2023 Jul 26;11(21):5023-5034. [doi: 10.12998/wjcc.v11.i21.5023] [Medline: 37583848]
64. Hwang S, Kwon AY, Jeong JY, et al. Immune gene signatures for predicting durable clinical benefit of anti-PD-1 immunotherapy in patients with non-small cell lung cancer. Sci Rep 2020;10(1):5721. [doi: 10.1038/s41598-019-57218-9]
65. Ravi A, Hellmann MD, Arniella MB, et al. Genomic and transcriptomic analysis of checkpoint blockade response in advanced non-small cell lung cancer. Nat Genet 2023 May;55(5):807-819. [doi: 10.1038/s41588-023-01355-5] [Medline: 37024582]
66. Lee K, Cha H, Kim J, et al. Dissecting transcriptome signals of anti-PD-1 response in lung adenocarcinoma. Sci Rep 2024 Sep 10;14(1):21096. [doi: 10.1038/s41598-024-72108-5] [Medline: 39256604]

Abbreviations

AUC: area under the receiver operating characteristics curve
DEG: differentially expressed gene
DL: deep learning
DNN: deep neural network
GEO: Gene Expression Omnibus
ICI: immune checkpoint inhibitor
LogFC: log fold changes
ML: machine learning
NSCLC: non–small cell lung cancer
PD-1: programmed cell death receptor-1
PD-L1: programmed cell death-ligand 1
RECIST: Response Evaluation Criteria in Solid Tumors
RQ: research question
SCLC: small cell lung cancer
SVM: support vector machine
TPM: transcripts per million
XGBoost: extreme gradient boosting

Edited by A Uzun; submitted 24.Dec.2024; peer-reviewed by KK Raja, PB Chandrashekar; revised version received 02.Sep.2025; accepted 04.Oct.2025; published 13.Jan.2026.

Please cite as:
Mubarak R, Anik FI, Rodriguez JT, Sakib N, Rahman MA
Unpacking Genomic Biomarkers for Programmed Cell Death Receptor-1 Immunotherapy Success in Non–Small Cell Lung Cancer Using Deep Neural Networks: Quantitative Study
JMIR Bioinform Biotech 2026;7:e70553
URL: https://bioinform.jmir.org/2026/1/e70553
doi:10.2196/70553

Random Survival Forest Versus Elastic-Net Regularized Cox Regression for Survival Prediction in Acute Myeloid Leukemia at Distinct Treatment Time Points: Model Performance Comparison Study

Oisín Brady[1,2], BSc; Sean Johnson[2], DPhil; Peter Giles[2], PhD; Caroline Alvares[2], PhD; Joanna Zabkiewicz[2], PhD; Carolina Fuentes[1], PhD

[1]School of Computer Science and Informatics, Cardiff University, Abacws, Senghennydd Road, Cardiff, United Kingdom

[2]School of Medicine, Cardiff University, Cardiff, United Kingdom

Corresponding Author:
Oisín Brady, BSc
School of Computer Science and Informatics, Cardiff University, Abacws, Senghennydd Road, Cardiff, United Kingdom

Abstract

Background: Risk group stratification based on the prediction of survival of patients with acute myeloid leukemia (AML) is complex. Despite common risk group categorization guidelines, the overall prognosis remains poor. Machine learning techniques have been shown to provide more accurate risk group stratification than conventional approaches using trial data. However, many time-to-event (TTE) models do not use training sets constrained to specific time windows, instead using aggregations of trial data.

Objective: This study aimed to evaluate the performance of (1) random survival forest (RSF) and (2) Cox proportional hazard regression with elastic net regularization (CoxNet) for survival prediction of patients with AML within a censoring window trained with available data recorded at discrete time points during the United Kingdom National Cancer Research Institute Acute Myeloid Leukaemia 17 randomized controlled trial (AML17).

Methods: For each stage in the AML17 trial, separate models were trained for each exhaustive k-choice combination of available AML17 data subsets. Data combinations for each model were further constrained according to the respective trial stage to avoid data leakage. Preliminary Pearson correlation methods were used to remove directly correlating features with the TTE prediction (time-to-death/5-y censoring point). Repeated k-fold stratified cross-validation was used on each dataset ablation to find candidate models. Permutation importance and elastic net regularization were used to monitor stability across validation folds and reduce the feature set of the highest performing stage RSF and Cox proportional hazard regression models, respectively. Finally, selected ablated models were re-evaluated using the nested, k-fold, stratified sampling cross-validation method with bootstrapping.

Results: Concordance index ranked the best models for data constricted up to the end of induction (RSF=0.68, CoxNet=0.67), stages 1 (RSF=0.69, CoxNet=0.68), 2 (RSF=0.68, CoxNet=0.66), and 3 (RSF=0.69, CoxNet=0.63) of the trial.

Conclusions: This study details the high prediction accuracy for time-to-survival-event predictions when training sets of CoxNet and RSF models, which are sequentially constricted to data measured up to the end of respective AML17 trial stages. The performance of these sequential TTE models is intended to justify their use as part of a wider digital twin system simulating multiple TTE outcomes for patients with AML.

(JMIR Bioinform Biotech 2026;7:e75678) doi:10.2196/75678

KEYWORDS

acute myeloid leukaemia; AML17; time-to-event; survival prediction; digital twin; random survival forest; cox proportional hazard regression; elastic net

Introduction

Cancer is an enigmatic issue with its pervasiveness seemingly as wide as the depth of its biological origins. As a group of diseases, cancer is highly variable in form, with roughly 200 types according to the National Cancer Institute [1]. One such type of cancer, acute myeloid leukemia (AML), occurs from genetic abnormalities in precursory cells responsible for differentiating into mature platelets, white blood cells, and red blood cells. Genetically, AML is highly stratified, with multiple potential mutation points in the lineage of precursory cells responsible for mature blood cell types, such as within the earliest hematopoietic progenitor cell stage, or by mutation of

immature intermediate blast cells [2]. Ultimately, AML causes uncontrolled proliferation of immature or nonfunctional blood cells, leading to systemic immune dysfunction and organ complications. AML is often subcategorized into primary (or de novo) AML and secondary AML; secondary AML is further subdivided into therapy-related AML or acute myeloid leukemia derived from antecedent hematological disorders, such as myelodysplastic syndrome [3]. Multiple genetic and environmental factors contribute to the cause and progression of AML, such as trisomy 21 (Down syndrome) [4], Fanconi anemia [5], Nucleophosmin 1 (NPM1) [6], or FMS-like (Feline McDonough Sarcoma [7]) tyrosine kinase 3 (FLT3) [8] mutations. External factors associated with AML include prolonged exposure to benzene [9], history of smoking [10], or cancer therapy–induced AML [11]. This heterogeneous nature of AML attributes to its difficulty to precisely diagnose and treat effectively [12].

Deriving accurate survival predictions of patients with AML is an important foundational step in establishing risk groups upon which accurate treatment methods can be produced. Resultingly, there is a great emphasis on risk group stratification of patients with AML as a precursory step for optimized resource allocations and identification of biomarkers contributing to treatment response, as seen in the European LeukemiaNet (ELN) project [13] or the World Health Organization (WHO) risk grading of AML [14]. Within recent years, the availability of genomic data through next-generation sequencing (NGS) [15] techniques combined with randomized controlled trial datasets, such as “the United Kingdom National Cancer Research Institute Acute Myeloid Leukaemia 17 (AML17) trial” [16-18], offers a wealth of information to make inferences on the disease and improved treatments. Indeed, 5-year overall survival rates for AML have improved, from 13% in 1970 to 55% (P<.001) in 2010 for patients younger than the age of 60 years at the MD Anderson Cancer Center hospital and from 8% in 1970 to 17% for those older than 60 years [19]. The Surveillance, Epidemiology, and End Results program based in the United States estimates the 5-year overall survival rate of patients with AML to be 31.9% based on survival data between 2014 and 2020, and estimates 20,800 new cases of the disease, constituting approximately 1% of new cancer cases in the United States in 2024, with 11,220 AML-related deaths [20].

Despite improvements in outcome, it is apparent that traditional hierarchical approaches for risk grouping are not able to capture the full complexity involved in stratification of AML [21], shown by the still dismal net survival rate of 13.6%, 5 years after diagnosis in England [22]. With the sheer quantity of data now available from NGS and randomized controlled trial databases, alternative machine learning (ML) techniques used within oncology [23] have been shown to capture complex features stratifying patients with AML [24,25]. The United Kingdom National Cancer Research Institute (UK-NCRI) AML17 trial contains detailed records of clinical, longitudinal minimal residual disease (MRD) reports and genetic sequence mutation profiles of approximately 3142 patients with AML younger than 60 years between 2007 and 2014. Such a large, time-based dataset offers an ideal training set for ML models to capture complex risk stratification of the disease. The original AML17 protocol used standard statistical methods, such as the log-rank test for time-to-event (TTE) outcomes (survival for all randomizations), Mantel-Haenszel tests for dichotomous outcomes, and Wilcoxon rank-sum and t tests for resource usage data. It has more recently been shown that data from AML17 can be used for highly accurate ML risk group stratification based on survival prediction. Tazi et al [26] applied several ML models trained on demographic, diagnostic, and genetic variables from several UK-NCRI AML trials, including AML17. By fitting models to predict overall survival via TTE of patient death up to censoring points, patients could be stratified by predicted survival risk measurements and separated into distinct groups based on delineating features. When compared with the ELN guideline, this new framework restratified 1 in 4 patients, with significantly improved prognostic accuracy. Another study using the following AML18 trial [27] used a random survival forest (RSF) model to update risk group stratification categories based on overall survival using age, sex, white blood cell count, gene mutations, and cytogenetic abnormalities of patients. Subsequently, numerous patients were restratified from risk groups in the standard 2022 ELN guideline, which could be used to retrospectively identify more optimal treatment paths [28].

Several ML models are specifically designed or adapted for TTE outcome prediction using right-censored data [29,30] as seen in AML trial datasets.

One such ML model, an adaptation of the random forest algorithm [31], RSF [32], as previously mentioned, has been used for time-dependent survival predictions, which excludes the proportional hazards assumption of statistical Cox proportional hazard regression (CPHR) models. CPHR is a statistical model commonly used across multiple scientific domains, including cancer research [33] for TTE predictions. In the case of AML, where multiple interacting and time-dependent biomarkers affect survival outcome [12], collinear features can negatively impact prediction accuracy when the independent features and proportional hazards assumptions of CPHR models are violated [34]. In such cases, the standard CPHR model can be adapted using regularization techniques such as the “Elastic Net” method (also known as “CoxNet” [Cox proportional hazard regression with elastic net regularization]) [35].

Performance between the 2 models varies depending on the datasets and implementation. Several instances in literature show that RSF prediction is comparable with or even outperforms CPHR models [36-38]. However, the converse is also referenced [39-41], suggesting that the application of these models is highly dependent on initial training datasets, preprocessing, model building methodologies, and the overall complexity of the predicted outcome. Pickett et al [42] conclude that RSF performs best when leveraging its nonlinear nature with multiple, longitudinal data points, many of which have unknown levels of significance.

None of the ML-based studies reviewed involved static and longitudinal training sets that were constricted according to trial time frames and sequentially exposed to more data, instead using an aggregation of data. This study seeks to investigate

the predictive performance of individual TTE models, beginning with 5-year survival status, which are sequentially exposed only to data available up to the conclusion of major time points in the AML17 trial. RSF and CPHR with Elastic Net have been chosen as TTE predictive models given their previous application in this context. A pipeline involving necessary data preprocessing, feature selection, and hyperparameter tuning used to build each will be detailed. Finally, after evaluation using the primary concordance index (c-index) metric alongside additional dynamic area under the receiver operating characteristic curve (also known as dynamic AUC) and Brier loss scores, the optimal models for survival prediction at select trial stages will be selected for future analysis. Future studies will analyze select model feature importance and significance with respect to state-of-the-art literature on AML risk stratification. The generalized pipeline will also serve as a template that can be adapted for additional TTE predictions other than death status. In the wider context of a "digital twin" [43] system, this multiple time-constrained model approach could provide accurate simulations of a wide variety of patient outcomes, not solely focused on risk stratification but also patient quality of life (QoL) and optimized care for additional comorbidities during treatment.

Methods

AML17 Trial Data

Data are sourced from the AML17 [17] drug trial for patients younger than 60 years, which includes 3142 clinical records. Patient clinical records are combined with MRD (n=2587), NGS mutation profiles (n=3579), and a separate collection of *NPM*1 [6] and *FLT*3 specific mutation profiles (n=3142) [44]. A pseudonymization process converted patient trial IDs into dummy IDs before data access, ensuring compliance with participant privacy and data protection regulations. Access to the pseudonymized dataset was stored on the Cardiff University Research Data Store [45] with access restricted to authorized researchers. Clinical records contain measures at induction, including previous blood disorders, height, weight, the French-American-British 8-category AML classification [46], WHO and Eastern Cooperative Oncology Group performance status [47], cytogenetic, karyotype, ethnic background, and more. Metadata references to all used data are available in Multimedia Appendices 1-4.

After early diagnostic and comorbidity measurements during the induction stage, the proceeding 4 stages contain longitudinal records on periodic treatment, response, toxicity, and supportive care. The MRD subset includes quantitative polymerase chain reaction on peripheral blood and bone marrow samples, and multiparameter flow cytometry using the leukemia-associated immunophenotype and "different from normal" techniques. MRD measurements are longitudinal; the trial protocol involved readings at the end of each major trial stage investigated in this study. AML17 collected FLT3 and NPM1 mutation profiles with automated flow cytometry and manual bone marrow and peripheral blood cytology measurements at days 2 - 3 from patient induction to the trial.

Ethical Considerations

Access to data from the UK-NCRI AML17 clinical trial was provided by the Cardiff University Centre for Trials Research [48], which curates and governs the trial database.

