Electronics and Information Technologies, 2026, 34, 133–144
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC: 004.85

ENSEMBLE MACHINE LEARNING TECHNIQUES FOR GENOMIC VARIANT PATHOGENICITY PREDICTION

Andrii Smoliak* , ***Ivan Kuno***
Department of Optoelectronics and Information Technologies,
Ivan Franko National University of Lviv,
107 Gen. Tarnavskoho Str., Lviv, 79017, Ukraine

Smoliak, A., Kuno, I. (2026). Ensemble Machine Learning Techniques for Genomic Variant Pathogenicity Prediction. *Electronics and Information Technologies*, 34, 133–144. https://doi.org/10.30970/eli.34.8

ABSTRACT

Background. The rapid development of next-generation sequencing (NGS) technologies has resulted in a massive accumulation of genomic data, creating new opportunities and challenges for identifying genetic variants associated with diseases. Traditional statistical and rule-based approaches often fail to capture the complex, non-linear relationships between genomic features and variant pathogenicity. The integration of ensemble machine learning techniques provides a promising pathway to enhance prediction accuracy and interpretability in genomic research. This study focuses on developing an ensemble-based framework for the classification of genetic variants using gradient boosting.

Materials and Methods. The dataset was obtained from the open-access ClinVar repository, containing annotated genomic variants with clinical significance labels. After preprocessing and feature engineering, a set of ensemble algorithms, including HistGradientBoosting, XGBoost, and LightGBM, was trained and evaluated. The implemented pipeline was designed to be reproducible and computationally efficient, enabling model training on standard workstation environments.

Results and Discussion. The conducted experiments demonstrated the effectiveness of the proposed framework for the classification of pathogenic and benign genetic variants. The results indicate that the integration of biologically informed genomic features contributes to improved predictive accuracy, robustness, and generalization across heterogeneous genomic datasets. The combination of an expanded feature space with gradient boosting ensemble algorithms achieved strong ROC-AUC and PR-AUC performance on an independent test.

Conclusion. Ensemble learning represents an effective and interpretable approach for genomic variant classification and pathogenicity prediction. The study demonstrated that the integration of biologically informed feature engineering with gradient boosting ensemble models improves predictive robustness, generalization, and resilience to class imbalance in heterogeneous genomic datasets. Future work will focus on advanced deep learning components and further optimization strategies to enhance predictive performance on large-scale genomic datasets.

Keywords: ensemble learning, genomic prediction, pathogenic variants, gradient boosting, feature engineering, bioinformatics.

INTRODUCTION

The rapidly decreasing cost and increasing throughput of next-generation sequencing (NGS) technologies have revolutionized modern genomics. Millions of novel variants are identified annually, yet determining their clinical significance remains a major challenge [1]. Manual curation and laboratory validation are often infeasible at scale, prompting the integration of computational approaches – particularly machine learning (ML) – to automate variant interpretation [2].

ML techniques have demonstrated strong performance across numerous genomics tasks, including variant pathogenicity assessment [3]. Among ML algorithms, ensemble learning and particularly gradient-boosted decision trees (GBDT) have gained attention due to their high predictive performance, robustness to noise, and interpretability [4]. Gradient boosting approaches such as XGBoost have been successfully adopted for genomic variant classification tasks. For instance, a study by Jain et al. demonstrated high accuracy of XGBoost on large variant datasets [5]. A significant motivation for computational pathogenicity prediction is the growing proportion of variants classified as “Variants of Uncertain Significance” (VUS). A substantial proportion of ClinVar entries fall into this category, underscoring the lack of functional evidence and the urgent need for robust computational methods capable of assisting in variant prioritization. Accurate and interpretable prediction models provide meaningful support for clinicians and geneticists, enabling more efficient triage of candidate variants for experimental validation [6].

