Joint Biotechnology Master Program

Palestine Polytechnic University

Deanship of Graduate Studies and Scientific Research

Bethlehem University

Faculty of Applied Sciences, Technology & Engineering

SpliceLoc-ML: A Location-Stratified Machine Learning Model for Non-Canonical Splice Variant Pathogenicity Prediction and Classification, with Application to Long QT Syndrome

By

Wedad Jawdi Yaqoub Murrar

In Partial Fulfillment of the Requirements for the Degree

Master of Science

June, 2026

The undersigned hereby certify that they have read and recommend to the Faculty of Scientific Research and Higher Studies at the Palestine Polytechnic University and the Faculty of Applied Sciences, Technology and Engineering at Bethlehem University for acceptance a thesis entitled:

SpliceLoc-ML: A Location-Stratified Machine Learning Model for Non-Canonical Splice Variant Pathogenicity Prediction and Classification, with Application to Long QT Syndrome

By: Wedad Jawdi Yaqoub Murrar

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Biotechnology.

Graduate Advisory Committee:

Committee Member (Student's Supervisor)	Date
Dr. Yaqoub Ashhab, Palestine Polytechnic University	
Committee Member (Internal Examiner)	Date
Dr. Hashem Tamimi, Palestine Polytechnic University	
Committee Member (External Examiner)	Date
Dr. Fouad Zahdeh	

Approved for the Faculties

Dean of Graduate Studies and Scientific Research	Dean of Graduate Studies and Scientific Research
Palestine Polytechnic University	Bethlehem University
Date	Date

Abstract

Non-canonical splice variants that occur outside the ±1/±2 GT-AG canonical splice site positions, represent a clinically important but computationally neglected class of genomic variation. Although splicing-disrupting variants account for approximately one-third of all disease-causing variants, the ACMG/AMP classification framework offers only weak computational evidence criteria for this class. Consequently, a large portion of pathogenic variants in these regions are excluded completely from analysis pipelines or in some cases classified as VUS, leaving patients undiagnosed despite carrying a disease-causing mutation that could be identified.

To address this gap, this study develops SpliceLoc-ML, a location-stratified ensemble meta-classifier that predicts the pathogenicity of non-canonical splice variants and maps its output directly onto the ACMG/AMP five-tier classification. Three separate Random Forest models were trained, for exonic, near-splice (3–20 bp), and deep intronic (>20 bp) variants on 27,994 variants: 6,828 experimentally validated pathogenic variants from SpliceVarDB and 21,166 benign controls from gnomAD v4.1, integrating scores from SpliceAI, Pangolin, MaxEntScan, and CADD-PHRED with the GC content and polypyrimidine tract density. Thresholds were anchored to the Tavtigian Bayesian posterior probability framework with the lower VUS boundary derived from the training data using Youden's J statistic.

All three models exceeded pre-defined performance thresholds, with AUROC values of 0.969, 0.975, and 0.936 and MCC values of 0.837, 0.853, and 0.811 for exonic, near-splice, and deep intronic variants respectively, outperforming every individual tool evaluated. SpliceLoc-ML was then applied to 4,349 non-canonical splice variants from Long QT Syndrome genes. The number of VUS decreased by 56.6% (from 489 in ClinVar to 212), and the model placed 54 variants in the Likely Pathogenic or Pathogenic tiers. These are computational assignments and need functional and clinical confirmation.

These results establish SpliceLoc-ML as a reliable framework that connects computational splice variant prediction directly to clinical variant reporting.

Keywords: Non-canonical splice variants; Splicing; Variant pathogenicity prediction; Machine learning; Long QT Syndrome

الملخص

تُمثّل المتغيرات الجينية التي تقع خارج المواقع الثابتة لعملية الربط, أي خارج الموقعين GT-AG 2±/1± فئةً من المتغيرات ذات أهمية سريرية واضحة، إلا أن الأدوات الحسابية التنبؤية المتاحة لم تُولِها الاهتمام الكافي حتى الآن. وعلى الرغم من أن المتغيرات التي تُعطّل عملية الربط تُسبّب ما يقارب ثلث حالات الأمراض الوراثية، يفتقر إطار التصنيف المعتمد دولياً ACMG/AMP إلى معايير حسابية كافية للتعامل مع هذه الفئة تحديداً، مما يؤدي إلى عدم تصنيف معظمها أو التصنَّيف كمتغير ذو أهمية سريرية غير مؤكدة، وهو تصنيف لا يُقدّم للطبيب أي قرار علاجي واضح.

انطلاقاً من هذه المشكلة، طوّرت هذه الدراسة نموذجSpliceLoc-ML، وهو نموذج تعلم آلي يُصنّف المتغيرات وفقاً لموقعها الجينومي، ويتنبأ بمدى تسبّبها في المرض، ثم يُخرج نتيجته مباشرةً ضمن منظومة التصنيف الخماسية المعتمدة ACMG/AMP .بُني النموذج على ثلاث خوارزميات مستقلة من نوع الغابة العشوائية Random forest، كل منها مدرَّب على فئة موقعية مختلفة: المتغيرات داخل الإكسون، والمتغيرات القريبة من مواقع الربط (3–20 نيوكليوتيد)، والمتغيرات البعيدة عنها داخل الإنترون (أكثر من 20 نيوكليوتيد). استُخدم لتدريب النماذج 27,994 متغيراً: 6,828 متغيراً مسبباً للمرض مؤكداً بالتجربة المخبرية من قاعدة SpliceVarDB، و21,166 متغيراً حميداً من قاعدةgnomAD . اعتمد النموذج على مخرجات أربعة أدوات حسابية متكاملة هي SpliceAI, Pangolin, MaxEntScan, CADD-PHRE ، مضافاً إليها خاصيتان تصفان تسلسل الحمض النووي المحلي للمتغير. وقد حُدّدت عتبات التصنيف استناداً إلى إطار الاحتمالية البايزية Bayesian approach الذي اقترحهTavtigian ، مع استنباط الحد الفاصل لفئة VUS من بيانات التدريب باستخدام إحصائيةJ Youden's .

أثبتت النماذج الثلاثة أداءً يتجاوز جميع الحدود الدنيا المحددة مسبقاً، إذ بلغت قيم مساحة تحت المنحنى 0.969 : للمتغيرات الإكسونية، و0.975 للمتغيرات القريبة من مواقع الربط، و0.936 للمتغيرات العميقة في الإنترون، فيما بلغت قيم معامل ماثيوز للارتباط: 0.837 و0.853 و0.811 على التوالي، متفوقةً على كل أداة منفردة خضعت للمقارنة. وحين طُبّق النموذج على 4,349 متغيراً غير قانوني الموقع في جينات مرتبطة بمتلازمة QT الطويلة، انخفض عدد متغيرات VUS بنسبة 56.6% (من 489 إلى 212 متغيراً)، وصنّف النموذج حسابياً 54 متغيراً ضمن فئتَي «مُرجَّح تسبّبها في المرض» و«مسبّبة للمرض»، وهي تصنيفات حسابية تحتاج إلى تأكيد تجريبي وسريري قبل اعتمادها.

تُقدّم هذه النتائج مجتمعةً نموذجَ SpliceLoc-ML بوصفه أداةً موثوقة تسدّ فجوةً حقيقية بين ما تستطيع الحوسبة التنبؤ به وما يحتاجه المختبر السريري فعلياً في تصنيف المتغيرات الجينية.

الكلمات المفتاحية: متغيرات الربط غير القانونية؛ عملية الربط؛ التنبؤ بإمراضية المتغيرات الجينية؛ التعلم الآلي؛ متلازمة QT الطويلة

Declaration

I declare that the master thesis entitled "SpliceLoc-ML: A Location-Stratified Machine Learning Model for Non-Canonical Splice Variant Pathogenicity Prediction and Classification, with Application to Long QT Syndrome" is my own original work, and thereby certify that unless stated, all work contained within this thesis is my own independent research and has not been submitted for the award of any other degree at any institution, except where due acknowledgment is made in the text.

In the preparation of this thesis, the AI assistant (claude.ai) was used in two specific areas, to assist in the development of Python code for data processing, model training, and result visualization, and to improve the clarity and phrasing of certain written sections. All codes were tested and verified, and all scientific interpretations, analytical decisions, and conclusions presented in this thesis are entirely my own.

Signature: *Wedad Jawdi Yaqoub Murrar*

Date: _____

Institution: *Palestine Polytechnic University*

Copyright © "Wedad Jawdi Yaqoub Murrar", 2026

All rights reserved.

Statement of Permission to Use

In presenting this thesis in partial fulfillment of the requirements for the joint master's degree in biotechnology at Palestine Polytechnic University and Bethlehem University, I agree that the library shall make it available to borrowers under rules of the library. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of the source is made.

Permission for extensive quotation from, reproduction, or publication of this thesis may be granted by my main supervisor, or in his absence, by the Dean of Higher Studies when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Signature: *Wedad Jawdi Yaqoub Murrar*

Date: _____

Dedication

I dedicate this work to myself, for every moment I chose not to give up,

To my family, whose support and prayers guided me at every step,

To my husband, for his patience and for standing by my side throughout it all,

To my son, Muhammad, who arrived in the middle of this journey and became, without even knowing it, my greatest reason to finish it,

And to every patient still waiting for an answer, and every family carrying the weight of uncertainty, this work was written with you in mind.

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Dr. Yaqoub Ashhab, for his invaluable guidance, patience, and for always pushing me to think more carefully. Your feedback shaped this work in more ways than I can count.

I am also incredibly grateful to the staff of the Biotechnology Research Center at Palestine Polytechnic University, Dr. Robin Abu Ghazaleh, Dr. Rami Arafeh, Asma Tamimi, Ayman Al Eadieh, Hassan Taradeh, and Zaid Taradeh, thank you for the support and for creating an environment where this kind of work is possible.

To my thesis committee, Dr. Hashem Tamimi and Dr. Fouad Zahdeh, thank you for taking the time to evaluate this work and for the valuable input you provide.

To my colleagues at work, Dr. Yaseen Qawasmi and Khaled Sadeh, thank you for your encouragement and for understanding on the days when this thesis took priority over everything else.

Contents

Abstract ..... **III**

الملخص ..... **III**

Declaration ..... **V**

Statement of Permission to Use ..... **VI**

Dedication ..... **VII**

Acknowledgement ..... **VIII**

List of Tables: ..... **XII**

Table of abbreviations ..... **XIII**

Chapter 1| Introduction **1**

1.1.The Diagnostic Gap in Rare Genetic Diseases 1

1.2.Mechanism and Biological Significance of RNA Splicing 5

1.3.Definition and Clinical Significance of Non-Canonical Splice Site Variants 7

1.4.Computational Prediction of Splicing Effects 8

1.5.Approaches to Variant Classification and the Challenge of Non-Canonical Variants 11

1.6.Machine Learning Approaches in Variant Pathogenicity Prediction 13

1.7.The Problem of Variants of Uncertain Significance 14

1.8.Long QT Syndrome as a Clinical Use Case 15

Chapter 2| Motivation and Objectives **18**

Chapter 3| Research Methodology **21**

3.1.Study Design 21

3.2.Data Sources 22

3.2.1.Positive controls: SpliceVarDB **22**

3.2.2.Negative controls: gnomAD v4.1 ..... **23**

3.3.Variant Annotation and Feature Extraction 24

3.4.Imputation of missing scores 25

3.5.Class balancing 26

3.6.Model Training and Cutoff Selection 27

3.7.The Five-Tier ACMG Classification 27

3.8.Model Evaluation 28

3.9.Feature importance 28

3.10.Comparison with SpliceVarDB Experimentally Validated Variants 29

3.11.Software and Reproducibility 29

3.12.Non-Canonical Splice Variants in LQTS Genes 30

Chapter 4| Results **31**

Part A: Model Building and Validation 31

4.1.Dataset Composition and Class Balancing 31

4.2.Model Performance on the Held-out Test Dataset 32

4.3.Receiver Operating Characteristic Curves 35

4.4.Score Distribution by True Label 36

4.5.Feature Importance 39

4.6.The five-Tier ACMG Classification Thresholds....41
4.7.Reclassification of the Full Dataset....41
4.8.Comparison with SpliceVarDB Experimentally Validated Variants....43
Part B| Case study – Long QT Syndrome....44
4.9.Reclassification of LQTS Variants....45
4.10.The five-Tier Classification Results per Gene....46
4.11.Variant of Uncertain Significance Analysis....47
4.12.Feature Score Distributions by ACMG classification....49
Chapter 5| Discussion....51
5.1.Overview....51
5.2.Dataset Design and Class Imbalance Management....52
5.3.Model Performance and Generalization....52
5.3.1.Absolute Performance....52
5.3.2.Train-Test Methodology and Absence of Overfitting....53
5.4.Comparison with Individual Splicing Tools....54
5.4.1. SpliceAI....54
5.4.2. Pangolin....55
5.4.3. MaxEntScan....55
5.5.Feature Importance and Biological Interpretability....56
5.6.ACMG Five-Tier Classification Framework....57
5.6.1.Threshold Derivation....57
5.6.2.Full Dataset Reclassification....57
5.7.The SpliceVarDB Laboratory Validation methods and Model Calibration....57
5.8.Clinical Application: Long QT Syndrome....59
5.8.1.VUS Resolution....59
5.8.2.Gene-Specific Findings....59
5.8.3.Features and Model Validation....60
5.9.Comparison with Published Frameworks....60
5.10. Limitations....62
5.11. Future work....63
Chapter 6| Conclusion....65
Chapter 7| Appendices....66
Appendix A| Software Packages and Libraries....66
Appendix B| Model Input Features....67
Appendix C| Hyperparameters....67
Chapter 8| References….....68

