# 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

### Original Papers

# 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.

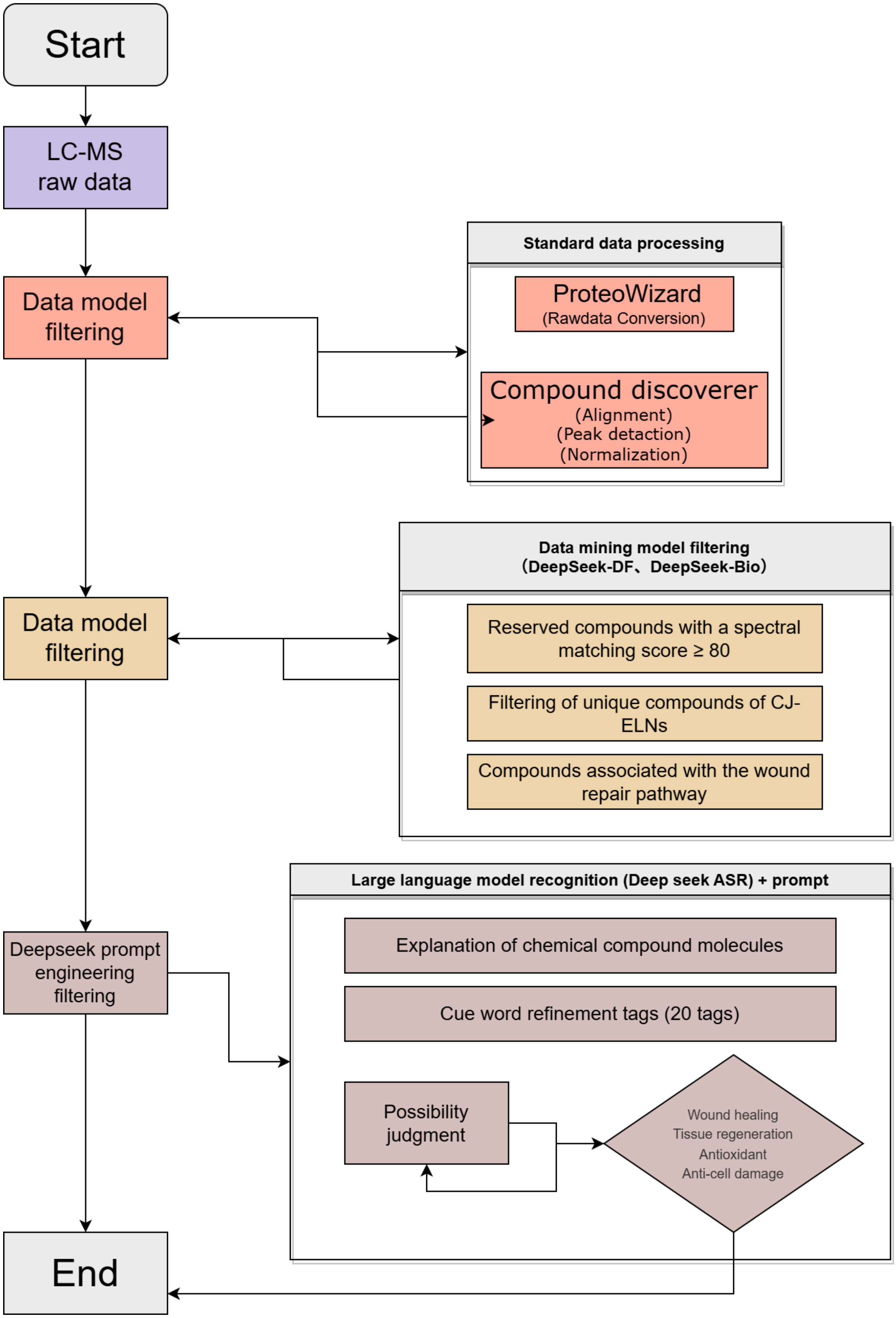

Start
LC-MS raw data
Data model filtering
Standard data processing
ProteoWizard
(Rawdata Conversion)
Compound discoverer
(Alignment)
(Peak detaction)
(Normalization)
Data model filtering
Data mining model filtering
（DeepSeek-DF、DeepSeek-Bio）
Reserved compounds with a spectral matching score ≥ 80
Filtering of unique compounds of CJ-ELNs
Compounds associated with the wound repair pathway
Deepseek prompt engineering filtering
Large language model recognition (Deep seek ASR) + prompt
Explanation of chemical compound molecules
Cue word refinement tags (20 tags)
Possibility judgment
Wound healing
Tissue regeneration
Antioxidant
Anti-cell damage
End

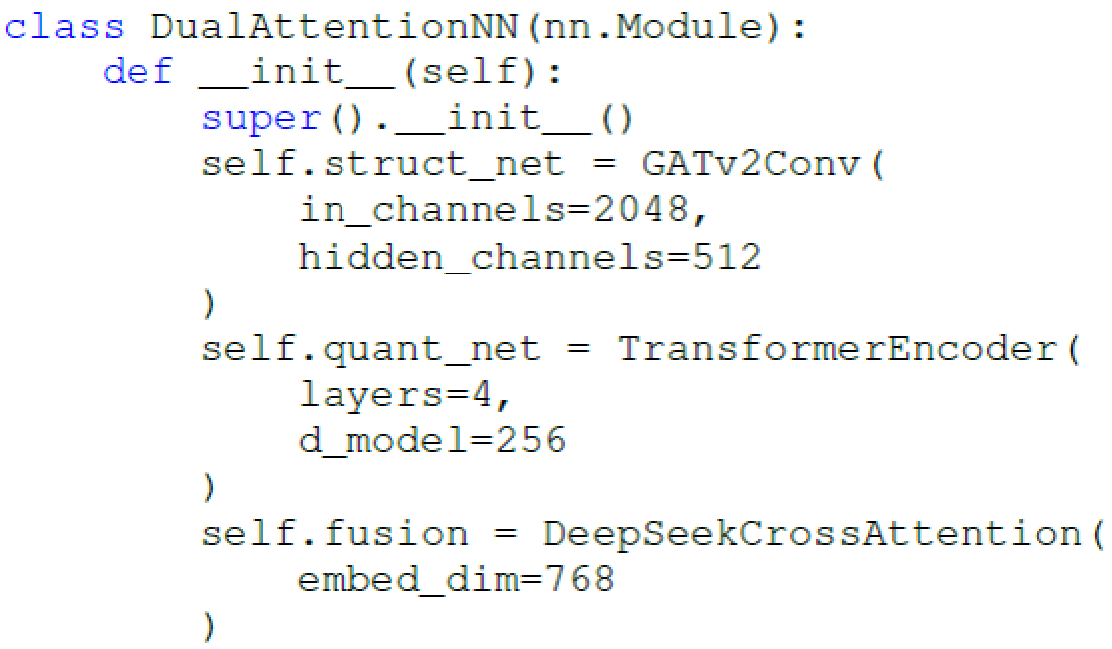
**Table .** Key terms, functions, tools and databases used in this study.

| Key terms | Functions | Tools/databases |
|---|---|---|
| mzML | Standardized data storage | ProteoWizard |
| DeepSeek-Bio | Biological pathway association analysis | Deepseek 671B Model Network Edition<br>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.