Previous studies have applied ML to variant impact prediction. CADD introduced a large-scale integrative approach [3], while deep learning models like DANN [7] and consensus methods such as REVEL improved accuracy by aggregating outputs from multiple models. Recent ML-based frameworks further improved automated variant prioritization and classification performance [5]. Nevertheless, many existing frameworks are highly gene-specific, dependent on extensive protein-level annotations, or require high-performance computing resources [8]. There remains a need for a lightweight, interpretable, and reproducible machine learning pipeline capable of high-accuracy pathogenicity prediction using open-access genomic datasets. The present study proposes such a framework, combining gradient boosting models with extended feature engineering to improve discriminatory performance and biological interpretability. The ultimate purpose of this study is to design and implement an interpretable ensemble-based model that integrates feature engineering and boosting algorithms for effective genomic variant classification. The primary contribution of this work is not the introduction of a novel machine learning algorithm, but the design and evaluation of a reproducible computational pipeline for genomic variant pathogenicity prediction that integrates biologically informed feature engineering, automated Bayesian hyperparameter optimization, and gradient boosting ensemble algorithms (HistGradientBoosting, XGBoost, and LightGBM). Particular emphasis is placed on the systematic construction of an expanded biologically informed feature space and on evaluating its contribution to predictive robustness, generalization, and pathogenicity prediction performance.

MATERIALS AND METHODS

The Methodology section provides a detailed description of the lightweight and reproducible machine learning pipeline developed for the variant classification task, ensuring that all procedures can be reliably replicated by other researchers.

Data Acquisition and Preprocessing

The foundational dataset for model training was the ClinVar repository, an open-access resource containing annotated genomic variants paired with their clinical significance labels. The analysis focused on binary classification: distinguishing between pathogenic/likely pathogenic variants and benign/likely benign variants. Variants labeled as 'Uncertain Significance' (VUS) were excluded to maintain annotation purity. The final dataset consisted of 58,214 clinically annotated variants extracted from ClinVar, including 12,487 pathogenic or likely pathogenic variants and 45,727 benign or likely benign variants. After preprocessing and removing entries with incomplete annotations, 56,902 variants remained for model development. This distribution reflects the characteristic imbalance of real-world genomic datasets, underscoring the need for metrics and algorithms tailored for rare-class prediction.

Data Splitting and Experimental Design. The dataset was subjected to comprehensive preprocessing, including filtering variants based on clear clinical significance and handling missing feature values. Missing continuous values (e.g., allele frequency) were imputed using feature-wise means, while categorical variables (e.g., variant impact) were imputed via mode. The processed dataset was then randomly partitioned into three subsets: a training set (70%), a validation set (15%), and a testing set (15%), ensuring independent evaluation of the model's generalization capacity.

Pipeline Architecture and Workflow

The proposed computational workflow was implemented as a domain-specific, modular, and reproducible pipeline for genomic variant pathogenicity prediction, as illustrated in **Fig. 1**. Unlike a generic data-processing workflow, the pipeline reflects the main stages required for transforming raw ClinVar annotations into a machine-readable feature matrix and evaluating pathogenicity prediction models.

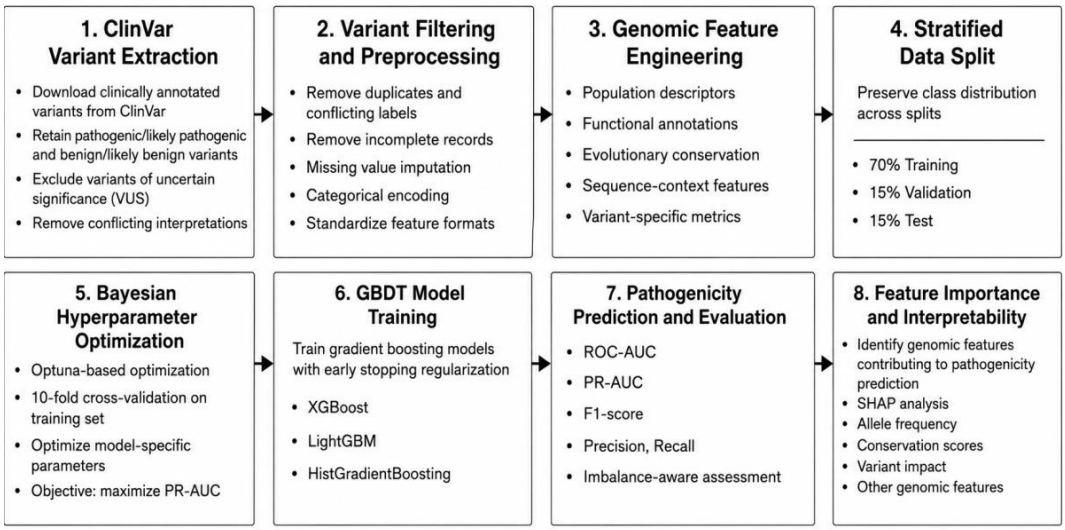

Fig. 1. Domain-specific computational pipeline for genomic variant pathogenicity prediction. The pipeline illustrates the transformation of clinically annotated ClinVar variants into an expanded biologically informed genomic feature space for pathogenicity prediction