Table of figures

Figure 1| The multi-step analytical pipeline used to identify and classify clinically relevant variants.....3
Figure 2|Splicing mechanism and regulation of pre-mRNA. 6
Figure 3| Canonical and non-canonical splice site variant positions.....7
Figure 4|Five-tier ACMG/AMP variant classification probability scale.....12
Figure 5| The QT interval in normal vs. Long QT Syndrome.. 16
Figure 6|Analytical workflow for the ensemble machine learning model for non-canonical splice variant pathogenicity prediction.....22
Figure 7| Filtering pipeline for positive and negative control variant datasets. 24
Figure 8| Performance Metrics of Models by Genomic Category. 35
Figure 9|ROC curves for the three Random Forest splicing models that were trained specifically on the test set (30%).....36
Figure 10|Violin plots of Model-Predicted Pathogenic Probability Scores by True Label of the Three Location Categories.. 37
Figure 11|SpliceLoc-ML vs individual splicing tools in exonic variants.....38
Figure 12| SpliceLoc-ML vs individual splicing tools in near-splice variants (3–20 bp). 38
Figure 13| SpliceLoc-ML vs individual splicing tools in deep intronic variants (>20 bp).....39
Figure 14| Gini-based feature importance rankings for the three category-specific Random Forest models.....41
Figure 15| Distribution of the five-tier ACMG classification for all variants (n = 27,994 variants) per each category.....42
Figure 16|Distribution of the Five-tier ACMG classification in the total 27,994 variants in the full dataset. 43
Figure 17|Model detection rate (ratio of laboratory-confirmed pathogenic variants classified as P/LP) by laboratory validation method.....44
Figure 18| Numbers of five-tier ACMG classification per LQTS gene shown as a bubble matrix. 47
Figure 19|In-depth analysis of the 212 VUS variants.....49
Figure 20|Violin plots of four key feature scores stratified by ACMG five-tier classification. 51

List of Tables:

Table 1|Summary of genomic variant categories identified by Next-Generation Sequencing, their detection by Whole-Exome Sequencing (WES) and Whole-Genome Sequencing (WGS), and their classification status in standard clinical pipelines. 1
Table 2| Distribution of disease-causing mutation types in the human genome. 5
Table 3|Five-tier ACMG/AMP classification thresholds. 28
Table 4| Dataset composition before and after class balancing. 31
Table 5|The balanced dataset after downsampling and the split (70 % /30 %, seed=42). 31
Table 6|Multi-threshold Diagnostic Performance Analysis by Genomic Category. 32
Table 7| Model performance on test set (30%).. 34
Table 8| Performance comparison of SpliceLoc-ML and individual splicing tools across genomic location categories. 39
Table 9| Five-tier ACMG classification thresholds for each category.. 41
Table 10|Model detection rate by laboratory validation method 44
Table 11| Summary of before and after reclassification of 4,349 LQTS non-canonical splice variants. 45
Table 12| Reclassified results with the five-tier ACMG reclassification by gene.. 46

Table of abbreviations

VUS	Variants of Uncertain Significance
HGMD	Human Gene Mutation Database
ESEs	Exonic Splicing Enhancers
ISEs	Intronic Splicing Enhancers
ESSs	Exonic Splicing Silencers
ISSs	Intronic Splicing Silencers
hnRNPs	heterogeneous nuclear Ribonucleoproteins
PPT	Polypyrimidine Tract
A	Branch Point Adenosine
NMD	Nonsense-Mediated Decay
PTC	Premature Termination Codon
PE	Cryptic Pseudoexon
ACMG	American College of Medical Genetics and Genomics
AMP	Association for Molecular Pathology
PVS1	Pathogenic Very Strong
PM2	Moderate Evidence of Pathogenicity
PP3	Supporting Evidence of Pathogenicity
PS3	Pathogenic Strong
BS3	Benign Strong
PWMs	Position Weight Matrices
MaxEntScan	Maximum Entropy Scan
CADD	Combined Annotation-Dependent Depletion
PDIVAS	Pathogenicity predictor for Deep-Intronic Variants causing Aberrant Splicing
gnomAD	Genome Aggregation Database
AUROC	Area Under the Receiver Operating Characteristic curve
PPV	Positive Predictive Values
NPV	Negative Predictive Values
MCC	Matthews Correlation Coefficient
LQTS	Long QT Syndrome
GRCh38	Genome Reference Consortium Human Build 38
SNVs	Single Nucleotide Variants

RNA-Seq	RNA sequencing
MFASS	Multiplexed Functional Assay of Splicing using Sort-seq
MaPSy	Massively Parallel Splicing Assay
VCF	Variant Call Format
AF	Allele Frequency
PHRED	Phred-like score
Ensembl VEP	Ensembl Variant Effect Predictor
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ROC	Receiver Operating Characteristic
AUC	Area Under the curve
RF	Random Forest

Chapter 1| Introduction

1.1.The Diagnostic Gap in Rare Genetic Diseases

Each rare genetic disease is uncommon on its own, but when considered together, approximately 6,172 known rare diseases categorized in the Orphanet database collectively affect around 263–446 million people worldwide, making them a significant global health problem. They are one of the main causes of death in young children and lifelong disability (Nguengang Wakap et al., 2020). These diseases put the families in difficult situations, patients frequently spend years consulting different specialists before a diagnosis is reached, so for many families the precise molecular diagnosis is not simply a solution, it determines whether a treatment is available and helps them to make decisions about having children and ends years of uncertainty (Zurynski et al., 2017).

The widespread availability of next-generation sequencing has changed the way of diagnosing the diseases. After the first clinical application of whole-exome sequencing in 2009, it became a routine clinical test instead of only a research tool because it can identify variants in all protein-coding regions of the human genome in a single test (Ng et al., 2010). Then whole-genome sequencing has followed, extending coverage to include intronic and regulatory regions. The variant categories identified by these technologies differ substantially in how well they are detected and classified in clinical practice, as summarized in Table 1.

Table 1|Summary of genomic variant categories identified by Next-Generation Sequencing, their detection by Whole-Exome Sequencing (WES) and Whole-Genome Sequencing (WGS), and their classification status in standard clinical pipelines. ✓ = detected, ✗ = not detected, Limited = WES capture kits vary, near-splice variants within ±10–50 bp of exons may be captured depending on the kit used.

Variant Category	Detected by WES	Detected by WGS	Standard Classification Criteria	Reference
Coding (missense, nonsense, frameshift)	✓	✓	Well established criteria (PVS1, PS1, PM1–PM5, PP2–PP3)	(Richards et al., 2015)
Canonical splice site (±1, ±2)	✓	✓	Well established Classified under PVS1	(Richards et al., 2015)
Non-canonical near splice	Limited	✓	Limited ACMG criteria Weak evidence only (PP3, PM2), frequently classified as VUS or excluded from analysis pipelines	(Brnich et al., 2020)
Non-canonical Deep intronic	✗	✓	Largely excluded from analysis pipelines or unclassified	(Barbosa et al., 2023)

Despite the technological advances, the comprehensive genomic sequencing still fails to identify a disease-causative variant in more than 50% of patients with suspected Mendelian disease (Wojcik et al., 2023). This diagnostic gap may be caused by three major interconnected problems. First, technological limitations in which certain genomic regions particularly non-coding and deep intronic regions remain difficult to sequence or analyze reliably (Wojcik et al., 2023). Second, incomplete gene-disease association knowledge means that many identified variants cannot be linked to a known disease because the responsible gene has not yet been characterized (Canavati et al., 2024). Third, even when a variant is detected in a known disease gene, proper variant classification remains a major problem, the typical human genome contains three to four million variants relative to the reference sequence, and determining which one of them causes disease from the millions that are entirely harmless is a fundamental challenge in clinical genomics and it is the central focus of this thesis (Claussnitzer et al., 2020).

The genomic variant that has been identified by the next-generation sequencing has to go through a structured multi-step pipeline before being clinically classified. The first step is filtration, in which common variants present at high frequency in population databases such as gnomAD are removed, as a variant causing severe Mendelian disease cannot persist at high frequency in the general population due to natural selection. The second step is prioritization in which the remaining variants are ranked according to predicted functional impact, inheritance pattern, and compatibility with the patient's phenotype. The third step is curation, in which each prioritized variant is reviewed against existing literature, functional studies, and curated databases such as ClinVar and OMIM to collect all available evidence. The fourth step is evidence collection, in which computational predictions, co-segregation data, population data, and if available functional assay results are systematically assembled. These four steps together lead to the final step of variant classification, in which all collected evidence is integrated to categorize the variant into one of five standardized tiers defined by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP), Pathogenic, Likely Pathogenic, Variant of Uncertain Significance (VUS), Likely Benign, or Benign (Richards et al., 2015). This tiered classification system provides the framework in which most computational tools and ensemble models operate.



Figure 1| The multi-step analytical pipeline used to identify and classify clinically relevant variants. Each step in this pipeline introduces a decision point where non-coding variants are more likely to be filtered out, deprioritized, or left without sufficient evidence for a definitive classification.

The major reason for the failure rate (more than 50%) is a significant systematic and consistent bias towards the classification and prioritization of clinical genomic variants within this pipeline. The focus on missense variants and protein-coding regions rather than non-coding variants. This bias started in the historical development as the first methods of sequencing focused only on the exome, and the earlier computational tools for predicting variant pathogenicity were designed exclusively for amino acid-changing mutations (Danis et al., 2021). As a result, variants in introns, regulatory regions, and splice positions outside the canonical ±1 and ±2 splicing sites have received less attention in both tool development and clinical practice because their functional consequences are harder to predict (Blakes et al., 2022). This bias is reflected in major variant databases such as ClinVar, where the proportion of non-coding variants with a definitive pathogenic or benign classification is significantly less frequent than coding variants. The consequence is that a large class of functionally important variants, particularly variants that affect RNA splicing through non-canonical mechanisms, continues to be insufficiently captured in diagnosis, even as whole-genome sequencing makes them technically detectable (Lord et al., 2019). Consequently, a large portion of pathogenic variants in these regions are excluded completely from analysis pipelines or in some cases classified as VUS, leaving patients undiagnosed despite carrying a disease-causing mutation that could be identified (Ellingford et al., 2022; Marwaha et al., 2022).

In order to understand why non-coding splicing variants are disproportionately affected by this classification gap, it may be useful to examine how much they actually contribute to human inherited disease, a contribution that has long been underestimated. Variants that cause human inherited diseases are categorized into several molecular classes, each one affecting gene function in distinct mechanism. According to the most recent release of Human Gene Mutation Database (HGMD Professional release 2024.4, n = 510,804 entries), the most common mutations

associated with diseases are the missense and nonsense mutations about 62.5%, followed by the small deletions 13.1%, then splicing mutations 8.3%, small insertions 5.7%, gross deletions 5.7%, gross insertions and duplications 1.5%, regulatory variants 1.4%, small indels 1.1%, complex rearrangements 0.6%, and repeat variations 0.1% (Cooper et al., 1998). While the missense and nonsense mutations collectively represent the dominant category, the contribution of splicing-disrupting variants to human disease has been underestimated, the 8.3% in HGMD only include the canonical splice site mutations that disrupt the invariant GT/AG dinucleotides at the ±1/±2 positions flanking each intron, it does not account for the large number of exonic and deep intronic variants that alter splicing through non-canonical mechanisms. Computational analyses have demonstrated that approximately 25% of exonic disease-associated disrupt exonic splicing enhancers or silencers, contributing to aberrant splicing independently of canonical splice site changes (Sterne-Weiler et al., 2011). Extending the scope of analysis, splicing-disrupting variants of all classes are estimated to account for approximately one-third of all disease-causing alleles (Clarke, 2024; Lim et al., 2011). This percent increases further when deep intronic variants that activate cryptic splice sites are considered, which is a class that remains largely invisible to standard clinical sequencing pipelines that focused on canonical splice site regions.

While when includes the synonymous coding variants, missense variants, and intronic variants at almost any position that alter the balance of splicing regulatory signals, not only the variants at canonical ±1 and ±2 splice sites, it is estimated that 50% of the disease-causing mutations exert their pathogenic effect by interfering with splicing (López-Bigas et al., 2005; Scotti & Swanson, 2016). The huge number of potential splicing disruption sites which exceeds the number of coding sequence, means that the intronic genome contains a high number of pathogenic variants that have hardly been studied in many diseases.

Taken together, these observations establish pre-mRNA splicing as one of the most frequently disrupted molecular processes in human genetic disease, and motivate the need for computational tools capable of identifying and classifying the all categories of splice-altering variants.

Table 2| Distribution of disease-causing mutation types in the human genome. Source: HGMD Professional release 2024.4 with total mutations: 510,804. Highlighted row: variants with a primary or substantial splicing mechanism.

