Iranian Journal of Soil and Water Research
Homepage: http://ijswr.ut.ac.ir
ISNN: 2423-7833

Comparing Machine Learning Algorithms for Identifying Antibiotic Resistance Genes (ARGs) in the Agricultural Soil Microbiome; Case Study in East Asia

Seyede Reyhaneh KeshikNevisRazavi[1] | Elham Farahani[2]✉ | Hojat Emami[3] | Narges Abedinzadeh[4] | Mohammad Abdolahi[5]

1. Department of Soil Science, Faculty of Agriculture, Ferdowsi University, Mashhad, Iran. E-mail: reyhaneh.razavi@mail.um.ac.ir,
2. Corresponding Author, Soil and Water Research Institute of Iran, Agricultural Research, Education and Extension Organization (AREEO), Karaj, Iran. E-mail: e_farahani@areeo.ac.ir
3. Department of Soil Science, Faculty of Agriculture, Ferdowsi University, Mashhad, Iran. E-mail: hemami@um.ac.ir
4. Department of Soil Science, Faculty of Agriculture, Ferdowsi University, Mashhad, Iran. E-mail: abedinzadeh.narges@alumni.um.ac.ir
5. Computer Department, Jihad Daneshgahi of Khorasan Razavi, Mashhad, Iran. E-mail: mabdolahi512@yahoo.com

Article Info

Article type: Research Article

Article history:

Received: Apr. 22, 2026

Revised: June. 3, 2026

Accepted: June. 7, 2026

Published online: June. 2026

Keywords:
Antibiotic resistance,
Machine learning,
Metagenomics,
Resistance genes,
Soil microbiome

ABSTRACT

With the escalating threat of antibiotic resistance, the accurate and comprehensive identification of antibiotic resistance genes (ARGs) in natural environments, particularly agricultural soils, has become a major concern in public and environmental health. In recent years, the application of machine learning algorithms has gained attention as a novel approach for analyzing complex metagenomic data and improving ARG detection. In this study, four machine learning algorithms—Random Forest, Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Multilayer Perceptron (MLP)—were compared for their ability to identify resistance genes in the agricultural soil microbiome in India and China. Metagenomic data were obtained from the NCBI database and processed using the ARGs-OAP tool. A set of biological features, including GC content, amino acid frequency, and codon usage, was extracted. Statistical differences between resistant and non-resistant genes were assessed using the Mann–Whitney U test, and only significant features were selected for model training. The results demonstrated that the models, particularly Random Forest (with 98% accuracy), were capable of identifying resistance genes with high performance, even under conditions of imbalanced data and limited training sample size. These findings highlight the effectiveness of the selected biological features and machine learning algorithms in detecting ARGs in the agricultural soil microbiome in East Asia. This approach could serve as an efficient tool for environmental monitoring and policy-making aimed at controlling the spread of antibiotic resistance.

Cite this article: KeshikNevisRazavi, S.R., Farahani, E., Emami, H Abedinzadeh, N., & Abdolahi, M. (2026). Comparing Machine Learning Algorithms for Identifying Antibiotic Resistance Genes (ARGs) in the Agricultural Soil Microbiome; Case Study in East Asia, *Iranian Journal of Soil and Water Research*, 57 (4), 929-947. https://doi.org/10.22059/ijswr.2026.412755.670117

Publisher: University of Tehran Press.

DOI: https://doi.org/10.22059/ijswr.2026.412755.670117

EXTENDED ABSTRACT

Introduction:

The spread of antibiotic resistance genes (ARGs) in agricultural soils has emerged as a critical public health threat, with soils serving as both reservoirs and transmission pathways for resistance determinants. Accurate identification of ARGs within complex soil metagenomic data is essential for monitoring resistance dissemination, yet conventional alignment-based methods remain limited to detecting known resistance genes and fail to identify novel or divergent variants. This study applies four machine learning algorithms, Random Forest, Support Vector Machine (SVM), XGBoost, and Multilayer Perceptron (MLP), to classify ARGs based on biologically informative sequence features, including GC content, codon usage, and amino acid composition. By systematically evaluating model performance under imbalanced data conditions and limited training samples, this work provides a comparative assessment of machine learning approaches for resistance gene detection and demonstrates the utility of interpretable features in distinguishing resistant from non-resistant sequences in agricultural soil microbiomes.

Objective(s)

The specific objectives of this study were to: (1) develop and train four machine learning models (Random Forest, SVM, XGBoost, and MLP) for binary classification of antibiotic resistance genes (ARGs) versus non-ARGs using sequence-derived features (GC content, codon frequency, and amino acid composition); (2) compare model performance using precision, recall, F1-score, and cross-validated AUC-ROC and AUC-PRC metrics under varying training/test splits (80/20, 70/30, 60/40, and 10/90); (3) identify the most influential biological features contributing to ARG classification through feature importance analysis; and (4) evaluate model robustness and generalization capability under class-imbalanced conditions representative of real-world metagenomic datasets.

Methods

This study utilized metagenomic sequence data from agricultural soil microbiomes obtained from the NCBI database and processed through the ARGs-OAP pipeline against the SARG database. A total of approximately 400,000 sequences were initially retrieved, with quality control performed using FastQC and Cutadapt. Biological features, including GC content, amino acid composition (21 amino acids), and codon usage frequency (64 codons), were extracted from each sequence. Given the highly imbalanced nature of the dataset (50 ARGs vs. hundreds of thousands of non-ARGs), a balanced subset was constructed by retaining all 50 ARG sequences and selecting 150 non-ARG sequences with statistically significant feature differences (Mann-Whitney U test, $p < 0.05$) and GC content restricted to 10-30%. Four machine learning algorithms, Random Forest, Support Vector Machine (SVM), XGBoost, and Multilayer Perceptron (MLP), were applied to the final dataset of 200 samples. Model performance was evaluated using precision, recall, and F1-score, with cross-validation performed under multiple training/test split ratios (80/20, 70/30, 60/40, and 10/90). Feature importance analysis was conducted to identify the most discriminative biological predictors.

Results

All four machine learning models demonstrated strong performance in ARG classification. Random Forest achieved the highest overall accuracy (0.9877) with perfect recall for the non-resistant class (1.000) and near-perfect recall for resistant genes (0.9524). MLP showed the highest cross-validated mean F1-score (0.9581 ± 0.0656) and achieved perfect recall for resistant genes (0.9524) with minimal false positives. SVM performed reliably with 0.9753 accuracy but showed lower recall for resistant genes (0.9048). XGBoost exhibited the weakest performance among the four models, with the lowest recall for resistant genes (0.8571) and lowest cross-validated F1-score (0.9379 ± 0.0791). AUC-ROC and AUC-PRC analyses confirmed Random Forest as the top performer (AUC-ROC = 0.987 ± 0.005; AUC-PRC = 0.984 ± 0.006), followed closely by MLP (AUC-ROC = 0.981 ± 0.006; AUC-PRC = 0.976 ± 0.007). Confusion matrices revealed that Random Forest and MLP misclassified only one sample each, while XGBoost misclassified three resistant genes. Feature importance analysis consistently identified specific codons, particularly CTG (Leucine), GCG (Alanine), and CGC (Arginine), along with GC content, as the most influential predictors across all models.

Conclusions

This study demonstrates that machine learning models, particularly Random Forest and Multilayer Perceptron, can effectively identify antibiotic resistance genes in agricultural soil microbiomes using biologically derived sequence features, even under class-imbalanced conditions and limited sample sizes. Key biological predictors, including specific codon usage patterns (CTG, GCG, CGC), amino acid composition (Leucine, Valine), and GC content, were identified as robust discriminators between resistant and non-resistant

sequences. These findings confirm that sequence-intrinsic features alone can provide sufficient signal for ARG detection independent of homology-based methods. The superior performance of ensemble and neural network approaches suggests their potential for integration into environmental monitoring pipelines. Future research should focus on validating these models on larger, geographically diverse datasets and incorporating additional feature types (e.g., genomic context, mobile genetic elements) to enhance generalizability and enable real-world deployment for antibiotic resistance surveillance in agricultural ecosystems.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authorship contribution

KeshikNevisRazavi, S.R. contributed to the conceptualization, methodology, software development, investigation, data curation, writing of the original draft, and manuscript review and editing. Farahani, E. was involved in validation, writing of the original draft, review and editing of the manuscript, supervision of the research process, and overall project administration. Emami, H. contributed to the validation, formal analysis, investigation, data curation, and writing of the original draft. Abedinzadeh, N. participated in software development, data curation, and writing of the original draft. Abdolahi, M. contributed to the methodology, software implementation, validation, formal analysis, investigation, data curation, writing of the original draft, review and editing and visualization of results. All authors have read and approved the final version of the manuscript.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors declare that no AI and AI-assisted technologies is used in this article.

Data availability statement

The data and materials used in the study can be available base on a reasonable request.

Acknowledgements

The authors would like to thank the contributors and maintainers of the NCBI database and the developers of the ARGs-OAP pipeline for providing open-access resources that made this study possible.

Ethical considerations

No human or animal subjects were involved, and therefore no ethical approval was required.

Conflict of interest

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

مجله تحقیقات آب و خاک ایران، دوره ۵۷، شماره ۴ شاپا: ۷۸۳۳-۲۴۲۳
Homepage: http://ijswr.ut.ac.ir

مقایسه الگوریتم‌های یادگیری ماشین برای شناسایی ژن‌های مقاوم به آنتی‌بیوتیک (ARGs) در میکروبیوم خاک‌های کشاورزی؛ مطالعه موردی در شرق آسیا

سیده ریحانه کشیک‌نویس رضوی[۱] | الهام فراهانی[۲] ✉ | حجت امامی[۳] | نرگس عابدین‌زاده[۴] | محمد عبداللهی[۵]

۱. گروه علوم خاک، دانشکده کشاورزی، دانشگاه فردوسی، مشهد، ایران. رایانامه: reyhaneh.razavi@mail.um.ac.ir

۲. نویسنده مسئول، سازمان تحقیقات، آموزش و ترویج کشاورزی، موسسه تحقیقات خاک و آب ایران، کرج، ایران. رایانامه: e_farahani@areeo.ac.ir

۳. گروه علوم خاک، دانشکده کشاورزی، دانشگاه فردوسی، مشهد، ایران. رایانامه: hemami@um.ac.ir

۴. گروه علوم خاک، دانشکده کشاورزی، دانشگاه فردوسی، مشهد، ایران. رایانامه: abedinzadeh.narges@alumni.um.ac.ir

۵. گروه کامپیوتر، جهاد دانشگاهی خراسان رضوی، مشهد، ایران. رایانامه: mabdolahi512@yahoo.com

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۵/۲/۲
تاریخ بازنگری: ۱۴۰۵/۳/۱۳
تاریخ پذیرش: ۱۴۰۵/۳/۱۷
تاریخ انتشار: تیر ۱۴۰۵

واژه‌های کلیدی:
ژن‌های مقاوم،
مقاومت آنتی‌بیوتیکی متاژنومیک،
میکروبیوم خاک،
یادگیری ماشین

چکیده

با گسترش روزافزون مقاومت آنتی‌بیوتیکی، شناسایی دقیق و جامع ژن‌های مقاومت آنتی‌بیوتیکی (ARGs) در محیط‌های طبیعی به‌ویژه خاک‌های کشاورزی، به یکی از دغدغه‌های مهم در حوزه سلامت عمومی و زیست‌محیطی تبدیل شده است. در سال‌های اخیر، بهره‌گیری از الگوریتم‌های یادگیری ماشین به‌عنوان رویکردی نوین برای تحلیل داده‌های پیچیده متاژنومیکی و بهبود شناسایی ARGs مورد توجه قرار گرفته است. در پژوهش حاضر، چهار الگوریتم یادگیری ماشین شامل جنگل تصادفی، تقویت گرادیانی، ماشین بردار پشتیبان و شبکه عصبی چند لایه با هدف شناسایی ژن‌های مقاوم در میکروبیوم خاک‌های کشاورزی دو کشور هند و چین مورد مقایسه قرار گرفتند. داده‌های متاژنومیکی از پایگاه داده NCBI استخراج و توسط ابزار ARGs-OAP پردازش شدند. مجموعه‌ای از ویژگی‌های زیستی شامل محتوای GC ، فراوانی آمینواسیدها و کدون‌ها استخراج گردید. تفاوت آماری میان ژن‌های مقاوم و غیرمقاوم با آزمون Mann–Whitney بررسی شد و تنها ویژگی‌های معنادار جهت آموزش مدل‌ها انتخاب شدند. نتایج نشان داد که مدل‌ها، به‌ویژه جنگل تصادفی (با دقت ۹۸٪)، قادر به شناسایی ژن‌های مقاوم با عملکرد بالا حتی در شرایط داده‌های نامتوازن و حجم آموزش محدود بودند. این یافته‌ها نشان‌دهنده کارایی بالای ویژگی‌های زیستی منتخب و الگوریتم‌های یادگیری ماشین در شناسایی ARGs در میکروبیوم خاک‌های کشاورزی شرق آسیا است، و می‌تواند به‌عنوان ابزاری کارآمد در سیاست‌گذاری‌های زیست‌محیطی و کنترل گسترش مقاومت آنتی‌بیوتیکی مورد استفاده قرار گیرد.

