People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
UNIVERSITY FERHAT ABBAS – SETIF-1



Faculty of Science
Computer Science Department

Field : Mathematics and Computer Science
Specialization : Computer Science
Specialty : data engineering and web technologies

DISTRIBUTED MULTIMODAL DEEP LEARNING FRAMEWORK FOR HIGH-PRECISION GENOMIC VARIANT CALLING

Academic year

2025 - 2026

Presented by:

OUARET Lamia

HAMOU Rahil

Supervised by:

Mr. NASRI Khaled

In front of the jury:

Mr. MEKROUD Noureddine

Ms. ZENBOUT Imene

Abstract

The rapid growth of next-generation sequencing technologies has transformed genomics into a highly data-intensive discipline, generating unprecedented volumes of biological data and creating new challenges for accurate and scalable genomic analysis. Among the various computational tasks in bioinformatics, variant calling plays a fundamental role in identifying genetic variations associated with diseases, personalized medicine, population genetics, and precision oncology. Although recent deep learning-based methods have significantly improved variant detection accuracy, many existing approaches rely on a single data representation and often fail to exploit the complementary information available within sequencing data. Furthermore, the increasing sensitivity and distributed nature of genomic datasets raise important concerns regarding scalability, privacy, and collaborative analysis.

This work presents a distributed multimodal deep learning framework for high-precision genomic variant calling. The proposed approach combines two complementary modalities: pileup tensor representations processed by a Convolutional Neural Network (CNN) and sequence-based representations analyzed by a Transformer encoder. A gated fusion mechanism is employed to integrate the features extracted from both modalities and generate robust variant classifications. The study also investigates the theoretical foundations of genomic data, reviews state-of-the-art variant calling techniques, and examines distributed computing paradigms and federated learning approaches applicable to genomics.

The implemented framework is designed to exploit the strengths of both spatial and sequential representations of genomic information, thereby providing a more comprehensive characterization of genetic variants. In addition, a federated extension is proposed as a future development to enable privacy-preserving collaborative learning across multiple institutions without requiring the exchange of sensitive genomic data.

By integrating multimodal deep learning with concepts from distributed computing, this work contributes to the ongoing convergence of artificial intelligence and genomics and provides a foundation for the development of scalable, accurate, and privacy-aware genomic analysis systems. Future extensions of the proposed framework may facilitate large-scale collaborative genomic studies and support the advancement of precision medicine.

Keywords: Genomic Variant Calling, Deep Learning, Multimodal Learning, Convolutional Neural Networks, Transformers, Feature Fusion, Distributed Computing, Federated Learning, Precision Medicine.

Résumé

La croissance rapide des technologies de séquençage de nouvelle génération a transformé la génomique en une discipline fortement intensive en données, générant des volumes sans précédent d'informations biologiques et créant de nouveaux défis en matière d'analyse génomique précise et évolutive. Parmi les nombreuses tâches computationnelles en bioinformatique, l'identification des variants génétiques occupe une place fondamentale dans la détection des variations génétiques associées aux maladies, à la médecine personnalisée, à la génétique des populations et à l'oncologie de précision. Bien que les méthodes récentes basées sur l'apprentissage profond aient considérablement amélioré la précision de la détection des variants, la plupart des approches existantes reposent sur une représentation unique des données et n'exploitent pas pleinement les informations complémentaires contenues dans les données de séquençage. En outre, la nature distribuée et la sensibilité des données génomiques soulèvent d'importantes problématiques liées à l'évolutivité, à la confidentialité et à l'analyse collaborative.

Ce travail présente un cadre distribué d'apprentissage profond multimodal destiné à l'identification de variants génétiques avec une haute précision. L'approche proposée combine deux modalités complémentaires : des représentations sous forme de tenseurs de pileup traitées par un réseau de neurones convolutif (CNN) et des représentations séquentielles analysées à l'aide d'un encodeur Transformer. Un mécanisme de fusion pondérée est utilisé afin d'intégrer les caractéristiques extraites à partir des deux modalités et de produire des classifications robustes des variants. Cette étude examine également les fondements théoriques des données génomiques, passe en revue les méthodes les plus récentes de détection des variants et analyse les paradigmes de calcul distribué et d'apprentissage fédéré applicables à la génomique.

Le cadre implémenté est conçu pour exploiter les avantages des représentations spatiales et séquentielles de l'information génomique, permettant ainsi une caractérisation plus complète des variations génétiques. Par ailleurs, une extension fondée sur l'apprentissage fédéré est proposée comme perspective future afin de permettre un apprentissage collaboratif respectueux de la confidentialité entre plusieurs institutions, sans nécessiter le partage direct des données génomiques sensibles.