Mutations type	Number of entries in HGMD	Percentage
Missense/nonsense	319,190	62.5 %
Splicing	42,559	8.3 %
Regulatory	6,953	1.4 %
Small deletions	67,065	13.1 %
Small insertions	29,310	5.7 %
Small indels	5,416	1.1%
Repeat variations	702	0.1 %
Gross insertions/duplications	7,580	1.5 %
Complex rearrangements	2,950	0.6%
Gross deletions	29,079	5.7 %
Total	**510,804**	100.0 %

1.2. Mechanism and Biological Significance of RNA Splicing

The RNA splicing process is complex and easily affected by sequence disruption, this is important to consider when studying the pathogenicity of non-coding variants. When a gene is transcribed from DNA into pre-messenger RNA, the non-coding introns are cut out and the coding exons are joined together to make a mature mRNA that can be translated into protein. This process is done by the spliceosome, which is a large and highly dynamic ribonucleoprotein complex assembled from five small nuclear RNAs (U1, U2, U4, U5, and U6) and more than 150 associated proteins (Vaz-Drago et al., 2017). The spliceosome recognizes each intron through a combination of short consensus sequences and regulatory sites distributed across the pre-mRNA, which include the 5' splice donor site located at the exon-intron junction with a highly conserved GT dinucleotide, the 3' splice acceptor site at the intron-exon junction with AG dinucleotide, and the branch point sequence that is usually located 18–40 nucleotides upstream of the acceptor and contains an

invariant adenosine residue, and finally the polypyrimidine tract that contains a sequence rich in pyrimidines between the branch point and the acceptor and allows the binding of U2AF (Cartegni et al., 2002; Yeo & Burge, 2003). These core elements alone are necessary but not sufficient, they are associated with series of more complex regulatory regions consisting of the presence of exonic and intronic splicing enhancers (ESEs and ISEs) and silencers (ESSs and ISSs), which are recognized by specific protein families called SR proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs) (Barash et al., 2010).

This system must have high accuracy because a single nucleotide change at the wrong position can shift the reading frame and introduce a premature stop codon, or activate a cryptic splice site that is an inactive sequence resembling the real one in composition but in normal situations is suppressed. When a pathogenic variant disrupts normal splicing, several outcomes are possible: the affected exon skipped entirely, or an intronic sequence incorporated into the mRNA as a pseudoexon, or an existing exon extended, or the mRNA degraded by nonsense-mediated decay (NMD) before it can be translated (Scotti & Swanson, 2016). Each of these outcomes would be enough to destroy protein activity, like in frameshift or nonsense mutation in the coding sequence, now none would be detected by an analysis design for exonic variants only.

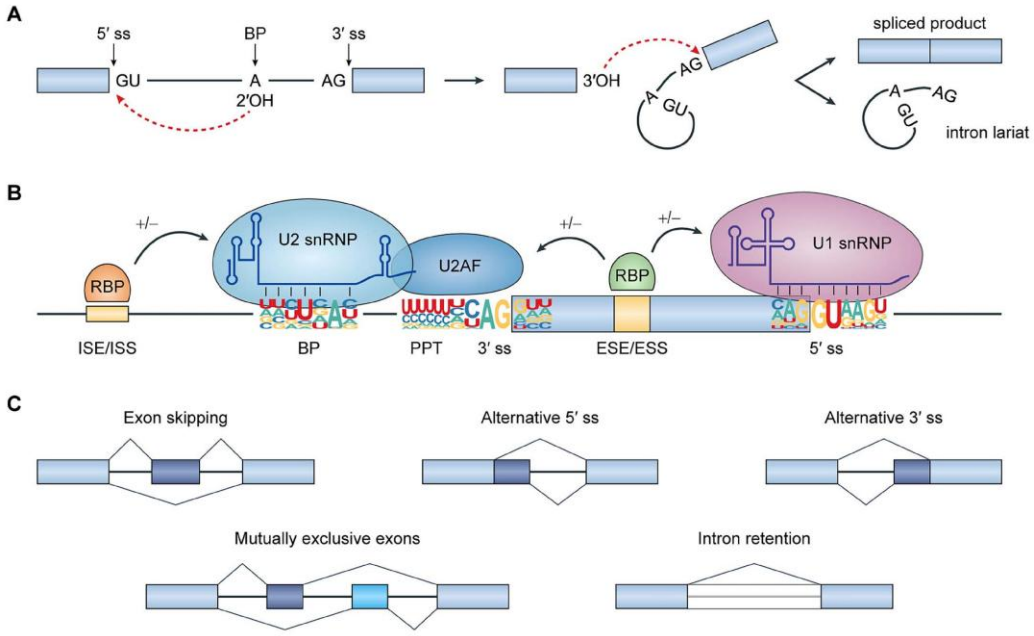


Figure 2|Splicing mechanism and regulation of pre-mRNA. (A) The two-step chemical pathway of splicing, which results in the removal of the intron in a lariat form and ligation of exons. (B) Assembly of the spliceosome and its regulatory network, core splicing signals (5' donor splice site, 3' acceptor splice site, Branch point, and

PPT) are recognized by core factors (U1 snRNP, U2 snRNP, and U2AF), while auxiliary RNA-binding proteins (RBPs) bind to exonic/intronic splicing enhancers or silencers (ESE/ISE/ESS/ISS). (C) The five outcomes of alternative splicing that increase the transcriptome complexity: exon skipping, alternative 5' or 3' splice sites, mutually exclusive exons, and intron retention. Reproduced from (Wang et al., 2023), Briefings in Bioinformatics, 24(5), bbad284. © Oxford University Press.

1.3.Definition and Clinical Significance of Non-Canonical Splice Site Variants

The term "canonical splice site variant" refers to substitutions at the invariant ±1 and ±2 (GT/AG) positions of the splice donor and acceptor sites. These variants are classified as pathogenic or likely pathogenic because disruption of the invariant dinucleotides affects the recognition signal for the spliceosome, and their functional consequence, like intron retention or exon skipping is highly predictable (Richards et al., 2015). Non-canonical variations include all other variations that occur outside the splice donor and acceptor sites, this is a broad and heterogeneous category that includes near-splice variants in the extended splice site region, variants in exons affecting splicing motifs without changing the codons, and deep intronic variants located more than 20 nucleotides from any splice site (Jung et al., 2021).

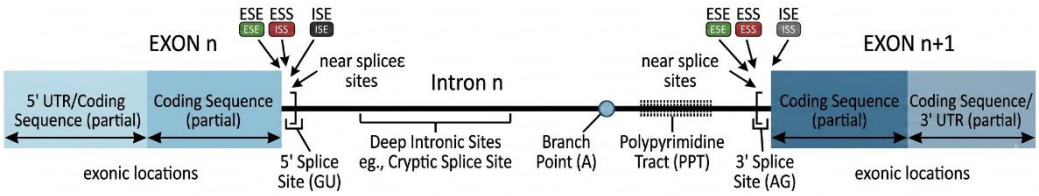


Figure 3| Canonical and non-canonical splice site variant positions. Illustration of an intron-exon boundary highlighting the 5' splice site (GT), 3' splice site (AG), branch point (A), polypyrimidine tract (PPT). it maps the location of exonic and intronic splicing enhancers and silencers (ESE, ESS, ISE, ISS) relative to near-splice and deep intronic (cryptic) sites. Brackets indicate partial untranslated regions (UTRs) and coding sequences within the flanking exons.

The clinical significance of non-canonical variants has been shown in many types of Mendelian diseases. For example, in the study of cystic fibrosis, a deep intronic variant in the *CFTR* gene (c.3849+10kbC>T) that is located more than 10,000 nucleotides from the closest exon, this mutation activates a cryptic splice site and creates a pseudoexon that causes loss of function of the CFTR protein, leading to a mild form of cystic fibrosis (Chiba-Falek et al., 1998). This variant was discovered in the early 1990s and shows that the relevant sequence space for disease-causing mutations was not limited to the exon-intron junction. Another example, in *DMD*, multiple deep intronic variants have been shown to create pseudoexons leading to disrupt the reading frame of the gene product dystrophin, causing Duchenne or Becker muscular dystrophy in patients originally thought to lack any causative

mutation in their genomes (Okubo et al., 2022). For inherited retinal diseases, deep intronic variants in *ABCA4*, *CEP290*, and other disease genes contribute significantly to heritability in patients whose diagnosis was unknown, cases where one pathogenic allele is identified but the second cannot be found in the coding sequence (Coppieters et al., 2010; Sangermano et al., 2019). In neurofibromatosis type I, intronic variants in *NF1* have been found to introduce premature stop codons that lead to produce pseudoexon activation (Pros et al., 2008). These examples highlight the fact that pathogenic non-canonical splice variants are not unique to a single disease, they represent a clinically important and systematically underdiagnosed class of pathogenic mutation in Mendelian disorders.

Despite these evidences, the non-canonical variants remain largely unclassified, whereas other are conservatively classified as variants of uncertain significance. In the ACMG/AMP variant classification guidelines that used in most clinical diagnostic laboratories around the world, the pathogenicity evidence can be applied to only canonical splice site variants, while much weaker evidence scores to non-canonical variants, for example, for canonical (±1, ±2) variants, the ACMG/AMP start at PVS1 (Pathogenic Very Strong), while for variants without established functional data or significant clinical history like non-canonical variants, they usually use PM2 for population rarity or PP3 for computational predictions, and neither of these alone, nor both combined, is sufficient to reach likely pathogenic classification (Richards et al., 2015). If there is functional evidence that is applied for the non-canonical variant, so PS3 (Pathogenic Strong) or BS3 (Benign Strong) are applied if the assay proves the variant causes a splicing error or shows the variant does not affect splicing respectively. The effect is that any variant at position ±3 from the consensus sequence or deeper would get an evidence score of being supportive of pathogenicity only (even when being fully disruptive for splicing), but not enough for a clinically significant conclusion on its own. What is needed in order to fill this classification gap is a computational tool that can predict the splicing impact of non-canonical variants with sufficient accuracy.

1.4.Computational Prediction of Splicing Effects

The development of computational tools to predict the effect of sequence variants on RNA splicing has a history of more than two decades ago, reflects the broader evolution of machine learning and genomics. Early tools depended on

position weight matrices (PWMs) derived from known splice site sequences to score how well a given sequence matched the consensus donor or acceptor motif. An essential improvement over this approach is MaxEntScan that was published by Yeo and Burge, where a maximum entropy model was utilized to capture non-adjacent dependencies between nucleotides within the splice site, generating predictive scores that accurately correlated with the experimental measurements of splice site strength (Yeo & Burge, 2003). MaxEntScan remains a standard reference tool in clinical variant analysis due to its biophysical basis, makes its scores interpretable because it is based on log-odds ratios, which means a specific decrease in score directly quantifies the reduction in the probability of the spliceosome recognizing that sequence compared to the genomic background, in other words, a significant decrease in the MaxEnt score for the mutant allele compared to the reference allele indicates a weakening of the splice site.

For many years, the tools that were designed with similar approaches to MaxEntScan (score-based, short-range, motif-centered) were the best tools available. They lacked one major feature, which at the same time made them special, their inability to predict variants located outside of the canonical splicing motifs. This means that deep intronic variants, whose positions are far from any annotated splice site, were predicted similarly to neutral intronic mutations because they were outside of the sequence region analyzed by the algorithms.

In 2019, Jaganathan and colleagues developed SpliceAI, a splicing variant prediction tool that was built using deep residual convolutional neural network trained on pre-mRNA sequences from the human genome and their corresponding splicing patterns as measured by RNA-sequencing. It analyzes a 10,000-nucleotide sequence surrounding the target region, so it can analyze long-range sequence context to predict the impact on known splicing regions and also any potential new splicing regions that might be created. This tool showed significant improvement over earlier tools in the ability to predict the impact of deep intronic variants, it identified pathogenic pseudoexon-activating variants that had been invisible to all previous computational analysis tools. The SpliceAI delta scores which measure the gain or loss probability at donor and acceptor positions have been widely used in research and clinical analyses, and the tool has been incorporated into major variant interpretation pipelines

including those used by ClinVar submitters and clinical genetics laboratories (Jaganathan et al., 2019).

Another interesting tool called Pangolin was published by Zeng and Li in 2022 and extended the deep learning approach in a different direction. Pangolin incorporated RNA-sequencing data from multiple mammalian species and used tissue-specific gene expression data in order to make splicing predictions based on tissues instead of training exclusively on human splicing data. This is biologically important because the alternative splicing is highly tissue-specific, and a variant that disrupts splicing in one specific tissue type like the heart, may not be effective at all in another tissue type, such as lymphoblastoid cells (Zeng & Li, 2022).

The general pathogenicity predictors have been incorporated into splicing variant interpretation. An example of these tools is the Combined Annotation-Dependent Depletion (CADD) method, developed by Kircher and colleagues and subsequently extended to CADD-Splice, which integrates various genomic features, including conservation scores, regulatory annotations, and splicing predictions into a single deleteriousness score that was specifically designed to reflect the evolution of humans (Rentzsch et al., 2021). However, the usefulness of the conservation score is not in its ability to predict splice sites but as a measure of evolutionary constraint. If a mutation exists at a highly conserved residue among all vertebrates, then there is a higher probability that it will have some functional effect compared to a residue that is poorly conserved.