استناد: کشیک نویس رضوی؛ ریحانه، فراهانی؛ الهام، امامی؛ حجت، عابدین‌زاده؛ نرگس، عبداللهی؛ محمد، (۱۴۰۵) مقایسه الگوریتم‌های یادگیری ماشین برای شناسایی ژن‌های مقاوم به آنتی‌بیوتیک (ARGs) در میکروبیوم خاک‌های کشاورزی؛ مطالعه موردی در شرق آسیا، *مجله تحقیقات آب و خاک ایران*، ۵۷ (۴)، ۹۲۹-۹۴۷. https://doi.org/10.22059/ijswr.2026.412755.670117

ناشر: مؤسسه انتشارات دانشگاه تهران.

DOI: https://doi.org/10.22059/ijswr.2026.412755.670117

مقدمه

مقاومت آنتی‌بیوتیکی یکی از چالش‌های جدی و رو به رشد در بهداشت عمومی جهانی است که تهدیدی فزاینده برای سلامت انسان‌ها، حیوانات و محیط زیست به شمار می‌رود. بر اساس گزارش سازمان جهانی بهداشت (WHO) در سال ۲۰۲۵، از هر شش عفونت باکتریایی در آزمایشگاه، یک مورد به درمان‌های آنتی‌بیوتیکی رایج مقاوم است. این گزارش که با استفاده از داده‌های بیش از ۲۳ میلیون عفونت از ۱۰۴ کشور جهان تهیه شده، نشان می‌دهد میزان مقاومت در منطقه آسیای جنوب شرقی و مدیترانه شرقی بالاترین سطح را دارد (World Health Organization, 2025). تخمین زده می‌شود که تا سال ۲۰۵۰، مرگ و میر ناشی از این پدیده سالانه به بیش از ۱۰ میلیون نفر خواهد رسید و هزینه‌های اقتصادی آن از مرز ۱۰۰ تریلیون دلار فراتر رود (O'Neill, 2016). این مقاومت زمانی رخ می‌دهد که باکتری‌ها پس از قرار گرفتن در معرض آنتی‌بیوتیک‌ها، که به‌طور معمول آن‌ها را از بین می‌برند یا رشدشان را متوقف می‌کنند، قادر به بقا و تکثیر باشند (Vuong et al., 2016; Gandhi et al., 2010; Mediavilla et al., 2016).

گزارش‌های اخیر حاکی از آن است که سالانه نزدیک به ۷۰۰ هزار نفر در سراسر جهان به دلیل عفونت‌های ناشی از باکتری‌های مقاوم به آنتی‌بیوتیک جان خود را از دست می‌دهند (Hu et al., 2016). سازمان جهانی بهداشت برای مقابله با این تهدید، یک برنامه اقدام جهانی را راه‌اندازی کرده است که هدف آن کنترل و کاهش مقاومت آنتی‌بیوتیکی در محیط‌های انسانی، حیوانی و زیست‌محیطی است. یکی از نگرانی‌های عمده، انتقال ژن‌های مقاوم به آنتی‌بیوتیک (ARGs) از خاک به اکوسیستم‌های انسانی، حیوانی و گیاهی است که می‌تواند پیامدهای جدی برای سلامت عمومی و امنیت غذایی به دنبال داشته باشد (Pehrsson et al., 2016; Apweiler et al., 2004; Berendonk et al., 2015).

خاک‌های کشاورزی به عنوان مخزن اصلی ژن‌های مقاومت به شمار می‌روند. استفاده گسترده از آنتی‌بیوتیک‌ها در دامپروری و کشاورزی، همراه با کاربرد کودهای دامی و آبیاری با فاضلاب، منجر به تجمع این ژن‌ها در خاک شده است (Delgado-Baquerizo et al., 2022). در یک پژوهش جامع ۲۸۵ ژن مقاومت در نمونه‌های خاک از ۱۰۱۲ نقطه جهان تحلیل شد و اولین نقشه جهانی توزیع ARGها در خاک‌های سطحی را ارائه گردید. نتایج نشان داد که اکوسیستم‌های خاکی به کانون‌های انتشار این ژن‌ها تبدیل شده‌اند (Delgado-Baquerizo et al., 2022). انتقال ژن‌های مقاومت از خاک به اکوسیستم‌های انسانی، حیوانی و گیاهی عمدتاً از طریق انتقال افقی ژن (HGT) و با استفاده از عناصر ژنتیکی متحرک مانند پلاسمیدها صورت می‌گیرد (Deng et al., 2025). با وجود پیشرفت‌های چشم گیر در فناوری‌های توالی‌یابی، شناسایی دقیق و جامع ARGها در محیط‌های پیچیده‌ای مانند خاک‌های کشاورزی همچنان با چالش‌هایی مواجه است. ازاین‌رو، توسعه روش‌های نوین و کارآمد برای شناسایی این ژن‌ها، گامی ضروری در جهت پایش و کنترل گسترش مقاومت آنتی‌بیوتیکی محسوب می‌شود.

هدف از این پژوهش، بررسی و ارزیابی کارایی چهار الگوریتم یادگیری ماشین شامل جنگل تصادفی، ماشین بردار پشتیبان، تقویت گرادیانی و شبکه عصبی چندلایه در شناسایی ژن‌های مقاوم به آنتی‌بیوتیک در میکروبیوم خاک‌های کشاورزی شرق آسیا است. این پژوهش با بهره‌گیری از ویژگی‌های زیستی استخراج‌شده از توالی‌ها (محتوای GC، فراوانی کدون‌ها و ترکیب اسیدهای آمینه)، به دنبال ارائه رویکردی مکمل برای روش‌های سنتی مبتنی بر هم‌ردیفی و کمک به درک بهتر توزیع و انتشار ژن‌های مقاوم در این اکوسیستم است.

پیشینه پژوهش

ظهور فناوری‌های پیشرفته توالی‌یابی DNA با توان عملیاتی بالا، ابزارهای قدرتمندی برای شناسایی و تحلیل ARGها در بخش‌های مختلف محیطی فراهم کرده است. این روش‌ها امکان نمایه‌سازی جامع ترکیب ژنتیکی جوامع میکروبی از نمونه‌هایی مانند کود دامی، کمپوست، فاضلاب، خاک و آب آلوده را فراهم می‌کنند (Pal et al., 2016; Forsberg et al., 2014; Pruden et al., 2013). شناسایی ARGها عمدتاً از طریق مقایسه توالی‌های DNA متاژنومی با پایگاه‌های داده مرجع صورت می‌گیرد. پایگاه (Comprehensive Antibiotic Resistance Database) CARD با ساختار هستی‌شناسی مقاومت آنتی‌بیوتیکی (ARO) و معیارهای دقیق ورود اطلاعات، یکی از معتبرترین منابع در این حوزه محسوب می‌شود (Jia et al., 2017). پایگاه (Structured Antibiotic Resistance Genes) SARG نیز به عنوان یک پایگاه تلفیقی، داده‌ها را از منابع متعدد گردآوری کرده و در ابزارهایی مانند ARGs-OAP مورد استفاده قرار می‌گیرد (Yin et al., 2022). مقایسه جامع پایگاه‌های داده ARG شامل CARD، ARDB، SARG، ResFinder، DeepARG-DB و HMD-ARG-DB در مطالعات

اخیر ارائه شده است. این فرآیند معمولاً از طریق هم‌ردیفی قرائت‌های خام یا خواندن فریم‌های باز پیش‌بینی‌شده (ORFs) با استفاده از ابزارهایی مانند BLAST (States & Agarwal, 1996)، Bowtie (Langmead et al., 2009) و DIAMOND (Buchfink et al., 2015) انجام می‌شود.

با این حال، روش‌های سنتی بیوانفورماتیکی به‌طور عمده به شناسایی ARGهای شناخته‌شده محدود هستند و توانایی شناسایی انواع جدید یا ناشناخته را ندارند (McArthur & Tsang, 2017ORFs). برخی ابزارها مانند ResFinder (Kleinheinz et al., 2014) و (Rowe et al., 2015) SEAR به‌طور خاص برای شناسایی ARGهای پلاسمیدی طراحی شده‌اند، در حالی که سامانه‌هایی مانند (Bradley et al., 2015) Mykrobe تنها بر روی ۱۲ نوع خاص از مقاومت‌های ضد میکروبی تمرکز دارند. بسیاری از این ابزارها بر رویکرد "بهترین تطابق" متکی هستند که دقت آن‌ها به میزان شباهت توالی‌ها به ARGهای موجود در پایگاه‌های داده بستگی دارد.

با توجه به محدودیت‌های روش‌های سنتی، یادگیری ماشین به عنوان یک روش جایگزین برای شناسایی ARGها معرفی شده است. مدل‌های یادگیری ماشین قادرند با تحلیل ویژگی‌های مختلف توالی‌های ژنتیکی، الگوهای جدیدی که در پایگاه‌های داده موجود نیستند را شناسایی کنند. مطالعات کاربردهای متنوع هوش مصنوعی را در شناسایی مقاومت آنتی‌بیوتیکی بررسی کرده و نشان داده است که الگوریتم‌های یادگیری ماشین می‌توانند در پیش‌بینی مقاومت از روی داده‌های ژنومی، استخراج ویژگی از پان‌ژنوم و تحلیل متاژنومی نقش مؤثری ایفا کنند (Scaglione et al., 2026). این پژوهش تأکید می‌کند که مدل‌های یادگیری ماشین، به‌ویژه رویکردهای مبتنی بر شبکه‌های عصبی عمیق و جنگل تصادفی، پتانسیل بالایی برای یکپارچه‌سازی با فناوری‌های نسل بعد توالی‌یابی (NGS) و تبدیل داده‌های خام به بینش‌های بالینی قابل اقدام دارند.

Davis et al., (۲۰۱۶) در پایگاه داده PATRIC از ترکیب روش‌های مبتنی بر هم‌ردیفی و یادگیری ماشین برای پیش‌بینی مقاومت آنتی‌بیوتیکی استفاده کردند. Arango-Argoty et al. (۲۰۱۸) مدل DeepARG را بر اساس یادگیری عمیق توسعه دادند. این ابزار که در پژوهش Scaglione et al. (۲۰۲۶) به عنوان یکی از ابزارهای پیشرفته معرفی شده، با استفاده از ویژگی‌های استخراج‌شده از توالی‌ها و رویکردهای مبتنی بر k-mer، قادر به شناسایی ARGها حتی با فراوانی کم است. DeepARG از دو مدل DeepARG-SS برای توالی‌های کوتاه و DeepARG-LS برای توالی‌های بلند استفاده می‌کند و می‌تواند ۳۰ دسته مقاومتی مختلف را پیش‌بینی کند. پژوهش Arango-Argoty et al. (۲۰۱۸) نشان داد که استفاده از ویژگی‌های استخراج‌شده از توالی و نه صرفاً هم‌ردیفی، می‌تواند موجب بهبود دقت شناسایی ARGها گردد.