All data were pseudonymized before being released to the research team. No direct patient identifiers were included in the dataset. All analyses were conducted on secure Cardiff University computing infrastructure. Ethical approval for the use of these data was granted by the Cardiff University School of Computer Science and Informatics Research Ethics Committee on February 28, 2025 (approval COMS/Ethics/2024/014). The research used secondary analysis of previously collected clinical trial data and did not involve direct contact with participants. Participants in the original AML17 trial provided written informed consent for their data to be used for research purposes. No compensation was provided to participants for this study, as it involved secondary analysis of previously collected trial data.

Data Cleaning

Overview

Paper Case Report Forms recorded data throughout the AML17 trial. An initial screening process of longitudinal data found value errors in the exported dataset, predominantly within date fields. The following sections detail the conditions for sample exclusions based on erroneous record entries.

Erroneous Record Removals

Most detectable erroneous values are date records, leading to the exclusion of 85 patient records with date values written outside of the trial time bounds between 2007 and 2014 (excluding annual follow-up dates that proceed after the official trial end date, ie, July 31, 2014) or with nonsequential or otherwise nonchronological trial stage entry times, likely due to data entry errors or outstanding queries with sites at trial closure. The exclusion of nonsequential course start dates removed 19 patient records from the study. After all initial exclusions, 3057 AML patients remained eligible for model training.

Feature Removals Before Feature Selection

The pseudonymized dummy ID was dropped before model training to avoid spurious correlation from potential protocol batch induction bias. Other exclusions include nonstandardized clinician notes, making them highly varied text fields, which are not immediately processable by RSF and CPHR models. Traditional preprocessing methodologies, such as dummy encoding, would introduce many Boolean representations of these features, most of which, given their variability, would have occurrences recorded seldomly, increasing data sparsity and potentially increasing model risk of overfitting [49]. Consequently, categorical or continuous features from these columns cannot be immediately cataloged. Information held in these fields is of potential clinical and ML model importance. However, additional preprocessing techniques are needed to scrape the potentially multiple continuous and categorical features existing in a single record. Data capture would also need to handle the detection of differently written versions of

the same category (eg, syntactic, spelling, or grammatical variations). Data mining using ML techniques, such as the usage of natural language processors [50], may be explored for text classification of these fields in future studies. A breakdown of such excluded features is available in Multimedia Appendix 5.

Preprocessing

Overview

Clinical, NGS, FLT-3, and MRD status records initially existed as separate pseudonymized comma-separated values (csv) files exported from the original AML17 trial database within the Cardiff University Centre for Trials Research. Each file was merged using the pseudonymized "DummyID" of each patient to produce unique patient records of all exported csv data. The training set was initialized using the Python "Pandas" [51] library "DataFrame" object [51], storing merged patient records. The data type was specified for each column and programmatically converted individual records that violate the expectation, if possible; otherwise, the patient record was dropped. This ensured that each feature vector is readable to the applied CoxNet and RSF models. The following sections define the data preparation steps necessary for model training.

Feature Set Data Representations

The AML17 dataset contains 2 general data types—continuous and categorical. Continuous features were scaled using Sci-Kit Learn's "StandardScaler" function [52], which computes a standard score of each sample based on its variance from the mean of the feature vector. The standardized continuous Unix Epoch time is used to represent all instances of date fields. Unix Epoch time denotes the total nonleap seconds elapsed since 00:00:00 UTC on January 1, 1970 [53]. While scaling is not a requirement for tree-based ensemble models, such as RSF, whose results are insensitive to the transformation [54], applying it to continuous variables avoids scenarios where features that are orders of magnitude higher than others influence the objective function disproportionately within CPHR models. This also standardizes data representations for training sets of both models. Dummy encoding is used to convert categorical features into discrete Boolean features for each level, which is readable to the CPHR model.

Given that the AML17 dataset includes many categorical features, dummy encoding drastically increases the total number of features fed into both models, increasing the potential complexity of the model and potentially introducing overfitting. Feature reduction techniques are used to attenuate this. For RSF, the permutation importance [55] technique is used to quantify feature importance. This involves randomly shuffling feature values a set number of times and measuring the effect on model performance each time through the c-index evaluation. Degradation of performance when changing these values indicates the RSF's relative reliance on a particular feature. The combination of 2 regularization methods, known as elastic net, was used to reduce the CPHR feature set. This combines $\ell 1$and $\ell 2$ regularization methods [35], simultaneously handling issues of collinearity within the dataset as well as feature reduction.

TTE Predictor Variable

CoxNet, RSF, and similar models used for time-based prediction, such as support vector machines for survival [56,57], use a target variable known as the "time-to-event" (TTE) variable, formalized by Equation 1 :

$$(1) y=\min(t,c)=\{t \text{ if } \delta=1, c \text{ if } \delta=0$$

where δ is a Boolean value representing event occurrence (in this instance, patient death), t being the time range from patient trial induction to the event, and c being the time range from patient trial induction to the censoring threshold.

Equation 1 is modified to shift the censoring window of patient records to 5 years from induction, formalized as Equation 2:

$$(2) y=\{t \text{ if } \delta=1 \wedge t<c, c \text{ otherwise}$$

This accounts for instances of patients who have died on either side of a 5-year censoring window, a threshold seen in follow-up analysis of AML17 [58]. A 5-year cutoff point provided a more even distribution of noncensored patients than lower thresholds for TTE prediction models. Clinical variables define the TTE variable as a tuple: $(\delta,(t|c))$, with inclusion of t or c depending on the censoring status of the event as described in equation 2. Both the ablation and final model tuning methods stratify each fold to ensure that the distributions of TTE indicators are approximately equally distributed with respect to the entire cohort to avoid nonrepresentative sampling issues.

Trial-Stage Sensitive Ablation Study

To determine what broad groups of initial training datasets were most important for each model, we conducted an ablation study that selected possible combinations of NGS, FLT3 and NPM1, clinical, and MRD data subsets. C-index evaluations of tuned and validated models ranked the combination relative importance. Permutation importance [55] and nonzero coefficient values determined individual feature importance for RSF and CoxNet models, respectively. Multiple RSF and CoxNet models predicted the TTE target after being trained on subsets of patients surviving up to the end of each AML17 course stage, that is, induction, C1, C2, and C3. We trained models at each stage on varying degrees of information based on precomputed data subset combinations. The data pulled from clinical and MRD subsets were constrained such that each model only had access to features recorded up to the end of their respective course stage to avoid potential data leakage and provide predictions using data measurements only available up to specific trial time points. We define the set of training data combinations as all possible K-choice, non–order-specific, nonrepeating items at each trial stage, formalized as:

$$(3) \sum_{c=1}^{5} \sum_{k=1}^{3} 3!/(3-k)!k!=35$$

where c represents the current trial stage (induction, C1, …, and C3), and k the number of selected data sources.

The protocol recorded FLT3, NPM1, and a broader collection of NGS mutation panel measurements for patients 2 - 3 days after their randomized induction to the trial; therefore, not necessitating further constriction, as all-time points are set after recording of these data.

Direct Pearson linear correlations to death status determined the exclusion of features from all data combinations. Likewise, we excluded categorical features with options that inferred patient death status. Excluded features are detailed in Multimedia Appendices 1-3. In total, with 4 time point stages (induction, C1, C2, and C3), the total k-choice combinations across each stage equaled 35 training sets. These trained a combined total of 70 CoxNet and RSF models.

Model Building

Overview

The model building process is divided into four major phases, that are (1) preprocessing; (2) ablation study; (3) final evaluation, based on the highest performing candidate ablations for RSF and CoxNet models at each trial stage; and (4) a baseline risk model comparison, which compares RSF and CoxNet c-indices against a standard Cox model used within the trial protocol to stratify patients post stage 1.

Preprocessing Phase

The following defines the overall steps involved in the preprocessing pipeline on data that are made consistent between both of the following phases. Cleaning methods detailed in earlier sections have been excluded for simplicity (refer to Data Cleaning section). The preprocessing pipeline is called and fit to data only available within the scope of the fold used within cross-validation (CV) of the ablation study and final model evaluation phases.

1. Drop all features with total missing entries >95%.
2. Create missing indicator features for each feature with at least 1 occurrence of a missing value.
3. Flatten the NGS data subset of gene mutation entries, dummy encode, and merge with the rest of the combined dataset (clinical, MRD, FLT3, and NPM1 subsets). Encoding is based only on available samples within the fold to avoid potential leakage of nonfold sample gene mutation entries.
4. Dummy encode karyotype features, labeling rare entries (n<5) to a “rare_class” category to avoid dimensionality explosions.
5. Dummy encode all other standard categorical features.
6. Preserve the ordinality of the identified ordinal record by keeping them as individual features, using predefined integer mappings from the protocol. Missing features are labeled with the sentinel value −999 consistently outside of all ordinal ranges.
7. Scale identified numerical features by removing the mean and scaling to unit variance (using Sci-Kit Learn’s StandardScaler)

Ablation Study Phase

The goal of this phase is to act as a sensitivity analysis of the major data sources available from the AML17 trial. By using every possible combination of these data sources (refer to equation 3), the most influential data can be determined using the average c-index performance. This, in turn, acts as a feature reduction step and ensures selected ablations for downstream analysis are using features of importance relative to the specific model and respective trial stage.

The following steps define the preliminary ablation study phase. For each ablation at each trial stage, the cohort applies repeated (n=3), stratified, 5-fold CV using a consistent random state seed for reproduction.

1. Feature preprocessing phase pipeline fit to the training set and transformed on the train and validation set within the fold scope.
2. Train a baseline RSF model on the fold’s training set.
3. Apply permutation importance on the baseline model (using 150 ensemble estimators and a consistent randomized state seed), recording feature stability per repeated fold.
4. Measure the c-index and inverse probability of censoring weights (IPCW) c-index of the baseline ablation RSF model.
5. For the same fold repetition, train a baseline CoxNet model and record feature stability using model coefficient values.

Final Evaluation Phase

The following steps define the final model-building and evaluation phase:

1. At each stage, select a CoxNet and RSF ablation model with the highest recorded average c-index across all repeated CV folds from the ablation study.
2. For each selected model, using their respective ablated dataset and trial cohort, use nested k-fold, stratified CV, with sample shuffling and the same randomization seed used consistently across all experiments.
    a. The outer loop is reserved for unbiased performance estimation across 10 folds of the training set.
    b. The inner loop is reserved for hyperparameter tuning across 3 folds. Validation samples are never included in model training in their respective loop, ensuring strict separation between training and validation data to avoid bias or overfitting. Grid search spaces for hyperparameter tuning of the models are:
        i. RSF: ‘n_estimators’ = [500, 750, 1000, 1250, 1500]
        ii. RSF: ‘max_features’ = [‘sqrt,’ ‘log2,’ 0.33, 0.5]
        iii. RSF: ‘max_features’ = [3, 5, 10, 15]
        iv. CoxNet: ‘l1_ratio’ = [0.01, 0.25, 0.5, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99, 1.0]
    c. Record fold’s bootstrapped performance estimates across 250 bootstrapped samples (250 samples chosen as a compromise between sample variance for CI precision and processing time constraints with the already expensive nested CV operations). Metrics include c-index, IPCW c-index, dynamic AUC, and dynamic Brier loss.

As fold event times vary across bootstraps, dynamic AUC and Brier loss metrics for fold bootstrap samples are interpolated to the nearest predefined “universal” time points for consistency between selected samples across the validation process.

Baseline Risk Model Comparison Phase

For comparison with the Cox linear regression model used in the AML17 trial protocol for risk assessment used on the cohort ending their first stage of treatment, an additional *non-nested* grid search CV assessment on a 90%:10% train-test split with 1000 bootstrapped samples was conducted. We recorded dynamic AUC, Brier loss, IPCW c-index, and c-index scores of both models at this stage. The purpose of this analysis is to suggest improved c-index performance of the respective CoxNet and RSF stage models against the protocols model. The rationale for using larger sample sizes for this assessment was to provide a more realistic indication of the model's performance relative to the protocol's model when trained with a sample size closer to real-world cases. As the test set is smaller and overlaps with hyperparameter tuning, this assessment is inherently optimistic. Therefore, it remains that the nested CV results with bootstrapped CIs illustrated in the Final Evaluation Phase section remain as the primary, more sensitive, and pessimistic evaluation of generalized model performance.

Model Evaluations

Overview

The c-index evaluated model performance from a held-out test set not used in previous training. The standard (Harrel) c-index used for survival model evaluation is dependent on the distribution of the censored events. Therefore, to avoid potential bias, we recorded an alternative adapted form of c-index based on the IPCW. However, likely due to the inner handling of censoring by the models, differences in IPCW c-index and standard c-index were none, or at most minuscule, so standard c-index remained the primary performance metric for model assessments. Both standard and IPCW adapted c-index results for all final models have been included for transparency. Secondary performance metrics involved:

1. Dynamic AUC [59] assessed predictive performances across patients selected at discrete time points from the end of the model's respective trial stage to the 5-year censoring point.
2. A time-dependent Brier loss score, measuring mean square difference between predicted and real TTEs at iterative time points, indicated the models' calibration.

Cumulative-Dynamic AUC

This performance metric assesses the model's ability to discriminate between patients who experience an event before a specific time period (t), and those who experience an event after [59]. AUC ranges from 0 to 1 inclusively, with higher values indicating better discrimination between patient events before and after t.

Dynamic Brier Loss Score

This performance metric assesses how well a model is calibrated, evaluating how closely model predictions match the real labeled TTE variable of a patient at time point t, typically referred to as the "ground truth." This is done by evaluating the difference in mean square predicted event times and the ground truth TTE at t. Brier loss at t is measured between 0, for models with perfect accuracy, and 1, for perfect inaccuracy. The integrated Brier loss score can also be measured to evaluate overall model calibration throughout the 5-year period since patient induction.

Pipeline Summary

A diagrammatic summary of the model training pipeline, briefly describing the 4-phase process detailed before, is presented in Figure 1. Descriptions of CoxNet and RSF are detailed in Multimedia Appendix 6 and Multimedia Appendix 7. Code for relevant experiments is available in Multimedia Appendix 8.