Although significant progress has been made in this area, one of the challenges faced by individual computational tools is that each capture only a portion of the relevant biological signal. The SpliceAI model shows superiority in predicting the long-range impact but has reduced sensitivity to detect variants inside splice sites motifs where MaxEntScan is more precise for these motifs (Smith & Kitzman, 2023). Pangolin provides additional knowledge related to tissue specificity that SpliceAI lacks, but its training on cross-species data means it may underweight human-specific splicing patterns (Zeng & Li, 2022). The CADD algorithm provides useful evolutionary information but it is not a splicing predictor. No single tool has been shown to dominate all others across all variant categories and genomic contexts, which motivates the development of ensemble approaches that integrate multiple sources of evidence (Danis et al., 2021; Moles-Fernández et al., 2021).

Several ensemble approaches have been developed to address the limitations of individual splicing prediction tools. ConSpliceML integrates SpliceAI, SQUIRLS, and a human splicing constraint score using a Random Forest framework and demonstrates improved specificity over individual tools (Cormier et al., 2022), but it was designed exclusively for deep intronic variants and cannot be applied across other genomic regions. SPiP integrates multiple splicing signals in a cascade pipeline that covers nearly all variant categories but shows reduced performance specifically for deep intronic variants (Leman et al., 2022). Introme combines predictions from several tools with gene features and achieves strong performance across variant categories, but produces a continuous pathogenicity score rather than a direct clinical classification, meaning its output must still be manually interpreted within the ACMG framework (Sullivan et al., 2023). The Pathogenicity predictor for Deep-Intronic Variants causing Aberrant Splicing (PDIVAS) was developed and represented a significant step in the identification of disease-causing mutations within non-coding regions. PDIVAS utilizes an ensemble machine-learning algorithm, a Random Forest model to classify variants as pathogenic or benign by integrating multiple splicing features and human splicing constraint metrics, but it is specifically designed to focus exclusively on deep-intronic variants (located at least 50 base pairs away from the nearest annotated splice site) and limiting its broader applicability across other genomic regions (Kurosawa et al., 2023). Critically, none of these approaches applies location-stratified training, the practice of building separate models calibrated to the biological realities of each genomic location category, and none produces an output directly mapped onto the ACMG five-tier posterior probability framework. This highlights the need for a more generalized ensemble approach.

1.5. Approaches to Variant Classification and the Challenge of Non-Canonical Variants

The transition from computational predictions to clinical variant classifications is primarily done by the ACMG/AMP classification system (Richards et al., 2015). This approach uses five classification tiers: Pathogenic, Likely Pathogenic, Variant of Uncertain Significance (VUS), Likely Benign, and Benign, each associated with specific evidence criteria that need to be met to reach each tier. The criteria are qualitative and categorical, a variant either meets a given criterion or it does not, and the tier is determined by combining the evidence weights across all applicable criteria.

One significant limitation of this approach is that it was primarily designed for coding variants and canonical splice site variants. The strongest pathogenicity criterion, PVS1 (Pathogenic Very Strong) is defined as a null variant in a gene where the loss of function is a known disease mechanism. It applies easily to frameshift, nonsense, and canonical splice site variants. In the case of non-canonical splice variants, the criteria provided by this approach are limited to PS3 (well-established functional studies showing a damaging effect) and the PP3 criterion (multiple lines of computational evidence supporting a deleterious effect). Neither of these two criteria alone can be used for a likely pathogenic assessment (Brnich et al., 2020). Therefore, many non-canonical splice variants that show a disrupted splicing in functional assays are classified as VUS simply because the model does not have a mechanism or algorithm to weight computational splicing evidence appropriately (Walker et al., 2023).

Tavtigian and colleagues set a more quantitatively robust approach for variant classification. They redefined the ACMG/AMP criteria based on a Bayesian paradigm (Tavtigian et al., 2018). Rather than treating evidence criteria as categorical checkboxes, they assign a likelihood ratio to each criterion that indicates the level of supportive evidence. By multiplying these values together with an initial probability, researchers calculate a posterior probability (final score of certainty). These Bayesian criteria have explicit posterior probability thresholds for each classification tier: Pathogenic ≥ 0.99, Likely Pathogenic ≥ 0.90, Likely Benign ≤ 0.10, and Benign ≤ 0.001. This probabilistic classification system allows computational prediction scores to be directly translated into a five-tiered classification based on the Bayesian criteria, assuming that the likelihood ratios for the predictions can be calibrated against known pathogenic and benign variants, providing a principled basis for translating a continuous model output probability into a five-tier clinical label (Tavtigian et al., 2018).

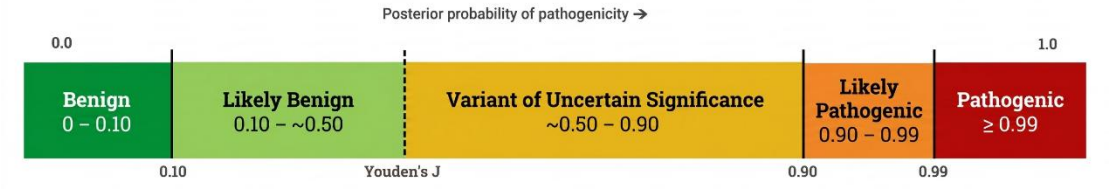


Figure 4|Five-tier ACMG/AMP variant classification probability scale. There are three thresholds values that are constant values published by Tavtigian et al. (2018). The lower limit of VUS (dashed line) is calculated based on locus-specific values using Youden's J index statistic (Youden, 1950). Figure created by the author.

The challenge of setting appropriate thresholds for computational tools used in splice variant classification has become a topic of interest for the ClinGen Sequence Variant Interpretation (SVI) working group. In the guidance published by Pejaver and colleagues in 2022, specific likelihood ratios have been set for PP3 and BP4 evidence for computational tools, enabling its implementation into the Tavtigian Bayesian model (Pejaver et al., 2022). This guidance was mainly developed to focus on missense variants and general pathogenicity predictors. The calibration of splicing-specific scores based on control data from well-known pathogenic and benign non-canonical splice variants remains an open problem, one that is directly addressed by the work presented in this thesis.

1.6.Machine Learning Approaches in Variant Pathogenicity Prediction

There has been an increase in utilizing machine learning in variant pathogenicity prediction with the increased availability of genomic variant databases and the progress in algorithms that can handle biological data. Distinguishing pathogenic from benign variants is a binary prediction problem at the computational level, distinct from the five-tier clinical classification process, and it has been approached with a range of algorithms from logistic regression and support vector machines in the early literature to gradient-boosted trees and deep neural networks in more recent work (Grimm et al., 2015).

The Random Forest classifier algorithm, introduced by Breiman in 2001, is one of the most widely used methods for the prediction of variant pathogenicity. Its suitability for this prediction is due to a number of practical considerations (Breiman, 2001). These algorithms show good tolerance and robust to variance in feature scales, an important feature since combining scores from different sources, each characterized by a different range of values (e.g., CADD [0–99], SpliceAI [0–1], MaxEntScan [-20 to +12]). Also, they handle class imbalance through cost-sensitive weighting, which is essential when pathogenic variants are significantly less than benign controls in the training dataset, they produce probability predictions instead of hard classifications making it possible to calibrate them against clinical thresholds. In addition, their measures of feature importance make it easier to understand which features have the most impact on the prediction (Chicco & Jurman, 2020).

One of the major issues in all machine learning methods for variant classification is the composition and quality of the training dataset. The positive control set must

consist of variants with experimental evidence of pathogenicity rather than database annotation, because database labels frequently contain errors and the label quality directly determines the upper bound of model performance (Harrison & Rehm, 2019). The negative control must be biologically appropriate, for example, the use of common variants in population datasets, including the variants in gnomAD with allele frequency greater than a predefined threshold, is justified from an evolutionary argument that a variant causing severe Mendelian disease cannot persist at high frequency in the general population due to natural selection (Karczewski et al., 2020). This hypothesis is well established and has been validated through several studies, but it carries an assumption that the population dataset does not have any systematic enrichment in pathogenic variants in specific genomic locations, an assumption that requires periodic re-examination as database sizes grow and variant annotations improve.

One persistent challenge for ensemble splicing models is canonical bias in both the training data and in the model calibration itself. This occurs because canonical splice site variants are easily identified and characterized with well-documented functional effects, they are more numerous in variant databases than non-canonical variants, which are harder to detect and require specialized functional assays to validate their functionality. A model that is trained using data dominated by canonical variants will use scoring thresholds based on the scoring distribution of canonical variants, which tend to have higher scores in general compared to other tools. As a result, non-canonical variants with intermediate scores will be categorized as uncertain even when those scores still have a biologically significant disruption depending on the genomic location of the variant. To reduce canonical bias when studying the non-canonical variants, training and evaluation should be done only by using non-canonical variants (Fraile-Bethencourt et al., 2019).

1.7.The Problem of Variants of Uncertain Significance

Variants of uncertain significance (VUS) have become one of the most significant challenges of modern genomics. VUS are variants whose evidence is insufficient to confidently classify them as pathogenic or benign. It is not a failure of sequencing technology but rather a result of the incomplete understanding of the particular case. In other words, by stating that the result is VUS, a clinical laboratory

states that a certain variation has been observed and the significance of this variation remains unknown (Balmaña et al., 2016).

The clinical consequences of VUS testing in patients are serious. The patients and their families are left in a state of uncertainty that can persist for years or even decades until enough evidence becomes available for reclassification. Clinicians face the difficult decision of whether to act on a VUS result, potentially initiating treatment or monitoring the patient for the presence of the variant that later turns out to be harmless, or ignore the problem and miss the diagnosis of a treatable disease. Research into the psychological impact of VUS results has documented higher levels of anxiety and depression in patients who receive them, especially if the genetic variant affects the gene responsible for a condition requiring special care (Makhnoon et al., 2019).

The VUS rates vary significantly depending on the gene, type of the variant, and level of expertise of the interpreting laboratory. However, a significant percentage of all VUSs in the rare disease correspond to non-canonical splice variants. The existing computational tools for predicting their effect produce scores that often fall in an intermediate range, and because the classification model does not have robust mechanisms for weighting this computational evidence, many of non-canonical splice variants end up classified as VUS regardless of their true pathogenicity (Soemedi et al., 2017). Reducing the VUS rate for this variant class through better-calibrated computational prediction is therefore not only a technical objective, it is a clinically meaningful goal with direct implications for patient care. To demonstrate this in a real clinical practice, Long QT Syndrome was selected as a proof-of-concept application, a condition where the consequences of misclassification are particularly serious and where non-canonical splice variants in the causative genes remain a notable source of unresolved diagnoses.

1.8.Long QT Syndrome as a Clinical Use Case

Long QT Syndrome (LQTS) is an inherited cardiac channelopathy in which dysfunction of ion channels delays ventricular repolarization, producing a prolonged QT interval on the electrocardiogram (ECG) as in Figure 5. This electrical change leads to serious consequences, it increases the risk of life-threatening ventricular arrhythmias such as torsades de pointes, which can progress to ventricular fibrillation and sudden cardiac death. In the United States alone, LQTS is estimated to contribute

to the deaths of 2,000-3,000 children and young adults every year with no prior symptoms (Vincent, 2000).

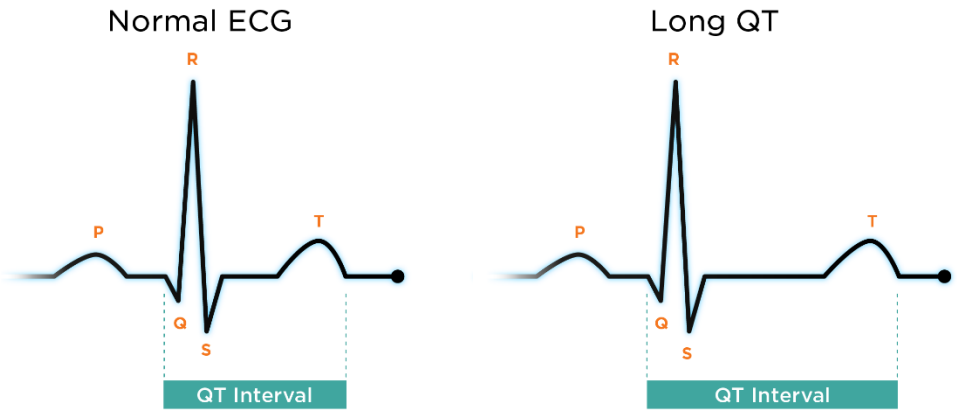


Figure 5| The QT interval in normal vs. Long QT Syndrome. The QT interval, measured from the onset of the Q wave to the end of the T wave, reflects the duration of ventricular depolarization and repolarization. In LQTS, delayed repolarization prolongs this interval and creates the electrophysiological substrate for life-threatening arrhythmias.

LQTS is a genetically heterogeneous disorder caused by pathogenic variants in genes that encode cardiac ion channels and regulatory proteins. To date, more than 17 genes have been associated with LQTS according to OMIM. The three most clinically significant subtypes, LQTS1, LQTS2, and LQTS3, are caused by variants in *KCNQ1*, *KCNH2*, and *SCN5A* genes respectively, accounting for about 75% of all genetically confirmed cases (Adler et al., 2020). Variants in these three genes cause an imbalance of the ion currents that control the repolarization phase.