در پژوهش دیگری، از الگوریتم جنگل تصادفی برای پیش‌بینی حداقل غلظت مهارکنندگی (MIC) سیپروفلوکساسین در اشریشیا کلی استفاده شده است (Pataki et al., 2024). آنها با استفاده از داده‌های توالی‌یابی کل ژنوم از مناطق جغرافیایی متنوع، مدلی توسعه دادند که قادر به شناسایی ویژگی‌های کلیدی مؤثر در مقاومت بود. چهار ویژگی برتر شناسایی‌شده توسط این مدل، برای آموزش مدل نهایی و پیش‌بینی MIC در ایزوله‌هایی با منشاء جغرافیایی ناشناخته مورد استفاده قرار گرفت. این پژوهش نشان‌دهنده قدرت جنگل تصادفی در انتخاب ویژگی و کاهش ابعاد داده‌های پیچیده ژنومی است.

پژوهش مهم دیگر توسط Her و Wu (۲۰۲۴) بر اهمیت ژن‌های غیرهسته‌ای (non-core genes) در پیش‌بینی مقاومت تأکید کرد. آنها با استفاده از رویکرد مبتنی بر پان‌ژنوم و الگوریتم ژنتیک، خوشه‌هایی از ویژگی‌ها را شناسایی کردند که قادر به تفکیک سویه‌های مقاوم از حساس در اشریشیا کلی بودند. مدل آنان برای آنتی‌بیوتیک‌های آمپی‌سیلین، جنتامایسین، تریمتوپریم–سولفامتوکسازول و سیپروفلوکساسین به سطح زیر منحنی (AUC) بالاتر از ۰/۹ دست یافت. جالب توجه اینکه این مدل، خوشه ژنی pmrC/pmrE را که به‌طور کلاسیک با مقاومت به پلی‌میکسین مرتبط است، به‌عنوان پیش‌بینی‌کننده مقاومت به آمپی‌سیلین و تریمتوپریم–سولفامتوکسازول نیز شناسایی کرد که نشان‌دهنده توانایی یادگیری ماشین در کشف الگوهای پنهان مقاومت است. یک چارچوب پان‌ژنوم برای استافیلوکوکوس اورئوس، سودوموناس آئروژینوزا و اشریشیا کلی توسعه داده شده است (Hyun et al., 2024). آنها با ایجاد یک ماتریس دودویی ژنوم–ویژگی و آموزش چندین مدل ماشین بردار پشتیبان به‌صورت گروهی، به دقت‌هایی بین ۷۹/۳ تا ۹۹/۵ درصد دست یافتند. مدل آنان نه تنها ۴۵ ژن مقاومت شناخته‌شده را بازیابی کرد، بلکه ۲۵ کاندید جدید نیز پیشنهاد داد. همچنین گزارش شده است که انتخاب بازه کنترل‌شده‌ای از محتوای GC می‌تواند دقت طبقه‌بندی و مقایسه بین ARGها و سایر ژن‌ها را بهبود دهد (Yang et al., 2013).

مطالعات متعددی به بررسی ژن‌های مقاومت آنتی‌بیوتیکی در خاک‌های کشاورزی با رویکردهای نوین پرداخته‌اند. در پژوهشی بنیادین، با تحلیل نمونه‌های خاک از نقاط مختلف جهان، نخستین نقشه جهانی توزیع ARGها در خاک‌های سطحی ترسیم و نشان داده

شده که فراوانی این ژن‌ها در خاک‌های کشاورزی به‌طور معنی‌داری بیشتر از سایر خاک‌هاست (Delgado-Baquerizo et al.,2022). در پژوهش دیگری، با ترکیب متاژنومیک شات‌گان و یادگیری ماشین، به بررسی ارتباط بین میکروبیوم انسان، دام و خاک در یک مزرعه مرغداری پرداخته شده است (Maciel-Guerra et al., 2022). نتایج آن‌ها نشان داد که ژن‌های مقاومت مشابه با اهمیت بالینی و عناصر ژنتیکی متحرک مرتبط، در نمونه‌های انسانی و دام مشترک است. همچنین، در پژوهشی جامع، توزیع و تنوع ARGها را در میکروبیوم انسان، طیور، خوک و خاک بررسی و همبستگی معنی‌داری بین جوامع باکتریایی و ژن‌های مقاومت، به‌ویژه در نمونه‌های خاک گزارش شده است (Bai et al., 2024). آن‌ها همچنین زیرمجموعه‌هایی از ARGها را به عنوان شاخصی برای ارزیابی سطح آلودگی مقاومت در نمونه‌ها معرفی کردند.

بررسی پژوهش‌های پیشین نشان می‌دهد که اگرچه پژوهش‌های متعددی به کارگیری یادگیری ماشین در شناسایی ARGها را بررسی کرده‌اند، اما شکاف‌های پژوهشی همچنان وجود دارد. نخست، اغلب مطالعات پیشین بر روی ARGهای بالینی متمرکز بوده‌اند و شناسایی ARGهای محیطی، به‌ویژه در خاک‌های کشاورزی، کمتر مورد توجه قرار گرفته است. دوم، مقایسه جامع و نظام‌مند چهار الگوریتم اصلی شامل جنگل تصادفی، تقویت گرادیانی، ماشین بردار پشتیبان و شبکه عصبی چندلایه بر روی یک مجموعه داده واحد با ویژگی‌های زیستی مشخص (محتوای GC، فراوانی کدون‌ها و ترکیب اسیدهای آمینه) به‌ندرت انجام شده است. سوم، چالش داده‌های نامتوازن که یکی از ویژگی‌های ذاتی مجموعه‌داده‌های متاژنومی واقعی است، در بسیاری از پژوهش‌ها نادیده گرفته شده و عملکرد مدل‌ها در چنین شرایطی کمتر مورد ارزیابی قرار گرفته است. پژوهش حاضر با هدف پر کردن این شکاف‌ها، به مقایسه عملکرد چهار الگوریتم یادگیری ماشین در شناسایی ARGهای موجود در میکروبیوم خاک‌های کشاورزی با استفاده از داده‌های خاک‌های دو کشور شرق آسیا، هند و چین، می‌پردازد.

روش‌شناسی پژوهش

جمع‌آوری و پردازش داده‌های متاژنومیک

در این پژوهش، توالی‌های ژنی مربوط به میکروبیوم خاک‌های کشاورزی از پایگاه داده مرکز ملی اطلاعات بیوتکنولوژی NCBI (National Center for Biotechnology Information) دریافت شد (https://ncbi.nlm.nih.gov/sra/). تمامی توالی‌های مورد استفاده مربوط به مناطق کشاورزی شرق آسیا، دو کشور چین و هند، می‌باشد. اطلاعات تفصیلی پروژه‌ها شامل شماره دسترسی (BioProject Accession) و موقعیت جغرافیایی، در جدول پیوست شماره ۱ ارائه گردیده است داده‌های متاژنومیک مورد استفاده در پژوهش حاضر با استفاده از توالی‌یابی شات‌گان کل ژنوم مستقیماً از DNA کل استخراج‌شده از نمونه‌های خاک کشاورزی، بدون هرگونه مرحله جداسازی، خالص‌سازی یا کشت میکروارگانیسم‌های خاص، حاصل گردیده‌اند. پس از دریافت داده‌ها، کیفیت توالی‌ها با استفاده از نرم‌افزار FastQC ارزیابی گردید (Andrews, 2010). این ابزار، اطلاعاتی درباره توزیع کیفیت در طول توالی و وجود احتمالی آداپتورهای اضافه‌شده را فراهم می‌سازد. توالی‌هایی که کیفیت پایین داشتند یا در آن‌ها آداپتور شناسایی شد، با استفاده از نرم‌افزار Cutadapt پاک‌سازی و حذف گردیدند (Martin, 2011). در مرحله بعد، توالی‌های پردازش‌شده (شامل ۴۰۰٬۰۰۰ توالی) با استفاده از ابزار ARGs-OAP مورد تجزیه و تحلیل قرار گرفتند و با پایگاه داده SARG مقایسه شدند تا ژن‌های مقاومت به آنتی‌بیوتیک شناسایی شوند (Yin et al., 2018). در پایگاه NCBI به برخی مقالات که توالی ژن‌های مذکور مورد استفاده قرار گرفته، اشاره شده است؛ Yang et al. (۲۰۲۱) تاثیر استفاده از کودهای آلی بر ARGهای خاک در سه منطقه باغی در کشور چین مورد مطالعه قرار دادند. Hinsu et al. (۲۰۲۱) با استفاده از هشت محیط کشت مختلف تنوع زیستی باکتریایی را در ریزوسفر گیاه بادام زمینی در هند مورد بررسی قرار دادند. Hinsu et al. (۲۰۲۱) تنوع جامعه میکروبی ریزوسفر گیاه بادام زمینی در دو حالت پیش از کاشت و پس از برداشت در هند مورد شناسایی و بررسی قرار دادند.

استخراج ویژگی‌ها

برای تحلیل داده‌های متاژنومیک و شناسایی ژن‌های مقاوم به آنتی‌بیوتیک، مجموعه‌ای از ویژگی‌های عددی از توالی‌ها استخراج گردید تا به‌عنوان ورودی برای مدل‌های یادگیری ماشین مورد استفاده قرار گیرند. این ویژگی‌ها بر اساس پژوهش‌های پیشین که کارایی آن‌ها را در تفکیک ژن‌های مقاوم نشان داده‌اند، انتخاب شدند (Arango-Argoty et al., 2018; Davis et al., 2016; Yang et al., 2013). ویژگی‌های مورد استفاده شامل میزان GC که نشان‌دهنده درصد بازهای گوانین و سیتوزین در توالی است (Yang et al., 2013)، ترکیب اسیدهای

آمینه که بیانگر فراوانی نسبی ۲۱ نوع آمینواسید در توالی‌های ترجمه‌شده می‌باشد (Arango-Argoty et al., 2018) و الگوی استفاده از کدون‌ها که فراوانی نسبی ۶۴ کدون استاندارد ژنتیکی را در بر می‌گیرد، بودند (Davis et al., 2016).

آماده‌سازی داده‌ها و ساخت زیرمجموعه هدفمند

با توجه به نامتوازن بودن مجموعه داده اولیه که تنها شامل ۵۰ ژن مقاوم و ۳۹۹۹۵۰ هزار ژن غیرمقاوم بود، یک زیرمجموعه هدفمند جهت آموزش و ارزیابی مدل‌ها طراحی شد. در این فرآیند، تمامی ژن‌های مقاوم (Label = 1) حفظ شدند. سپس، از میان ژن‌های غیرمقاوم (Label = 0)تعداد ۱۵۰ ژن انتخاب گردید که در ویژگی‌های انتخاب شده شامل فراوانی ۶۴ کدون، ۲۱ آمینواسیدها و میزان GC، تفاوت معنادار آماری با ژن‌های مقاوم داشتند.

یکی از معیارهای کلیدی در انتخاب این ژن‌ها، محدود کردن میزان GC به بازه‌ای بین ۱۰ تا ۳۰ درصد بود. این کار به منظور حذف توالی‌های غیرطبیعی و کاهش اثر نویز زیستی انجام شد، چرا که پژوهش‌های قبلی نشان داده‌اند انتخاب بازه کنترل‌شده‌ای از GC می‌تواند دقت طبقه‌بندی و مقایسه بین ARGs و سایر ژن‌ها را بهبود دهد (Yang et al., 2013). دلیل انتخاب فراوانی کدون‌ها به عنوان ویژگی، این است که الگوی استفاده از کدون‌ها در ژن‌های مقاوم اغلب با ژن‌های میزبان تفاوت دارد و این تفاوت می‌تواند به عنوان یک امضای تشخیصی برای شناسایی ARGs مورد استفاده قرار گیرد (Davis et al., 2016; Arango-Argoty et al., 2018). همچنین ترکیب آمینواسیدی پروتئین‌های مقاوم (مانند پمپ‌های افلاکس و آنزیم‌های تخریب‌کننده آنتی‌بیوتیک) اغلب حاوی مقادیر بیشتری از آمینواسیدهای آبگریز و باردار مثبت است که می‌تواند به تفکیک ژن‌های مقاوم از غیرمقاوم کمک کند (Poole, 2005; Forsberg et al., 2014). در نهایت، مجموعه داده نهایی با ۲۰۰ نمونه (۵۰ ژن مقاوم و ۱۵۰ ژن غیرمقاوم) تهیه شد.