En combinant l'apprentissage profond multimodal et les concepts du calcul distribué, ce travail contribue à la convergence croissante entre l'intelligence artificielle et la génomique et fournit une base pour le développement de systèmes d'analyse génomique évolutifs, précis et respectueux de la confidentialité. Les développements futurs du cadre proposé pourraient favoriser les études génomiques collaboratives à grande échelle et soutenir les avancées de la médecine de précision.

Mots-clés : Identification des variants génétiques, Apprentissage profond, Apprentissage multimodal, Réseaux de neurones convolutifs, Transformers, Fusion de caractéristiques, Calcul distribué, Apprentissage fédéré, Médecine de précision.

ملخص

أدى التطور السريع لتقنيات التسلسل الجيني من الجيل الجديد(NGS) إلى تحويل علم الجينوم إلى مجال يعتمد بشكل كبير على البيانات، حيث يتم إنتاج كميات هائلة وغير مسبوقة من المعلومات البيولوجية، مما أوجد تحديات جديدة تتعلق بإجراء تحليلات جينومية دقيقة وقابلة للتوسع. ومن بين مختلف المهام الحاسوبية في المعلوماتية الحيوية، تحتل عملية اكتشاف المتغيرات الجينية مكانة أساسية لما لها من دور مهم في تحديد الاختلافات الجينية المرتبطة بالأمراض، والطب الشخصي، وعلم الوراثة السكانية، وعلم الأورام الدقيق. وعلى الرغم من أن الأساليب الحديثة المعتمدة على التعلم العميق قد ساهمت بشكل كبير في تحسين دقة اكتشاف المتغيرات الجينية، فإن العديد من الطرق الحالية تعتمد على تمثيل واحد للبيانات، وبالتالي لا تستفيد بصورة كاملة من المعلومات التكاملية المتاحة في بيانات التسلسل الجيني. بالإضافة إلى ذلك، فإن الطبيعة الموزعة والحساسة للبيانات الجينومية تثير تحديات مهمة تتعلق بقابلية التوسع، وحماية الخصوصية، وإجراء التحليلات التعاونية.

يقدم هذا العمل إطاراً موزعاً للتعلم العميق متعدد الأنماط من أجل تحقيق اكتشاف عالي الدقة للمتغيرات الجينية. ويعتمد النهج المقترح على دمج نمطين متكاملين من البيانات، يتمثل الأول في تمثيلات المصفوفات الناتجة عن الPileup والتي تتم معالجتها بواسطة الشبكات العصبية الالتفافية (CNN)، بينما يعتمد النمط الثاني على التمثيلات التسلسلية التي تتم معالجتها باستخدام مشفر قائم على بنية المحولات (Transformer). كما يتم استخدام آلية دمج موجهة لدمج الخصائص المستخرجة من كلا النمطين بهدف إنتاج تصنيفات أكثر موثوقية للمتغيرات الجينية. إضافة إلى ذلك، تستعرض هذه الدراسة الأسس النظرية للبيانات الجينومية، وتناقش أحدث تقنيات اكتشاف المتغيرات الجينية، كما تتناول مفاهيم الحوسبة الموزعة وأساليب التعلم الفيدرالي المطبقة في مجال الجينوميات.

وقد صُمم الإطار المقترح للاستفادة من مزايا كل من التمثيلات المكانية والتسلسلية للمعلومات الجينومية، مما يسمح بوصف أكثر شمولاً للمتغيرات الجينية. كما تم اقتراح امتداد قائم على التعلم الفيدرالي باعتباره توجهاً مستقبلياً يهدف إلى تمكين التعلم التعاوني بين المؤسسات المختلفة مع الحفاظ على خصوصية البيانات الجينومية وعدم الحاجة إلى تبادلها بشكل مباشر.

ومن خلال الجمع بين التعلم العميق متعدد الأنماط ومفاهيم الحوسبة الموزعة، يساهم هذا العمل في تعزيز التكامل المتزايد بين الذكاء الاصطناعي وعلم الجينوم، ويوفر أساساً لتطوير أنظمة تحليل جينومي دقيقة وقابلة للتوسع وتحافظ على خصوصية البيانات. ومن المتوقع أن تساهم التوسعات المستقبلية للإطار المقترح في دعم الدراسات الجينومية التعاونية واسعة النطاق وتعزيز تطبيقات الطب الدقيق.

الكلمات المفتاحية: اكتشاف المتغيرات الجينية، التعلم العميق، التعلم متعدد الأنماط، الشبكات العصبية الالتفافية، المحولات، دمج الخصائص، الحوسبة الموزعة، التعلم الفيدرالي، الطب الدقيق.