Despite decades of research and the availability of clinically comprehensive genetic tests, the diagnostic rate of genetic testing in LQTS remains incomplete, genetic testing identifies a causative variant in approximately 75% of patients with a definitive clinical diagnosis of LQTS, with a significant number of cases where the genetic cause of their condition still remains unknown (Ackerman et al., 2011). Among patients who receive a genetic result, a significant fraction carries variants classified as VUS. ClinVar entries for LQTS-associated genes contain a high number of VUS relative to other well-characterized Mendelian disease genes.

This combination of factors makes LQTS an ideal clinical case for the ensemble model developed in this study. It is a well-defined monogenic disease with a limited set of disease-causing genes, making it tractable for targeted variant reclassification. Its disease genes carry a documented but incompletely classified splicing variants.

The clinical consequences of misclassification are severe, an undiagnosed LQTS patient remains at risk of sudden cardiac death, while an incorrectly classified VUS prevents cascade screening of at-risk family members. Applying the ensemble model to reclassify non-canonical splice-region variants in LQTS disease genes would therefore constitute a direct and clinically meaningful contribution to both molecular diagnosis and patient management in this condition.

Chapter 2| Motivation and Objectives

Despite the remarkable progress in genomic medicine over the past two decades, a significant proportion of patients who undergo genetic testing still leave without a clear diagnosis. This is not always because their condition lacks a genetic basis, in many cases, the responsible variant exists but falls outside the genomic regions that routine diagnostic pipelines are designed to examine. Non-canonical splicing variants, those that do not occur at the well-characterized boundaries of exons or at the two nucleotide positions of canonical splice sites, are particularly likely to be missed. When these variants are detected, they are frequently assigned uncertain classifications or missed altogether, leaving clinicians unable to make informed decisions and families without answers. The challenge is partly technical and partly conceptual. Several computational tools have been developed to predict whether a given variant affects splicing, and each of them brings genuine value. However, these tools were designed independently, and their outputs are expressed on different scales, with different reference assumptions. More importantly, they were not built with the recognition that variants in different parts of the genome affect splicing through fundamentally different biological mechanisms. A deep intronic variant that creates a cryptic splice site works very differently from an exonic variant that weakens an exon splicing enhancer, and both differ from a near-splice variant that directly disrupts a donor or acceptor motif. Applying the same scoring logic to all three is not just imprecise, it risks systematically underestimating the pathogenicity of variants in categories that generate weaker or more context-dependent signals.

Existing tools also share a more practical limitation: their scores cannot be directly translated into clinical variant classifications without essential manual interpretation. The ACMG/AMP five-tier system, which classifies variants as pathogenic, likely pathogenic, uncertain significance, likely benign, or benign, is the internationally accepted standard for clinical reporting, but none of the available splicing prediction tools produce output calibrated to this framework. As a result, clinicians must bridge the gap themselves, often relying on institutional guidelines or expert judgment rather than a principled, reproducible method.

In this work, we developed SpliceLoc-ML to address these limitations directly. The core idea is straightforward: rather than adding yet another independent prediction tool, this framework integrates the outputs of four established tools, SpliceAI,

Pangolin, MaxEntScan, and CADD, and trains a location-stratified ensemble model on top of them. Each tool contributes a distinct biological perspective: SpliceAI captures long-range sequence context, Pangolin provides tissue-specific splicing patterns, MaxEntScan evaluates short-range splice motif strength, and CADD reflects evolutionary constraint. Together, they offer a more complete picture of variant impact than any single tool can provide alone. This ensemble approach is consistent with the broader evidence that integration of complementary predictors improves classification performance over individual tools in variant pathogenicity assessment (Nicora et al., 2022). The stratification by genomic location is equally central to the design. Rather than applying a universal classifier to all variants, three separate models were built, one for near-splice variants (located 3 to 20 bp from a splice site), one for exonic variants, and one for deep-intronic variants (located more than 20 bp from a splice site). This approach ensures that each classifier is calibrated to the score distributions of its own variant class, preventing the stronger and more predictable signals of canonical splice-site disruption from dominating the model and hide the signals of non-canonical variants.

Critically, the model's output is expressed as a posterior probability that maps directly onto the ACMG/AMP five-tier classification system. This means the results are not random numbers on a proprietary scale, they are clinically interpretable values grounded in a framework that laboratories and clinicians already use and trust. This design choice directly addresses the translational gap identified above, enabling the model's output to be adopted into clinical reporting workflows without requiring additional manual score conversion.

The specific primary objectives of this study are:

1. To develop and validate SpliceLoc-ML, a location-stratified Random Forest ensemble for predicting the pathogenicity of non-canonical splice-region variants, trained on experimentally validated pathogenic variants from SpliceVarDB and benign controls from gnomAD. Model performance is evaluated on a held-out test set using AUROC, sensitivity, specificity, PPV, NPV, and MCC, with a target AUROC ≥ 0.85 and sensitivity $\geq 80\%$ as the minimum threshold for clinically acceptable performance.
2. To apply the validated SpliceLoc-ML model to reclassify a real-world dataset of non-canonical splice variants in eight genes associated with autosomal

dominant long QT syndrome (LQTS): KCNQ1, KCNH2, SCN5A, CALM1, CALM2, CALM3, KCNJ2, and TRDN, as a proof-of-concept demonstration of the model's clinical utility.

Chapter 3| Research Methodology

3.1.Study Design

SpliceLoc-ML is a location-stratified ensemble meta-classifier built to predict the clinical pathogenicity of genetic variants with potential splicing consequences. Its design based on two foundational principles: that multiple complementary prediction tools, used together, outperform any single tool used alone, and that variants occurring in different locations on the genome require separate models, calibrated to the biological realities of their respective locations. The model operates by taking the scores produced by four established splicing and conservation predictors, SpliceAI, Pangolin, MaxEntScan, and CADD with the GC content and PPT density, and using them as input features for a Random Forest meta-classifier. This meta-model is trained to output a posterior probability that aligns with the ACMG/AMP five-tier classification framework, making its results directly usable in clinical variant reporting without the need for manual score translation.

Three separate classifiers were trained, one for each of the following variant categories: near-splice variants, exonic variants, and deep-intronic variants. This stratified architecture was chosen because these three classes not only differ in their genomic location but also in the molecular mechanisms through which they disrupt normal splicing. A single classifier applied across all three would be dominated by the strong, consistent signals of near-canonical disruption, effectively reducing sensitivity for the more subtle effects that characterize deep intronic and exonic splicing variants. By training separately on each category, each classifier learns the score distributions specific to that class, enabling more balanced and accurate predictions across the full spectrum of non-canonical splice variants.

The model was developed using data drawn from a broad range of human genes and disease contexts, with features selected for their applicability across any protein-coding gene. This design makes the framework potentially generalizable well beyond the genes studied here. To evaluate its clinical utility, SpliceLoc-ML was applied to a curated set of variants from eight genes all associated with autosomal dominant long QT syndrome as a proof-of-concept. All analyses were conducted using the GRCh38/hg38 human reference genome assembly.

The overall analytical workflow is summarized in Figure *6*. Briefly, data collection yielded 6,828 pathogenic variants from SpliceVarDB and 21,166 benign variants from

gnomAD. Variants were categorized by genomic location and annotated with the six features described above (SpliceAI, Pangolin, MaxEntScan, CADD, GC content, and PPT density). The dataset was then split into 70% training and 30% test sets using a stratified approach to preserve class balance. A separate Random Forest ensemble was trained for each variant category, and the resulting models were calibrated and validated before application to the LQTS gene set.

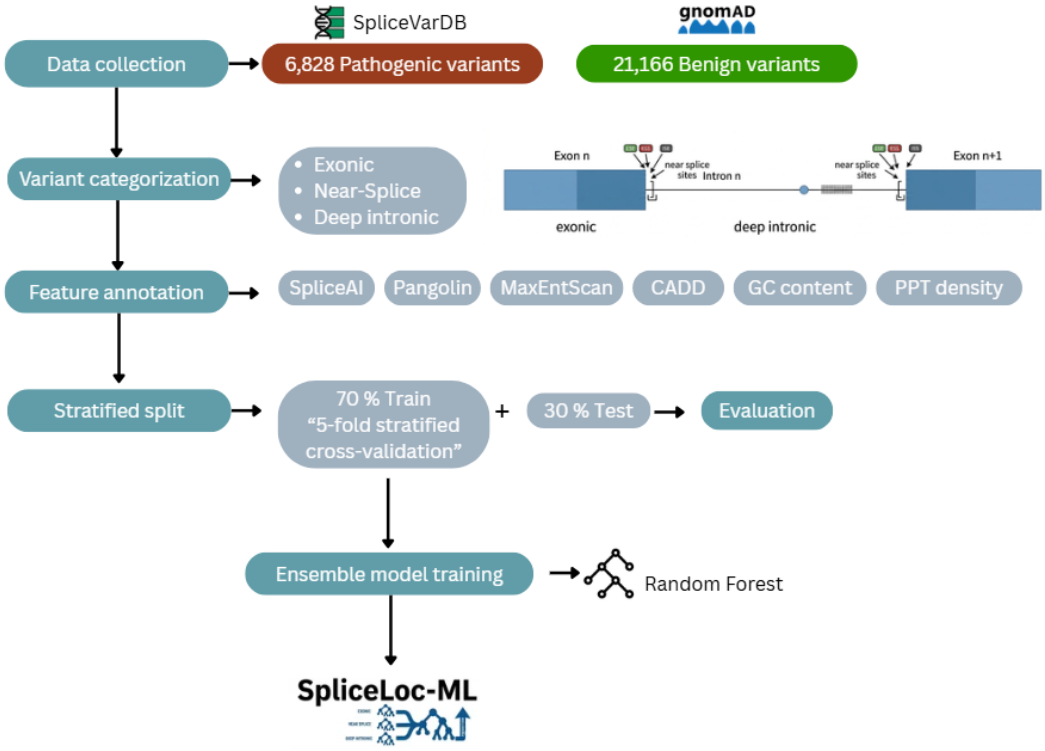


Figure 6|Analytical workflow for the ensemble machine learning model for non-canonical splice variant pathogenicity prediction.

3.2.Data Sources

3.2.1. Positive controls: SpliceVarDB

Pathogenic splice-altering variants were obtained from SpliceVarDB, a curated database of experimentally validated splice-altering variants across a broad range of human disease genes (Sullivan et al., 2024). An initial set of 13,100 variants was obtained. Variants were filtered as in **Figure** *7*.

The SpliceVarDB records the validation assay used for each variant, which include RNA sequencing (RNA-Seq), minigene reporter assays, Multiplexed Functional Assay of Splicing using Sort-seq (MFASS), and Massively Parallel Splicing Assay MaPSy (Sullivan et al., 2024).

3.2.2. Negative controls: gnomAD v4.1

The benign variants were extracted from the Genome Aggregation Database (gnomAD) version 4.1 genome sequencing dataset, which combines the data of whole-genome and whole-exome sequencing from 807,162 individuals based on the release announcement from gnomAD v4.0 (Guez et al., 2026). The variants were downloaded directly from Google Cloud Storage using (gsutil), which has access to the gnomAD VCF files, and filtered via (BCFtools "bcftools -v snps") by using the Google Colab. The filters were applied as in **Figure** *7*. The main criterion for obtaining benign variants was the AF > 0.001 because according to (Jaganathan et al., 2019) a variant that causes a severe Mendelian disease must be eliminated by natural selection before reaching an allele frequency higher than 1 in 1,000 individuals.

The ACMG/AMP criterion BA1 uses an allele frequency of 5% or higher as stand-alone evidence for a benign classification (Ghosh et al., 2018; Richards et al., 2015). This cutoff is meant for classifying single variants in diagnosis, and it is kept high to avoid calling low-penetrance pathogenic alleles benign. Here the frequency filter was used only to build a set of control variants for training. For most rare Mendelian diseases, a fully penetrant pathogenic variant is not expected to reach a frequency above 0.001 (Whiffin et al., 2017), so we used this value. A 5% cutoff would leave fewer benign variants in each category. The disadvantage of our choice is that the control set may include variants with a frequency between 0.001 and 0.05 that are pathogenic with reduced penetrance (Section 5.10). We did not test how the results change with a different threshold.