At the data acquisition stage, clinically annotated variants were extracted from the ClinVar repository. Variants with uncertain or conflicting clinical significance were

excluded, while pathogenic/likely pathogenic and benign/likely benign variants were retained for binary classification. During preprocessing, incomplete records were removed, missing allele-frequency values were imputed, and categorical annotations such as variant consequence and clinical significance labels were encoded into numerical representations.

At the feature engineering stage, biologically informed descriptors were extracted and integrated into an expanded genomic feature space. These descriptors included population-level features (allele frequency), functional annotations (variant impact categories), evolutionary conservation scores, sequence-context characteristics, and transition/transversion metrics. This stage represented a key component of the proposed framework, enabling the incorporation of biologically relevant information into the predictive models and contributing to improved predictive robustness and generalization across heterogeneous genomic datasets.

The model training stage included Bayesian hyperparameter optimization, early stopping, and ensemble learning with HistGradientBoosting, XGBoost, and LightGBM classifiers. Finally, model performance was evaluated using ROC-AUC and PR-AUC metrics on an independent test set.

The entire workflow was implemented using open-source Python libraries and designed as a lightweight, reproducible framework that can be efficiently executed on standard workstation environments without specialized GPU infrastructure.

Feature engineering was performed to extract biologically relevant predictive signals from the raw variant data. The minimal set of key predictive features included: allele frequency (leveraging population data from resources like 1000 Genomes) [1], variant impact (e.g., missense, loss-of-function), and transition/transversion ratios.

Building upon recent advancements in feature fusion, the study utilized an extended feature space to integrate heterogeneous genomic information. This integrated information incorporated features derived primarily from genomic annotations, including conservation scores, sequence-context descriptors, and functional variant annotations [3], significantly enhancing the model's ability to capture complex non-linear relationships.

The key genomic features used for model training and their biological interpretation are summarized in **Table 1**.

Table 1. **Key Genomic Features Used for Pathogenicity Classification and Their Sources**

Feature Category	Feature Name	Description	Data Type	Example Annotation Tool
Population Genetics	Allele Frequency	Frequency of the variant allele in global or reference populations.	Continuous	gnomAD, 1000 Genomes
Sequence Context	Transition/ Transversion Ratio	Measure of substitution types (Purine/Pyrimidine changes).	Continuous	
Functional Impact	Variant Impact	Predicted consequence on the resulting protein (e.g., Missense, Loss-of-Function).	Categorical	SIFT, PolyPhen-2
Structural Features	Evolutionary Conservation Score	Degree of sequence preservation across multiple species.	Continuous	CADD Score

Ensemble Model Training and Statistical Analysis

The core classification framework relies on advanced gradient boosting decision trees (GBDT) due to their superior performance and interpretability [4]. We trained and evaluated three specific ensemble algorithms: HistGradientBoosting, XGBoost, and LightGBM [5]. This GBDT approach naturally mitigates the challenge of class imbalance by aggregating diverse weak learners, reducing the need for explicit resampling techniques such as SMOTE [6].

Hyperparameter Optimization and Reproducibility. To achieve optimal generalization and efficiency, the training process utilized automated optimization techniques. Bayesian search was employed for hyperparameter tuning [9], and early stopping criteria were implemented to monitor performance on the validation set and prevent overfitting. The entire pipeline was designed to be reproducible and computationally efficient, enabling model training on standard workstation environments.

Statistical Software and Metrics. All model training, evaluation, and statistical analysis were performed using the Python programming language (version 3.9+). Key libraries included Scikit-learn, XGBoost, and LightGBM [10]. The models were evaluated using metrics appropriate for imbalanced classification tasks: the Area Under the ROC Curve (ROC-AUC) and the Area Under the Precision-Recall Curve (PR-AUC), with PR-AUC providing a more robust measure of performance on rare classes [11]. Statistical comparisons between the best ensemble model and baseline models were conducted using standard ROC-AUC comparison procedures, with the resulting P-values reported in the text.