Figure 1. Summary of the 4-phase pipeline process for survival time-to-event prediction model building at the AML17 trial stages. AML17: United Kingdom National Cancer Research Institute Acute Myeloid Leukaemia 17 randomized controlled trial; AUC: area under the receiver operating characteristic curve; c-index: concordance index; CoxNet: Cox proportional hazard regression with elastic net regularization; CV: cross-validation; FLT3: Feline McDonough sarcoma-Like Tyrosine kinase 3; MRD: minimal residual disease; NGS: next-generation sequencing; NPM1: Nucleophosmin 1; RSF: random survival forest.

Results

Overview

Results are composed of, first, event (patient death within censoring window) status and timing distributions are measured, describing the spread of target events across cohorts and time for RSF and CoxNet models. Second, a Kaplan-Meier survival curve is shown for each cohort, including the full AML17 cohort, after cleaning for erroneous TTE indicators described in the Data Cleaning section. These curves provide a visual summary of the baseline survival patterns between cohorts, which the stage-specific survival models are tasked with capturing. Third, feature missingness correlations are visualized using heatmaps for data sources across AML17 cohorts, highlighting potential missingness mechanisms and motivating the usage of missing indicator variables for models. Fourth, Table 1 reports feature set sizes before and after reduction steps for RSF and CoxNet models, showing how reduction methods remove redundant features, focusing on the most informative and stable predictors quantitatively selected during phase 3 of the methodology. Fifth, c-index, dynamic AUC, and dynamic Brier quantify ranking accuracy, time-dependent discrimination, and overall predictive accuracy for each stage, providing a suite of metrics for cross-reference with similar work and reproducibility. Finally, feature importance is visualized using Venn diagrams of overlapping top-ranked features between stage-specific RSF and CoxNet models, and vertical bar charts illustrate the relative importance of the top 30 highest-ranking predictors in each model.

Table . Feature set reductions for each RSF[a] and CoxNet[b] model at their corresponding trial stage.

Trial stage	Feature reduction (n before, n after, n after encoding)
Post induction	• RSF: 78, 70, 392 • CoxNet: 56, 45, 205
Post-C1	• RSF: 156, 126, 479 • CoxNet: 115, 85, 225
Post-C2	• RSF: 228, 173, 596 • CoxNet: 228, 168, 463
Post-C3	• RSF: 290, 142, 478 • CoxNet: 211, 174, 452

[a]RSF: random survival forest.

[b]CoxNet: Cox proportional hazard regression with elastic net regularization.

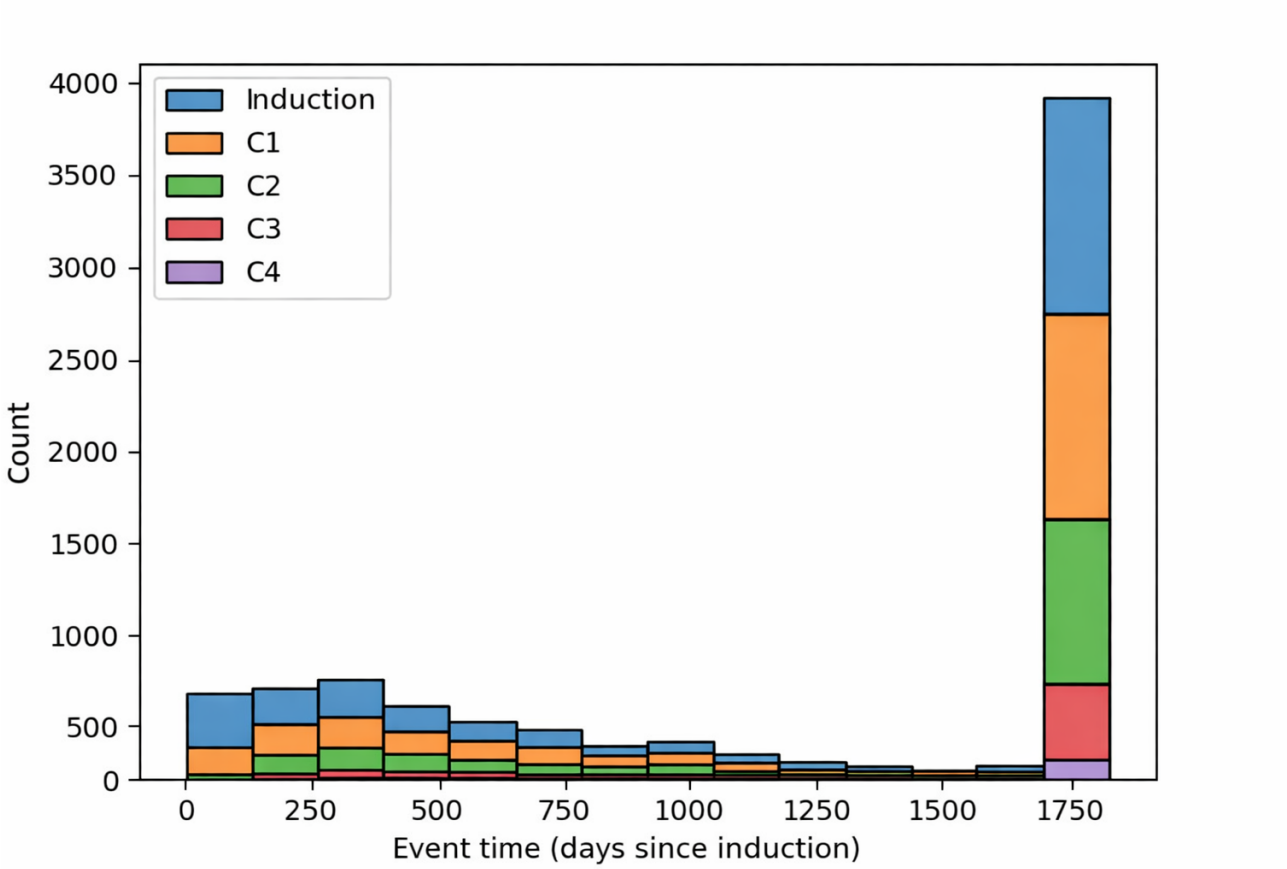
Target Class Distributions

TTE (event=death status) of both models is the predictor variable of both models. Class imbalances influence the overall performances of a model [60]. There were no major class distribution imbalances for each of the 5 trial stages used for model training. No training sets had minority class percentages <41% and the average minority class percentage was approximately 44% (induction=42.8%, C1=46.9%, C2=48.9%, C3=41.3%, C4=42.9%). In literature, there is no definitive threshold at which imbalances are considered severe enough to affect the performance of ML models. A rule of thumb is that an imbalance is considered “moderate” when minority classes are 1% - 20% of the dataset [61]. Since this was not the case, the use of over- or undersampling techniques or synthetic data generation techniques such as the Synthetic Minority Oversampling Technique [62] was deemed not necessary.

Event Time Distributions

Analysis of event time distributions in Figure 2 shows a disproportionate number of right-censored events for patient sets used for each of the 5 selected trial stages. This initially justified the usage of a c-index based on IPCW, which is specifically adapted for this situation [63] rather than the standard c-index, which is dependent on TTE distributions. However, it was found that differences between c-index and adapted IPCW c-index readings of CoxNet and RSF models were identical for stage models (eg, postinduction stage mean c-index=0.6760, mean IPCW c-index=0.6760). Full performance metric results across all stage ablation models are included in Multimedia Appendix 9.

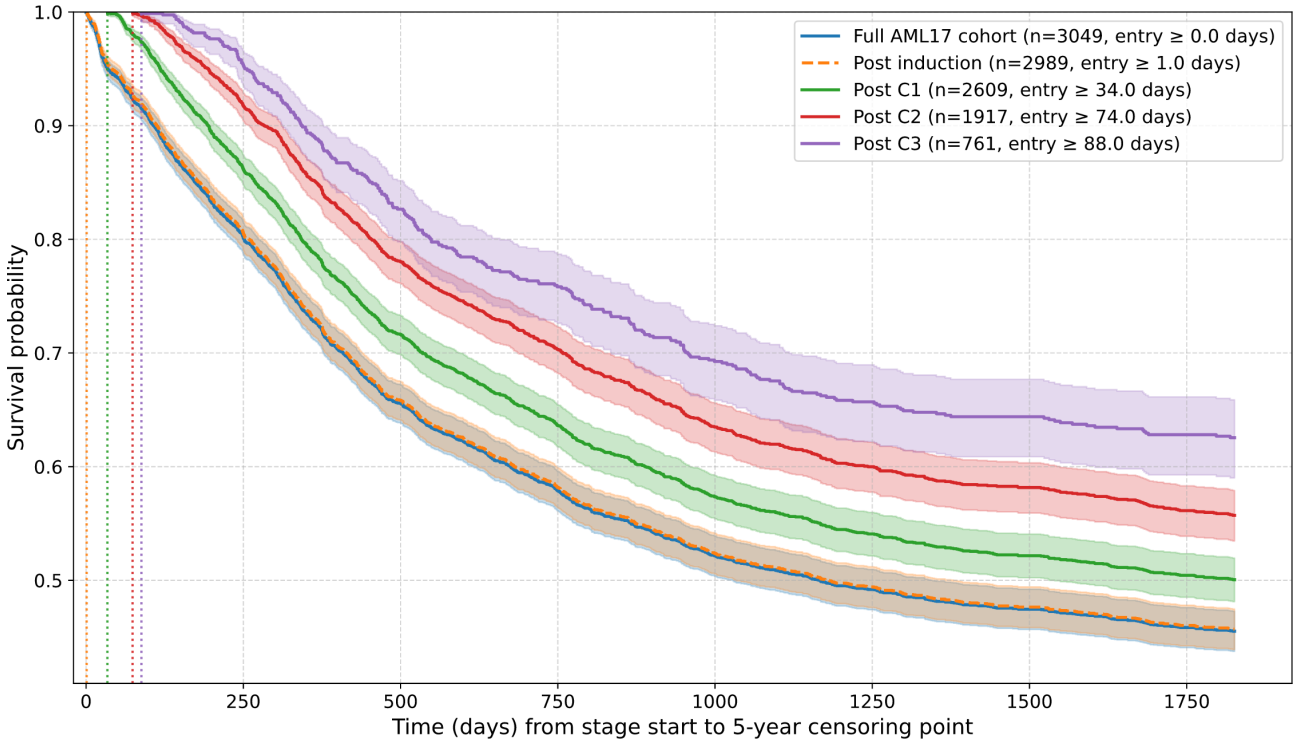
Figure 2. Distribution of time-to-event target variables in each of the four trial stages with censoring 5 years from patient induction. The distribution is irregular for all stages with the final histogram bin of induction, C1, C2, and C3 stages (C4 was excluded from further experiments because of small sample size concerns [n=177]).



Cohort Kaplan-Meier Survival Curves

Kaplan-Meier survival curves for each patient cohort for modeling show distinct survival patterns between trial stage cohorts, excluding the full cohort upon trial entry and the postinduction stage, as it began within 1 - 3 days according to the AML17 protocol. This gives an indication of the distinguishable baseline survival patterns that vary between AML17 stage cohorts, which each model is tasked to predict. For visual clarity, the posttreatment stage 4 cohort was excluded from Figure 3, as it overlaps across many cohorts. Note that the posttreatment stage 4 was excluded from further modeling due to small sample size concerns (n=177).

Figure 3. Kaplan-Meier survival curves for each cohort used by random survival forest and Cox proportional hazard regression with elastic net regularization models after data cleaning. The full AML17 cohort (post data cleaning) is also shown (blue). Shaded regions represent upper and lower CIs, solid lines (made dashed orange for postinduction for readability) represent average survival probability at the time point. Vertical dashed lines show trial stage cohort earliest start times since induction, with exact values shown in the legend.