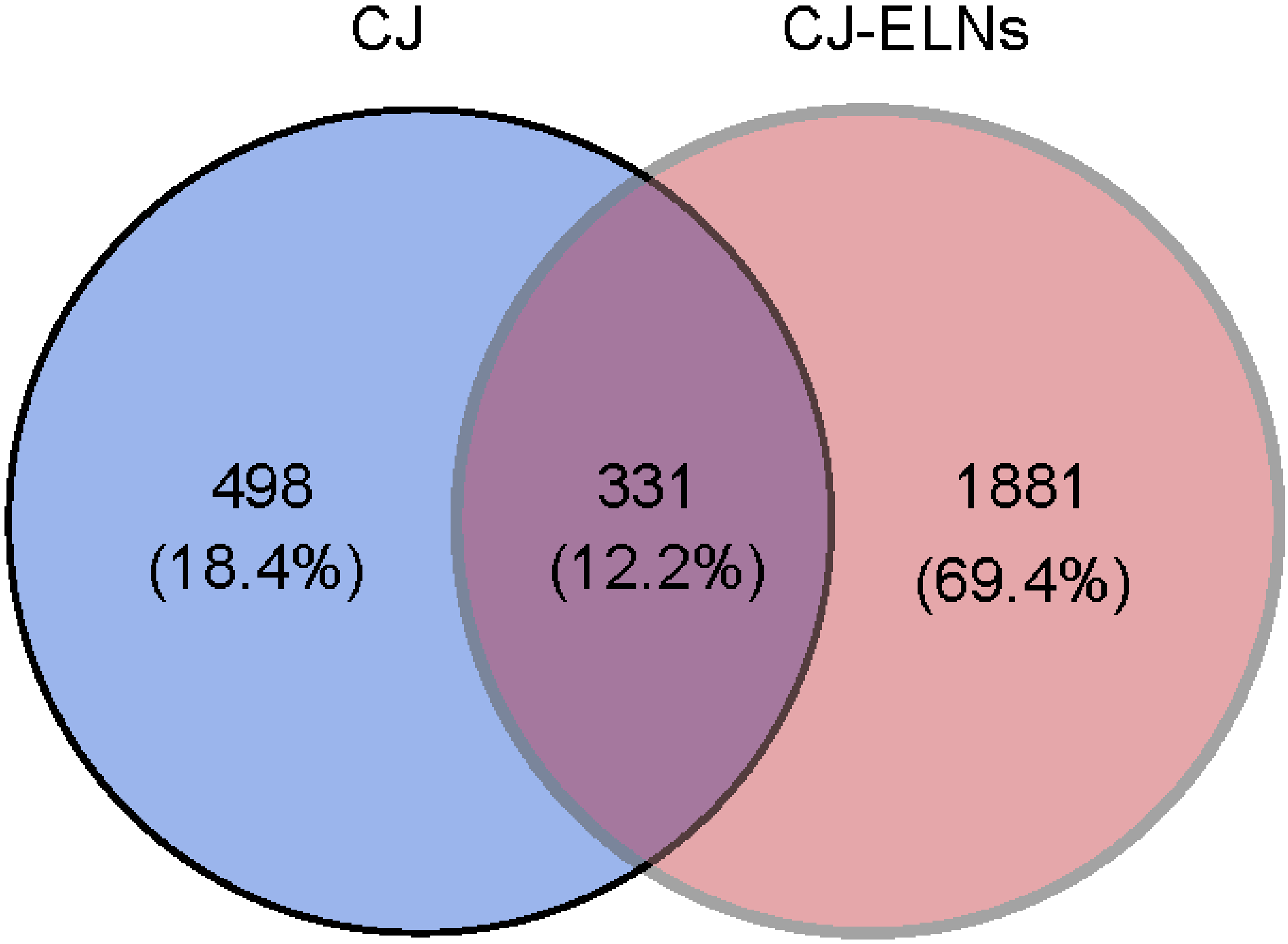
| Compound | Functions | Categories | Possibility |
|---|---|---|---|
| (-)-Epicatechin 3-O-gallate | Antioxidant, anti-inflammatory, anti-cancer, cardiovascular protection, glucose and lipid metabolism regulation. | Organic compound, antioxidant factor, anti-inflammatory factor, energy metabolism, phenolic 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.