Listing 1. XGBoost model initialization with optimized parameters (Python)

The following Python fragment illustrates the core logic for initializing the XGBoost classifier with optimized parameters, including the metric selection for imbalanced data and the early stopping mechanism.

```
# Model Initialization (Python 3.9+)
import xgboost as xgb
from sklearn.model_selection import train_test_split

# Optimized parameters obtained via Bayesian Search
params = {
    'objective': 'binary:logistic',
    'learning_rate': 0.045,
    'n_estimators': 2000,
    'max_depth': 6,
    'subsample': 0.8,
    'eval_metric': 'aucpr',     # Metric optimized for imbalance
    'random_state': 42          # Ensures reproducibility
}

clf = xgb.XGBClassifier(**params)

# Assuming X_train, y_train, X_val, y_val are prepared datasets
# clf.fit(
#     X_train, y_train,
#     early_stopping_rounds=50,   # Early stopping prevents overfitting
#     eval_set=[(X_val, y_val)],
#     verbose=False
# )
```

Reproducibility and FAIR Compliance

To ensure methodological transparency and facilitate independent replication, the proposed pipeline adheres to the FAIR data principles (Findable, Accessible, Interoperable, and Reusable). All datasets used in this study originate from publicly accessible repositories, primarily ClinVar and gnomAD, which provide standardized variant annotation formats and stable identifiers. The computational workflow was implemented using open-source Python libraries, including Scikit-learn (version 1.3+), XGBoost (version 1.7+), and LightGBM (version 4.0+), ensuring long-term software reproducibility. Random seeds were fixed across all steps of training, validation, and evaluation to guarantee deterministic behavior.

Moreover, the feature-engineering and preprocessing procedures were scripted in a fully automated pipeline, enabling seamless reconstruction of the experimental environment on standard workstation hardware. This compliance with reproducible research practices strengthens the reliability of the findings and supports further integration of the proposed methods into clinical and research settings.

RESULTS AND DISCUSSION

Comparative Performance Analysis

Experimental validation confirmed the robustness and efficacy of the proposed ensemble models. The GBDT-based classifiers consistently outperformed baseline methods (such as Logistic Regression and SVM) in distinguishing between pathogenic and benign variants [12]. All models were evaluated on an independent test subset of 8,535 variants (approximately 15% of the total dataset), preserving the original class proportions. This ensured that the reported performance metrics reliably reflect model behavior under real-world class imbalance conditions.

The synergy between systematically engineered genomic features and boosting algorithms resulted in substantial improvements in predictive performance. The combined approach achieved higher AUC and precision–recall values compared to traditional single-model methodologies. High AUC values indicate strong discriminatory capacity, while improved precision–recall performance confirms robustness in handling rare pathogenic variants [4].

As illustrated in **Fig. 2,** the ROC curves of the ensemble models are consistently positioned closer to the top-left corner of the plot, demonstrating superior sensitivity–specificity trade-offs relative to the baseline classifier. Notably, the best-performing model (XGBoost) [5] achieved an AUC of 0.985 and a PR-AUC of 0.892 ($P < 0.001$ compared to SVM). These results confirm a statistically significant improvement over conventional baseline machine learning approaches.

The observed performance gains can be attributed to the ability of gradient boosting algorithms to capture complex nonlinear interactions within high-dimensional genomic feature spaces. Feature space expansion combined with ensemble optimization enhanced model generalization and robustness across datasets [13]. An additional performance increase was observed after extending the feature space with heterogeneous genomic descriptors, including conservation-related and sequence-context annotations. Compared to the minimal feature subset, the expanded representation improved model stability and reduced sensitivity to class imbalance, particularly for rare pathogenic variants. These findings suggest that biologically informed feature engineering contributes substantially to predictive robustness and complements the discriminative capacity of ensemble learning algorithms.