الگوریتم‌های یادگیری ماشین

در این پژوهش، چهار الگوریتم یادگیری ماشین شامل جنگل تصادفی، ماشین بردار پشتیبان، درخت‌های تقویت‌شده گرادیانی و شبکه عصبی چندلایه برای شناسایی ژن‌های مقاوم به آنتی‌بیوتیک مورد ارزیابی قرار گرفتند. ماشین بردار پشتیبان یکی از الگوریتم‌های قدرتمند در طبقه‌بندی داده‌هاست که به‌ویژه در فضاهایی با ابعاد بالا عملکرد بالایی دارد (Cortes and Vapnik, 1995). جنگل تصادفی یک مدل یادگیری گروهی است که با ترکیب چندین درخت تصمیم، یک طبقه‌بندی قوی و مقاوم نسبت به نویز ایجاد می‌کند. هر درخت بر اساس یک زیرمجموعه تصادفی از داده‌های آموزشی ساخته می‌شود و در هر گره از درخت تنها زیرمجموعه‌ای تصادفی از ویژگی‌ها برای تقسیم در نظر گرفته می‌شود. این رویکرد باعث کاهش همبستگی بین درخت‌ها، کاهش بیش‌برازش و افزایش تعمیم‌پذیری مدل نهایی می‌شود (Breiman, 2001). مدل تقویت گرادیانی یک الگوریتم یادگیری قدرتمند است که با استفاده از درخت‌های تصمیم پیاپی، خطاهای مدل را در هر مرحله کاهش می‌دهد. این مدل به دلیل کارایی بالا، مقیاس‌پذیری و قابلیت تنظیم دقیق شاخص‌ها، در مسائل طبقه‌بندی به‌طور گسترده مورد استفاده قرار می‌گیرد (Chen and Guestrin, 2016). استفاده از ارزیابی متقابل داخلی (cross-validation) و قابلیت اصلاح خطاهای مرحله‌به‌مرحله، از جمله مزایای کلیدی این مدل محسوب می‌شوند (Aydın et al., 2023). شبکه عصبی چندلایه نیز به عنوان یکی از ساختارهای کلاسیک یادگیری عمیق، از چندین لایه متصل به‌هم شامل لایه ورودی، لایه‌های پنهان و لایه خروجی تشکیل شده است. این شبکه‌ها قادر به مدل‌سازی روابط پیچیده و غیرخطی بین متغیرها هستند و در بسیاری از مسائل طبقه‌بندی عملکرد قابل توجهی از خود نشان می‌دهند. برای بهبود دقت و جلوگیری از بیش‌برازش، مراحل پیش‌پردازش شامل نرمال‌سازی ویژگی‌ها و تنظیم دقیق شاخص‌هایی مانند تعداد لایه‌ها، تعداد نورون‌ها، نرخ یادگیری و نوع تابع فعال‌سازی انجام شد (Novielli et al., 2024). برای بررسی عملکرد مدل‌ها در شرایط مختلف، مجموعه داده نهایی در نسبت‌های متفاوتی به داده‌های آموزش و آزمون تقسیم گردید، تقسیمات شامل ۲۰/۸۰، ۳۰/۷۰، ۴۰/۶۰ و ۹۰/۱۰ بودند تا پایداری و تعمیم‌پذیری مدل‌ها در مواجهه با داده‌های نادیده مورد ارزیابی قرار گیرد.

ارزیابی عملکرد مدل‌ها

برای سنجش دقت و کارایی الگوریتم‌های یادگیری ماشین در شناسایی ژن‌های مقاوم به آنتی‌بیوتیک، از سه شاخص کلیدی شامل دقت (Precision)، بازخوانی (Recall) و امتیاز F1 (F1-score) استفاده شد. این معیارها به‌ویژه در شرایطی که داده‌ها نامتوازن هستند (تعداد نمونه‌های مثبت و منفی برابر نیست)، اهمیت دوچندان پیدا می‌کنند؛ زیرا بررسی معیار دقت کلی (Accuracy) می‌تواند تصویری غیرواقعی از عملکرد مدل ارائه دهد.

معیار دقت نشان می‌دهد چند درصد از نمونه‌هایی که مدل به‌عنوان "مثبت" پیش‌بینی کرده، واقعاً مثبت بوده‌اند (رابطه ۱).

$$\text{percision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$ رابطه ۱)

معیار بازخوانی درصدی از نمونه‌های واقعی مثبت را نشان می‌دهد که مدل آن‌ها را به‌درستی شناسایی کرده است. (رابطه ۲)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$ رابطه ۲)

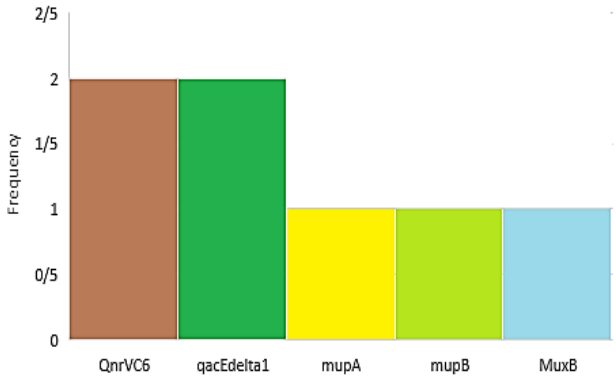
معیار امتیاز F1 میانگین هماهنگ دقت و بازخوانی است که توازنی بین این دو شاخص برقرار می‌کند و در شرایط داده‌های نامتوازن، معیاری مناسب‌تر از دقت کلی محسوب می‌شود (رابطه ۳).

$$\text{F1} - \text{score} = \frac{\text{Percision} \times \text{Recall} \times 2}{\text{Percision} + \text{Recall}}$$ رابطه ۳)

در این روابط TP (True Positive) تعداد نمونه‌های مثبت هستند که به درستی به‌عنوان مثبت پیش‌بینی شده‌اند. FP (False Positive) تعداد نمونه‌های منفی هستند که به اشتباه به‌عنوان مثبت پیش‌بینی شده‌اند و FN (False Negative) تعداد نمونه‌های مثبت هستند که به اشتباه به‌عنوان منفی پیش‌بینی شده‌اند.

نتایج

بر اساس نتایج حاصل از تحلیل داده‌های متاژنومی با استفاده از پایگاه داده SARG و ابزار ARGs-OAP، از بین ۴۰۰ هزار ژن، تنها پنج ژن با بیشترین مقاومت و فراوانی در میکروبیوم خاک‌های کشاورزی شناسایی شدند (شکل ۱). این ژن‌ها شامل QnrVC6، qacEdelta1، mupA، mupB و MuxB بودند. ژن QnrVC6 یکی از اعضای خانواده ژن‌های مقاوم به فلوروکینولون‌ها است که با محافظت از آنزیم‌های توپوایزومراز II و IV، موجب کاهش اثربخشی این دسته از آنتی‌بیوتیک‌ها می‌شود (Strahilevitz et al., 2009). ژن qacEdelta1 با مقاومت به ترکیبات ضدعفونی‌کننده، به‌ویژه آمونیوم‌های چهارظرفیتی، مرتبط بوده و اغلب در پلاسمیدهای کلاس ۱ به‌همراه سایر ژن‌های مقاوم یافت می‌شود (Gillings, 2014). ژن‌های mupA و mupB در مقاومت نسبت به آنتی‌بیوتیک موپیروسین نقش دارند؛ این ژن‌ها با تغییر آنزیم ایزولوسین tRNA سنتتاز موجب کاهش حساسیت باکتری به این آنتی‌بیوتیک موضعی می‌شوند (Seah et al., 2012). همچنین ژن MuxBیکی از اجزای سیستم پمپ افلاکس نوع RND است که در خارج‌سازی آنتی‌بیوتیک‌ها از سلول و کاهش غلظت داخل‌سلولی آن‌ها نقش دارد (Poole, 2005). شناسایی این ژن‌ها در خاک‌های کشاورزی، نشان‌دهنده وجود مخزنی از مقاومت‌های متنوع در محیط است که ممکن است منشأ انسانی یا دامی داشته باشد و در گسترش مقاومت آنتی‌بیوتیکی در اکوسیستم‌های طبیعی نقش ایفا کند (Berendonk et al., 2015). از نظر توزیع میزبانی، ژن‌های QnrVC6 و MuxB عمدتاً در جنس Pseudomonas وPseudomonas aeruginosa به‌ویژه گونه شناسایی شدند. ژن qacEdelta1 بیشتر در خانواده Enterobacteriaceae (از جمله اشریشیا کلی) یافت شدند. همچنین ژن‌های mupAو mupB عمدتاً در گونه استافیلوکوکوس اورئوس و همچنین در برخی اکتینوباکترهای خاکزی مشاهده شدند.



شکل ۱. فراوانی (بر حسب درصد) پنج ژن مقاوم به آنتی‌بیوتیک شناسایی‌شده با بیشترین فراوانی در میکروبیوم خاک‌های کشاورزی.