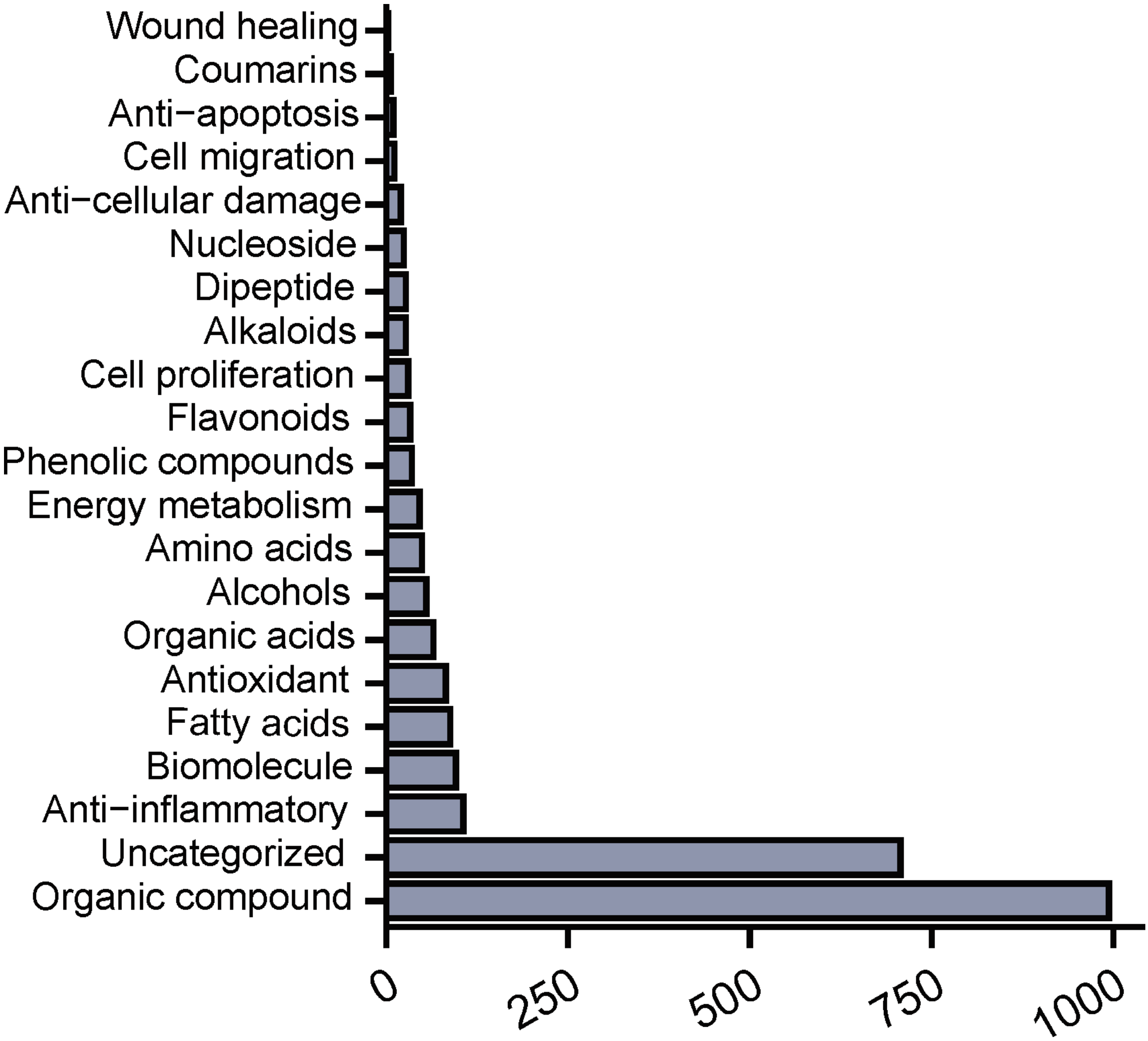


### Rapid and Accurate Data Mining of Compounds in CJ-ELNs With Functional Properties

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.

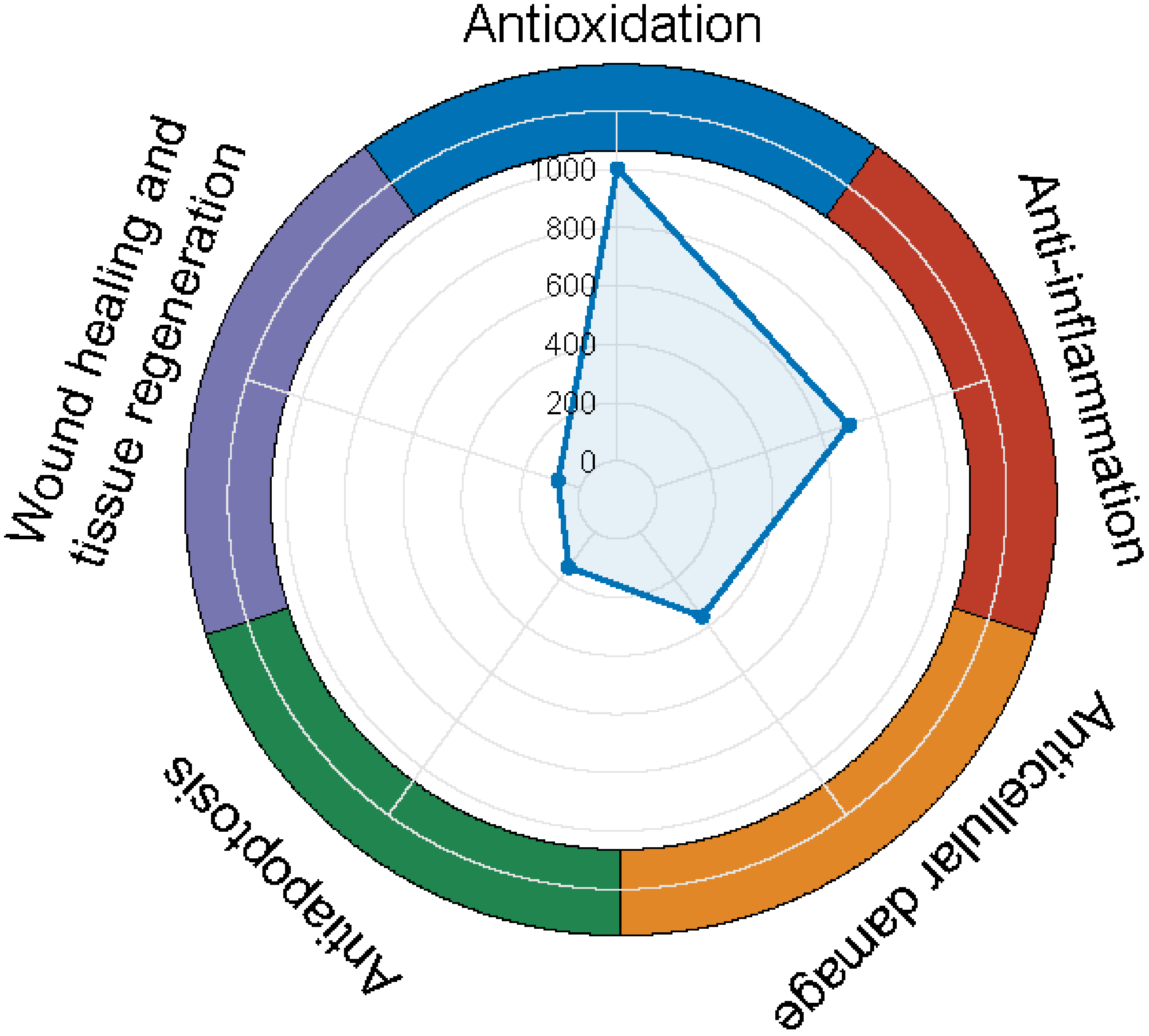
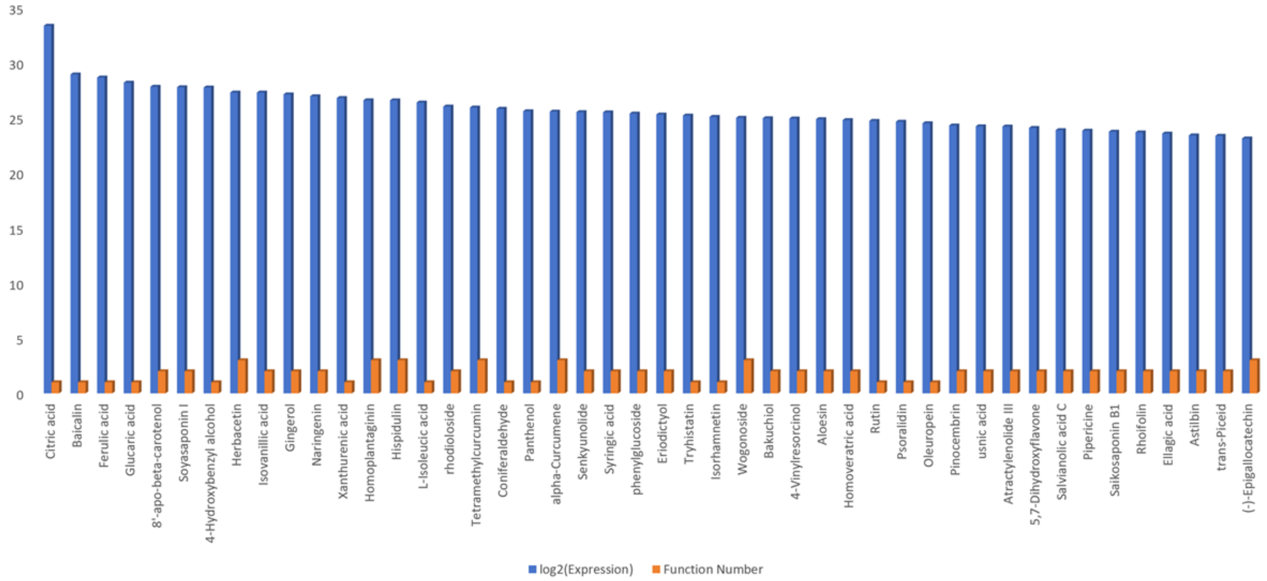