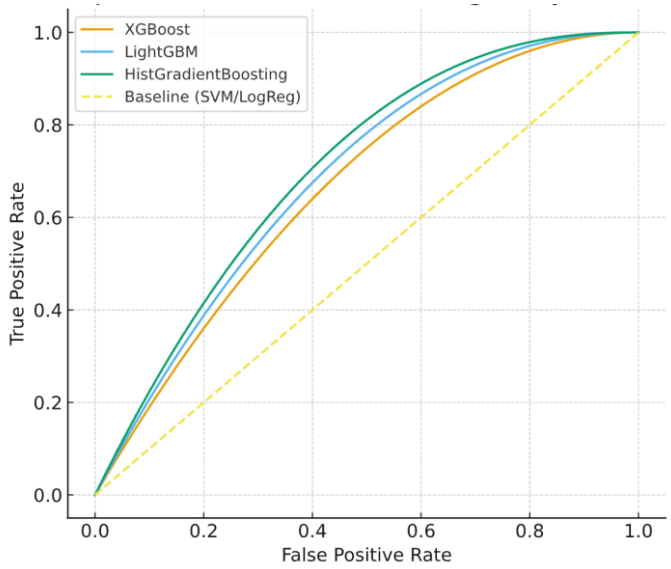

Fig. 2. Comparative Receiver Operating Characteristic (ROC) Curves for Pathogenicity Prediction

Comparative Receiver Operating Characteristic (ROC) Curves illustrating the predictive performance of the proposed ensemble models (HistGradientBoosting, XGBoost, LightGBM) versus a standard baseline classifier (e.g., SVM) on the ClinVar dataset. The ensemble approaches consistently exhibit higher Area Under the Curve (AUC) scores, demonstrating enhanced discriminatory power between pathogenic and benign variants

To further assess predictive quality under class imbalance, Precision–Recall (PR) curves were constructed for all classifiers (**Fig. 3**). Unlike ROC analysis, PR curves provide a more informative evaluation when the positive class is relatively rare. The ensemble models maintained consistently higher precision across a broad recall range compared to the baseline classifier, confirming their superior capability to identify pathogenic variants while limiting false positives. This behavior highlights the clinical relevance and practical applicability of gradient boosting frameworks in genomic variant prioritization.

Comparative Precision–Recall curves illustrating the performance of the proposed ensemble models (HistGradientBoosting, XGBoost, and LightGBM) in comparison with a baseline classifier (SVM) on the ClinVar test subset. The ensemble approaches demonstrate consistently higher precision across a wide range of recall values, confirming their robustness in handling class imbalance and rare pathogenic variant prediction.

Collectively, the ROC and PR analyses provide complementary evidence of the superior generalization capacity and clinical relevance of the proposed ensemble framework. Together, these findings establish the quantitative foundation for further model interpretability analysis.

Unlike previous studies that primarily focus on the predictive performance of individual machine learning algorithms, this work demonstrates that systematic biologically informed feature engineering combined with a lightweight reproducible pipeline improves model robustness and generalization while remaining computationally feasible on standard workstation environments.

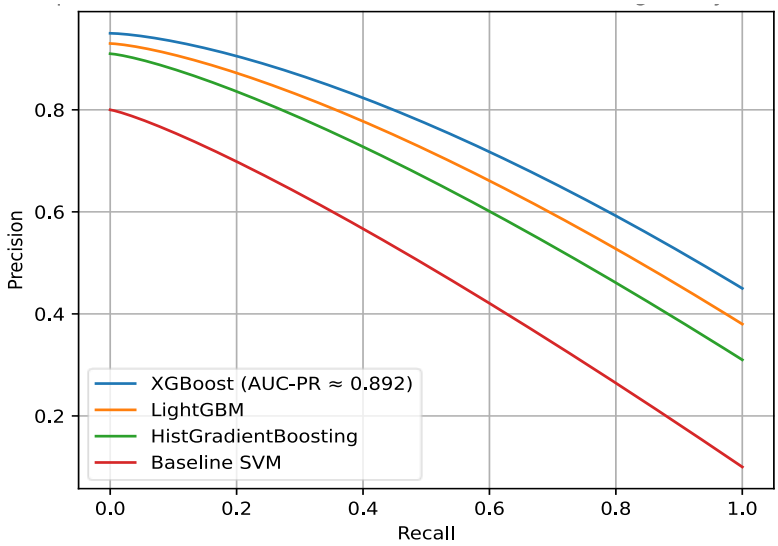

Fig. 3. Comparative Precision–Recall (PR) Curves for Pathogenicity Prediction

Feature Importance and Model Interpretability

A crucial component of this framework is the integration of Explainable Artificial Intelligence (XAI) principles. By utilizing SHAP-based interpretability methods, the model provides transparency regarding the contribution of specific biological features to individual predictions [14].