Contents

List of Abbreviations **vi**

1 Theoretical Background of Genomic Data **2**
1 Introduction . . . 2
2 Genomic Data . . . 2
2.1 Core Components . . . 2
3 Big Genomic Data acquisition . . . 3
3.1 Sample Collection and Preparation . . . 3
3.2 DNA Sequencing (The Digitization Step) . . . 4
3.3 Base Calling . . . 5
3.4 Raw Data Storage – FASTQ . . . 5
3.5 Sequence Alignment and Assembly . . . 6
4 File Format . . . 6
4.1 Raw Data Acquisition (FASTQ) . . . 7
4.2 Sequence Alignment and Mapping (SAM/BAM/CRAM) . . . 8
4.3 Variant Identification (VCF) . . . 9
5 Data Resources . . . 9
5.1 The 1000 Genomes Project (1kGP) . . . 9
5.2 The Genome Aggregation Database (gnomAD) . . . 10
5.3 ClinVar . . . 11
5.4 Genome in a Bottle (GIAB) . . . 12
6 Existing Technologies . . . 13
6.1 Variant Calling & Interpretation . . . 13
6.2 Protein Structure Prediction . . . 14
6.3 Genomic Foundation Models & Transformers . . . 14
6.4 AI-Enhanced CRISPR & Genome Editing . . . 14
6.5 Generative AI Models for Genomics . . . 15
6.6 Genome Annotation & Functional Element Identification . . . 15
6.7 Clinical Genomics & Precision Medicine . . . 15
6.8 Data Harmonization & Quality Improvement . . . 16
6.9 Multi-Omics Integration & Spatial Transcriptomics . . . 16
7 Conclusion . . . 16

2 State of the Art **18**
1 Introduction . . . 18
2 Genetic Variant Types . . . 18
2.1 Single Nucleotide Polymorphisms (SNPs) . . . 18
2.2 Insertions and Deletions (Indels) . . . 18

2.3 Multiple Nucleotide Polymorphisms (MNPs) . . . 19
2.4 Copy Number Variants (CNVs) . . . 20
2.5 Structural Variants (SVs) . . . 20
3 Variant calling Definition . . . 21
4 Standard Variant Calling Pipeline . . . 22
4.1 Data Pre-processing and Quality Control . . . 22
4.2 Variant Discovery (Calling) . . . 24
4.3 Filtering, Refinement, and Quality Control . . . 28
5 Challenges in Variant Calling . . . 31
5.1 Repetitive and Low-Complexity Genomic Regions . . . 31
5.2 Homopolymer-Induced Sequencing Errors . . . 32
5.3 Sequencing Artifacts and Strand Bias . . . 32
5.4 Structural Variant Detection . . . 33
5.5 Computational Scale and Data Management . . . 33
6 Advantages and Features of Variant Calling . . . 33
6.1 Comprehensive Variant Detection . . . 33
6.2 High Accuracy and Sensitivity . . . 34
6.3 Joint Calling and Population-Scale Analysis . . . 34
6.4 Scalability and Reproducibility . . . 34
6.5 Clinical Applications . . . 35
6.6 Flexibility Across Sequencing Strategies . . . 35
6.7 Benchmarking and Continuous Improvement . . . 35
7 Applications of Variant Calling . . . 35
7.1 Clinical Diagnostics and Rare Disease . . . 36
7.2 Cancer Genomics and Precision Oncology . . . 36
7.3 Pharmacogenomics and Personalized Medicine . . . 36
7.4 Population Genetics and Complex Disease (GWAS) . . . 37
7.5 Infectious Disease and Public Health Surveillance . . . 37
7.6 Agricultural Genomics and Crop Improvement . . . 37
7.7 Conservation Genetics and Wildlife Biology . . . 38
8 DeepVariant (Modern Iterations) . . . 38
8.1 Overview and Core Innovation . . . 38
8.2 Model Architecture and Technical Development . . . 39
8.3 Training Data and Benchmarking . . . 39
8.4 Pipeline Architecture and Workflow . . . 39
8.5 Performance Metrics and Comparative Advantages . . . 40
8.6 Strengths, Limitations, and Special Features . . . 40
9 Clair3 . . . 40
9.1 Overview and Core Innovation . . . 40
9.2 Model Architecture and Input Design . . . 40
9.3 Training Data and Sequencing Context . . . 41
9.4 Pipeline Workflow and Computational Infrastructure . . . 41
9.5 Performance and Comparative Benchmarks . . . 41
9.6 Strengths, Limitations, and Special Features . . . 42
10 DeepTrio . . . 42
10.1 Overview and Core Innovation . . . 42
10.2 Model Architecture and Input Design . . . 42
10.3 Training Data and Benchmarking . . . 43