**Figure 6.** Expression levels ($\log_2$-transformed) of naturally active compounds derived from *Cayratia japonica* exosome-like nanovesicle identified by an integration of liquid chromatography-mass spectrometry and DeepSeek models.



# 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 12.07.25; peer-reviewed by V Alevizos, X Kang; revised version received 14.10.25; accepted 19.10.25; published 08.01.26.*

*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*

# 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.

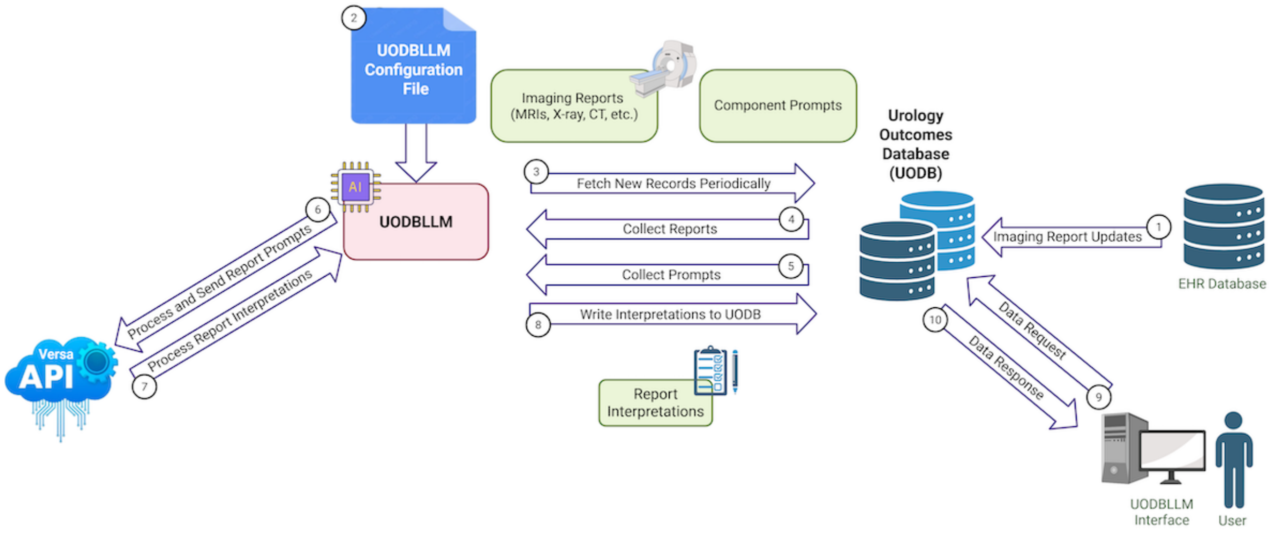
# Methods

## 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).


2
UODBLLM Configuration File
Imaging Reports (MRIs, X-ray, CT, etc.)
Component Prompts
Urology Outcomes Database (UODB)
AI
UODBLLM
3
Fetch New Records Periodically
4
Collect Reports
5
Collect Prompts
8
Write Interpretations to UODB
1
Imaging Report Updates
EHR Database
6
Process and Send Report Prompts
Process Report Interpretations
7
Versa API
Report Interpretations
10
Data Response
Data Request
9
UODBLLM Interface
User

## 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 30.12.24; peer-reviewed by A Abe, JGD Ochoa, S Mohanadas, T Ekundayo; revised version received 29.08.25; accepted 21.10.25; published 06.01.26.*