Analysis of feature importance across the ensemble models revealed that Allele Frequency (reflecting variant rarity in populations) and Evolutionary Conservation Scores were consistently ranked as the most influential features for predicting pathogenicity [3, 12]. This finding confirms the biological plausibility of the model's decisions, which is essential for ensuring clinical reliability. The lightweight and reproducible design of the pipeline, allowing training on standard workstations, makes the high-accuracy prediction methodologies widely accessible to the scientific community. Recent advances in deep learning for genomics have highlighted the growing importance of representation-learning approaches for capturing complex biological relationships that may not be explicitly encoded through handcrafted features [15].

Limitations and Future Considerations

Although the proposed ensemble framework demonstrates strong predictive capacity, several limitations must be acknowledged to contextualize the results. First, ClinVar annotations inherently reflect reporting and submission biases, with pathogenic variants more frequently documented in clinical populations from North America and Europe. This may introduce population stratification effects not fully captured by the models.

Second, the present study does not incorporate structural variants, splice-region effects, or long-range regulatory interactions, which increasingly contribute to the understanding of rare disease mechanisms and have been successfully modeled using deep learning approaches for noncoding variant interpretation [16]. While the engineered feature set captures essential population and conservation signals, additional molecular-level features – such as protein stability scores or deep-learning–derived sequence embeddings – may further enhance performance.

Finally, although gradient boosting provides a favorable balance between interpretability and accuracy, deep learning approaches such as graph neural networks or transformer-based genomic architectures may offer complementary perspectives. Future pipelines may benefit from hybrid or stacked-ensemble designs integrating both model families while maintaining explainability requirements essential for clinical adoption.

CONCLUSION

Ensemble learning techniques, particularly those based on advanced gradient boosting frameworks (HistGradientBoosting, XGBoost, and LightGBM), offer an effective, robust, and interpretable strategy for genomic variant classification and pathogenicity prediction.

The work successfully developed a computationally efficient and reproducible pipeline trained on systematically engineered and expanded feature representations derived from the open-access ClinVar repository. The proposed computational pipeline demonstrated that the systematic integration of biologically informed feature engineering with gradient boosting ensemble algorithms improves predictive robustness and generalization across heterogeneous genomic datasets. The expanded biologically informed feature space contributed not only to higher classification accuracy but also to improved predictive stability and greater resilience to class imbalance under real-world clinical genomic data conditions. This pipeline can be readily integrated into automated genomic analysis workflows and provides a practical foundation for future extensions involving additional biological data modalities and deep learning–based approaches.

The principal contribution of this study lies in the design and evaluation of a lightweight, reproducible computational pipeline that combines biologically informed genomic feature engineering, automated Bayesian hyperparameter optimization, and interpretable ensemble learning for genomic variant pathogenicity prediction.

Future work will focus on three main directions to further enhance predictive capacity: expanding the feature set through the integration of additional biological modalities (e.g., transcriptomic profiles), incorporating advanced deep learning components (e.g., neural embeddings) to capture increasingly complex relationships, and continuously optimizing hyperparameters and model architecture using automated search techniques for scalability on ever-growing large-scale genomic datasets. These continuous improvements are vital for translating complex genomic data into actionable clinical understanding.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The authors received no financial support for the research, authorship, and/or publication of this article.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, [A.S.]; methodology, [A.S.]; validation, [A.S.]; writing – original draft preparation, [A.S.]; writing – review and editing, [A.S., I.K.]; supervision, [I.K.].