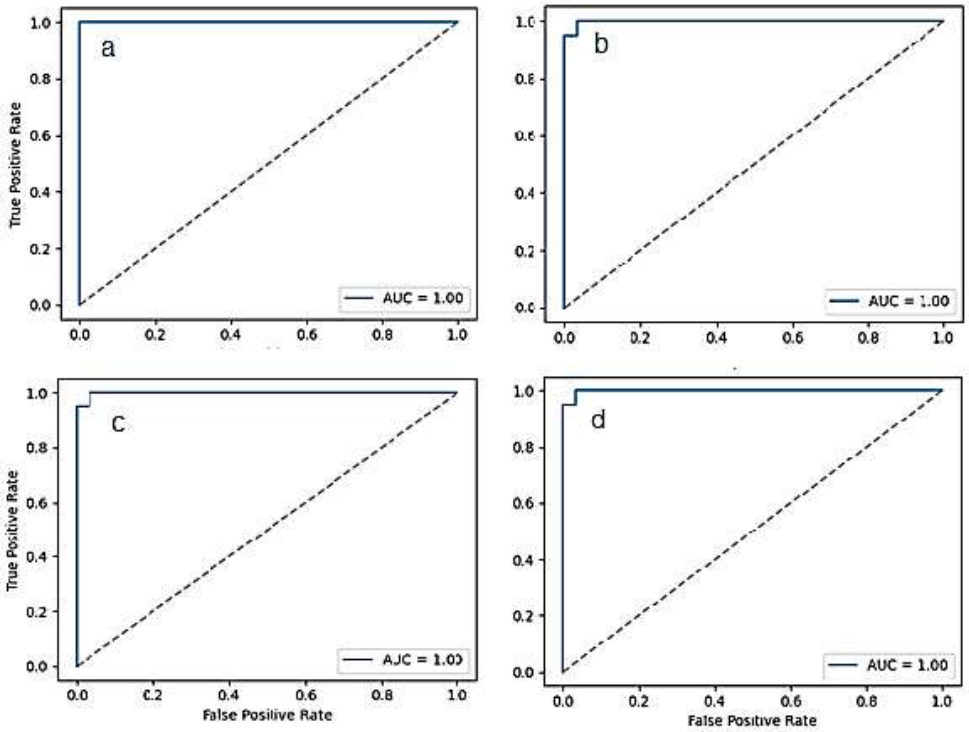
نتایج ارائه‌شده در جدول ۱، عملکرد چهار الگوریتم یادگیری ماشین شامل جنگل تصادفی، ماشین بردار پشتیبان، تقویت گرادیانی و شبکه عصبی چندلایه را در شناسایی ژن‌های مقاوم به آنتی‌بیوتیک نشان می‌دهد. ارزیابی مدل‌ها بر اساس معیارهای دقت، امتیازF1 ، بازخوانی کلاس مقاوم (کلاس ۱) و بازخوانی کلاس غیرمقاوم (کلاس ۲) انجام شده است. بر اساس نتایج، مدل جنگل تصادفی با دقت ۰/۹۸۷۷، بالاترین عملکرد را در میان چهار الگوریتم مورد بررسی داشته است. امتیاز F1 این مدل ۰/۹۷۵۶ محاسبه شد که نشان‌دهنده تعادل مناسب بین دقت و بازخوانی است. بازخوانی این مدل برای کلاس مقاوم ۰/۹۵۲۴ و برای کلاس غیرمقاوم ۱/۰۰۰۰ گزارش شده است که بیانگر توانایی بالای آن در شناسایی صحیح هر دو دسته ژن‌ها می‌باشد. مدل ماشین بردار پشتیبان با دقت ۰/۹۷۵۳، عملکردی نزدیک به جنگل تصادفی از خود نشان داد. امتیاز F1 این مدل ۰/۹۵۰۰ به‌دست آمد. بازخوانی مدل ماشین بردار پشتیبان برای کلاس مقاوم ۰/۹۰۴۸ و برای کلاس غیرمقاوم ۱/۰۰۰۰ بود. اگرچه این مدل در شناسایی ژن‌های غیرمقاوم عملکرد کاملی داشته، اما در شناسایی ژن‌های مقاوم نسبت به جنگل تصادفی ضعیف‌تر عمل کرده است. مدل شبکه عصبی چندلایه با دقت ۰/۹۶۳۱، عملکرد قابل قبولی در طبقه‌بندی ژن‌ها ارائه داد. امتیاز F1 این مدل ۰/۹۳۰۲ محاسبه شد. بازخوانی مدل شبکه عصبی برای کلاس مقاوم ۰/۹۵۲۴ و برای کلاس غیرمقاوم ۰/۹۶۶۷ بود که نشان می‌دهد این مدل در شناسایی ژن‌های مقاوم عملکردی هم‌تراز با جنگل تصادفی داشته است. در میان مدل‌های مورد بررسی، تقویت گرادیانی با دقت ۰/۹۶۳۰ و امتیاز F1 برابر با ۰/۹۲۳۱، پایین‌ترین عملکرد را در شناسایی ژن‌های مقاوم به خود اختصاص داد. بازخوانی این مدل برای کلاس مقاوم ۰/۸۵۷۱ و برای کلاس غیرمقاوم ۱/۰۰۰۰ بود. این نتایج نشان می‌دهد که اگرچه مدل تقویت گرادیانی در شناسایی ژن‌های غیرمقاوم عملکرد کاملی دارد، اما در تشخیص ژن‌های مقاوم نسبت به سایر مدل‌ها ضعیف‌تر عمل می‌کند و این ضعف می‌تواند منجر به نادیده‌گرفتن تعداد قابل توجهی از ژن‌های مقاوم در کاربردهای عملی شود.

جدول ۱. میانگین عملکرد مدل‌های یادگیری ماشین به روش اعتبارسنجی متقاطع ۱۰ مرحله‌ای با ۲۰ تکرار

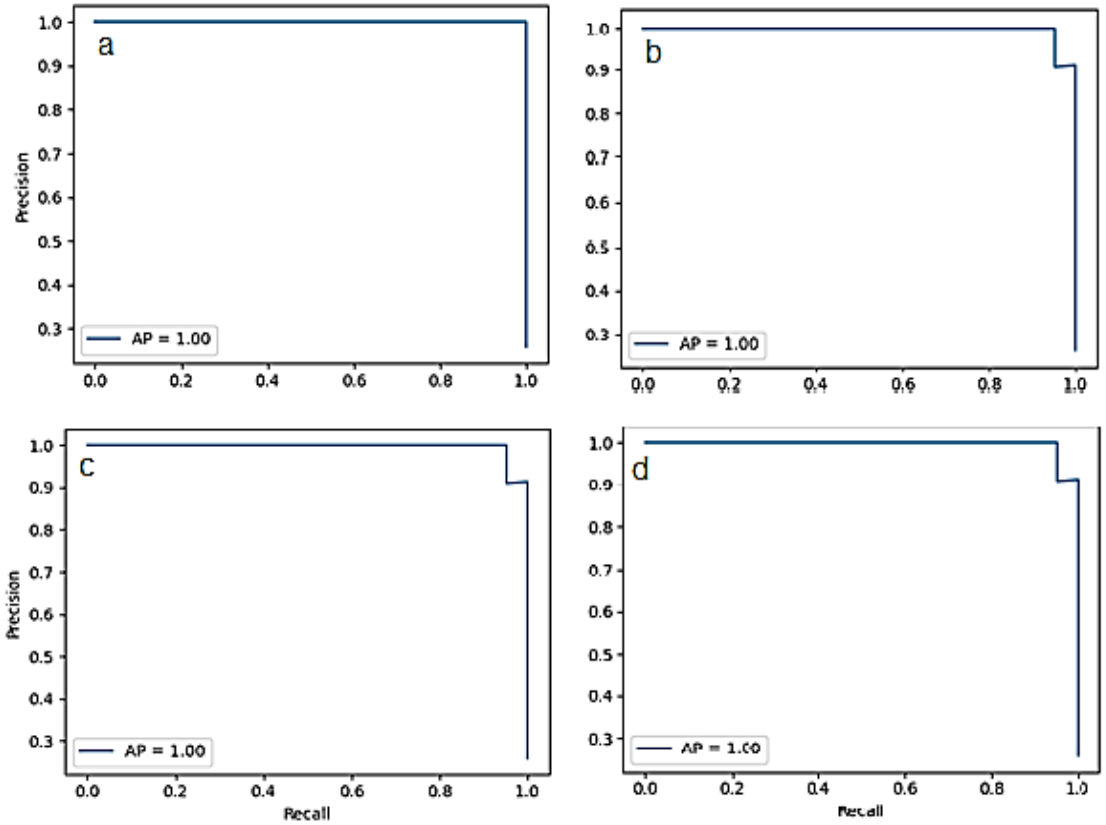
مدل	دقت	F1 امتیاز	بازخوانی کلاس ۱	بازخوانی کلاس ۲
ماشین بردار پشتیبان	۰/۹۷۵۳	۰/۹۵۰۰	۰/۹۰۴۸	۱/۰۰۰۰
جنگل تصادفی	۰/۹۸۷۷	۰/۹۷۵۶	۰/۹۵۲۴	۱/۰۰۰۰
تقویت گرادیانی	۰/۹۶۳۰	۰/۹۲۳۱	۰/۸۵۷۱	۱/۰۰۰۰
شبکه عصبی چندلایه	۰/۹۶۳۱	۰/۹۳۰۲	۰/۹۵۲۴	۰/۹۶۶۷

بر اساس استانداردهای رایج در ارزیابی مدل‌های طبقه‌بندی زیستی، معیارهای اعتبارسنجی به سه دسته عملکرد ضعیف (F1-score < 0.60)، عملکرد متوسط (F1-score 0.90-0.60) و عملکرد قوی (F1-score ≥ 0.90) تقسیم می‌شوند. نتایج نشان می‌دهد که مدل‌های جنگل تصادفی و ماشین بردار پشتیبان با بازخوانی کامل برای کلاس غیرمقاوم، توانایی بالایی در شناسایی صحیح ژن‌های غیرمقاوم دارند. از سوی دیگر، مدل‌های جنگل تصادفی و شبکه عصبی چندلایه با بازخوانی ۰/۹۵۲۴ برای کلاس مقاوم، بهترین عملکرد را در شناسایی ژن‌های مقاوم ارائه داده‌اند. این یافته‌ها حاکی از آن است که انتخاب مدل مناسب برای شناسایی ژن‌های مقاوم به آنتی‌بیوتیک باید با توجه به اهمیت نسبی شناسایی صحیح ژن‌های مقاوم در مقایسه با ژن‌های غیرمقاوم صورت گیرد.

نتایج تحلیل منحنی مشخصه عملیاتی گیرنده (ROC) (شکل ۲) و منحنی دقت–یادآوری (PRC) (شکل ۳) نشان داد که مدل جنگل تصادفی بالاترین کارایی را در تشخیص ژن‌های مقاوم به آنتی‌بیوتیک داشته است؛ به‌طوری‌که میانگین مساحت زیر منحنی (AUC) برای منحنی ROC معادل ۰/۹۷۶ با انحراف معیار ۰/۰۰۵ و برای منحنی PRC معادل ۰/۹۸۴ با انحراف معیار ۰/۰۰۶ به‌دست آمد. این مقادیر نشان‌دهنده دقت و قابلیت تفکیک بالای این مدل در طبقه‌بندی صحیح ژن‌های مقاوم و غیرمقاوم می‌باشد. در رتبه دوم، مدل شبکه عصبی چندلایه عملکرد نزدیکی به جنگل تصادفی داشت و توانست میانگین AUC برابر با ۰/۹۸۱ ± ۰/۰۰۶ در منحنی ROC و ۰/۹۷۶ ± ۰/۰۰۷در منحنی PRC را کسب کند. این نتایج بیانگر تعادل مناسب بین دقت و یادآوری در عملکرد این مدل است. مدل ماشین بردار پشتیبان نیز با ثبت مقادیر AUC برابر با ۰/۹۶۸ ± ۰/۰۰۸ در منحنی ROC و ۰/۹۶۲ ± ۰/۰۱۰ در منحنی PRC، عملکرد قابل قبولی از خود نشان داد و در جایگاه سوم قرار گرفت.



شکل ۲. منحنی مشخصه عملیاتی گیرنده ROC (a) جنگل تصادفی، (b) ماشین بردار پشتیبان (c) تقویت گرادیانی (d) شبکه عصبی چندلایه

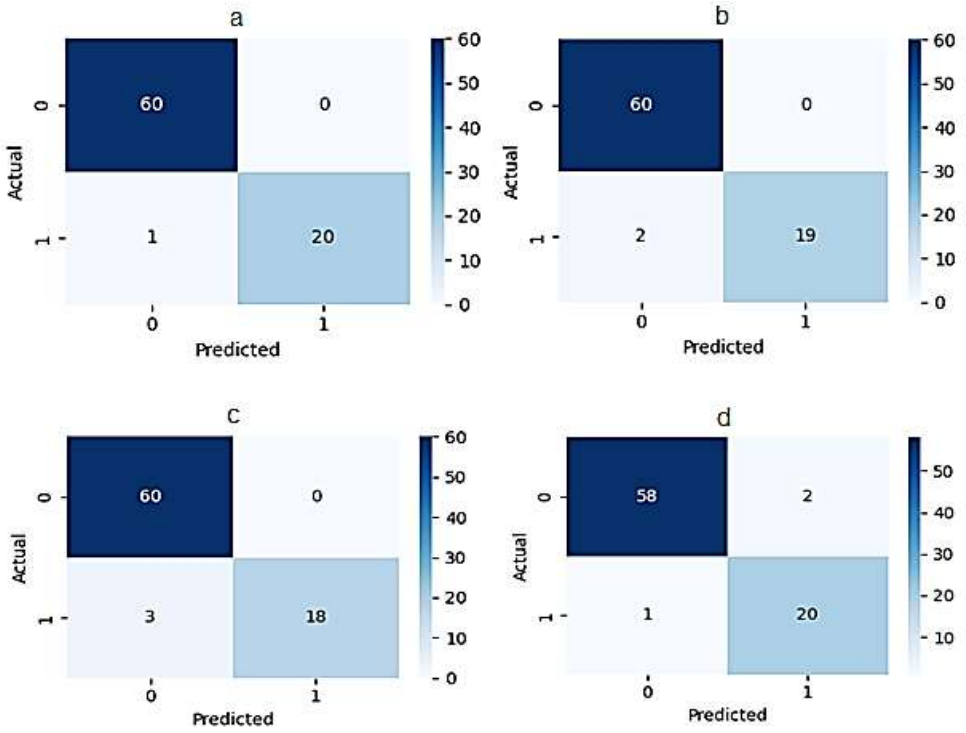


شکل ۳. منحنی دقت-یادآوری PRC (a) جنگل تصادفی، (b) ماشین بردار پشتیبان (c) تقویت گرادیانی (d) شبکه عصبی چندلایه

در مقابل، مدل تقویت گرادیانی نسبت به سایر الگوریتم‌ها عملکرد ضعیف‌تری داشت و کمترین مقادیر AUC را در هر دو منحنی ارائه داد؛ به‌گونه‌ای که میانگین AUC در منحنی ROC برابر با ۰/۹۵۱ ±۰/۰۰۹ و در منحنی PRC معادل ۰/۹۴۳ ±۰/۰۱۱ بود. این امر نشان‌دهنده توانایی پایین‌تر این مدل در تشخیص دقیق ژن‌های مقاوم، در مقایسه با مدل‌های دیگر است.