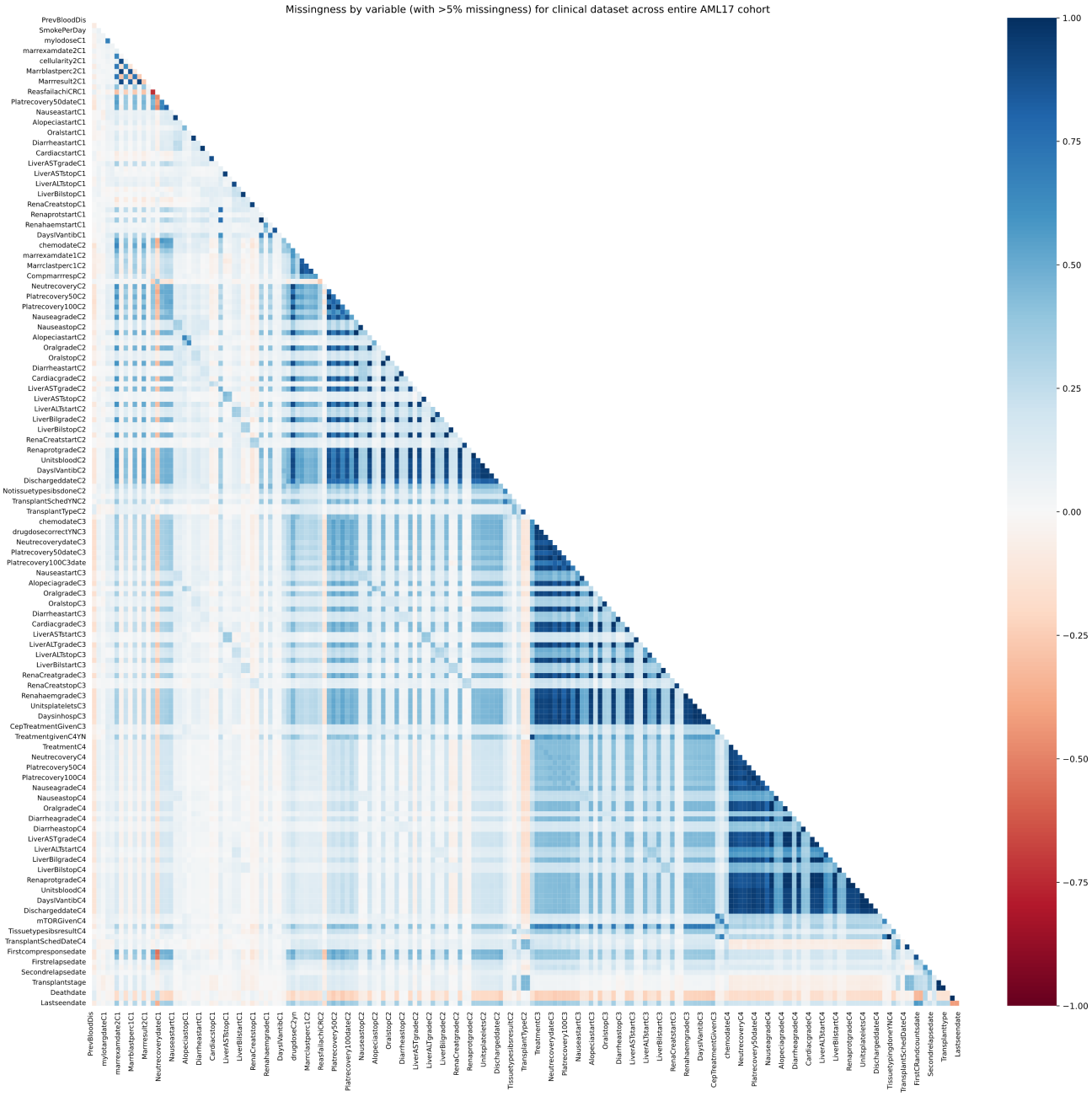
Data Missingness

Data missingness, nonexistent trial records, per trial stage cohort, and full AML17 cohort (post–data cleaning record exclusions) have been visualized using heatmaps for missingness correlation between features and frequency missingness plots for overall record missingness. Given the volume of total features, each figure has been further divided by data subset (eg, Clinical or NGS). The following matrix shows the missingness correlation between all clinical variables across the whole AML17 cohort used in this study. This provides justification for the usage of the missingness indicator features supplied as additional predictors for each model, which can be used to assess the informativeness of missingness and additional follow-up analysis. All additional figures have been included in Multimedia Appendix 10.

Across the entire AML17 cohort, the clinical subset has strong missingness correlations, suggesting a possible missing not at random mechanism for clinical data recorded at specific trial stages, indicated by the dark blue blocks of longitudinal features recorded at each trial stage in Figure 4. This is expected for any patients failing to proceed to a stage due to death or exclusion criteria.

When restricted to individual stage cohorts, the missingness correlation of clinical data shows a less pronounced block of strong positive relations, as this stage excludes patients who died beforehand (and thus have missing records) or who were otherwise no longer eligible based on trial protocols. Correlations that remain often are clinically explainable, for instance, records for comorbidity timings, such as those indicating nausea durations (“NauseastartC1” and “NauseastopC1”), are obviously both missing if the patient did not have such a symptom. However, to avoid possible assumptions, such variables, even if frequently missing across entire cohorts, were investigated for informativeness to model survival predictions by including an additional missing indicator variable for each; informativeness was then identified through stable feature importance analysis of final models illustrated in the Feature Importances section.

Figure 4. Missingness correlation per feature across the full AML17 cohort for the clinical data subset.



Final Optimized Trial Stage Model Measurements

The highest performing models selected via c-index and feature stability selection defined in phase 3 (refer to Final Evaluation Phase in Methods section) were retrained for hyperparameter tuning under nested CV and re-evaluated according to c-index. Key statistics, such as training sample size, the ablation components, and feature set pre- and postreduction, are measured. Feature reduction was model-specific. RSF pruned features with average relative permutation importance. Nonzero feature coefficients determined by elastic net regularization remained in the CPHR model. Tables 1 and 2 show feature reductions, best performing data source ablations, training sample sizes, and average c-index with 95% CIs of each stage-specific model. Raw JSON metric files can be found in Multimedia Appendix 11.

RSF shows higher c-index readings across all stages. However, CoxNet had smaller CIs across all stages, suggesting it may be more generalizable, which requires further analysis using an external dataset. Given that the bootstrapping across each nested fold was set to 150, which was used as a compromise between precision and computational runtimes, it may be the case that a higher number of sample iterations could yield more precise and potentially smaller intervals in both models. All recorded evaluation metrics for final models trained with nested CV are available in Multimedia Appendix 11.

Table . The highest-performing ablation components, training set sample sizes, and respective c-indices for each model after average performance from the ablation analysis phase. Subscripts below longitudinal ablation components indicate the cohort data subset is additionally constrained to only recorded prior to the respective stage.

Trial stage	Ablation components	Training sample size	Model average c-index (95% CI)
Postinduction	• RSF[a]: $Clinical_i$, MRD_i[b], NGS[c], *FLT3*[d], and *NPM1*[e] • CoxNet[f]: $Clinical_i$, *FLT3*, and *NPM1*	2989	• RSF: 0.68 (0.62 - 0.74) • CoxNet: 0.67 (0.62 - 0.72)
Post-C1	• RSF: $Clinical_{C1}$, MRD_{C1}, *FLT3*, and *NPM1* • CoxNet: $Clinical_{C1}$, *FLT3*, and *NPM1*	2609	• RSF: 0.69 (0.63 - 0.76) • CoxNet: 0.68 (0.62 - 0.74)
Post-C2	• RSF: $Clinical_{C2}$, MRD_{C2}, *FLT3*, and *NPM1* • CoxNet: $Clinical_{C2}$, MRD_{C2}, *FLT3*, and *NPM1*	1917	• RSF: 0.68 (0.61 - 0.74) • CoxNet: 0.66 (0.58 - 0.74)
Post-C3	• RSF: $Clinical_{C3}$, MRD_{C3}, *FLT3*, and *NPM1* • CoxNet: $Clinical_{C3}$, *FLT3*, and *NPM1*	761	• RSF: 0.69 (0.56 - 0.81) • CoxNet: 0.63 (0.49 - 0.77)

[a]RSF: random survival forest.

[b]MRD: minimal residual disease.

[c]NGS: next-generation sequencing.

[d]FLT3: Feline McDonough sarcoma-Like Tyrosine kinase 3.

[e]NPM1: Nucleophosmin 1.

[f]CoxNet: Cox proportional hazard regression with elastic net regularization.

Evaluation of the AML17 Protocol Model

Using the full posttrial stage C1 cohort, the AML17 trial protocol risk assessment Cox linear regression model was compared with the RSF and CoxNet models at the same stage using a 0.9:0.1 train-test split and 1000 bootstrapped samples by c-index.

These results show the potential performance gain between trial risk-based models, such as the one used in AML17 for patient treatment stratification, and the RSF and CoxNet models used within this study. The reader is reminded that this specific protocol comparative analysis was exploratory and excluded a nested CV process, unlike those used to produce results highlighted in Table 2. The larger training sample set used for these specific models, while suggesting nonpessimistic performance with more realistic training set sizes, is at risk of overfitting against an external dataset and should be interpreted with caution. It should also be noted that CIs between all models overlap. Future work will include obtaining additional trial data from subsequent AML18 and AML19 trials as an external validation source.

Dynamic AUC

Dynamic AUC was computed for each stage-specific model (RSF and CoxNet) from the beginning of the stage to the 5-year censoring point. Dynamic AUC quantifies a model's discriminative ability at a given point in time *t*, representing how well the model distinguishes between patients who experience an event before *t* and those who remain event-free beyond *t*. AUC scores are bounded between values 0 and 1 inclusively, with higher values indicating better discrimination. A value of 0.5 corresponds to random chance, where a model cannot distinguish between patients. Values below 0.5 indicate a model makes predictions in the opposite ordering, effectively inverting the risk estimate. Note that an AUC of 0 indicates that the model perfectly ranks patients in the reverse order of risk; in such cases, inverting predicted risk scores would yield an AUC of 1, corresponding to perfect discrimination.

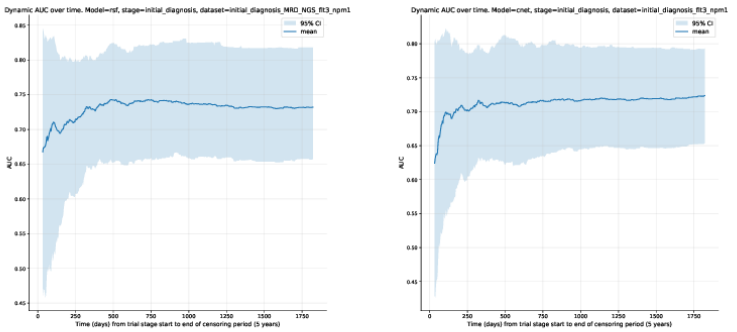
Following the ablation study, candidate models were evaluated by their mean c-index across all repeated, stratified folds of the CV process. At each trial stage, the highest performing RSF and CoxNet models were selected for a final, more robust nested CV with additional bootstrapping per fold for performance estimates. As observed, event time points differ across bootstrapping samples, a universal set of equally intervaled time points was first predefined. Each fold bootstrap sample AUC measurement was then interpolated to its nearest universal time point.

AUC shows an initial period of instability immediately after the stage's baseline. This behavior was expected—early time points will have a smaller cumulative observed event count than the remaining time window of the trial stage. Additionally, risk distributions shift rapidly in these early time windows as patients transition into new treatments based on the existing protocol risk assessment and randomized allocation. As time progresses,

all models tend to a more stabilized AUC score, shown by the plateau of the mean AUC and narrowing of CIs.

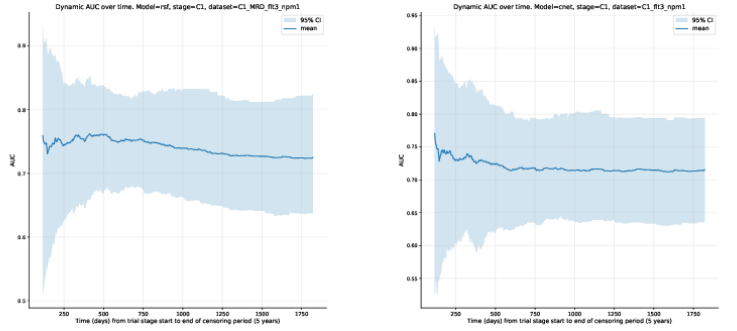
Later stages, (particularly poststage 3) exhibit visibly wider CIs and flatter trajectories. This reflects the considerably lower sample sizes of this stage ("Post-C3" n=761, Table 2) and higher censoring proportions (Figure 2) at this deeper trial stage. Consequently, this limits statistical power and increases uncertainty in time-dependent discrimination estimates. Therefore, dynamic AUC curves at poststage 3 (Figures 5G and 5H) should be interpreted with greater caution.

Figure 5. Dynamic AUC performance of each stage-specific model plotting average dynamic AUC performance as a dark blue line and 95% CI as the light blue shaded region. AUC: area under the receiver operating characteristic curve.



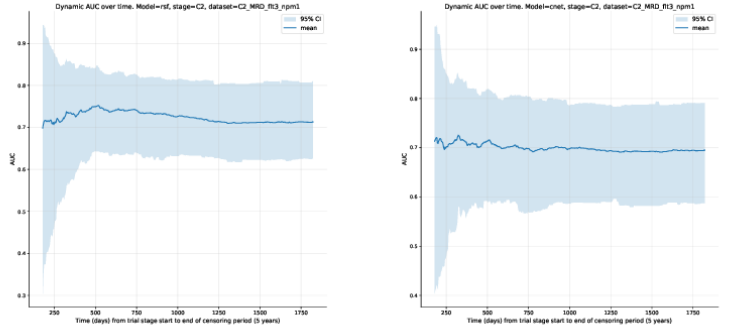
(a) RSF – postinduction

(b) CoxNet – postinduction



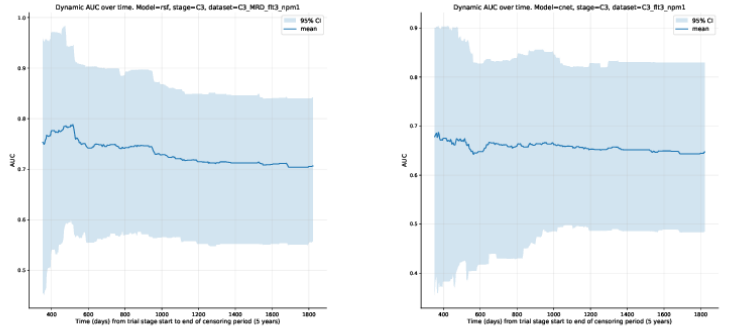
(c) RSF – poststage 1

(d) CoxNet – poststage 1



(e) RSF – poststage 2

(f) CoxNet – poststage 2



(g) RSF – poststage 3

(h) CoxNet – poststage 3

Dynamic Brier Loss

Dynamic Brier loss was computed for each stage-specific RSF and CoxNet model selected via the ablation study to evaluate prediction accuracy over time. At each time point *t*, the Brier score measures the mean square difference between the predicted survival probability at *t* and the observed event status. Score values are bounded between 0 and 1 inclusively, where 0

represents a perfectly accurate model and 1 a completely inaccurate model.