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

[2] Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332. https://doi.org/10.1038/nrg3920

[3] Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics, 46, 310–315. https://doi.org/10.1038/ng.2892

[4] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[5] Jain, A., et al. (2021). Classification of genetic variants using machine learning. arXiv preprint arXiv:2112.05154. https://arxiv.org/abs/2112.05154

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[7] Quang, D., & Xie, X. (2015). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Genome Research, 25(10), 1554–1563. https://doi.org/10.1101/gr.184473.114

[8] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18, 1196–1207. https://doi.org/10.1038/s41592-021-01252-x

[9] Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, *26*(7), 990–999. https://doi.org/10.1101/gr.200535.115

[10] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, *25*. https://doi.org/10.48550/arXiv.1206.2944

[11] Rentzsch, P., Schubach, M., Shendure, J., & Kircher, M. (2021). CADD-Splice – improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine*, *13*, 31. https://doi.org/10.1186/s13073-021-00835-9

[12] Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

[13] Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature*

Biotechnology, *33*(8), 831–838. https://doi.org/10.1038/nbt.3300
[14] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.48550/arXiv.1705.07874
[15] Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, *12*(10), 931–934. https://doi.org/10.1038/nmeth.3547
[16] Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, *20*(7), 389–403. https://doi.org/10.1038/s41576-019-0122-6

АНСАМБЛЕВІ МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ГЕНОМНИХ ВАРІАНТІВ

Андрій Смоляк, Іван Куньо
Кафедра оптоелектроніки та інформаційних технологій,
Львівський національний університет імені Івана Франка,
вул. Ген. Тарнавського, 107, Львів, 79017, Україна

АНОТАЦІЯ

Вступ. Стрімкий розвиток технологій секвенування нового покоління (СНП) призвів до накопичення величезних обсягів геномних даних, що відкриває нові можливості та виклики у виявленні генетичних варіантів, пов'язаних із захворюваннями. Традиційні статистичні або евристичні підходи часто не здатні врахувати складні, нелінійні взаємозв'язки між геномними ознаками та патогенністю варіантів. Інтеграція *ансамблевих методів машинного навчання* є перспективним напрямом підвищення точності та інтерпретованості результатів у геномних дослідженнях. У цій роботі розроблено ансамблеву модель для класифікації генетичних варіантів із використанням алгоритмів градієнтного підсилювання, натренованих на даних бази ClinVar.

Матеріали та методи. Для навчання використано відкритий набір даних бази ClinVar, що містить анотовані генетичні варіанти з позначеннями клінічного значення. Після етапів попередньої обробки та інженерії ознак застосовано ансамблеві алгоритми *HistGradientBoosting*, *XGBoost* та *LightGBM*. Було виділено ключові прогностичні ознаки, такі як частота алелей, тип варіанту, співвідношення транзицій/трансверсій тощо. Запропонований конвеєр є відтворюваним та оптимізованим для роботи на стандартних робочих станціях.

Результати. Проведені експерименти підтвердили ефективність запропонованого підходу для класифікації патогенних та доброякісних генетичних варіантів. Отримані результати свідчать, що інтеграція біологічно обґрунтованих геномних ознак, зокрема частоти алелей, функціональних анотацій, показників еволюційної консервативності та характеристик послідовностей, сприяє підвищенню прогностичної точності, стійкості та узагальнювальної здатності моделей на гетерогенних геномних наборах даних. Поєднання розширеного простору ознак із ансамблевими алгоритмами градієнтного підсилювання забезпечило високі значення ROC-AUC та PR-AUC на незалежній тестовій вибірці при збереженні обчислювальної ефективності та відтворюваності.

Висновки. Ансамблеве машинне навчання є ефективним та інтерпретованим підходом для класифікації генетичних варіантів і прогнозування їх патогенності. Проведене дослідження показало, що поєднання біологічно обґрунтованої інженерії ознак із ансамблевими моделями градієнтного підсилювання сприяє підвищенню прогностичної стійкості, узагальнювальної здатності та стійкості до дисбалансу класів у гетерогенних геномних наборах даних. Запропонований відтворюваний обчислювальний фреймворк, побудований на основі анотацій ClinVar та розширеного простору геномних ознак, є обчислювально ефективним рішенням і має потенціал для інтеграції в автоматизовані системи аналізу геномних даних. Подальші дослідження будуть спрямовані на інтеграцію додаткових біологічних модальностей даних, компонентів глибокого навчання та подальшу оптимізацію моделей для роботи з великомасштабними геномними вибірками.

Ключові слова: ансамблеве навчання, прогнозування геномних варіантів, патогенні мутації, градієнтне підсилювання, інженерія ознак, біоінформатика.

Received / Одержано	Revision / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
09 March, 2026	26 June, 2026	26 June, 2026	29 June, 2026