10.4 Pipeline Workflow and Computational Requirements 43
10.5 Performance and Comparative Benchmarks 43
10.6 Strengths, Limitations, and Special Features 43
11 Octopus 44
11.1 Overview and Core Innovation 44
11.2 Model Architecture and Algorithmic Components 44
11.3 Training Data and Benchmarking Context 44
11.4 Pipeline Workflow and Computational Architecture 44
11.5 Performance and Comparative Benchmarks 45
11.6 Strengths, Limitations, and Special Features 45
12 Strelka2 45
12.1 Overview and Core Innovation 45
12.2 Model Architecture and Input Design 46
12.3 Training Data and Benchmarking Context 46
12.4 Pipeline Workflow and Computational Architecture 46
12.5 Performance and Comparative Benchmarks 47
12.6 Strengths, Limitations, and Special Features 47
13 GATK HaplotypeCaller (Modern Usage v4.2+) 47
13.1 Overview and Core Innovation 47
13.2 Training Data and Population Databases 48
13.3 Pipeline Workflow and Infrastructure Requirements 48
13.4 Performance and Comparative Benchmarks 49
13.5 Strengths, Limitations, and Special Features 49
14 Medaka 49
14.1 Overview and Core Innovation 49
14.2 Model Architecture and Input Design 49
14.3 Training Data and Platform Specificity 50
14.4 Pipeline Workflow and Computational Requirements 50
14.5 Performance and Comparative Benchmarks 50
14.6 Strengths, Limitations, and Special Features 51
15 PEPPER-Margin-DeepVariant 51
15.1 Overview and Core Innovation 51
15.2 Model Architecture and Multi-Component Design 51
15.3 Training Data and Sequencing Context 52
15.4 Pipeline Workflow and Computational Infrastructure 52
15.5 Performance and Comparative Benchmarks 52
15.6 Strengths, Limitations, and Special Features 52
16 DeepSomatic 53
16.1 Overview and Core Innovation 53
16.2 Model Architecture and Input Design 53
16.3 Training Data and Benchmarking Context 53
16.4 Pipeline Workflow and Computational Infrastructure 54
16.5 Performance and Comparative Benchmarks 54
16.6 Strengths, Limitations, and Special Features 54
17 SVision 55
17.1 Overview and Core Innovation 55
17.2 Model Architecture and Visual Encoding Strategy 55
17.3 Training Data and Benchmarking Context 55

17.4 Pipeline Workflow and Computational Infrastructure 56
17.5 Performance and Comparative Benchmarks 56
17.6 Strengths, Limitations, and Special Features 56
18 Final Comparative Analysis 57
18.1 Critical Discussion 58
19 Conclusion 59

3 Distributed Computing for Genomic Variant Calling 61
1 Introduction 61
2 The Genomic Data Challenge 61
2.1 Data Volume 61
2.2 Data Sensitivity 61
2.3 The Fragmentation Problem 62
3 Distributed Computing Foundations 62
3.1 Definition and Core Concepts 62
3.2 Parallel vs. Distributed Computing 62
3.3 Types of Distributed Computing 63
3.4 Key Frameworks 65
4 Federated Learning 65
4.1 Definition and Motivation 65
4.2 How Federated Learning Works 65
4.3 Types of Federated Learning 66
4.4 Online Federated Learning Frameworks in Genomics 67
4.5 Challenges in Federated Learning 69
5 Privacy and Security in Federated Genomics 70
5.1 Why Genomic Privacy Matters 70
5.2 Privacy Techniques Used With Federated Learning 71
6 Conclusion 72

4 Implementation of the Proposed Pipeline 74
1 Introduction 74
2 System Overview 74
3 Dataset and Ground Truth 75
4 Label Scheme and Class Design 76
5 Data Preparation Pipeline 77
5.1 Manifest 77
5.2 Pileup Tensors (CNN Input) 77
5.3 Read Windows (Transformer Input) 78
5.4 Modality Alignment 79
6 Model Architecture 79
6.1 CNN Encoder (PileupCNN) 79
6.2 Transformer Encoder (DNASequenceTransformer) 81
6.3 Gated Fusion 84
6.4 Classifier and Auxiliary Heads 85
7 Training Methodology 85
8 Experimental Evaluation 86
8.1 Test Setup 86
8.2 Evaluation Protocol 87
8.3 Results 88

8.4 Discussion 88
8.5 Future Work 89
9 Proposed Federated Extension 90
9.1 Motivation 90
9.2 Federated Training Procedure 91
9.3 Server-Side Weight Aggregation 91
9.4 Adaptation of the Loss Function 92
9.5 Federated Inference for Genomic Variant Detection 93
10 Conclusion 94

References **95**