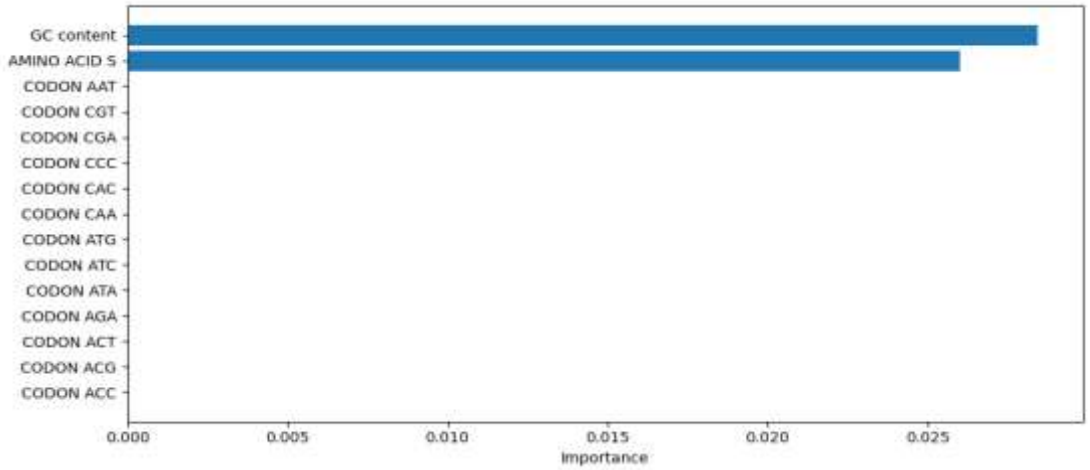
نتایج تحلیل ماتریس درهم‌آمیختگی (شکل ۴) نشان داد که مدل جنگل تصادفی با شناسایی صحیح تمام ۶۰ نمونه متعلق به کلاس

غیرمقاوم (True Negative) و ۲۰ نمونه از ۲۱ نمونه مقاوم (True Positive)، بالاترین دقت را در هر دو کلاس داشته است. این مدل تنها یک خطای نوع دوم (False Negative) ثبت کرده و فاقد هرگونه خطای نوع اول (False Positive) بوده است. مدل شبکه عصبی چندلایه نیز عملکرد بسیار مطلوبی ارائه داد. این مدل با شناسایی درست ۵۹ نمونه غیرمقاوم (TN) و تمام ۲۱ نمونه مقاوم (TP)، بدون بروز خطای نوع دوم (FN) و تنها با یک مورد خطای نوع اول (FP) عمل کرد که نشان‌دهنده دقت بالا و تعادل مناسب در تفکیک دو کلاس می‌باشد. مدل ماشین بردار پشتیبان تمام نمونه‌های غیرمقاوم را به‌درستی طبقه‌بندی کرد (TN)، اما در شناسایی نمونه‌های مقاوم، دو مورد خطای نوع دوم (TP, FN) ثبت شد این موضوع موجب کاهش حساسیت مدل در شناسایی کلاس مثبت شده است. در مقابل، مدل تقویت گرادیانی ضعیف‌ترین عملکرد را در طبقه‌بندی ژن‌های مقاوم نشان داد. اگرچه تمام ۶۰ نمونه غیرمقاوم را به‌درستی شناسایی کرد (TN)، اما تنها ۱۸ مورد از ۲۱ نمونه مقاوم را به‌درستی پیش‌بینی نمود (TP) که با سه خطای نوع دوم (FN) همراه بود. این خطاها منجر به کاهش قابلیت اطمینان این مدل در تشخیص ژن‌های مقاوم شده است.

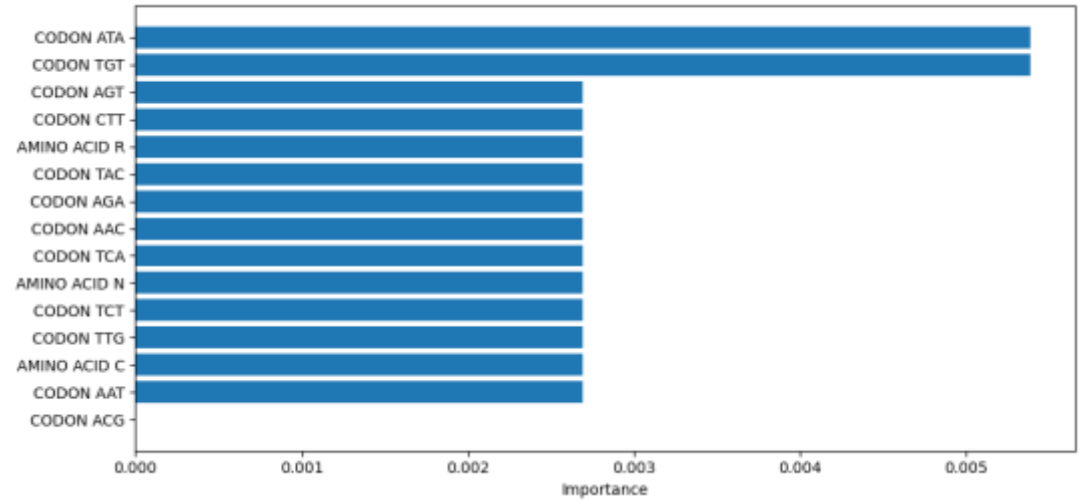


شکل ۴. نمودار ماتریس درهم آمیختگی (a) جنگل تصادفی، (b) ماشین بردار پشتیبان (c) تقویت گرادیانی (d) شبکه عصبی چندلایه

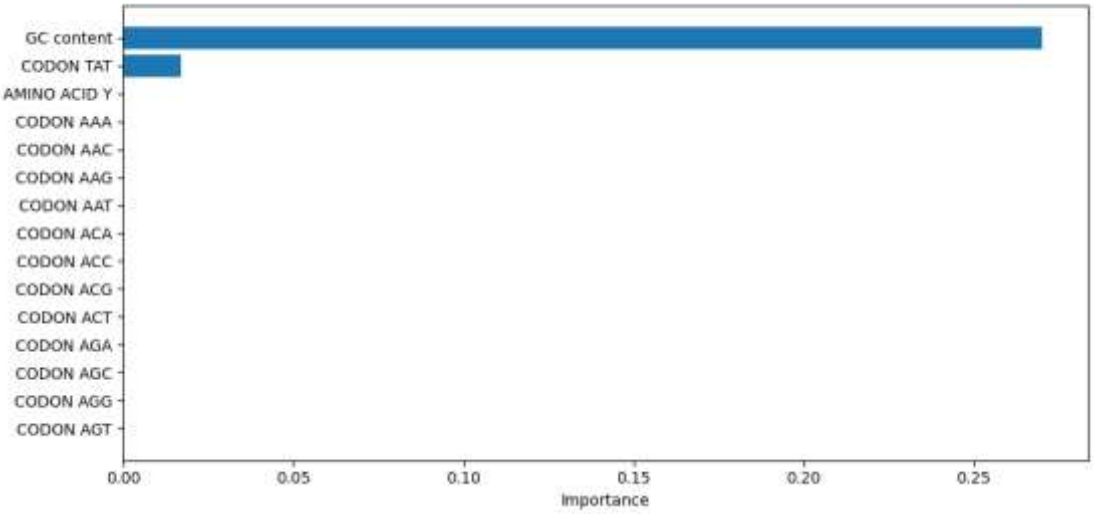
نتایج تحلیل نمودار اهمیت ویژگی‌ها (شکل‌های ۵، ۶، ۷ و ۸) نشان داد که میزان GC در تفکیک ژن‌های مقاوم و غیرمقاوم در مدل‌های جنگل تصادفی و تقویت گرادیانی مهم‌ترین نقش را داشته است. در واقع این دو مدل تنها با در نظر گرفتن این ویژگی ژن‌های مقاوم را از غیر مقاوم تفکیک کردند و سایر ویژگی‌ها کمترین سهم را در عملکرد مدل جنگل تصادفی و تقویت گرادیانی داشتند. در مقابل، برای مدل‌های ماشین بردار پشتیبان و شبکه عصبی چند لایه توزیع اهمیت ویژگی‌ها به‌صورت یکنواخت‌تری میان متغیرها پخش شده است. در واقع این دو مدل تنها به یک ویژگی مهم برای شناسایی ژن‌ها اکتفا نکردند و با در نظر گرفتن ترکیبی از کدون‌های مربوط به آمینواسید ژن‌های مقاوم را از غیرمقاوم تفکیک کردند. در ماشین بردار پشتیبان کدون ATA (مرتبط با ایزولوسین)، کدون TGT (مرتبط با سیستئین)، کدون AGT (مرتبط با سرین)، کدون CTT و TTG (مرتبط با لوسین)، کدون TAC (مرتبط با متیونین)، کدون AGA (مرتبط با آرژنین)، کدون AAC (مرتبط با آسپاراژین)، کدون TCA و TCT (مرتبط با سرین) درمیان ۱۰ ویژگی مهم قرار داشتند. با مقایسه نمودار اهمیت ویژگی‌های مدل بردار و شبکه عصبی چند لایه، کدون‌های مربوط به اسیدهای آمینه لوسین و آرژنین نقش برجسته‌ای در تفکیک ژن‌های مقاوم از غیرمقاوم داشته‌اند.



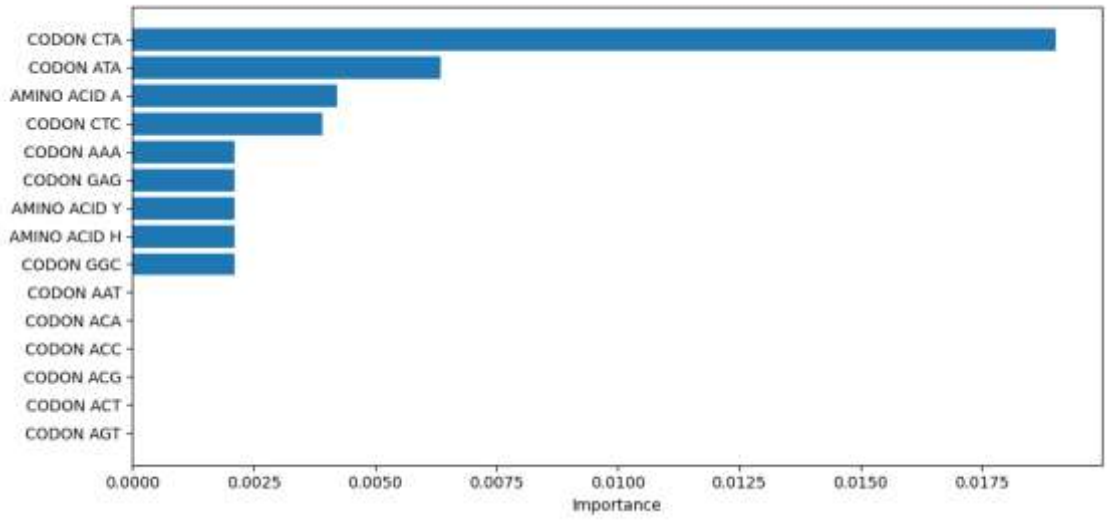
شکل ۵. نمودار اهمیت ویژگی‌ها در جنگل تصادفی،



شکل ۶. نمودار اهمیت ویژگی‌ها در ماشین بردار پشتیبان



شکل ۷. نمودار اهمیت ویژگی‌ها در تقویت گرادیانی



شکل ۸. نمودار اهمیت ویژگی‌ها در شبکه عصبی چندلایه

در مجموع، همگرایی قابل توجهی میان ویژگی‌های منتخب در مدل‌های مختلف وجود دارد. این همگرایی نشان می‌دهد که برخی ویژگی‌های زیستی مشترک نظیر کدون‌های پرتکرار و الگوهای خاص ترکیب اسیدهای آمینه، به‌طور مؤثری اطلاعات تشخیصی مهمی را برای مدل‌های یادگیری ماشین فراهم کرده‌اند و می‌توانند به‌عنوان نشانگرهای زیستی برای شناسایی ژن‌های مقاوم به آنتی‌بیوتیک مورد استفاده قرار گیرند.