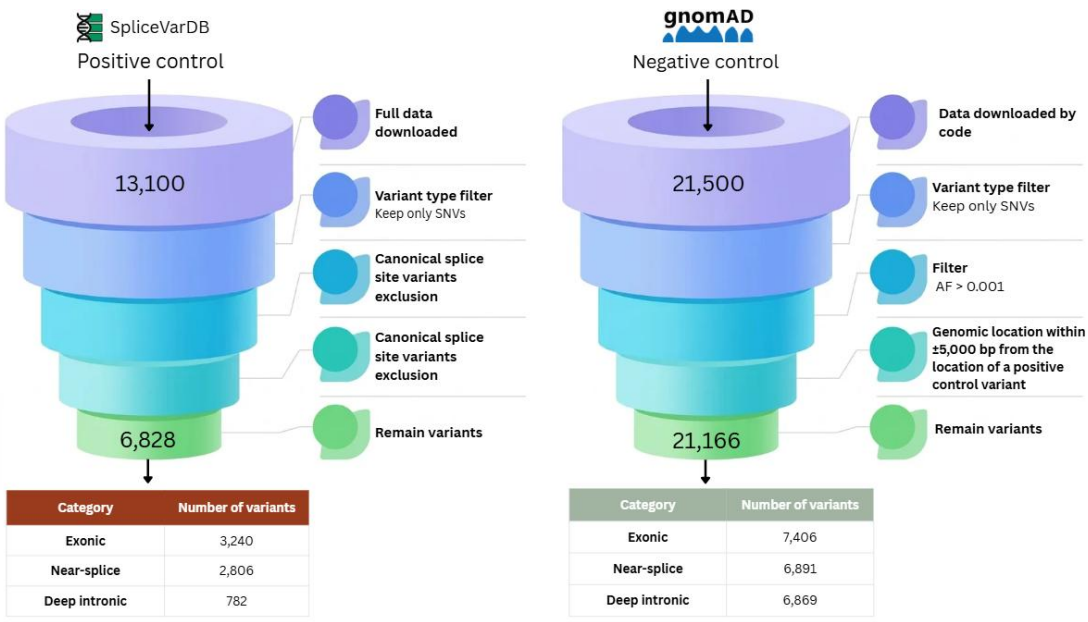


Figure 7| Filtering pipeline for positive and negative control variant datasets. Funnel diagrams illustrating the sequential filtering steps applied to (left) positive control variants from SpliceVarDB and (right) negative control variants from gnomAD v4.1. Starting from 13,100 and 21,500 variants respectively. Final datasets comprised 6,828 positive control variants and 21,166 negative control variants, each categorized into exonic, near-splice, and deep intronic variants.

3.3.Variant Annotation and Feature Extraction

Each variant was annotated by using the Ensembl Variant Effect Predictor (VEP) tool, using GRCh38 as the reference assembly. Four computational splicing prediction tools were chosen to be used as features, selected to provide complementary biological information and cover potential blind spots of individual tools.

SpliceAI was run via VEP (SpliceAI plugin) by using a custom implementation that outputs the raw scores. Raw scores of SpliceAI were used as they improve the sensitivity for predicting non-canonical variants outside canonical splice site motifs. Four delta scores (Δ acceptor gain, Δ acceptor loss, Δ donor gain, Δ donor loss) and the maximum of the four (max_splice_val) were retained as separate features. SpliceAI was selected for its long-range context modeling, reading up to 10,000 nucleotides around each variant being assessed (Jaganathan et al., 2019).

The Pangolin model was run locally on the Ubuntu operating system through the Pangolin command-line tool, which produces splicing probability changes across 4 human tissues (brain, heart, liver and testis). The run was done by the mask option parameter (--mask True) and distance option (-d 300). The maximum absolute score among all scored positions was extracted from the raw output and used as a single numeric feature (Pangolin_score). The Pangolin tool was chosen for its tissue-specific

predictions of splice site usage, trained on RNA-seq data from four tissues (heart, liver, brain and testis) of four species, which gives a splicing signal that complements SpliceAI (Zeng & Li, 2022). SpliceAI and Pangolin are themselves trained models, and the effect of using their scores as features is discussed in Section 5.10.

MaxEntScan was run via VEP. Three variables were extracted: the reference allele splice site score (MaxEntScan_ref), the alternate allele splice site score (MaxEntScan_alt), and the difference between them (MaxEntScan_diff). MaxEntScan provides biophysical modeling of short-range splice site motif strength, offering a mechanistically distinct signal to the deep learning-based tools (Yeo & Burge, 2003).

CADD-PHRED scores were obtained via VEP annotation. CADD integrates evolutionary conservation across species with diverse genomic features to produce a single deleteriousness score on a PHRED-scaled 0–99 range. Within this ensemble, CADD provides a pathogenicity signal independent of splicing mechanics. A variant at a highly conserved site that is also given a high splicing score is considered more likely to be pathogenic (Kircher et al., 2014).

Two sequence-context features were computed based on the sequence itself using Python scripts: GC content and polypyrimidine tract (PPT) density, each calculated in a 50 bp window centered on the variant position. These features add secondary splicing signals that important for splicing regulation (Amit et al., 2012; Coolidge et al., 1997).

3.4.Imputation of missing scores

Missing scores were handled by using a biologically justified two-group strategy, following the framework established by (Petrazzini et al., 2021) and the framework of each tool. When a missing score has a clear biological interpretation, such as those obtained by SpliceAI delta scores, Pangolin score, ada_score, rf_score, and MaxEntScan_diff, then missing values were imputed with zero, representing the biological null state that there is no splicing effect. However, for MaxEntScan_alt, MaxEntScan_ref, and CADD_PHRED, zero imputation would be biologically misleading. A MaxEntScan score of zero represents a very weak splice site, which is a real measurement, not an absent one. These features were therefore imputed using the median for each category that computed separately for each of the three location categories to prevent cross-category contamination of the imputed values.

MaxEntScan was designed for short splice site motifs, so it gives no useful signal for deep intronic variants. We could not leave these values missing, because the Random Forest in scikit-learn v1.0.2 does not accept missing inputs, so they were imputed. The imputed value is the same for every variant with a missing score in a category, so it cannot help the model separate pathogenic from benign variants. In the SHAP analysis (Section 4.5) the three MaxEntScan features had a mean |SHAP| of 0.001 or less in the deep intronic model, compared with 0.226 for Pangolin and 0.092 for CADD PHRED. We did not retrain the deep intronic model without these features.

3.5.Class balancing

An imbalance between positive and negative examples was identified in each of the three variant-locus categories, highlighting the need to account for class distribution during model training and performance evaluation. The deep intronic category showed the greatest class imbalance with benign variants substantially outnumbering pathogenic variants (6,869 vs. 782; benign-to-pathogenic ratio ≈ 8.8:1). So, prior to training, random downsampling of the majority class was applied. Exonic benign variants were downsampled to 3,300 (resulting ratio ~1.02:1) and near-splice benign variants to 3,000 (~1.07:1). For deep intronic variants, benign variants were downsampled to 3,000, yielding a residual ratio of (3.84:1).

This class imbalance was addressed using a cost-sensitive learning approach through class weighting, implemented via the *compute_class_weight()* function from Scikit-learn machine learning library (Pedregosa et al., 2011) in Python, using the balanced mode parameter. This configuration automatically assigns weights to each class that are inversely proportional to their frequencies in the training data. For the deep intronic variant model, this resulted in a pathogenic class weight of approximately 2.42 and a benign class weight of approximately 0.63, meaning that misclassifying a pathogenic variant was penalized 2.42 times more heavily than misclassifying a benign variant during decision tree construction.

3.6.Model Training and Cutoff Selection

Three separate Random Forest models were trained using scikit-learn library (v. 1.0.2), one per variant-locus. The dataset was split into a training set (70%) and a held-out testing set (30%) using stratified sampling to preserve class proportions. The entire process of developing the model, from hyperparameter optimization to cutoff determination was performed exclusively on the training set. The testing set was used only once, for final evaluation, to prevent overfitting and data leakage.

Every Random Forest followed these hyperparameters, 500 decision trees with maximum depth 10, minimum samples 5 per leaf, and the square root of the number of features considered at every split. Also, the random seed was fixed at 42 across all operations to ensure reproducibility. The hyperparameters were optimized via RandomizedSearchCV with 5-fold stratified cross-validation. Two unique factors made the Random Forest architecture the most suitable for the study: (1) Random Forest is naturally scaling invariant and does not require feature normalization despite the heterogeneous scales of the input features, for example, CADD: 0–99, SpliceAI: 0–1, MaxEntScan: –20 to +12. (2) Cost-sensitive class weighting is incorporated through amplified vote contributions from weighted samples in each decision tree.

Cutoff values for binary classifications were obtained from the probability distributions of the training sets using Youden's J index, where (J = Sensitivity + Specificity – 1). Ten possible candidate cutoff methods were evaluated and ranked by Youden's J on the training set, the highest value of J was then used for determining the cutoff value for that category.

3.7. The Five-Tier ACMG Classification

Posterior probability values of the Splice-loc-ML model were transformed to a five-category clinical classification based on recommendations of the ACMG in terms of the Bayesian posterior probability model proposed by Tavtigian et al. (2018). Three classification thresholds, namely T_benign, T_likely_pathogenic, and T_pathogenic, are published constants from the Bayesian framework, which are not derived from the data, the thresholds values are determined in Table 3. The remaining threshold, T_vus_lower is calculated per category by the Youden's J threshold. Variants

classified as VUS represent cases where computational evidence alone is insufficient for classification, these require additional functional evidence.

Table 3|Five-tier ACMG/AMP classification thresholds. Three thresholds are internationally validated constants from Tavtigian et al. (2018), one is data-derived from the training set (J: Youden's J threshold).

Threshold	Probability Range	ACMG Tier Assigned	Source
$T_1 = 0.10$	≤ 0.10	Benign	Tavtigian et al. (2018) — Bayesian derivation from ACMG/AMP evidence codes
$T_2 = 0.10 – J^*$	$0.10 < p \leq J^*$	Likely Benign	Tavtigian lower boundary / Youden's J upper boundary
$T_3 = J^*$ (data-derived)	$J^* < p < 0.90$	VUS	Youden's J statistic on training ROC curve
$T_4 = 0.90$	$0.90 \leq p < 0.99$	Likely Pathogenic	Tavtigian et al. (2018) — posterior probability ≥ 0.90
$T_5 = 0.99$	≥ 0.99	Pathogenic	Tavtigian et al. (2018) — posterior probability ≥ 0.99

3.8.Model Evaluation

The performance of the trained model was evaluated on the 30% held-out test set using six metrics: AUROC as the key criterion to quantify discrimination capacity, sensitivity and specificity computed using Youden's J cutoff, positive predictive value (PPV) and negative predictive value (NPV) for evaluating prediction validity of positive and negative classifications results, and Matthews Correlation Coefficient (MCC) as the primary metric of classification effectiveness in class-imbalance cases, which incorporates all four parameters of the confusion matrix and has been demonstrated to be more informative than accuracy or F1 score, especially for imbalanced binary classification. Minimum acceptable performance thresholds were set as AUROC ≥ 0.85, sensitivity ≥ 0.80, specificity ≥ 0.80 (Jaganathan et al., 2019; Zeng & Li, 2022), and MCC ≥ 0.50 (Chicco & Jurman, 2020).

3.9.Feature importance

The feature importance was quantified using SHAP (SHapley Additive exPlanations) values, computed on the held-out test set using the tree_path_dependent

perturbation method. Mean absolute SHAP values were calculated per feature to represent overall importance for each variant-locus model.

3.10. Comparison with SpliceVarDB Experimentally Validated Variants

The performance of Splice-loc-ML model was evaluated using variants data that had been experimentally validated for splicing impact using four different laboratory approaches reported in SpliceVarDB: RNA-sequencing (n = 5,163), minigene assay (n = 333), MaPSy (n = 203), and MFASS (n = 1,060). The detection rate was calculated as the percentage of validated disease-causing variants rated by the model as pathogenic or likely pathogenic. All four approaches have different levels of relevance to the native biology since RNA-sequencing identifies splicing in patients' samples, whereas MFASS relies on non-native cell lines and artificial reporter constructs.

3.11. Software and Reproducibility

All analyses were implemented in Python 3.10 within Google Colaboratory (Google Colab). Core libraries used were: scikit-learn (v. 1.0.2) for Random Forest training, metrics, class weights, and imputation, imbalanced-learn 0.11 for random downsampling, pandas 2.0 and NumPy 1.25 for data processing, matplotlib 3.7 and seaborn 0.12 for plotting, and shap 0.42 for feature importance. Pangolin v.1.1 was run in Ubuntu environment. Ensembl VEP tool version 110 was used to annotate variants using plugins SpliceAI, MaxEntScan, and CADD. A fixed random seed 42 was used throughout to ensure reproducibility during data splitting, downsampling, and model training. All trained model objects, binary cutoffs, and five-tier classification thresholds were saved using the joblib library. A custom prediction function, predict_splicing_effect(), was developed to accept a variant dataframe as input and return three outputs for each variant: the splicing effect probability score, the binary pathogenicity label, and the five-tier ACMG-style classification label along with its corresponding thresholds. This approach was designed to enable direct future application of the pipeline to real patient-derived variant datasets without the need for retraining. All software packages and libraries are listed in the Appendix A. All code,

analysis scripts, and trained model files are publicly available at: https://github.com/wedadmurar/SpliceLoc-ML.

3.12. Non-Canonical Splice Variants in LQTS Genes

To assess the clinical applicability of the trained ensemble Splice-loc-ML model using a real-world patient dataset, the three variant-locus category-specific models were applied to non-canonical splice-region variants across eight Long QT Syndrome (LQTS)-associated genes: KCNQ1, KCNH2, SCN5A, CALM1, CALM2, CALM3, KCNJ2, and TRDN. The variants were downloaded from ClinVar and annotated using VEP with SpliceAI, MaxEntScan, and CADD plugins, then run in Pangolin via Ubuntu, the GC content and PPT density were calculated via code in Google Colab. Variants filtering was conducted by limiting the analysis to the canonical MANE Select transcript for each gene and keeping only SNVs. Missense, nonsense, frameshift, stop-codon, and canonical ±1/±2 splice site variants were excluded, followed by the removal of variants previously classified as pathogenic or likely pathogenic.