*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

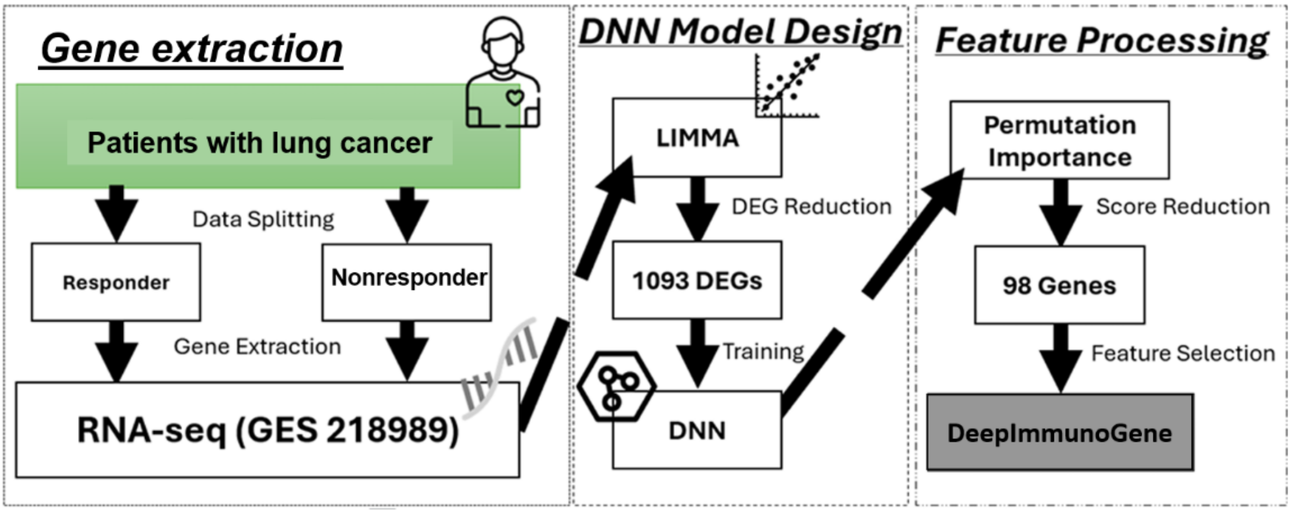
PD-1 immunotherapy, and how do they contribute to model performance?

## 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.

$$(1)\quad L_{BCE}=-\frac{1}{N}\sum_{i=1}^{N} y_i\times\log(p(y_i))+(1-y_i)\times\log(1-p(y_i))$$

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.

$$(2)\text{Importance score}=\text{accuracybaseline}-\text{accuracypermutation}$$

### 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.

$$(3)\text{Accuracy}=\text{TP+TNTP+TN+FP+FN} \times \in [0,1]$$

$$(4)\text{Recall}=\text{TPTP+FN}\times\in 0,1$$

$$(5)\text{Specificity}=\text{TNTN+FP}\times\in 0,1$$

$$(6)\text{Precision}=\text{TPTP+FP}\times\in 0,1$$

$$(7)\text{F1}=2\ \times\text{Precision}\ \times\text{Recall Precision+Recall}\ \times \in [0,1]$$

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

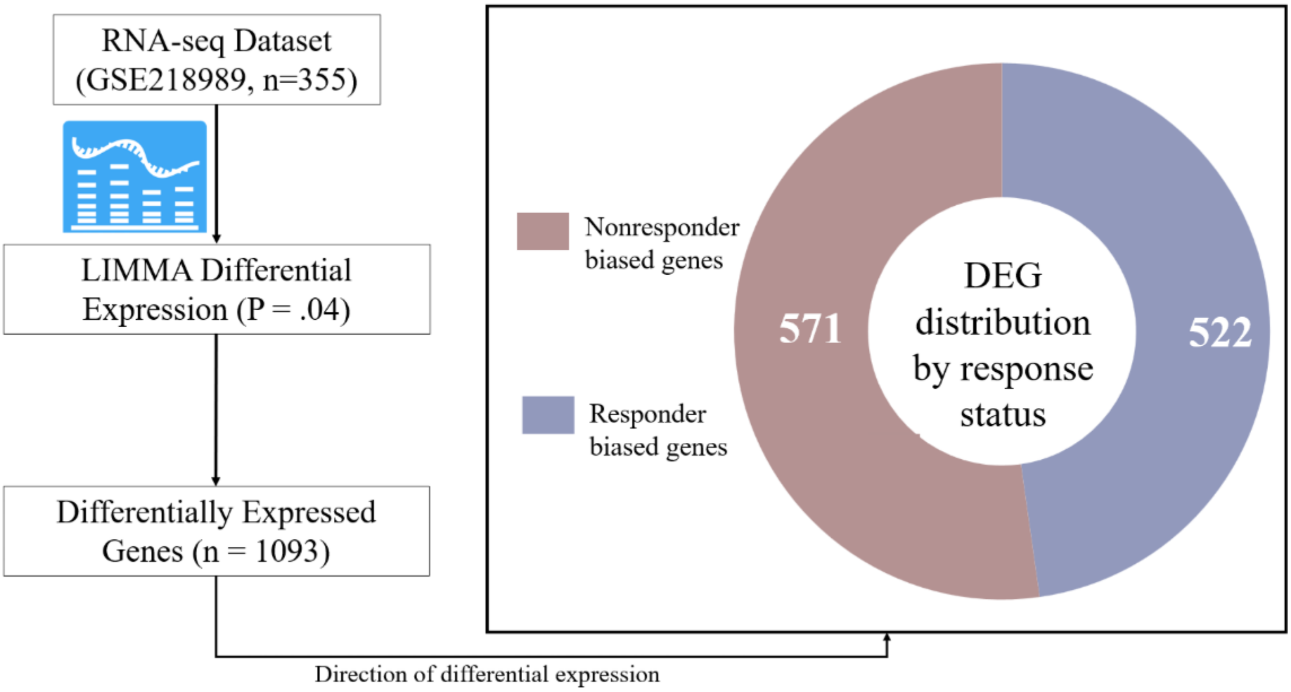
This study used only publicly available or fully deidentified secondary data; therefore, institutional review board approval and informed consent were not required. No personal identifiers were accessed, and privacy and confidentiality were strictly maintained.

## 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.


RNA-seq Dataset (GSE218989, n=355)
LIMMA Differential Expression (P = .04)
Differentially Expressed Genes (n = 1093)
Direction of differential expression
Nonresponder biased genes
Responder biased genes
571
522
DEG distribution by response status

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.