بحث

نتایج این پژوهش نشان داد که الگوریتم‌های یادگیری ماشین به‌ویژه جنگل تصادفی و شبکه عصبی چندلایه، عملکرد چشم گیری در شناسایی ژن‌های مقاوم به آنتی‌بیوتیک در میکروبیوم خاک‌های کشاورزی شرق آسیا دارند. مدل جنگل تصادفی با دقت کلی ۰/۹۸۷۷ و میانگین مساحت زیر منحنی ROC معادل ۰/۹۸۷، بالاترین توان تفکیک کلاس‌های مقاوم و غیرمقاوم را نشان داد. همچنین پس از جنگل تصادفی، ماشین بردار پشتیبان با مقدار بالا میانگین F1 در اعتبارسنجی متقاطع (۰/۹۵۸۱±۰/۶۵۶) توانست تعادل بسیار مطلوبی میان دقت و بازخوانی فراهم کند و از این منظر قابل توجه بود. عملکرد مطلوب مدل جنگل تصادفی را می‌توان به ساختار تجمیعی آن نسبت داد؛ زیرا با ترکیب چندین درخت تصمیم‌گیری و انتخاب تصادفی از ویژگی‌ها، از بیش‌برازش جلوگیری کرده و پایداری مدل را در شرایط داده‌های نامتوازن افزایش می‌دهد (Breiman, 2001). این موضوع با یافته‌های پژوهش‌های مشابه در حوزه شناسایی ژن‌های مقاوم همخوانی دارد..Pataki et al (۲۰۲۴) نیز از جنگل تصادفی برای پیش‌بینی حداقل غلظت مهارکنندگی با موفقیت استفاده کردند و قدرت این الگوریتم را در انتخاب ویژگی‌های کلیدی نشان دادند. همچنین، یافته‌های پژوهشی در محیط ریشه گیاه نشان داد که جنگل تصادفی در میان پنج الگوریتم مختلف برتر بوده است که با نتایج پژوهش حاضر همسو است (Ma et al., 2025).

مدل ماشین بردار پشتیبان اگرچه عملکرد نسبتاً پایداری داشت، ولی بازخوانی پایین‌تر آن در طبقه‌بندی ژن‌های مقاوم (۰/۹۰۴۸) نسبت به جنگل تصادفی و شبکه عصبی چندلایه، حاکی از کاهش حساسیت در شناسایی کلاس مثبت بود. علت این مسئله را می‌توان در وابستگی بالای ماشین بردار پشتیبان به موقعیت دقیق مرز تصمیم‌گیری و حساسیت آن به توزیع داده‌ها، به‌ویژه در شرایط کلاس نامتوازن، جست‌وجو کرد (Cortes and Vapnik, 1995). در تحقیقی با وجود دستیابی به دقت‌های بالا با استفاده از ماشین بردار پشتیبان در یک چارچوب پان‌ژنوم، به حساسیت این الگوریتم به پارامترهای تنظیمی و توزیع داده‌ها اشاره کرده‌اند که با یافته‌های این پژوهش قابل تطبیق است (Hyun et al., 2024). در مقابل، الگوریتم تقویت گرادیانی با وجود شهرت بالا در مسائل طبقه‌بندی، ضعیف‌ترین عملکرد را در بین مدل‌ها داشت. بازخوانی پایین آن برای کلاس مقاوم (۰/۸۵۷۱) و مقدار مساحت زیر منحنی کمتر در هر دو منحنی ROC و PRC می‌تواند ناشی از حساسیت این الگوریتم به حجم داده، تنظیمات پیچیده شاخص‌ها و احتمال بیش‌برازش باشد (Chen and Guestrin, 2016; Aydın et al., 2023). این یافته نشان می‌دهد که عملکرد برتر تقویت گرادیانی در برخی حوزه‌ها لزوماً به مسائل زیستی با داده‌های محدود و نامتوازن قابل تعمیم نیست.

یکی از نقاط قوت این پژوهش، شناسایی و تحلیل دقیق ویژگی‌های زیستی مؤثر در عملکرد مدل‌هاست. بررسی نمودار اهمیت

ویژگی‌ها نشان داد که کدون‌هایی نظیر CTG (کدگذار لوسین)، GCG (کدگذار آلانین) و CGC (کدگذار آرژنین) در میان مهم‌ترین فاکتورها در هر چهار مدل قرار گرفتند. این کدون‌ها که به اسیدهای آمینه آب‌گریز یا باردار مثبت مربوط هستند، در ساختار پروتئین‌های مقاومتی مانند پمپ‌های افلاکس یا آنزیم‌های تخریب‌کننده آنتی‌بیوتیک، فراوانی بیشتری دارند (Poole, 2005; Forsberg et al., 2014). همچنین (Her and Wu, 2024) در پژوهش خود به اهمیت ژن‌های غیرهسته‌ای و الگوهای کدونی در پیش‌بینی مقاومت اشاره کرده‌اند که با یافته‌های پژوهش حاضر همخوانی دارد. ویژگی میزان GC نیز در تمامی مدل‌ها از اهمیت بالایی برخوردار بود و در بین ۱۰ ویژگی برتر قرار گرفت. انتخاب محدوده کنترل‌شده‌ای از میزان GC (۱۰ تا ۳۰ درصد) در مرحله آماده‌سازی داده‌ها، منجر به کاهش نویز و افزایش دقت مدل‌ها شد. این انتخاب با پژوهش Yang et al., (2013) که بیان کرده‌اند توالی‌های ARG معمولاً دارای میزان GC متفاوت از ژنوم‌های میزبان هستند، هم‌راستا است، به ویژه زمانی که ARGها از طریق انتقال افقی کسب شده باشند. در نقشه جهانی ARGها نیز به نقش عوامل محیطی و ویژگی‌های ذاتی توالی‌ها در توزیع ژن‌های مقاوم اشاره شده است (Delgado-Baquerizo et al., 2022).

مقایسه نتایج این پژوهش با پژوهش‌های مشابه نشان‌دهنده اعتبار بالای رویکرد اتخاذشده است. به‌طور خاص، در پژوهش Arango-Argoty et al. (۲۰۱۸) که از یادگیری عمیق برای توسعه مدل DeepARG استفاده شد، مشخص گردید که استفاده از ویژگی‌های استخراج‌شده از توالی و نه صرفاً هم‌ردیفی، می‌تواند موجب بهبود دقت شناسایی ARGها گردد. هرچند آن مدل بر پایه یادگیری عمیق بنا شده بود، اما در این پژوهش نشان داده شد که حتی مدل‌هایی سبک‌تر مانند جنگل تصادفی و شبکه عصبی چندلایه نیز با انتخاب هدفمند ویژگی‌ها می‌توانند دقتی هم‌تراز ارائه دهند. همچنین، با ترکیب متاژنومیک شات‌گان و یادگیری ماشین، ارتباط بین میکروبیوم انسان، دام و خاک بررسی شده که بر پیچیدگی و اهمیت رویکردهای یکپارچه در پژوهش مقاومت تأکید دارد (Maciel-Guerra et al., 2022).

در مجموع، این پژوهش نشان می‌دهد که ترکیب روش‌های آماری برای انتخاب ویژگی (مانند آزمون Mann–Whitney) با الگوریتم‌های یادگیری ماشین، می‌تواند چارچوب مؤثری برای شناسایی دقیق ARGها در خاک‌های کشاورزی ارائه دهد. الگوریتم‌هایی مانند جنگل تصادفی و شبکه عصبی چندلایه، به‌ویژه در شرایط نامتوازن و داده‌های کوچک، قابلیت پیاده‌سازی عملی در سامانه‌های پایش محیطی و ابزارهای بیوانفورماتیکی را دارند. با این حال، محدود بودن حجم داده‌ها و تمرکز بر منطقه جغرافیایی خاص، از جمله عوامل محدودکننده در تعمیم‌پذیری نتایج هستند. پیشنهاد می‌شود در پژوهش‌های آتی، مجموعه داده‌های بزرگ‌تر و متنوع‌تر از مناطق مختلف جغرافیایی مورد استفاده قرار گیرد و امکان توسعه سامانه‌های پیش‌بینی ARG با کاربرد عملی در پایش زیست‌محیطی فراهم شود.

نتیجه‌گیری و پیشنهاد

پژوهش حاضر با هدف مقایسه عملکرد چهار الگوریتم یادگیری ماشین شامل جنگل تصادفی، ماشین بردار پشتیبان، تقویت گرادیانی و شبکه عصبی چندلایه در شناسایی ژن‌های مقاوم به آنتی‌بیوتیک در میکروبیوم خاک‌های کشاورزی در شرق آسیا انجام شد. یافته‌ها نشان داد که الگوریتم‌های جنگل تصادفی و ماشین بردار پشتیبان با دستیابی به دقت بیشتر از ۹۷ درصد و امتیاز F1 بزرگتر یا مساوی ۹۵، عملکرد برتری در تفکیک ژن‌های مقاوم از غیرمقاوم داشتند. این دو مدل حتی در شرایط داده‌های نامتوازن و حجم نمونه محدود، توانایی بالایی در شناسایی صحیح ژن‌های مقاوم از خود نشان دادند. اگرچه تمامی مدل‌ها عملکرد قوی نشان دادند، اما جنگل تصادفی با کمترین خطا در شناسایی ژن‌های مقاوم، ریسک اشتباه کمتری داشته و برای کاربردهای پایش محیطی ارجح است. تحلیل اهمیت ویژگی‌ها، نقش کلیدی کدون‌های لوسین، آلانین، آرژنین و همچنین میزان GC را در تمایز ژن‌های مقاوم آشکار ساخت. این یافته‌ها می‌تواند مبنایی برای توسعه نشانگرهای زیستی مبتنی بر ویژگی‌های توالی در پایش مقاومت آنتی‌بیوتیکی در محیط‌های کشاورزی فراهم آورد.

با توجه به محدودیت‌های پژوهش حاضر به عنوان یک مطالعه موردی و تمرکز بر منطقه جغرافیایی خاص، شرق آسیا، پیشنهاد می‌شود در پژوهش‌های آتی از مجموعه داده‌های متنوع‌تر از مناطق مختلف دنیا استفاده شود تا تعمیم‌پذیری مدل‌ها افزایش یابد. همچنین اعتبارسنجی مدل‌های توسعه‌یافته بر روی داده‌های مستقل و واقعی میدانی می‌تواند گامی مهم در جهت کاربرد عملی آن‌ها باشد. تلفیق ویژگی‌های توالی با داده‌های محیطی نظیر pH، دما و نوع کاربری خاک می‌تواند به بهبود دقت پیش‌بینی کمک کند. علاوه بر این، توسعه ابزارهای کاربرپسند مبتنی بر مدل‌های بهینه برای پایش زیست‌محیطی ژن‌های مقاومت و بررسی امکان استفاده از معماری‌های پیشرفته‌تر یادگیری عمیق با حفظ قابلیت تفسیرپذیری، از دیگر زمینه‌های قابل پیگیری در تحقیقات آینده است.