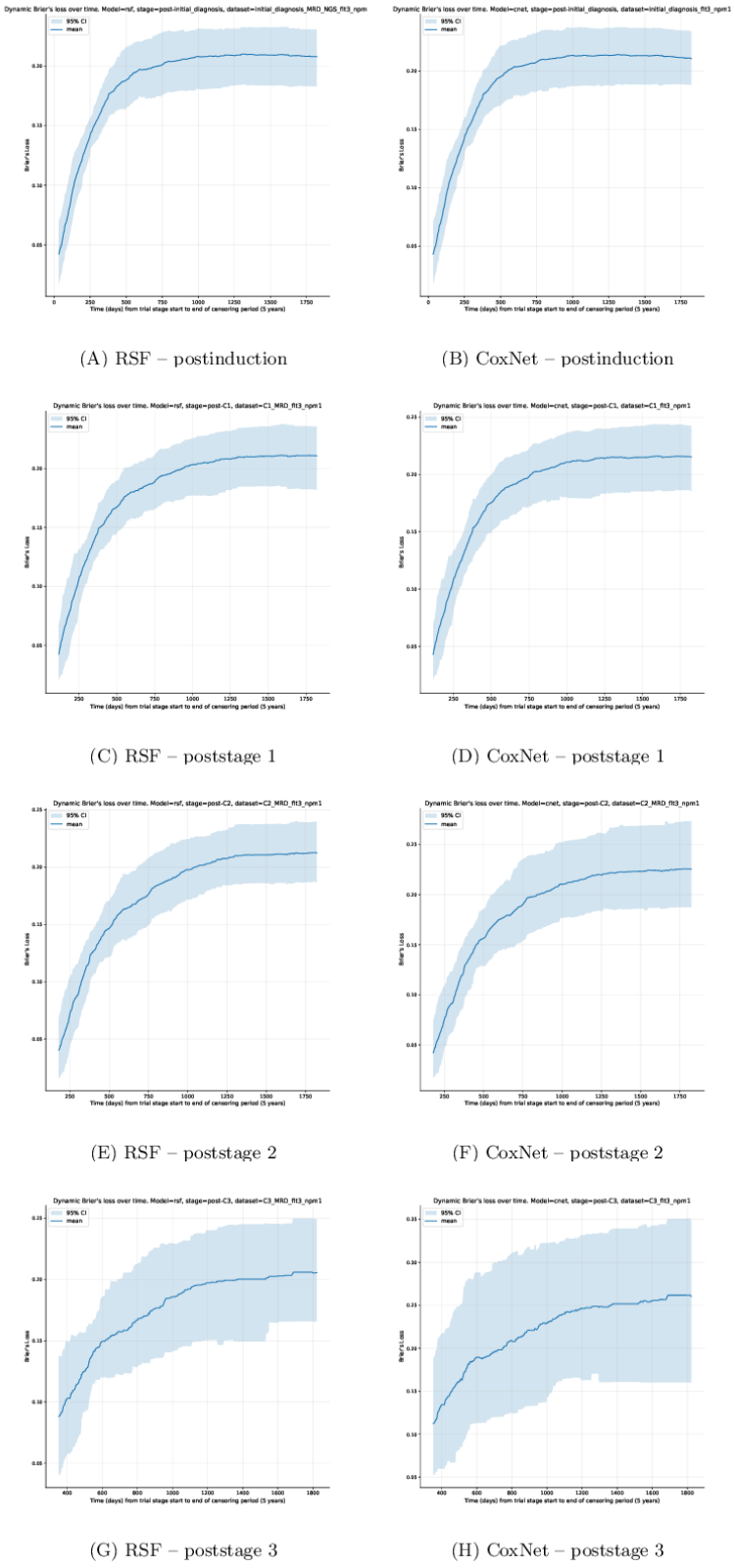
To enable consistent comparison across CV fold bootstrapped samples, dynamic Brier loss curves were evaluated over the same set of predefined, equally intervaled time points used for the dynamic AUC assessment.

Across all stages, dynamic Brier loss scores begin with very low loss values and narrow CIs. This behavior is expected—a smaller proportion of events has been observed at these baseline windows, most patients remain event-free, resulting in highly accurate short-term predictions at low variance. As time progresses throughout each stage model, the Brier loss score increases and the CIs widen. This reflects both the increasing difficulty of long-term survival prediction and the gradual decrease of the at-risk population contributing to the estimate (as the number of patients experiencing an event before t increases, decreasing the remaining subset of at-risk patients available from t).

Similarly, to dynamic AUC results, later stages (particularly poststage 3) show visibly wider CIs. This pattern arises from the reduced sample size available within the cohort (761 patients; Table 2) and higher censoring proportions (Figure 2), which limit the precision of time-dependent accuracy estimates and increase variance at later t evaluation points. Therefore, poststage 3 (Figure 6G and 6H) should be interpreted with caution.

Full resolution dynamic Brier loss and AUC plots are available in Multimedia Appendix 11.

Figure 6. Dynamic Brier loss of each stage-specific model plotting average Brier's loss as a dark blue line and 95% CI as the light blue shaded region.



(A) RSF – postinduction

(B) CoxNet – postinduction

(C) RSF – poststage 1

(D) CoxNet – poststage 1

(E) RSF – poststage 2

(F) CoxNet – poststage 2

(G) RSF – poststage 3

(H) CoxNet – poststage 3

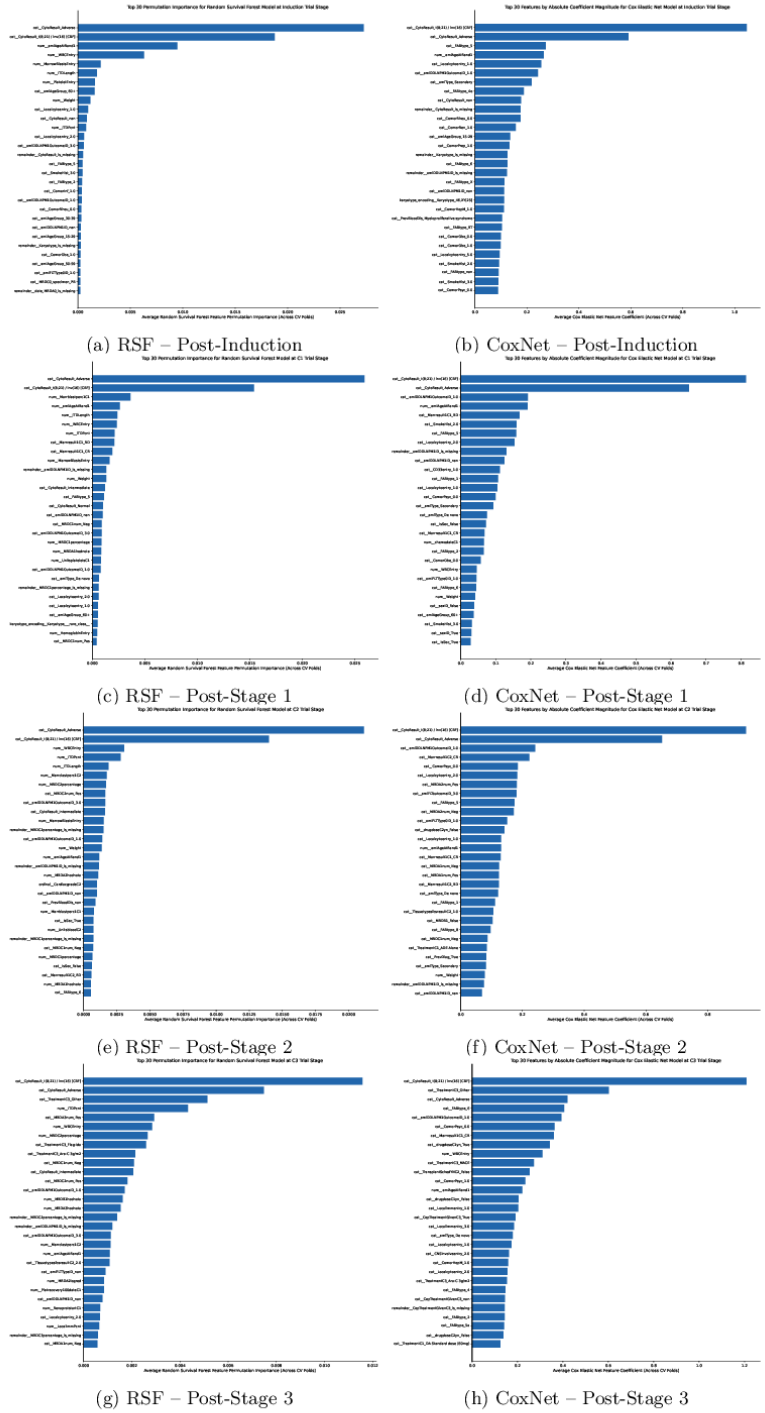
Feature Importances

The following figures show the feature importance of the best-performing tuned models at each select stage of the trial. Importance was ranked according to relative importance scores calculated from permutation importance for RSF and by coefficient values from CoxNet. For postinduction, post-C1, post-C2, and post-C3 models, RSF's feature selection method included more features in the final model. Given the small difference in c-index and dynamic AUC performance between

RSF and CoxNet during trial stage predictions, it is unlikely that the increased feature pool holds strong predictive candidates, suggesting that the increased RSF feature dimensionality from the feature selection process could pose a risk to overfitting and lack of generalizability.

Feature importance for each trial stage model has been ranked to highlight the strongest predictors. For figure readability below, only the top 30 highest-ranked features are included; the total number of features for all stage models exceeds this number, as illustrated in Table 1. While the principal goal of this study concerns the performance of time-sequential TTE prediction models through the trial, future works investigating and explaining generalizations of these models are necessary. Therefore, as a precursory step for future work, we have highlighted the most important features of the best models for each stage in Figure 7. Features that consistently mapped over all or the majority of (often excluding the early postinduction stage as longitudinal records did not exist at this time point) trial stage models include known AML prognosticators—age, white blood cell count, marrow blast values, cytogenetic risk groups (such as the core-binding factor t(8;21)/inv(16)), *NPM1*, and *FLT3* markers [13,64,65]. There are also several consistently important predictors unreferenced as biological risk factors. Missing indicators (eg, blast marrow, *FLT3* mutation type, and internal tandem duplication length value entries) suggest predictive informativeness of missing measurements, which may reflect selection bias, measurement practices from clinical sites, or the severity of a patient's condition. Other predictors include administrative timing fields, such as marrow blast test dates or chemotherapy timings, which, while not strictly biological risk factors, suggest the importance of treatment timing or intensity. It is possible that more precise dosage levels and treatment timings could be recommended from such a system on a per-patient basis. Predictors measured as important but not referenced within the literature most likely reflect measurement practice bias or even data leakage, not necessarily causal disease biology, and warrant further analysis.

Figure 7. Top 30 ranked relative feature importance of trial stage specific models.



Discussion

Principal Findings

The primary performance metric, c-index, of the best models for each trial stage shows that RSF outperforms CoxNet throughout all major stages of the trial, with differences in performance becoming more apparent throughout successive trial stages. This suggests the optimal simulation of TTE outcome prediction would incorporate RSF or additional

nonlinear model types that can capture nonlinear, complex relationships.

Brier loss of stage models appears similar in mean square error trends for all stage models except at stage 3, likely due to having the smallest sample size for training. In the instance of stage 3, the difference in accuracy is larger between RSF and its lower accuracy counterpart, CoxNet. However, it must be noted that the sample size for training of this model is the lowest, as final trial arm branches become increasingly fractional, using 761 samples. Therefore, overfitting is increasingly more likely at this stage, and conclusions made for the outperforming model should be made with caution.

Cumulative AUC shows that both models tend to perform similarly throughout time, being most unstable at the initial stages of the model's available prediction window before censoring (which spans from the end of the respective stage to the 5-year censoring point since patient induction). Notable inflection points occur at similar early time frame windows, which eventually plateau into a stable time-dependent prediction. The degree of change during the initial window of each stage model before stabilization appears more drastic for both models, showing instances of under- and overperformance relative to their mean AUC score. This indicates that the initial periods of all trial stage models are the most sensitive zones in terms of predictive performance. While it cannot necessarily be concluded that the large performance differences before plateaus are solely a result of inadequate sample sizes, there are naturally fewer total event instances in the earliest sample periods with respect to the aggregate of events throughout the entire 5-year window. With low event counts at the earliest AUC time points, variability may be higher, resulting in an overestimate of performance as seen across dynamic AUC plots of all stage models. This would explain the noticeable fluctuations before stabilization of the mean AUC value for each stage model. This initial window would also be the most dynamic point in time with respect to patient treatment selection, where the trial protocol outline determined the treatment stratification of patients into one of multiple arms and so was highly influential on overall survival. This initial unstable period approximately coincides with the worst-case trial length for patients from induction to the end of stage 4, roughly 230 days [16]. Aside from effective sample sizes, it seems intuitive that the most sensitive prediction period of models trained on longitudinal data would be the initial time window after their measurements, as they more accurately reflect the real state of the patient, in the sense that there has been less time for longitudinal measurements to differ from their latest recorded value.

The ablation study stresses the importance of data held in the clinical dataset as well as *FLT3* and *NPM1* and longitudinal MRD records (metadata available in Multimedia Appendices 1-3), which were used for all the highest-ranked RSF models for each stage. The differences in ablated data sources will provide contextual pointers for future analysis of model features with traditional techniques (such as survival curves) or ML techniques (such as clustering), prioritizing features ranked most important by their respective model (Figure 1). Surprisingly, the NGS subset was only included in the postinduction stage RSF model, indicating that the strongest mutational markers came from the FLT3 and NPM1 dataset and karyotype information held in the clinical dataset, both of which remained consistently important across all stage CoxNet and RSF models, as seen in Table 2. NGS mutational biomarkers remained sparse per cohort and most likely effectively acted as noise to both models. This suggests further sensitivity analysis with a wider range of available features via composite NGS feature engineering. The preprocessing of NGS data in these experiments only used available gene mutation indicators to avoid catastrophic explosions in feature set dimensionality by the creation of composite variables. Some excluded features include tumor variant allele frequency and gene mutation base start and end locations, which are candidates for future analysis. A list of all available NGS features can be found in Multimedia Appendices 1-3.