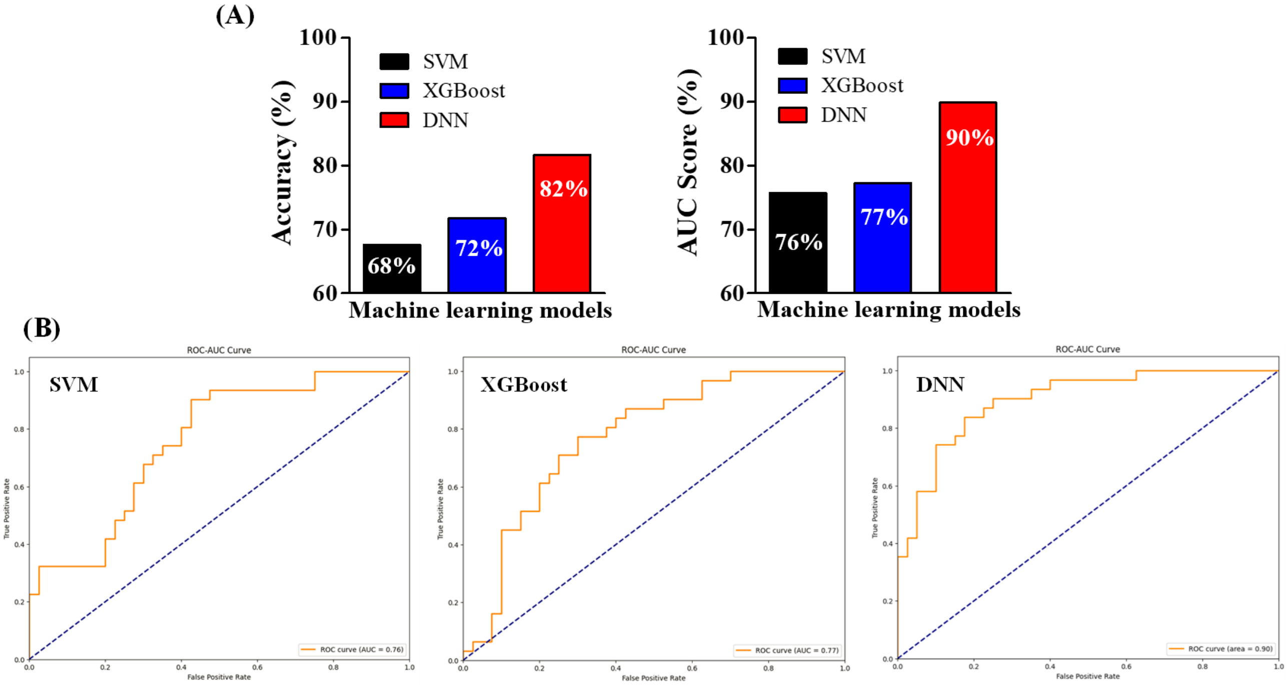


**Table .** Performance comparison of machine learning models for predicting response to programmed cell death receptor-1 immunotherapy.

| 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.

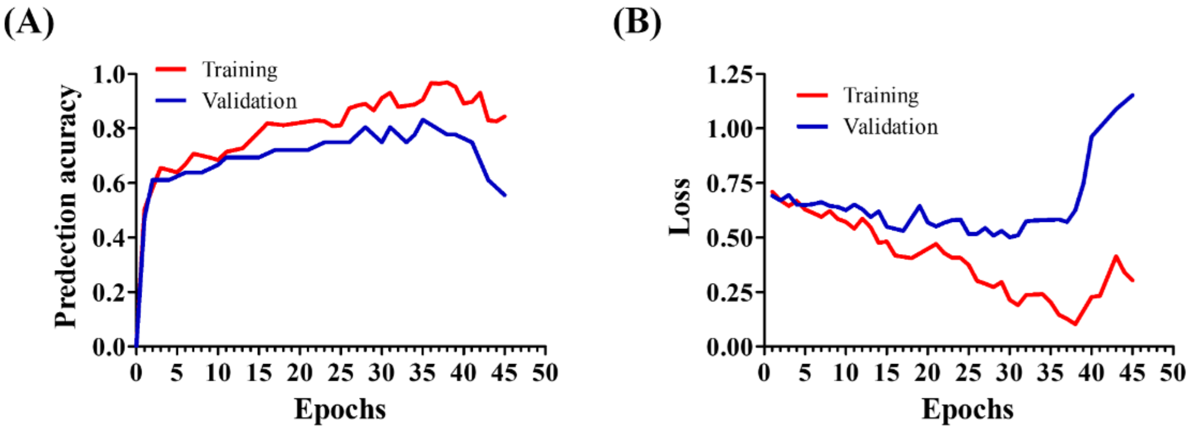
[e] A statistically significant difference from DeepImmunoGene when compared to SVM or XGBoost.