پیوست شمارهٔ ۱

جدول پیوست ۱. اطلاعات تکمیلی پروژه‌های مورد استفاده در پژوهش حاضر

شماره دسترسی	کشور	منطقه جغرافیایی	نام دانشگاه/موسسه تحقیق مربوطه	موضوع پروژه	عنوان مقالات مرتبط	doiمقاله
PRJNA744181	چین	گوانشی	Chinese Academy of Sciences	The effect of application of biofertilizer on farmland. ITS	Effects of biofertilizer on soil microbial diversity and antibiotic resistance genes (Yang et al., 2021)	10.1016/j.scitotenv.2022.153170
PRJNA738647	چین	لیشو	China Agricultural University	Metagenomic analyses of soil microbial communities and functions across nitrogen gradients	مقاله مرتبطی یافت نشد	
PRJNA665712	هند	گجرات	Saurashtra University	This project contains sequencing data from rhizosphere and bulk soil samples. It also contains data from cultured organisms on various media.	Characterizing rhizosphere microbiota of peanut (Arachis hypogaea L.) from pre-sowing to post-harvest of crop under field conditions (Hinsu et al., 2021)	10.1038/s41598-021-97071-3
					To culture or not to culture: a snapshot of culture-dependent and culture-independent bacterial diversity from peanut rhizosphere (Hinsu et al., 2021)	10.7717/peerj.12035
PRJNA411903	هند	گوراکپور	VIT University	Paddy cultivated Tarai Soil rhizosphere Metagenomics	مقاله مرتبطی یافت نشد	

ملاحظات اخلاقی

حامی مالی

حمایت مالی از طرف سازمان یا دانشگاهی برای این مقاله صورت نگرفته است.

مشارکت نویسندگان

نویسنده اول: مشارکت در مفهوم سازی، روش شناسی، توسعه نرم افزاری، تحقیق و تحلیل داده ها، نوشتن پیش نویس مقاله، ویرایش مقاله

نویسنده دوم: مشارکت در اعتبار سنجی، تهیه پیشنویس مقاله، ویرایش و بازنویسی مقالهٍ، راهنمایی در پروسه پژوهشی مقاله

نویسنده سوم: مشارکت در اعتبار سنجی، آنالیز داده ها، تحلیل و تفسیر اطلاعات و نتایج، تهیه پیشنویس مقاله

نویسنده چهارم : مشارکت در بخش نرم افزاری و روش شناسی، نوشتن پیش نویش مقاله

نویسنده پنجم : مشارکت در بخش نرم افزاری و روش شناسی، اعتبارسنجی داده ها، آنالیز داده، نوشتن پیش نویش مقاله، ویرایش مقاله

اعلامیه هوش مصنوعی مولد و فناوری‌های مبتنی بر هوش مصنوعی در فرایند نگارش

از هوش مصنوعی در این مقاله استفاده نشده است.

بیانیه دسترسی به داده‌ها

داده‌هایی پژوهش حاضر از طریق درخواست از نویسندگان قابل دسترسی است.

سپاسگزاری

نویسندگان بدین وسیله مراتب تقدیر و تشکر خود را از پایگاه داده NCBI و توسعه دهندگان ARGs-OAP برای امکان استفاده رایگان از منابع و داده ها ابراز میدارند.

پیروی از اصول اخلاق پژوهش

نویسندگان اصول اخلاقی را در انجام و انتشار این پژوهش علمی رعایت نموده‌اند و این موضوع مورد تأیید همه آنهاست.

تعارض منافع

بنا بر اظهار نویسندگان این مقاله تعارض منافع ندارد.

REFERENCES

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Arango-Argoty, G., Garner, E., Pruden, A., et al. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1), 23.

Aydın, Y., Işıkdağ, U., Bekdaş, G., Nigdeli, S. M., & Geem, Z. W. (2023). Use of machine learning techniques in soil classification. *Sustainability*, 15(3), 2374.

Bai, Y., Ruan, X., Li, R., et al. (2024). Distribution and diversity of antibiotic resistance genes across human, poultry, pig, and soil microbiomes. *Science of the Total Environment*, 912, 168901.

Berendonk, T. U., Manaia, C. M., Merlin, C., et al. (2015). Tackling antibiotic resistance: The environmental framework. *Nature Reviews Microbiology*, 13(5), 310-317.

Bradley, P., Gordon, N. C., Walker, T. M., et al. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nature Communications*, 6, 10063.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59-60.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Davis, J. J., Boisvert, S., Brettin, T., et al. (2016). Antimicrobial resistance prediction in PATRIC and RAST. *Scientific Reports*, 6, 27930.

Delgado-Baquerizo, M., Hu, H. W., Maestre, F. T., et al. (2022). The global distribution and environmental drivers of the soil antibiotic resistome. *Microbiome*, 10(1), 219.

Deng, Z., et al. (2025). Ecological distribution, dissemination potential, and health risks of antibiotic resistance genes and mobile genetic elements in soils across diverse land-use types in China. *Environmental Research*, 285(Pt 2), 122459.

Forsberg, K. J., Patel, S., Gibson, M. K., et al. (2014). Bacterial phylogeny structures soil resistomes across habitats. *Nature*, 509, 612-616.

Gandhi, N. R., Nunn, P., Dheda, K., et al. (2010). Multidrug-resistant and extensively drug-resistant tuberculosis: A threat to global control of tuberculosis. *Lancet*, 375, 1830-1843.

Gillings, M. R. (2014). Integrons: past, present, and future. *Microbiology and Molecular Biology Reviews*, 78(2), 257-277.

Her, H. L., & Wu, Y. W. (2024). Pan-genome approach with genetic algorithm identifies non-core gene clusters for antibiotic resistance prediction in Escherichia coli. *Briefings in Bioinformatics*, 25(2), bbae078.

Hinsu, A. T., Panchal, K. J., Pandit, R. J., Koringa, P. G., & Kothari, R. K. (2021). Characterizing rhizosphere microbiota of peanut (Arachis hypogaea L.) from pre-sowing to post-harvest of crop under field

conditions. *Scientific reports*, 11(1), 17457.
Hinsu, A., Dumadiya, A., Joshi, A., Kotadiya, R., Andharia, K., Koringa, P., & Kothari, R. (2021). To culture or not to culture: a snapshot of culture-dependent and culture-independent bacterial diversity from peanut rhizosphere. *PeerJ*, 9, e12035.
Hu, Y., Liu, F., Lin, I. Y., et al. (2016). Dissemination of the mcr-1 colistin resistance gene. *Lancet Infectious Diseases*, 16, 146-147.
Hyun, J. C., et al. (2024). A pan-genome framework for antibiotic resistance prediction in Staphylococcus aureus, Pseudomonas aeruginosa, and Escherichia coli. *mSystems*, 9(1), e00982-23.
Jia, B., Raphenya, A. R., Alcock, B., et al. (2017). CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45, D566-D573.
Kleinheinz, K. A., Joensen, K. G., & Larsen, M. V. (2014). Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage*, 4(2), e27943.
Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10, R25.
Ma, R. A., et al. (2025). A machine learning approach to predict phyllosphere resistome abundance across urbanization gradients. *Environment International*, 202, 109655.
Maciel-Guerra, A., et al. (2022). Dissecting microbial communities and resistomes for interconnected humans, soil, and livestock. *The ISME Journal*, 16, 21-32.
Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 10–12.
Martínez, J. L., Coque, T. M., & Baquero, F. (2015). What is a resistance gene? Ranking risk in resistomes. *Nature Reviews Microbiology*, 13(2), 116-123.
McArthur, A. G., & Tsang, K. K. (2017). Antimicrobial resistance surveillance in the genomic age. *Annals of the New York Academy of Sciences*, 1388, 78-91.
Mediavilla, J. R., Patrawalla, A., Chen, L., et al. (2016). Colistin- and carbapenem-resistant Escherichia coli harboring mcr-1 and blaNDM-5. *mBio*, 7, e01191-16.
Novielli, P., Romano, D., Magarelli, M., Bitonto, P. D., Diacono, D., Chiatante, A., Lopalco, G., Sabella, D., Venerito, V., Filannino, P., et al. (2024). Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification. *Frontiers in Microbiology*, 15, 1348974.
O'Neill, J. (2016). *Tackling drug-resistant infections globally: Final report and recommendations*. Review on Antimicrobial Resistance.
Pal, C., Bengtsson-Palme, J., Kristiansson, E., & Larsson, D. G. (2016). The structure and diversity of human, animal and environmental resistomes. *Microbiome*, 4, 54.
Pataki, B. Á., et al. (2024). Random Forest-based prediction of ciprofloxacin minimum inhibitory concentration in Escherichia coli using whole-genome sequencing data. *Journal of Antimicrobial Chemotherapy*, 79(3), 512-521.
Pehrsson, E. C., Tsukayama, P., Patel, S., et al. (2016). Interconnected microbiomes and resistomes in low-income human habitats. *Nature*, 533, 212-216.
Poole, K. (2005). Efflux-mediated antimicrobial resistance. *Journal of Antimicrobial Chemotherapy*, 56(1), 20-51.
Pruden, A., Larsson, D. J., Amézquita, A., et al. (2013). Management options for reducing the release of antibiotics and antibiotic resistance genes to the environment. *Environmental Health Perspectives*, 121, 878.
Rowe, W., Baker, K. S., Verner-Jeffreys, D., et al. (2015). SEAR: A cloud-compatible web pipeline for detecting antimicrobial resistance genes. *PLOS One*, 10, e0133492.
Scaglione, G., Mastroianni, N., Rizzo, A., et al. (2026). Integrating artificial intelligence with genome sequencing against antimicrobial resistance: a narrative review. *Frontiers in Public Health*, 14, 1757161.
Seah, C., Alexander, D. C., Louie, L., Simor, A., Low, D. E., Longtin, J., & Melano, R. G. (2012). MupB, a new high-level mupirocin resistance mechanism in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, 56(4), 1916-1920.
States, D. J., & Agarwal, P. (1996). Compact encoding strategies for DNA sequence similarity search. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* (Vol. 4, pp. 211-217).
Strahilevitz, J., Jacoby, G. A., Hooper, D. C., & Robicsek, A. (2009). Plasmid-mediated quinolone resistance:

a multifaceted threat. *Clinical Microbiology Reviews*, 22(4), 664-689.

Vuong, C., Yeh, A. J., Cheung, G. Y., & Otto, M. (2016). Investigational drugs to treat methicillin-resistant Staphylococcus aureus. *Expert Opinion on Investigational Drugs*, 25, 73-93.

Wang, T., Hansen, K. R., Loving, J., Paschalidis, I. C., van Aggelen, H., & Simhon, E. (2021). Predicting antimicrobial resistance in the intensive care unit. *arXiv preprint*, arXiv:2111.03575.

World Health Organization. (2025). *Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report 2025*. Geneva: WHO.

Yang, L., Heckmann, D., Monk, J. M., Kavvas, E., & Palsson, B. O. (2020). A biochemically-interpretable machine learning classifier for microbial GWAS. *Nature Communications*, 11(1), 1-11.

Yang, Y., Li, B., Ju, F., & Zhang, T. (2013). Exploring variation of antibiotic resistance genes in activated sludge over a four-year period through a metagenomic approach. *Environmental Science & Technology*, 47(18), 10197-10205.

Yang, L. Y., Lin, C. S., Huang, X. R., Neilson, R., & Yang, X. R. (2022). Effects of biofertilizer on soil microbial diversity and antibiotic resistance genes. *Science of the Total Environment*, 820, 153170.

Yin, X., Jiang, X., Chai, B., Li, L., Yang, Y., Cole, J. R., Tiedje, J. M., & Zhang, T. (2018). ARGs-OAP v2.0 with an expanded SARG database and hidden Markov models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics*, 34, 2263–2270.

Yin, X., Zheng, X., Li, L., et al. (2022). ARGs-OAP v3.0: Antibiotic-Resistance Gene Database Curation and Analysis Pipeline Optimization. *Engineering*.

Zheng, D., Yin, G., Liu, M., Chen, C., Jiang, Y., Hou, L., & Chen, H. (2023). A systematic review of antibiotic resistance genes and their associations with bacterial pathogens in aquatic ecosystems. *Ecotoxicology and Environmental Safety*, 251, 114521.