Chapter 4| Results

Part A: Model Building and Validation

4.1. Dataset Composition and Class Balancing

The original dataset comprised 27,994 variants, which were categorized into three distinct genomic-location classes. The positive control set, which was obtained from SpliceVarDB, contained 6,828 experimentally validated pathogenic splice-altering variants: 3,240 exonic, 2,806 near-splice, and 782 deep intronic. The negative control set contained 21,166 benign variants: 7,406 exonic, 6,891 near-splice, and 6,869 deep intronic, as in Table 4.

After downsampling, exonic benign variants became 3,300 (ratio ~1.02:1), while near-splice benign variants were 3,000 (~1.07:1), and deep intronic benign variants were downsampled to 3,000 (3.84:1). This ratio was balanced by using the cost-sensitive method by assigning the weight for the benign class was 0.63 while the pathogenic class was weight 2.42.

Table 4| Dataset composition before and after class balancing. Deep intronic class imbalance was managed using a combination of random downsampling and cost-sensitive class weights (pathogenic weight ≈ 2.42, benign weight ≈ 0.63).

Category	Benign (Original)	Benign (After downsampling)	Pathogenic	Resulting ratios
Exonic	7,406	3,300	3,240	~ 1.02:1
Near-Splice	6,891	3,000	2,806	~ 1.07:1
Deep Intronic	6,869	3,000	782	~ 3.84:1
Total	**21,166**	**9,300**	**6,828**	—

The total number of variants after downsampling became 16,128 variants both benign and pathogenic variants. The 70% training dataset and 30% held-out test set were as in **Table *5***.

Table 5|The balanced dataset after downsampling and the split (70 % /30 %, seed=42).

Category	Training set (70%)	Test set (30%)	Total
Exonic	4,578	1,962	**6,540**
Near-splice	4,064	1,742	**5,806**
Deep intronic	2,647	1,135	**3,782**
Total	**11,289**	**4,839**	**16,128**

The remaining benign to pathogenic ratio in the deep intronic category (3.84:1) was handled during training with balanced class weights (0.63 for benign and 2.42 for pathogenic), computed on the training set only. The test set was not resampled, so it keeps the same ratio (about 900 benign and 235 pathogenic variants). Sensitivity and specificity do not depend on this ratio, and they are reported for each category in Table 7 (0.866 and 0.966 for exonic, 0.904 and 0.948 for near-splice, and 0.872 and 0.953 for deep intronic variants). The deep intronic sensitivity (95% confidence interval 0.82–0.91) is therefore similar to the exonic one. At the tier level, fewer pathogenic deep intronic variants reached the Likely Pathogenic or Pathogenic tiers (489 of 782, 62.5%) than near-splice (2,221 of 2,806, 79.2%) or exonic variants (2,376 of 3,240, 73.3%) (Section 4.7).

4.2. Model Performance on the Held-out Test Dataset

The performance of the three RF models based on specific categories were evaluated on the held-out test set (30% of balanced data, withheld during all training and cutoff derivation steps). Performance was determined using six evaluation metrics: AUROC, sensitivity, specificity, PPV, NPV, and MCC. Pre-defined minimum thresholds (AUROC ≥ 0.85, sensitivity ≥ 0.80, specificity ≥ 0.80, MCC ≥ 0.50) were met by all three models.

The optimal binary thresholds were determined by evaluating ten different cutoff methodologies on the training set as shown in the Table 6. Once the optimal cutoffs were identified via Youden's J index, they were applied as fixed parameters to the held-out test set. The results in

Table 7 represent the final unbiased performance of the model on the test dataset, ensuring that the metrics are not artificially inflated by over-fitting to the training distribution.

Table 6|Multi-threshold Diagnostic Performance Analysis by Genomic Category. This table show the balance between sensitivity and specificity across a probability (0.0 to 1.0). For each category, exonic, near-splice, and deep intronic. diagnostic metrics including PPV, NPV, F1-score, and MCC are provided. The "OPTIMAL" label identifies the decision boundary that maximizes the Youden's J statistic, serving as the basis for final model classification.

Exonic												
Threshold	Sensitivity	Specificity	PPV	NPV	Youden-J	F1	MCC	TP	TN	FP	FN	Label
0	1	0	0.4954	0	0	0.6626	0	2268	0	2310	0	Standard

0.1	0.9766	0.6684	0.743	0.9668	0.645	0.844	0.6767	2215	1544	766	53	Standard
0.2	0.9555	0.8152	0.8354	0.9491	0.7706	0.8914	0.7775	2167	1883	427	101	Standard
0.3	0.9356	0.9004	0.9022	0.9344	0.8361	0.9186	0.8363	2122	2080	230	146	Standard
0.4	0.9175	0.9342	0.9319	0.9203	0.8517	0.9247	0.852	2081	2158	152	187	Standard
0.5	**0.8981**	**0.9593**	**0.9559**	**0.9056**	**0.8575**	**0.9261**	**0.8595**	**2037**	**2216**	**94**	**231**	**OPTIMAL**
0.6	0.8704	0.9792	0.9763	0.885	0.8496	0.9203	0.8554	1974	2262	48	294	Standard
0.7	0.8382	0.9892	0.987	0.8616	0.8274	0.9065	0.8379	1901	2285	25	367	Standard
0.8	0.8011	0.9948	0.9934	0.8359	0.796	0.887	0.8125	1817	2298	12	451	Standard
0.9	0.7372	1	1	0.7949	0.7372	0.8487	0.7655	1672	2310	0	596	Standard
1	0.0582	1	1	0.5196	0.0582	0.11	0.1739	132	2310	0	2136	Standard
Near-splice												
Threshold	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Youden-J**	**F1**	**MCC**	**TP**	**TN**	**FP**	**FN**	**Label**
0	1	0	0.4833	0	0	0.6516	0	1964	0	2100	0	Standard
0.1	0.9873	0.7281	0.7725	0.9839	0.7154	0.8668	0.7356	1939	1529	571	25	Standard
0.2	0.9776	0.8671	0.8731	0.9764	0.8447	0.9224	0.8471	1920	1821	279	44	Standard
0.3	0.9567	0.9205	0.9184	0.9579	0.8772	0.9372	0.8767	1879	1933	167	85	Standard
0.4	0.9379	0.949	0.9451	0.9423	0.8869	0.9415	0.8872	1842	1993	107	122	Standard
0.5	**0.9201**	**0.9676**	**0.9637**	**0.9283**	**0.8877**	**0.9414**	**0.8898**	**1807**	**2032**	**68**	**157**	**OPTIMAL**
0.6	0.8987	0.98	0.9768	0.9118	0.8787	0.9361	0.8836	1765	2058	42	199	Standard
0.7	0.8722	0.9881	0.9856	0.8921	0.8603	0.9254	0.869	1713	2075	25	251	Standard
0.8	0.8447	0.9948	0.9934	0.8726	0.8395	0.913	0.8526	1659	2089	11	305	Standard
0.9	0.7821	1	1	0.8307	0.7821	0.8777	0.806	1536	2100	0	428	Standard
1	0.0107	1	1	0.5194	0.0107	0.0212	0.0745	21	2100	0	1943	Standard
Deep intronic												
Threshold	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Youden-J**	**F1**	**MCC**	**TP**	**TN**	**FP**	**FN**	**Label**
0	1	0	0.2066	0	0	0.3425	0	547	0	2100	0	Standard
0.1	0.9909	0.4838	0.3333	0.9951	0.4747	0.4988	0.3948	542	1016	1084	5	Standard
0.2	0.9415	0.8562	0.6304	0.9825	0.7977	0.7551	0.6992	515	1798	302	32	Standard
0.3	0.9287	0.9057	0.7195	0.9799	0.8344	0.8109	0.764	508	1902	198	39	Standard
0.4	0.9141	0.9286	0.7692	0.9765	0.8426	0.8354	0.7927	500	1950	150	47	Standard
0.5	**0.9031**	**0.9567**	**0.8444**	**0.9743**	**0.8598**	**0.8728**	**0.839**	**494**	**2009**	**91**	**53**	**OPTIMAL**
0.6	0.8629	0.9786	0.913	0.9648	0.8415	0.8872	0.8594	472	2055	45	75	Standard

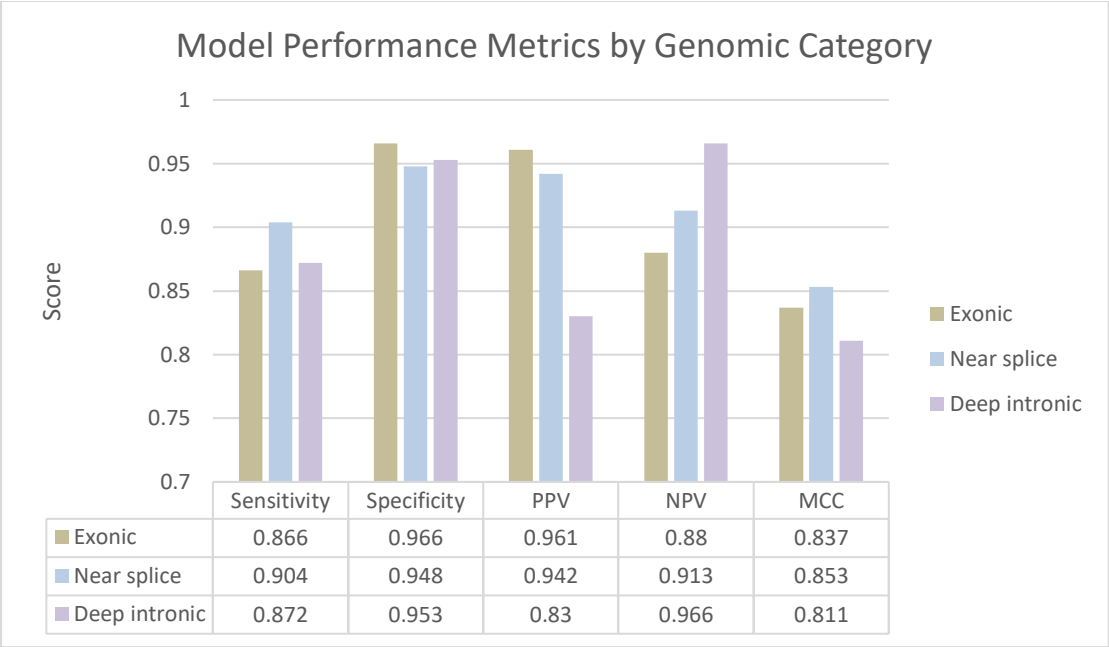
0.7	0.8263	0.9862	0.9397	0.9561	0.8125	0.8794	0.8532	452	2071	29	95	Standard
0.8	0.7879	0.9929	0.9664	0.9473	0.7808	0.8681	0.8446	431	2085	15	116	Standard
0.9	0.6033	0.9976	0.9851	0.9061	0.6009	0.7483	0.7318	330	2095	5	217	Standard
1	0	1	0	0.7934	0	0	0	0	2100	0	547	Standard

Table 7| Model performance on test set (30%). Cutoff = Youden's J binary threshold derived on training set only. All values evaluated on test dataset. **Green bold** = best value per column. All three categories pass all thresholds.

Category	Cutoff	AUROC	Sensitivity	Specificity	PPV	NPV	MCC
Exonic	**0.527**	0.969	0.866	**0.966**	0.961	0.880	0.837
Near-Splice	**0.474**	**0.975**	**0.904**	0.948	**0.942**	0.913	**0.853**
Deep Intronic	**0.512**	0.936	0.872	0.953	0.830	**0.966**	0.811

The thresholds in table 3 utilize a 0.1-step increment to visualize broad performance trends, while the final classification thresholds (T_2) were determined using an exhaustive comparison for all unique predicted probabilities. This high-resolution optimization identified specific cutoffs, 0.527 for exonic, 0.474 for near-splice, and 0.512 for deep intronic, which represent the absolute mathematical maximum for Youden's J statistic.

The near-splice model achieved the highest overall performance, with AUROC = 0.975, sensitivity= 0.904, specificity = 0.948, and MCC = 0.853. The exonic model performed approximately similar results, with an AUROC = 0.969, sensitivity = 0.866, specificity = 0.966, and MCC = 0.837. Deep intronic model that originally had the most imbalanced category, achieved an AUROC = 0.936, sensitivity = 0.872, specificity = 0.953, and MCC = 0.811. Although the deep intronic AUROC was the lowest of the three, it is still much higher than the acceptable threshold and represents a strong result given the inherent biological complexity of classifying variants located far from annotated splice sites and the severe class imbalance in that category.



	Sensitivity	Specificity	PPV	NPV	MCC
Exonic	0.866	0.966	0.961	0.88	0.837
Near splice	0.904	0.948	0.942	0.913	0.853
Deep intronic	0.872	0.953	0.83	0.966	0.811

Figure 8| Performance Metrics of Models by Genomic Category. Shows the performance metrics graphically for all three models. The almost identical height of the bars confirms that none of the model favors one metric over another, all metrics are higher than 0.80 for all categories.

4.3. Receiver Operating Characteristic Curves

Figure 9 shows the receiver operating characteristic (ROC) curves for the three models, evaluated on the held-out test set. Each curve plots the true positive rate (sensitivity) against the false positive rate (1 − specificity) across all probability thresholds. All three curves are strongly elevated toward the top left corner of the ROC, which indicates that all the three models significantly overcome randomness in their predictions. Among the three ROC curves, the one produced by the near splice method is closer to the perfect point (AUC = 0.975), followed by the exonic model (AUC = 0.969) and at the last the deep intronic model (AUC = 0.936). The closeness of all three curves in the high-performance region shows that the stratified modelling approach produces strong discrimination across all three genomic location categories, despite the differences in class imbalance and biological complexity between them.