## 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.

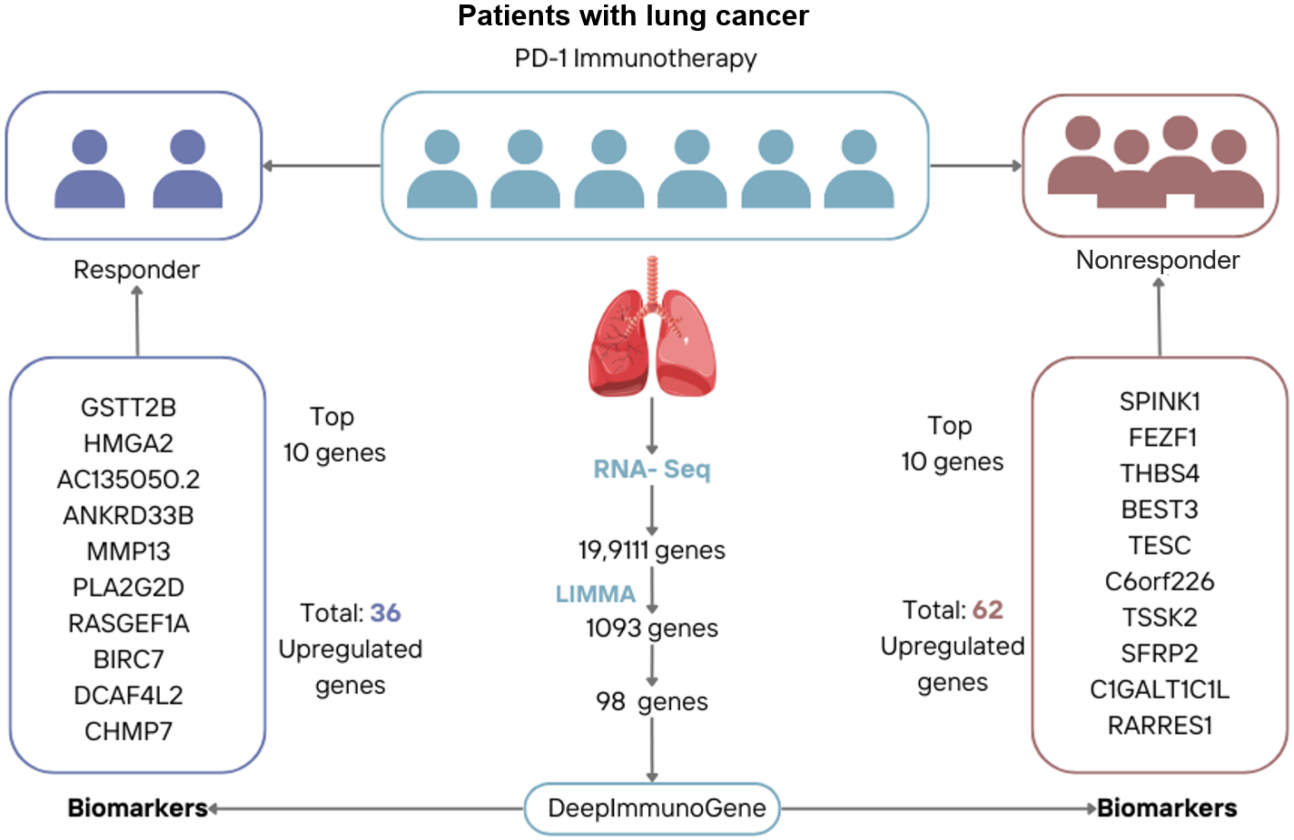


**Figure 6.** Gene importance ranking using permutation analysis. Permutation importance applied to the 1093 differentially expressed genes using the DeepImmunoGene model identified 98 genes with nonzero importance scores. These genes were ranked based on their contribution to model prediction performance,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

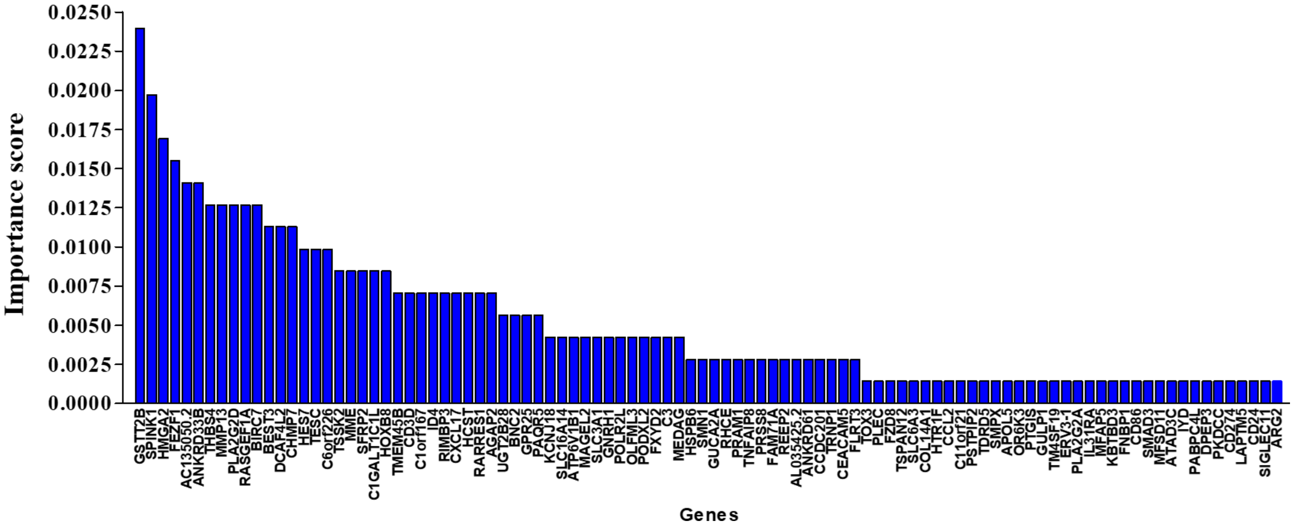
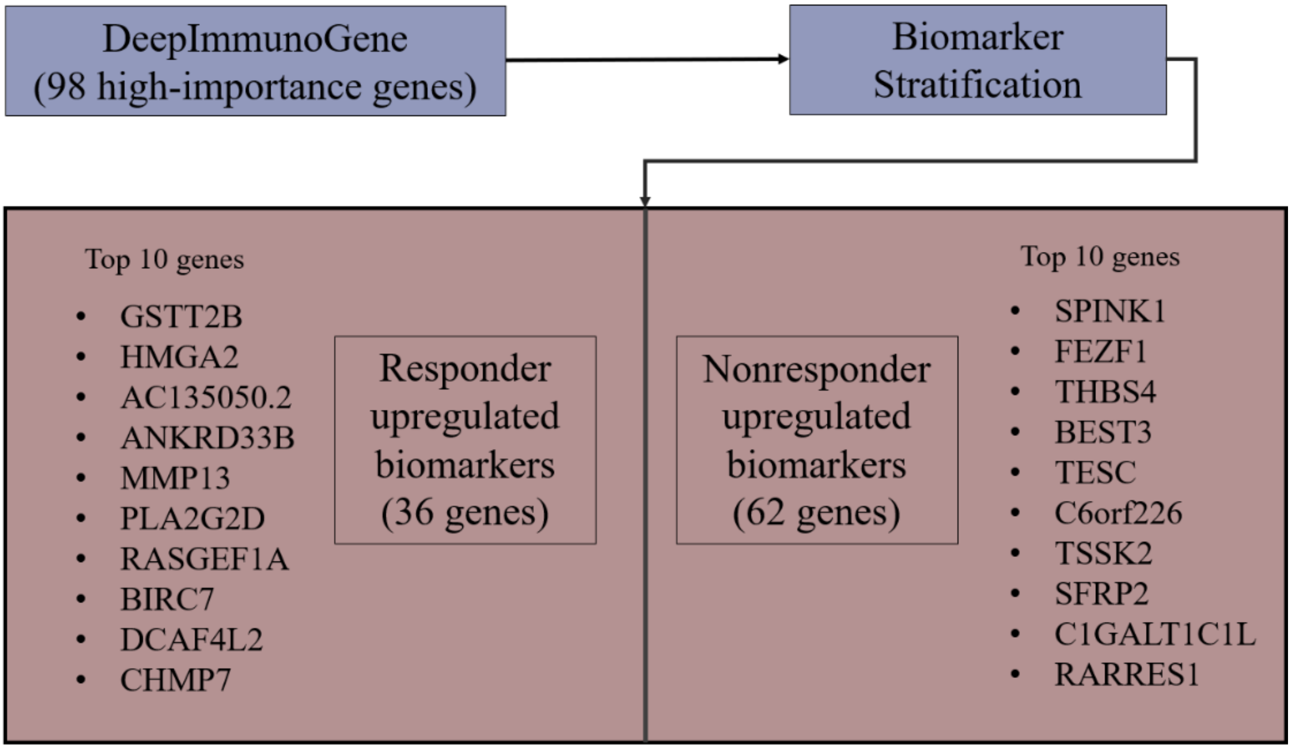