Performance analysis of models constrained to data available up to trial stages approximately equates to performances in literature for a Cox proportional hazard model used for risk group stratification based on aggregate non–time point constrained data [26]. In the wider context of digital twins [43], the results in performance in this study, particularly at the earlier stage time points, suggest that it is possible to provide accurate simulations on survival risk based on models trained at iterative stages of treatment. A digital twin would simulate multiple AML patient outcomes that are relevant to patient well-being; this invites further study on other time-based outcome predictions using the generalized method of this study, such as for QoL or comorbidity prediction.

Approaching a digital twin system with a core prediction layer using ML models should have access to longitudinal data with a temporal resolution, optimal for AML-based predictions. Given longitudinal records, such as MRD, were used by the highest performing stage models (excluding the initial postinduction stage), future work should also involve sensitivity analyses on model prediction accuracy using more frequent, intrastage measurement. This is a practically challenging area of research as trial-based data, such as AML17, even with longitudinal records, is often restricted to stage-wise updates which last several weeks or more depending on chemotherapy regimen. Such a system also demands an automated, digitalized data collection scheme, which is not the norm in trial protocols, where many records are paper-based. Currently existing ML models trained on stage-wise model, such as those shown in these experiments, indicate the promising practicality of their usage in clinical environments when used as the core stratification method within a digital twin system. However, the surrounding architecture necessary to feed models with accurate patient data with high temporal resolution is lacking. The total man-hours to preprocess trial-specific records and validate models with comparable features across trial or real-world data sources for such a system would greatly compromise its applicability, unless a standardized data capture process which records data in an ML model–friendly manner (eg, with improved error handling – particularly for critical date-time entries, mandatory classification levels instead of clinician written note fields, stricter numerical field unit standardizations, a higher volume of QoL records, and higher resolution of longitudinal record entries) is developed for

patients with AML. While RSF can model complex nonlinear relationships, they do not explicitly account for sequential dependencies; when higher resolution longitudinal data are available, researchers should evaluate the performance and feasibility of ML models designed to exploit temporal structures, such as recurrent or long short-term memory neural networks.

In the context of an AML digital twin, the presented RSF and CoxNet models here would act as the core prediction layer, updating patient-specific risk estimates whenever new measurements become available. As AML data are collected at discrete protocol-defined intervals (per trial treatment stages), this framework does not make use of real-time *streamed* data (which typically do not exist in trial-based datasets) but instead "real-time upon update" recalculation of risk as new clinical or MRD entries are recorded for a patient. Generalizability across clinical sites would be supported by a standardized preprocessing pipeline like what has been developed in this study. The monitoring of feature distribution drift (eg, via Population Stability Index) would allow for precise trigger points for model retraining when necessary. The planned external validation of RSF and CoxNet models as core digital twin predictive layers on AML18 and 19 datasets will further quantify cross-site robustness.

Run Times

Model training and validation were computed on the Cardiff University "Hawk" high-performance computing cluster. Jobs were submitted via Simple Linux Utility for Resource Management, using 1 node, 1 task, and 14 CPU cores with the high-throughput partition. Processing time measurements show that RSF is much more intensive to validate than CoxNet, primarily due to the time complexity of the exhaustive feature importance method used—permutation importance. For example, when comparing runtimes of the largest cohort and ablation set—postinduction stage, with ablation components (clinical, MRD, NGS, *FLT3*, and *NPM1*)—averaged across the 3 repeated, k=5 split CV loop, RSF took on average 147 seconds per fold, while CoxNet on average took 1 second. This difference is made more apparent after the more robust internal nested CV process, where the same RSF and CoxNet models and ablations took 7.9 hours and 1.4 hours, respectively, for the *entire* validation process across all folds. A caveat arises, particularly with the usage of the RSF model in clinical applications with large sample sizes; while predictions from trained models are near instantaneous, the full training and validation times of such models may be costly for critical patients relying on fast treatment delivery. In practice, it may be necessary to use CoxNet (or a similar "fast" baseline model) for predictions and counterfactual simulations as a preliminary tool while more sensitive, albeit slower models, such as RSF, are trained. With respect to identifying when models should be retrained, a quantifiable approach to determine the threshold should be calculated. For example, the Population Stability Index can be used to measure if there is a significant drift in a feature between the reference (trained) sample set and the new dataset. This would minimize redundancy and processing constraints in retraining by otherwise using an arbitrary threshold for determining when a model should be retrained, in turn minimizing potential performance or generalizability loss.

Comparisons With Previous Work

Using a Cox proportional hazards model with ridge regression, Tazi et al [26] recorded an IPCW c-index of approximately 0.7 with a combined total of 26 clinical, demographic, FLT3 (ITD), and other molecular class features aggregated from UK-NCRI, such as AML11, 12, 14, 15, 16, and 17 trials. Models constrained to data at each trial stage in this study have comparable IPCW c-index scores of around 0.69 using RSF. The number of features retained in this study after model reduction processes is variable per stage, highlighting the significance of specific features at select time points for TTE predictions. Many features consistently shared across stages are noted in existing ELN risk classification, and those not referenced indicate the potential importance of timing and informativeness from both administrative and testing fields. The total number of retained features is higher for each stage model than in the study by Tazi et al [26], which used 26 features, as opposed to over 160 features across all RSF stage models. Future analysis will involve all features used at each stage in comparison with existing literature and guidelines for risk stratification, in addition to external validation to assess model generalizability.

RSF and CoxNet models outperform the AML17 protocol's Cox model for risk stratification based on c-index ("Evaluation of AML17 Protocol Model" in Table 3, showing c-index readings of AML17 Protocol Cox Linear Regression=0.66, CoxNet=0.68, and RSF=0.70), suggesting both implementations can more accurately predict TTE. However, CIs overlap, suggesting larger datasets may be necessary to conclude an absolute improvement in generalized performance.

Table . C-indices for each model trained on a larger cohort sample size of data available from post stage.

Model	C-index (95% CI)
AML17[a] Protocol Cox Linear Regression	0.66 (0.63 - 0.68)
CoxNet[b]	0.68 (0.63 - 0.73)
RSF[c]	0.70 (0.64 - 0.76)

[a]AML17: United Kingdom National Cancer Research Institute Acute Myeloid Leukaemia 17 randomized controlled trial.

[b]CoxNet: Cox proportional hazard regression with elastic net regularization.

[c]RSF: random survival forest.

Limitations

Since the completion of the AML17 trial, substantial progress has been made in patient therapy options through the identification of prognostic, predictive, and targetable molecular abnormalities [19]. While the analysis of models used in this study provides insight into survival prediction performance using longitudinally restricted data, future work would benefit from the inclusion of more recent trial datasets using newly approved first-line treatment options, such as Midostaurin and CPX-351. Additionally, AML17 patient-reported outcomes, including QoL, were excluded from model training due to concerns on overall sample size within trial stage time points. The future inclusion of datasets with larger pools of patient-reported outcomes data is of particular interest for the prediction of additional outcome responses that may reflect on the social and psychological health of patients at different stages of disease treatment and progression. Further research into additional outcome predictions can be integrated as part of a generalized AML digital twin that can inform patients and provide accurate recommendations to health care practitioners, particularly where survival risk between treatments is marginal, but there are clear differences in secondary predictions, such as patients' QoL.

Conclusion

This study shows the practicality of time-to-survival-event predictions when training sets of CoxNet and RSF models, which are sequentially constricted to data measured up to the end of respective AML17 trial stages. The performance of these sequential TTE models is intended to justify their use as part of a wider digital twin system simulating multiple TTE outcomes for patients with AML. The primary c-index metric shows comparable scores to the literature that uses similar models on aggregate sets of similar trial data. Consistent and stable important features relative to each stage-specific model are supported by ELN literature on AML classification, and additional nonreferenced predictors suggest the importance of stage-specific administrative and timing fields. Additional cumulative-dynamic AUC and Brier loss metrics have been provided. The most immediate future work includes feature analysis of the best models at each stage, further comparison with existing risk group stratification guidelines such as ELN, external validation with follow-up AML18 and 19 trial programs, the implementation of a minimally adapted pipeline for different outcome measurements, and the inclusion of patient self-assessed QoL form records alongside a broader collection of longitudinal MRD data, which have been recorded after trial treatment stages as detailed in the AML17 protocol and follow-up AML18 and 19 trials.

Acknowledgments

We thank the clinicians, research nurses, and laboratory scientists who enrolled patients and provided samples for the AML17 trial. We acknowledge and thank all the patients and families for their participation in, and support of, the trial.

This research was undertaken using the supercomputing facilities at Cardiff University operated by Advanced Research Computing at Cardiff (ARCCA) on behalf of the Cardiff Supercomputing Facility and the Supercomputing Wales (SCW) project. We acknowledge the support of the latter, which is part-funded by the European Regional Development Fund (ERDF) via the Welsh Government.

Funding

The AML17 trial received research support from Cancer Research UK (CRUK/08/025, A29806).This work was supported by the Engineering and Physical Sciences Research Council Doctoral Training Partnership under grant EP/W524682.

Data Availability

The datasets analyzed during the current study are not publicly available due to the sensitive nature of individual participant clinical data, but are available from the trial sponsor on reasonable request. Access requires submission of a research proposal, a Statistical Analysis Plan, and execution of a Data Sharing Agreement. Requests can be made by contacting the UK-NCRI AML17 trial sponsor via ctr@cardiff.ac.uk or through the Centre for Trials Research data request webpage [66].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical data dictionary.
[XLSX File, 48 KB - bioinform_v7i1e75678_app1.xlsx]

Multimedia Appendix 2
MRD data dictionary.
[XLSX File, 28 KB - bioinform_v7i1e75678_app2.xlsx]

Multimedia Appendix 3
FLT3NPM1 data dictionary.
[XLSX File, 18 KB - bioinform_v7i1e75678_app3.xlsx]

Multimedia Appendix 4
NGS metadata.
[DOCX File, 16 KB - bioinform_v7i1e75678_app4.docx]

Multimedia Appendix 5
Feature exclusion summary.
[XLSX File, 15 KB - bioinform_v7i1e75678_app5.xlsx]

Multimedia Appendix 6
Cox proportional hazard regression and regularization techniques.
[DOCX File, 19 KB - bioinform_v7i1e75678_app6.docx]

Multimedia Appendix 7
Random survival forest.
[DOCX File, 18 KB - bioinform_v7i1e75678_app7.docx]

Multimedia Appendix 8
All relevant Python files used for preprocessing, building, and validation of RSF/CoxNet models.
[ZIP File, 53 KB - bioinform_v7i1e75678_app8.zip]

Multimedia Appendix 9
Feature importance values and c-index scores for each ablated model across the repeated k-fold cross-validation from the ablation study phase.
[ZIP File, 17639 KB - bioinform_v7i1e75678_app9.zip]

Multimedia Appendix 10
Missingness correlation per feature heatmaps and missingness frequency bar charts across AML17 treatment stage cohorts for each data source subset used to train RSF and CoxNet models.
[ZIP File, 4906 KB - bioinform_v7i1e75678_app10.zip]

Multimedia Appendix 11
Contains (1) feature importance results for final treatment stage-wise model builds both pre- and postfeature reduction, (2) run-time logs of final CoxNet and RSF models for each fold of nested cross-validation, (3) full resolution dynamic AUC and dynamic Brier loss plots for final stage models, and (4) A JSON file of average values across all bootstrapped samples of each fold for all evaluation metrics: c-index, IPCW c-index, dynamic Brier loss, and dynamic AUC.
[ZIP File, 13739 KB - bioinform_v7i1e75678_app11.zip]

References

1. Cancers by body location/system. National Cancer Institute. URL: https://www.cancer.gov/types/by-body-location [accessed 2025-03-25]
2. Saultz JN, Garzon R. Acute myeloid leukemia: a concise review. J Clin Med 2016 Mar 5;5(3):33. [doi: 10.3390/jcm5030033] [Medline: 26959069]
3. Kim S, Yoon SS, Hong J, et al. Characterization and prognosis of secondary acute myeloid leukemia in an Asian population: AML with antecedent hematological disease confers worst outcomes, irrespective of cytogenetic risk. Anticancer Res 2020 May;40(5):2917-2924. [doi: 10.21873/anticanres.14269] [Medline: 32366443]
4. Baruchel A, Bourquin JP, Crispino J, et al. Down syndrome and leukemia: from basic mechanisms to clinical advances. haematol 2023 Jul;108(10):2570-2581. [doi: 10.3324/haematol.2023.283225]
5. Alter BP. Fanconi anemia and the development of leukemia. Best Pract Res Clin Haematol 2014;27(3-4):214-221. [doi: 10.1016/j.beha.2014.10.002] [Medline: 25455269]
6. Sportoletti P, Grisendi S, Majid SM, et al. Npm1 is a haploinsufficient suppressor of myeloid and lymphoid malignancies in the mouse. Blood 2008 Apr 1;111(7):3859-3862. [doi: 10.1182/blood-2007-06-098251] [Medline: 18212245]

