DOI: 10.25205/978-5-4437-1843-9-7

DEVELOPMENT OF A METHOD FOR PREDICTING CHROMOTHRIPSIS IN THE HUMAN GENOME USING MACHINE LEARNING ALGORITHMS

P. E. Karitskaia, Y. V. Vyatkin

Novosibirsk State University

✉ p.karitskaya@g.nsu.ru

Abstract

A comparison was conducted between classical machine learning algorithms and a BiLSTM-based neural network for predicting chromothripsis events in the human genome using structural variation data. The BiLSTM model, which processes sequential input, demonstrated significantly higher classification performance (PR AUC 0.94 vs. 0.78 for the best classical model), highlighting its potential for clinical diagnostics.

Introduction

Chromothripsis is a phenomenon involving multiple chromosomal rearrangements that occur as a result of a single catastrophic event. It is associated with poor prognosis in cancer and has high diagnostic value. Current chromothripsis detection algorithms are often based on statistical approaches that require manual parameter tuning [1, 2], or they are highly specialized — for example, trained only on copy number variation (CNV) data or specific cancer types [3].

Aim — to develop a method for predicting chromothripsis events in the human genome based on structural variation (SV) data using machine learning and deep learning algorithms.

Materials and methods

The study was conducted using data from ChromothripsisDB, the largest annotated repository of chromothripsis cases [3]. After preprocessing and filtering, the final dataset included 19713 events.

For classical machine learning, aggregated features were used, such as the number and types of structural variations (SVs) within a genomic region. Logistic regression, support vector classifier (SVC), random forest, and XGBoost models were tested [4]. Class imbalance (11.4 % positive cases) was addressed using the SMOTE method and class weighting [5]. Ensemble methods such as Voting and Stacking were applied to improve performance.

For the deep learning approach, a custom architecture was developed to work directly with SV sequences. Each genomic region was represented as a sequence of 200 events, where each event was described by 10 features (coordinates, type, read support, etc.). A two-layer bidirectional LSTM network (BiLSTM) with an attention mechanism was implemented [6]. To handle class imbalance, a WeightedRandomSampler and a combined loss function (Focal Loss + F1 Loss) were used. The model was trained using 10-fold stratified cross-validation.

Results

Comparison of classical machine learning models on a hold-out test set showed that the XGBoost algorithm achieved the best performance, with ROC AUC = 0.93 and PR AUC = 0.78. Ensemble methods yielded similar results but did not provide a significant improvement (see table).

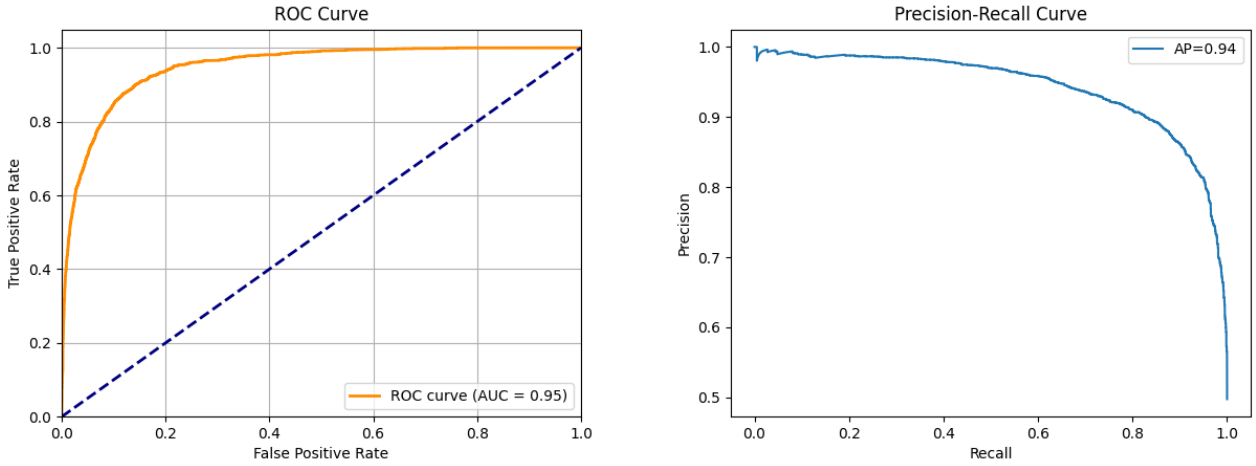
The developed BiLSTM-based neural network outperformed all classical models across key metrics. On validation folds, the model achieved average scores of ROC AUC = 0.95 and PR AUC = 0.94 (see figure). Training dynamics analysis indicated a stable decrease in loss without signs of overfitting, confirming the effectiveness of the applied regularization methods and the model's ability to generalize.

Conclusions

The proposed architecture based on a bidirectional LSTM network with an attention mechanism demonstrated superior predictive performance compared to classical machine learning algorithms. This is attributed to the model's ability to capture complex spatial and contextual dependencies in structural variation sequences, which are lost during feature aggregation. The results confirm the potential of deep learning for analyzing raw genomic data and lay the groundwork for developing more accurate and automated tools for clinical diagnostics.

Comparative metrics of models on the hold-out test set

Model	ROC AUC	PR AUC	F1-score
LogReg	0.92	0.74	0.62
SVM	0.91	0.74	0.67
Random Forest	0.93	0.74	0.58
XGBoost	0.93	0.78	0.73
Voting	0.93	0.77	0.69
Stacking	0.93	0.78	0.73
BiLSTM + Attention	0.95	0.94	0.70



ROC (*left*) and Precision-Recall (*right*) curves for the BiLSTM network in the chromothripsis event classification task

References

1. Govind S. K. et al. ShatterProof: operational detection and quantification of chromothripsis // BMC Bioinformatics. Springer Science and Business Media LLC. 2014. Vol. 15, No. 1.

2. Cortés-Ciriano I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing // Nat. Genet. Springer Science and Business Media LLC. 2020. Vol. 52, No. 3. P. 331–341.

3. Yu J. et al. Chromothripsis detection with multiple myeloma patients based on deep graph learning // Bioinformatics / ed. Lu Z. Oxford University Press (OUP). 2023. Vol. 39, No. 7.

4. James G. et al. An Introduction to Statistical Learning // Springer Texts in Statistics. Springer International Publishing, 2023.

5. Chawla N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique // JAIR. AI Access Foundation, 2002. Vol. 16. P. 321–357.