**Figure 7.** DeepImmunoGene-based stratification of predictive biomarkers associated with programmed cell death receptor-1 (PD-1) immunotherapy response. Using permutation importance and deep neural network modeling, 98 high-importance genes were identified and stratified based on direction of differential 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.



DeepImmunoGene (98 high-importance genes)

Biomarker Stratification

Top 10 genes

- GSTT2B
- HMGA2
- AC135050.2
- ANKRD33B
- MMP13
- PLA2G2D
- RASGEF1A
- BIRC7
- DCAF4L2
- CHMP7

Responder upregulated biomarkers (36 genes)

Nonresponder upregulated biomarkers (62 genes)

Top 10 genes

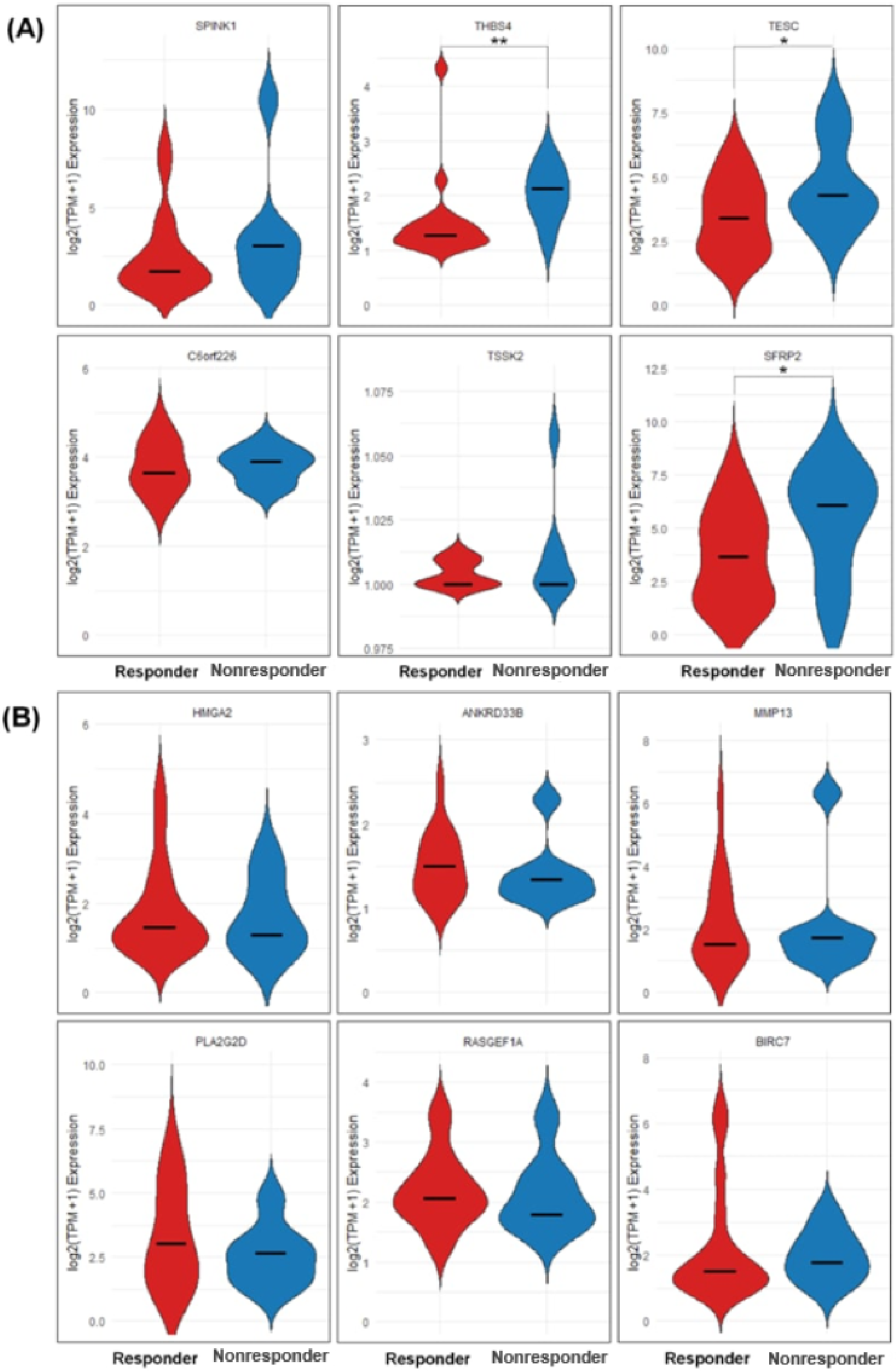
- SPINK1
- FEZF1
- THBS4
- BEST3
- TESC
- C6orf226
- TSSK2
- SFRP2
- C1GALT1C1L
- RARRES1

## 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.


(A)
SPINK1
log2(TPM+1) Expression
0
5
10
THBS4
**
0
1
2
3
4
TESC
*
0.0
2.5
5.0
7.5
10.0
C6orf226
0
2
4
6
TSSK2
0.975
1.000
1.025
1.050
1.075
SFRP2
*
0.0
2.5
5.0
7.5
10.0
12.5
Responder
Nonresponder
(B)
HMGA2
0
2
4
6
ANKRD33B
0
1
2
3
MMP13
0
2
4
6
8
PLA2G2D
0.0
2.5
5.0
7.5
10.0
RASGEF1A
0
1
2
3
4
BIRC7
0
2
4
6
8
Responder
Nonresponder

# 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 24.12.24; peer-reviewed by KK Raja, PB Chandrashekar; revised version received 02.09.25; accepted 04.10.25; published 13.01.26.*

*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*

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