7. Donner L, Fedele LA, Garon CF, Anderson SJ, Sherr CJ. McDonough feline sarcoma virus: characterization of the molecularly cloned provirus and its feline oncogene (v-fms). J Virol 1982 Feb;41(2):489-500. [doi: 10.1128/JVI.41.2.489-500.1982] [Medline: 6281462]
8. Kiyoi H, Kawashima N, Ishikawa Y. FLT3 mutations in acute myeloid leukemia: therapeutic paradigm beyond inhibitor development. Cancer Sci 2020 Feb;111(2):312-322. [doi: 10.1111/cas.14274] [Medline: 31821677]
9. Shallis RM, Weiss JJ, Deziel NC, Gore SD. A clandestine culprit with critical consequences: benzene and acute myeloid leukemia. Blood Rev 2021 May;47:100736. [doi: 10.1016/j.blre.2020.100736] [Medline: 32771228]
10. Fircanis S, Merriam P, Khan N, Castillo JJ. The relation between cigarette smoking and risk of acute myeloid leukemia: an updated meta-analysis of epidemiological studies. Am J Hematol 2014 Aug;89(8):E125-E132. [doi: 10.1002/ajh.23744] [Medline: 24753145]
11. Strickland SA, Vey N. Diagnosis and treatment of therapy-related acute myeloid leukemia. Crit Rev Oncol Hematol 2022 Mar;171:103607. [doi: 10.1016/j.critrevonc.2022.103607] [Medline: 35101585]
12. Desai RH, Zandvakili N, Bohlander SK. Dissecting the genetic and non-genetic heterogeneity of acute myeloid leukemia using next-generation sequencing and in vivo models. Cancers (Basel) 2022 Apr 27;14(9):2182. [doi: 10.3390/cancers14092182] [Medline: 35565315]
13. Döhner H, Wei AH, Appelbaum FR, et al. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. Blood 2022 Sep 22;140(12):1345-1377. [doi: 10.1182/blood.2022016867] [Medline: 35797463]
14. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 2016 May 19;127(20):2391-2405. [doi: 10.1182/blood-2016-03-643544] [Medline: 27069254]
15. Qin D. Next-generation sequencing and its clinical application. Cancer Biol Med 2019 Feb;16(1):4-10. [doi: 10.20892/j.issn.2095-3941.2018.0055] [Medline: 31119042]
16. AML 17 protocol for patients aged under 60. Cardiff University Centre for Trials Research. URL: https://trials.cardiff.ac.uk/aml/17/web/files/new3/AML%2017%20Protocol%20June11%20v7.1%20.pdf [accessed 2024-10-12]
17. AML17. Cardiff University | Centre for Trials Research. URL: https://www.cardiff.ac.uk/centre-for-trials-research/research/studies-and-trials/view/aml17 [accessed 2024-10-14]
18. Burnett AK, Russell NH, Hills RK, et al. A randomized comparison of daunorubicin 90 mg/m2 vs 60 mg/m2 in AML induction: results from the UK NCRI AML17 trial in 1206 patients. Blood 2015 Jun 18;125(25):3878-3885. [doi: 10.1182/blood-2015-01-623447] [Medline: 25833957]
19. Kantarjian H, Kadia T, DiNardo C, et al. Acute myeloid leukemia: current progress and future directions. Blood Cancer J 2021 Feb 22;11(2):41. [doi: 10.1038/s41408-021-00425-3] [Medline: 33619261]
20. Cancer stat facts: leukemia — acute myeloid leukemia (AML). National Cancer Institute — Surveillance, Epidemiology and End Results Program. URL: https://seer.cancer.gov/statfacts/html/amyl.html [accessed 2024-10-22]
21. Boscaro E, Urbino I, Catania FM, et al. Modern risk stratification of acute myeloid leukemia in 2023: integrating established and emerging prognostic factors. Cancers (Basel) 2023 Jul 6;15(13):3512. [doi: 10.3390/cancers15133512] [Medline: 37444622]
22. Acute myeloid leukaemia (AML) statistics. Cancer Research UK. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-aml [accessed 2024-05-12]
23. Yaqoob A, Musheer Aziz R, verma NK. Applications and techniques of machine learning in cancer classification: a systematic review. Hum-Cent Intell Syst 2023 Dec;3(4):588-615. [doi: 10.1007/s44230-023-00041-3]
24. Chen S, Wang Y, Chi P, et al. Survival prediction optimization of acute myeloid leukemia based on T-cell function-related genes and plasma proteins. Blood 2022 Nov 15;140(Supplement 1):6300-6302. [doi: 10.1182/blood-2022-163201]
25. Gal O, Auslander N, Fan Y, Meerzaman D. Predicting complete remission of acute myeloid leukemia: machine learning applied to gene expression. Cancer Inform 2019;18:1176935119835544. [doi: 10.1177/1176935119835544] [Medline: 30911218]
26. Tazi Y, Arango-Ossa JE, Zhou Y, et al. Unified classification and risk-stratification in acute myeloid leukemia. Nat Commun 2022 Aug 8;13(1):4622. [doi: 10.1038/s41467-022-32103-8] [Medline: 35941135]
27. AML18. Cardiff University | Centre for Trials Research. URL: https://www.cardiff.ac.uk/centre-for-trials-research/research/studies-and-trials/view/aml18 [accessed 2024-05-09]
28. Versluis J, Metzner M, Wang A, et al. Risk stratification in older intensively treated patients with AML. JCO 2024 Dec;42(34):4084-4094. [doi: 10.1200/JCO.23.02631]
29. Vock DM, Wolfson J, Bandyopadhyay S, et al. Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. J Biomed Inform 2016 Jun;61:119-131. [doi: 10.1016/j.jbi.2016.03.009] [Medline: 26992568]
30. Suresh K, Severn C, Ghosh D. Survival prediction models: an introduction to discrete-time modeling. BMC Med Res Methodol 2022 Jul 26;22(1):207. [doi: 10.1186/s12874-022-01679-6] [Medline: 35883032]
31. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]

32. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008 Sep;2(3):3. [doi: 10.1214/08-AOAS169]
33. Qin J, Shen Y. Statistical methods for analyzing right-censored length-biased data under Cox model. Biometrics 2010 Jun;66(2):382-392. [doi: 10.1111/j.1541-0420.2009.01287.x] [Medline: 19522872]
34. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 2005 Jul 1;21(13):3001-3008. [doi: 10.1093/bioinformatics/bti422] [Medline: 15814556]
35. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B 2005 Apr 1;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
36. Leger S, Zwanenburg A, Pilz K, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Sci Rep 2017 Oct 16;7(1):13206. [doi: 10.1038/s41598-017-13448-3] [Medline: 29038455]
37. Adeoye J, Hui L, Koohi-Moghadam M, Tan JY, Choi SW, Thomson P. Comparison of time-to-event machine learning models in predicting oral cavity cancer prognosis. Int J Med Inform 2022 Jan;157:104635. [doi: 10.1016/j.ijmedinf.2021.104635] [Medline: 34800847]
38. Kurt Omurlu I, Ture M, Tokatli F. The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer. Expert Syst Appl 2009 May;36(4):8582-8588. [doi: 10.1016/j.eswa.2008.10.023]
39. Cygu S, Seow H, Dushoff J, Bolker BM. Comparing machine learning approaches to incorporate time-varying covariates in predicting cancer survival time. Sci Rep 2023 Jan 25;13(1):1370. [doi: 10.1038/s41598-023-28393-7] [Medline: 36697455]
40. Aivaliotis G, Palczewski J, Atkinson R, Cade JE, Morris MA. A comparison of time to event analysis methods, using weight status and breast cancer as a case study. Sci Rep 2021 Jul 7;11(1):14058. [doi: 10.1038/s41598-021-92944-z] [Medline: 34234154]
41. Qiu X, Gao J, Yang J, et al. A comparison study of machine learning (random survival forest) and classic statistic (Cox proportional hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. Front Oncol 2020;10:551420. [doi: 10.3389/fonc.2020.551420] [Medline: 33194609]
42. Pickett KL, Suresh K, Campbell KR, Davis S, Juarez-Colunga E. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. BMC Med Res Methodol 2021 Oct 17;21(1):216. [doi: 10.1186/s12874-021-01375-x] [Medline: 34657597]
43. Qi Q, Tao F, Hu T, et al. Enabling technologies and tools for digital twin. Journal of Manufacturing Systems 2021 Jan;58:3-21. [doi: 10.1016/j.jmsy.2019.10.001]
44. Garciaz S, Hospital MA. FMS-like Tyrosine Kinase 3 inhibitors in the treatment of acute myeloid leukemia: an update on the emerging evidence and safety profile. Onco Targets Ther 2023;16:31-45. [doi: 10.2147/OTT.S236740] [Medline: 36698434]
45. Hawk Cardiff Research Datastore (RDS) access VM. Supercomputing Wales Portal. URL: https://portal.supercomputing.wales/index.php/hawk-cardiff-research-datastore-rds-access-vm/ [accessed 2025-04-08]
46. Bennett JM, Catovsky D, Daniel MT, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. Br J Haematol 1976 Aug;33(4):451-458. [doi: 10.1111/j.1365-2141.1976.tb03563.x] [Medline: 188440]
47. ECOG performance status scale. ECOG-ACRIN Cancer Research Group. URL: https://ecog-acrin.org/resources/ecog-performance-status/ [accessed 2025-03-08]
48. Centre for Trials Research | Home. Cardiff University. URL: https://www.cardiff.ac.uk/centre-for-trials-research [accessed 2024-10-12]
49. Ying X. An overview of overfitting and its solutions. J Phys: Conf Ser 2019 Mar 1;1168(2):022022. [doi: 10.1088/1742-6596/1168/2/022022]
50. Lauriola I, Lavelli A, Aiolli F. An introduction to deep learning in natural language processing: models, techniques, and tools. Neurocomputing 2022 Jan;470:443-456. [doi: 10.1016/j.neucom.2021.05.103]
51. Pandas.DataFrame. pandas. URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html [accessed 2024-10-14]
52. StandardScaler. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html [accessed 2024-10-12]
53. General concepts. The Open Group. URL: https://pubs.opengroup.org/onlinepubs/9699919799/basedefs/V1_chap04.html#tag_04_16 [accessed 2024-10-12]
54. de Amorim LBV, Cavalcanti GDC, Cruz RMO. The choice of scaling technique matters for classification performance. Appl Soft Comput 2023 Jan;133:109924. [doi: 10.1016/j.asoc.2022.109924]
55. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010 May 15;26(10):1340-1347. [doi: 10.1093/bioinformatics/btq134] [Medline: 20385727]
56. Belle V, Pelckmans K, Suykens J, Huffel S. Presented at: Survival SVM: a practical scalable algorithm; Apr 23-25, 2008 URL: https://www.esann.org/sites/default/files/proceedings/legacy/es2008-95.pdf
57. Sksurv.svm.fastsurvivalsvm. scikit-survival 0271. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html [accessed 2025-03-27]

58. Russell NH, Burnett AK, Hills RK, et al. Long term follow up from the NCRI AML17 trial of attenuated arsenic trioxide and ATRA therapy for newly diagnosed and relapsed acute promyelocytic leukaemia. Blood 2016 Dec 2;128(22):897-897. [doi: 10.1182/blood.V128.22.897.897]
59. van Geloven N, He Y, Zwinderman AH, Putter H. Estimation of incident dynamic AUC in practice. Comput Stat Data Anal 2021 Feb;154:107095. [doi: 10.1016/j.csda.2020.107095]
60. Napierala K, Stefanowski J. Types of minority class examples and their influence on learning classifiers from imbalanced data. J Intell Inf Syst 2016 Jun;46(3):563-597. [doi: 10.1007/s10844-015-0368-1]
61. Datasets: imbalanced datasets. 'Machine Learning', Google for Developers. URL: https://developers.google.com/machine-learning/crash-course/overfitting/imbalanced-datasets [accessed 2025-01-14]
62. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. jair 2002 Jun;16:321-357. [doi: 10.1613/jair.953]
63. Sksurv.metrics.concordance_index_ipcw. scikit-survival 0270. URL: https://scikit-survival.readthedocs.io/en/stable/api/generated/sksurv.metrics.concordance_index_ipcw.html [accessed 2025-03-11]
64. Feng S, Zhou L, Zhang X, et al. Impact Of ELN risk stratification, induction chemotherapy regimens and hematopoietic stem cell transplantation on outcomes in hyperleukocytic acute myeloid leukemia with initial white blood cell count more than 100 × 109/L. Cancer Manag Res 2019;11:9495-9503. [doi: 10.2147/CMAR.S225123] [Medline: 31807075]
65. Juliusson G, Jädersten M, Deneberg S, et al. The prognostic impact of FLT3-ITD and NPM1 mutation in adult AML is age-dependent in the population-based setting. Blood Adv 2020 Mar 24;4(6):1094-1101. [doi: 10.1182/bloodadvances.2019001335] [Medline: 32203582]
66. Centre for trials research - data requests. Cardiff University. URL: https://www.cardiff.ac.uk/centre-for-trials-research/collaborate-with-us/data-requests/ [accessed 2026-04-06]

Abbreviations

AML: acute myeloid leukemia
AML17: United Kingdom National Cancer Research Institute Acute Myeloid Leukaemia 17 randomized controlled trial
AUC: area under the receiver operating characteristic curve
c-index: concordance index
CoxNet: Cox proportional hazard regression with elastic net regularization
CPHR: Cox proportional hazard regression
CV: cross-validation
ELN: European LeukemiaNet
FLT3: Feline McDonough sarcoma-Like Tyrosine kinase 3
IPCW: inverse probability of censoring weights
ML: machine learning
MRD: minimal residual disease
NGS: next-generation sequencing
NPM1: Nucleophosmin 1
QoL: quality of life
RSF: random survival forest
TTE: time-to-event
UK-NCRI: United Kingdom National Cancer Research Institute
WHO: World Health Organization

Edited by E Uzun; submitted 08.Apr.2025; peer-reviewed by H Yan, VF Calsavara; revised version received 15.Nov.2025; accepted 30.Dec.2025; published 29.Apr.2026.

Please cite as:
Brady O, Johnson S, Giles P, Alvares C, Zabkiewicz J, Fuentes C
Random Survival Forest Versus Elastic-Net Regularized Cox Regression for Survival Prediction in Acute Myeloid Leukemia at Distinct Treatment Time Points: Model Performance Comparison Study
JMIR Bioinform Biotech 2026;7:e75678
URL: https://bioinform.jmir.org/2026/1/e75678
doi:10.2196/75678

Readability of AI-Generated Patient Information on Glucagon-Like Peptide-1 Receptor Agonists

Tyler Williams[1]; Ines Bilic-Curcic[1], MD; Jonathan Hurley[2], MD; Harisankeerth Mummareddy[3], MD; Maja Cigrovski Berkovic[4], MD; Silvija Canecki Varzic[5], MD; Marina Gradiser[6], MD

[1]Faculty of Medicine, Tulane University, New Oreleans, LA, United States

[2]School of Medicine, Tulane University, New Orelans, LA, United States

[3]University of Tennessee Health Science Center, Menphis, TN, United States

[4]Faculty of Kinesiology, University of Zagreb, Horvacanski zavoj 15, Zagreb, Croatia

[5]Faculty of Dental Medicine and Health, University of Osijek, Osijek, County of Osijek-Baranja, Croatia

[6]Faculty of Medicine, University of Split, Split, Split-Dalmatia, Croatia

Corresponding Author:
Maja Cigrovski Berkovic, MD
Faculty of Kinesiology, University of Zagreb, Horvacanski zavoj 15, Zagreb, Croatia

Abstract

Artificial intelligence (AI)–generated content on glucagon-like peptide-1 receptor agonists (GLP-1RAs) gave informationally detailed responses, but its readability remains suboptimal for many patients. Incorporating literacy-sensitive design principles into AI health communication is essential to ensure equitable access to digital medical information.

(JMIR Bioinform Biotech 2026;7:e90572) doi:10.2196/90572

KEYWORDS

health literacy; GLP-1RA; AI; readability; ChatGPT; Google Gemini; artificial intelligence

Introduction

Glucagon-like peptide-1 receptor agonists (GLP-1RAs) have become cornerstone therapies for type 2 diabetes mellitus and obesity, with additional benefits extending to cardiovascular risk reduction and metabolic health [1,2]. Their rapid clinical uptake has been paralleled by a surge in public interest and online information-seeking [3-5]. Artificial intelligence (AI)–based conversational tools are increasingly used by patients as informal sources of medical guidance. However, engagement with digital health information is strongly influenced by health literacy. Patient education materials are generally recommended to be written at or below an eighth-grade reading level to ensure comprehension across diverse populations [6,7]. Data from the OECD Survey of Adult Skills indicate that a substantial proportion of adults demonstrate limited literacy skills, raising concerns that complex digital health content may exacerbate existing health disparities [8]. Given the expanding role of AI-generated medical information, evaluating whether such content is accessible to patients is crucial. This study assessed the readability of AI-generated responses to common patient questions regarding GLP-1RAs, hypothesizing that language complexity would exceed recommended thresholds for patient-facing materials.

Methods

A cross-sectional descriptive analysis was conducted using two large language models: ChatGPT (GPT-4.1; OpenAI) and Google Gemini (Gemini 2.5 Flash; Google DeepMind). Ten frequently asked patient questions related to GLP-1RAs were identified through review of online patient forums and routine clinical encounters. Topics included dosing, side effects, safety, insurance coverage, off-label use, and expected outcomes.

Each question was submitted to both models using identical prompts requesting a “clear, patient-friendly response.” Responses were collected verbatim between June and August 2025. All queries were submitted using the models’ web-based interfaces. Each query was entered in a new session to avoid retention of prior conversational context. No system-level prompts, role instructions, or parameter adjustments (eg, temperature, top_p) were applied; therefore, outputs reflect default model behavior as experienced by typical users. Identical prompt instructions were used for all queries across both platforms.

Readability was evaluated using three validated measures: (1) Flesch Reading Ease Score, where higher scores indicate easier readability; (2) Flesch-Kincaid Grade Level (FKGL), estimating required US school grade level; and (3) mean sentence length and word count as proxies for textual complexity.

Analyses were performed using ReadablePro software and cross-validated with Microsoft Word readability tools. In addition, two independent health communication experts qualitatively reviewed responses for tone, sentence structure, and technical language use. Paired *t* tests were used to compare mean readability scores between ChatGPT and Gemini responses. In addition to *P* values, paired mean differences with 95% CIs and effect sizes (Cohen d_z) were calculated to assess the magnitude of differences.

Before readability analysis, all outputs were standardized to plain text format. Formatting elements such as bullet points, headings, and line breaks were removed to ensure consistent processing across readability tools.

The full list of patient questions, verbatim model outputs, and per-question readability metrics are provided in Multimedia Appendices 1-3.

Results

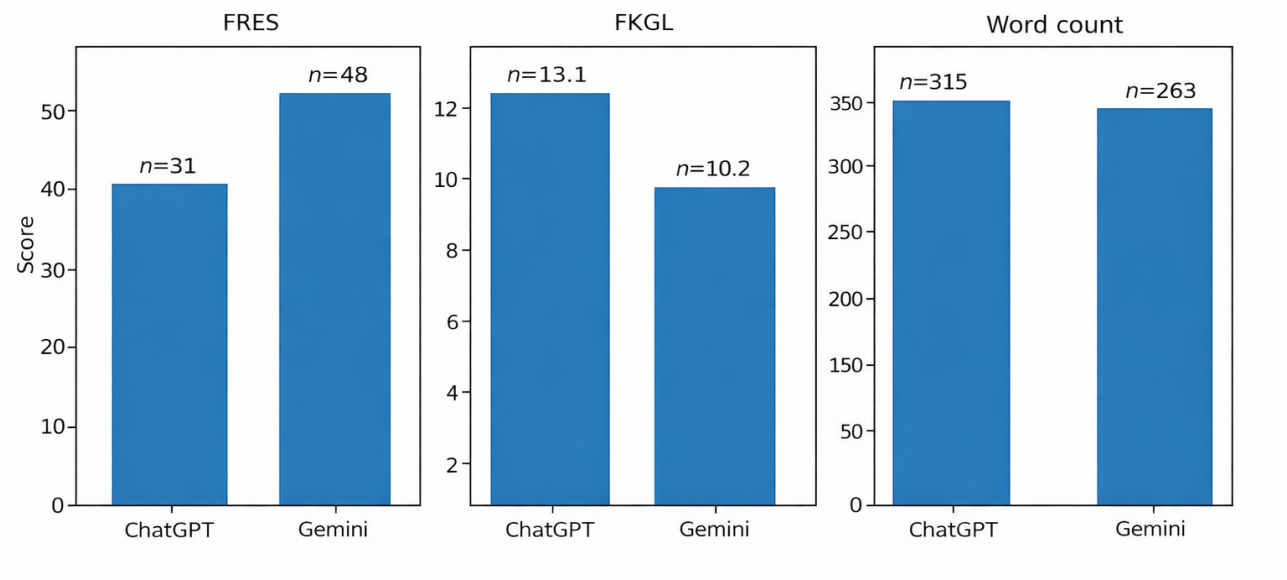
Across all 10 questions, both models generated detailed responses ranging from 150 to 450 words. Gemini produced shorter sentences and more concise explanations, whereas ChatGPT responses were longer and more technically detailed.

Mean Flesch Reading Ease Score was significantly higher for Gemini than for ChatGPT (47.97 vs 31.65; *P*=.004), indicating superior readability (Figure 1). Similarly, mean FKGL was lower for Gemini (10.2) compared with ChatGPT (13.1), although both exceeded the recommended eighth-grade level.

Using FKGL as the operational measure, both models exceeded the recommended ≤8th-grade level for patient education materials by approximately 2 - 5 grade levels on average.

Qualitative analysis revealed that ChatGPT frequently used complex sentence structures and specialized terminology, while Gemini responses adopted a more conversational tone with simpler vocabulary.

Figure 1. Comparison of readability metrics between ChatGPT and Google Gemini responses. ChatGPT responses demonstrated lower Flesch Reading Ease Scores (FRES) and higher Flesch-Kincaid Grade Levels (FKGLs) compared to Gemini, indicating greater linguistic complexity. Both models exceeded recommended readability thresholds for patient education materials.



Discussion

This study demonstrates that AI-generated patient information on GLP-1RAs varies significantly in readability depending on the model used. Although Gemini responses were more accessible than those generated by ChatGPT, neither consistently met recommended readability standards for patient education. These findings align with prior evidence showing that online health information frequently exceeds the literacy capacity of the general population [9-12]. These findings also highlight limitations of general-purpose AI models in health care contexts. Domain-specific, medically trained AI agents may provide more consistent, context-aware, and clinically appropriate outputs. Such systems could incorporate structured medical knowledge, terminology control, and longitudinal context, which are critical for reliable patient communication. In clinical applications, AI systems must also include mechanisms for uncertainty management and safety. When confidence in generated responses is low, fallback strategies, such as retrieval from validated medical sources or escalation to health care professionals, are essential to ensure reliability and patient safety.

Given that AI tools prioritize completeness and accuracy, readability may be unintentionally deprioritized. Without deliberate incorporation of plain-language principles, AI-generated health content risks widening disparities among individuals with limited literacy or digital health skills.

It is important to note that outputs from large language models are highly sensitive to prompting strategies. This study intentionally used identical prompts to simulate a real-world patient scenario, in which users typically enter similar queries across platforms without advanced prompt engineering. However, more sophisticated prompting approaches or

orchestration layers may yield responses with substantially different readability profiles. Future studies should explore the impact of prompt design on health communication outcomes.

AI systems also offer the potential for interactive, bidirectional communication and personalization, which may enhance patient engagement. However, without appropriate attention to readability and comprehension, these advantages may not translate into meaningful improvements in patient outcomes.

An important opportunity for future development is the integration of literacy-adaptive mechanisms within AI systems. Tailoring responses to a user's health literacy level, through simplified language or dynamic adjustment of complexity, could significantly enhance accessibility and patient understanding.

This study has several limitations. First, AI-generated outputs are subject to variability over time due to model updates and changes in system behavior; therefore, the findings reflect the specific models, settings, and time frame studied. Second, results may differ with alternative prompting strategies. Third, this study evaluated readability but did not assess factual accuracy or patient comprehension outcomes, which should be addressed in future research.

ChatGPT and Gemini provide clinically relevant responses on GLP-1RAs, but responses remain insufficiently readable for many patients. Integrating literacy-sensitive frameworks into AI health communication is essential to ensure equitable access to digital medical information and maximize the clinical benefits of emerging therapies.

Implementation of standardized readability frameworks within AI systems may represent a key step toward equitable digital health communication.

Funding

The authors declared no financial support was received for this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patient questions used in the analysis.
[DOCX File, 14 KB - bioinform_v7i1e90572_app1.docx]

Multimedia Appendix 2
Artificial intelligence–generated responses (verbatim).
[DOCX File, 16 KB - bioinform_v7i1e90572_app2.docx]

Multimedia Appendix 3
Supplementary table.
[DOCX File, 14 KB - bioinform_v7i1e90572_app3.docx]

References

1. Nauck MA, Meier JJ. Incretin hormones: their role in health and disease. Diabetes Obes Metab 2018 Feb;20 Suppl 1:5-21. [doi: 10.1111/dom.13129] [Medline: 29364588]
2. Wilding JPH, Batterham RL, Calanna S, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med 2021 Mar 18;384(11):989-1002. [doi: 10.1056/NEJMoa2032183] [Medline: 33567185]
3. Javaid A, Baviriseaty S, Javaid R, et al. Trends in glucagon-like peptide-1 receptor agonist social media posts using artificial intelligence. JACC Adv 2024 Sep;3(9):101182. [doi: 10.1016/j.jacadv.2024.101182] [Medline: 39372460]
4. Auerbach N, Liu VN, Huang DR, Clift AK, Al-Ammouri M, El-Osta A. What are community perspectives and experiences around GLP-1 receptor agonist medications for weight loss? A cross-sectional survey study in the UK. BMJ Public Health 2025;3(2):e002519. [doi: 10.1136/bmjph-2024-002519] [Medline: 40734969]
5. Lipari M, Berlie H, Saleh Y, Hang P, Moser L. Understandability, actionability, and readability of online patient education materials about diabetes mellitus. Am J Health Syst Pharm 2019 Jan 25;76(3):182-186. [doi: 10.1093/ajhp/zxy021] [Medline: 31408087]
6. Kincaid JP, et al. Derivation of new readability formulas. : Naval Technical Report; 1975 URL: https://apps.dtic.mil/sti/tr/pdf/ADA006655.pdf [accessed 2026-04-15]
7. Doak CC, Doak LG, Root JH. Teaching Patients with Low Literacy Skills, 2nd edition: J.B. Lippincott; 1996.
8. Survey of adult skills 2023: country note – Croatia. OECD - Education GPS. 2023. URL: https://gpseducation.oecd.org/CountryProfile?primaryCountry=HRV&treshold=5&topic=AS [accessed 2026-04-07]
9. Bernard S, Cooke T, Cole T, Hachani L, Bernard J. Quality and readability of online information about type 2 diabetes and nutrition. JAAPA 2018 Nov;31(11):41-44. [doi: 10.1097/01.JAA.0000546481.02560.4e] [Medline: 30358679]

10. Yeung AWK, et al. Online information about semaglutide: mixed methods study. JMIR Infodemiology 2025;5:e59767. [doi: 10.2196/59767] [Medline: 40198905]
11. Osborn CY, Bains SS, Egede LE. Health literacy, diabetes self-care, and glycemic control in adults with type 2 diabetes. Diabetes Technol Ther 2010 Nov;12(11):913-919. [doi: 10.1089/dia.2010.0058] [Medline: 20879964]
12. Sørensen K, Pelikan JM, Röthlin F, et al. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). Eur J Public Health 2015 Dec;25(6):1053-1058. [doi: 10.1093/eurpub/ckv043] [Medline: 25843827]

Abbreviations

AI: artificial intelligence
FKGL: Flesch-Kincaid Grade Level
GLP-1RA: glucagon-like peptide-1 receptor agonist

Edited by Z Yue; submitted 30.Dec.2025; peer-reviewed by A Bethanabatla, H Vundavalli, Y Hu; revised version received 23.Mar.2026; accepted 25.Mar.2026; published 05.May.2026.

Please cite as:
Williams T, Bilic-Curcic I, Hurley J, Mummareddy H, Cigrovski Berkovic M, Canecki Varzic S, Gradiser M
Readability of AI-Generated Patient Information on Glucagon-Like Peptide-1 Receptor Agonists
JMIR Bioinform Biotech 2026;7:e90572
URL: https://bioinform.jmir.org/2026/1/e90572
doi:10.2196/90572

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org