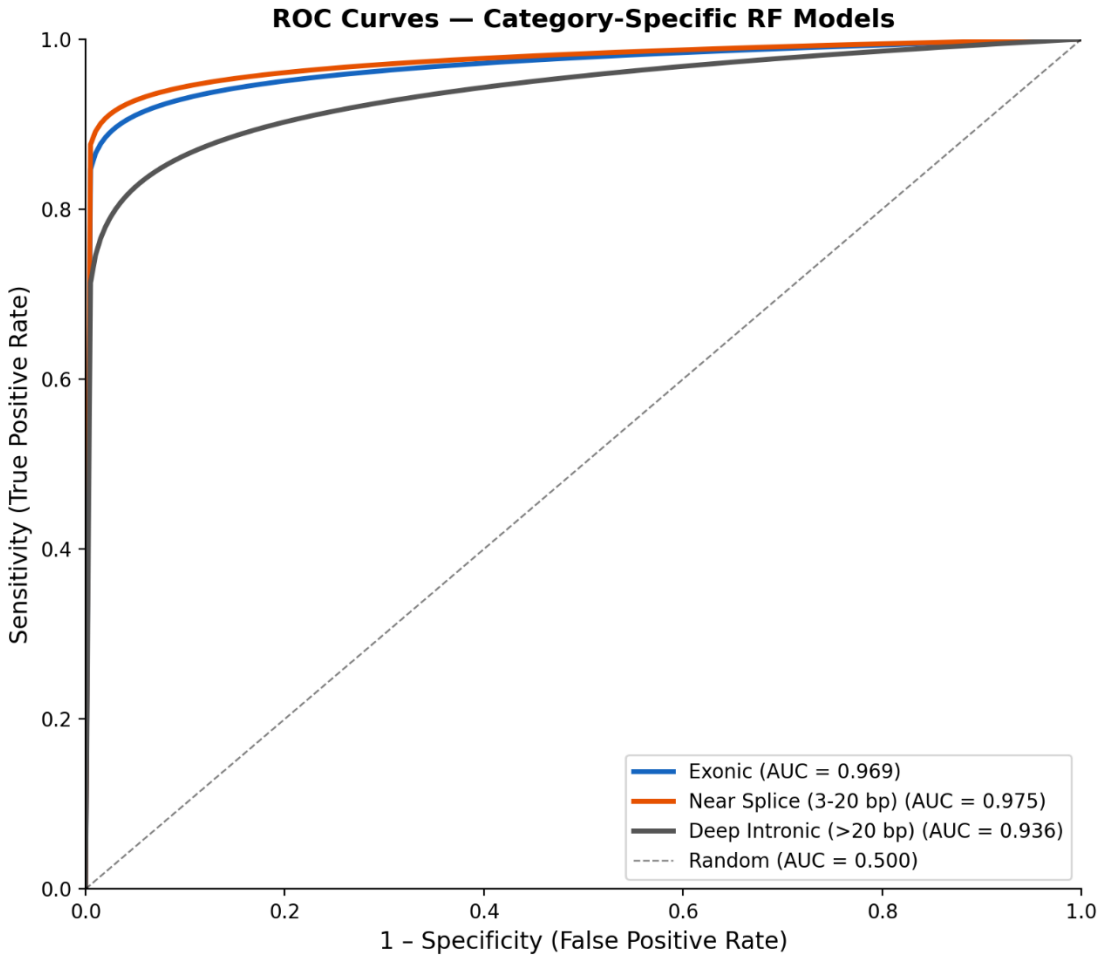


Figure 9|ROC curves for the three Random Forest splicing models that were trained specifically on the test set (30%). AUC values: Exonic = 0.969, Near Splice = 0.975, Deep Intronic = 0.936. The dashed line represents the random classification (AUC = 0.500). All three models outperform random classification.

4.4. Score Distribution by True Label

Violin plots that classify probability scores based on true labels (benign and pathogenic) were shown in Figure 10 for each of the three models. These distributions show how effectively each model separates the two classes across the probability range.

In the three categories, the benign variants showed an extremely tight cluster of scores close to 0, with the dense mass of the distribution clustered below 0.25. The pathogenic variants produced a distribution concentrated near 1, with most scores above 0.75. These two distinct distributions provide the graphical evidence of the high AUROC values and prove that the probabilities generated by the model have real discriminative power.

Another distinguishing feature in the distribution of benign deep intronic variants, which shows an extended tail on the high side extending up to 1 compared to the exonic and near-splice categories. This indicates that a subset of benign deep intronic variants received higher pathogenic probability scores which is the primary source of false positives in that category. This observation is consistent with the lower specificity and AUROC of the deep intronic model because of the intrinsic uncertainty of the deep intronic region due to the presence of certain common variants within regions of the genome similar to cryptic splice sites.

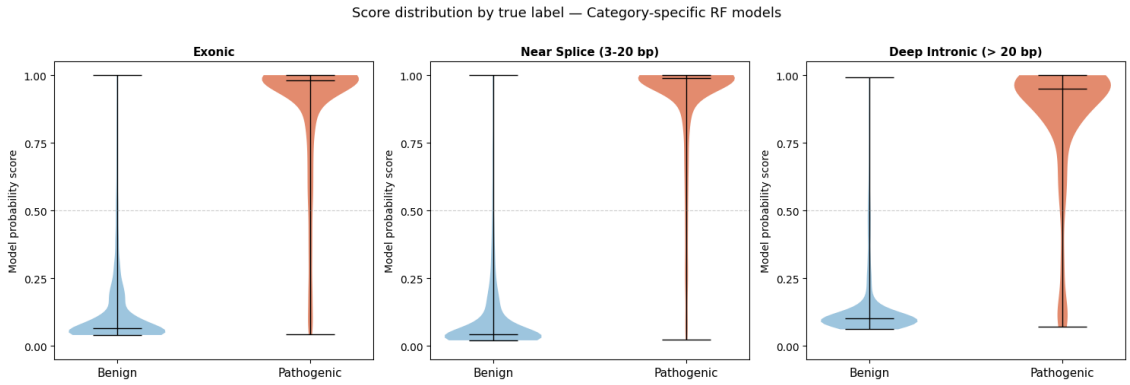


Figure 10|Violin plots of Model-Predicted Pathogenic Probability Scores by True Label of the Three Location Categories. **Blue** = benign variants, **Orange** = pathogenic variants. The dashed horizontal line represents the probability = 0.50.

In Figure 11, Figure 12, and Figure 13 the violin plots show the distribution of pathogenicity scores calculated by five different tools: our Ensemble model (SpliceLoc-ML), SpliceAI (maximum delta score), Pangolin, CADD PHRED, and MaxEntScan (reference–alternate difference). Variants are classified across three location categories: exonic, near-splice, and deep intronic. Each panel is scaled independently to its own y-axis to maintain an interpretable range for each scoring system. Discriminative performance is quantified for each panel using the AUROC, where a value of 1.0 indicates perfect separation and 0.5 indicates random classification. Statistical significance was tested by using two-sided Mann–Whitney U test. The best AUROC score was obtained by the ensemble SpliceLoc-ML model across all location categories (Exonic: 0.971, Near-Splice: 0.980, Deep Intronic: 0.958), outperforming all individual tools.

Between the individual tools, the best discriminative ability was shown by Pangolin with AUROC range of 0.932–0.934, then MaxEntScan 0.531–0.888, and at the last SpliceAI 0.686–0.692. For MaxEntScan showed statistically significant but negligible

discriminative ability for deep intronic variants AUROC = 0.531, $p < 0.001$, consistent with its design limitations outside the boundaries of canonical splice site.

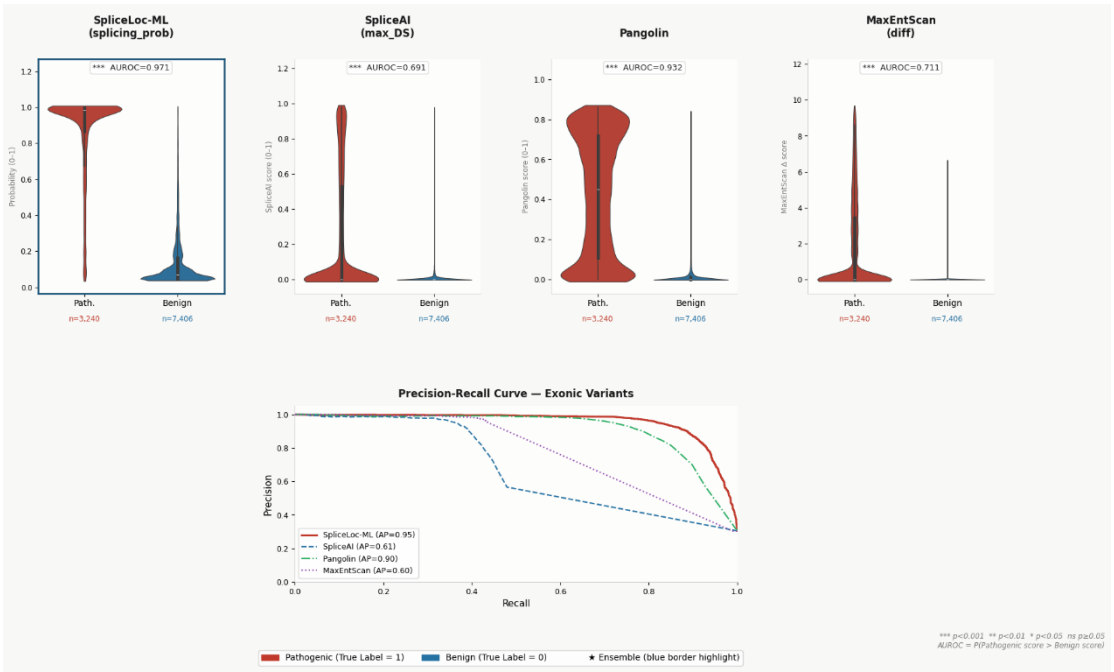


Figure 11|SpliceLoc-ML vs individual splicing tools in exonic variants. Violin plots showing pathogenicity score distributions for SpliceLoc-ML, SpliceAI, Pangolin, and MaxEntScan between pathogenic (n=3,240) and benign (n=7,406) exonic variants. AUROC values were calculated using the Mann–Whitney U test on the full dataset. The precision-recall curve compares all four tools on the same subset. Violin shapes are based on the 1st–99th percentile range for visual clarity. Significance thresholds: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns $p \geq 0.05$.

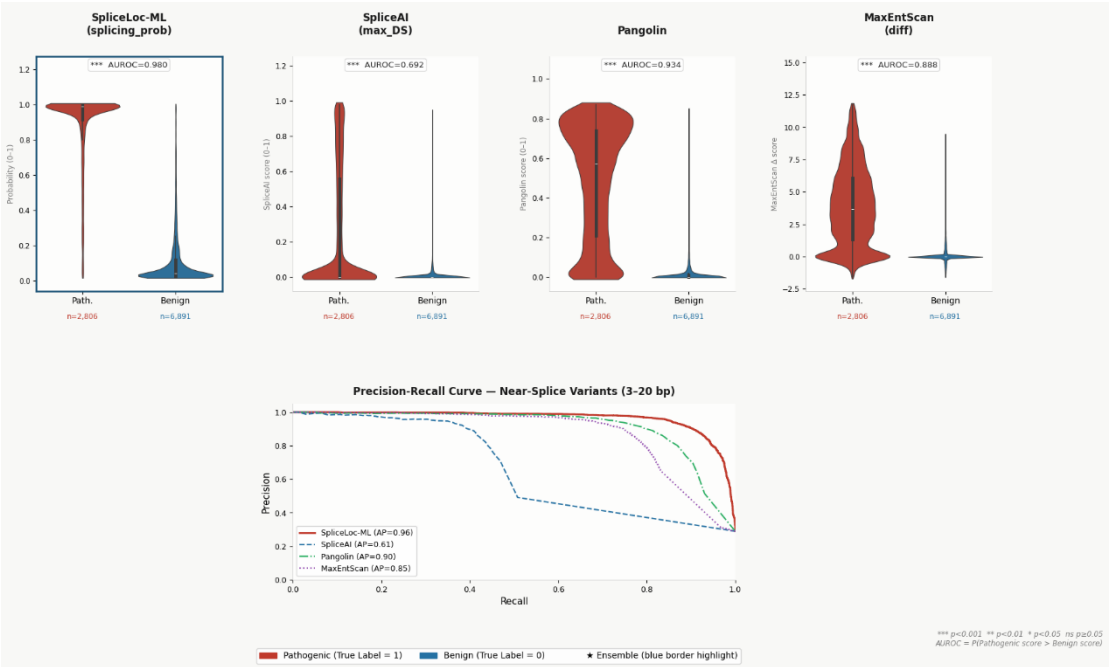


Figure 12| SpliceLoc-ML vs individual splicing tools in near-splice variants (3–20 bp). Violin plots showing pathogenicity score distributions between pathogenic (n=2,806) and benign (n=6,891) near-splice variants, with corresponding precision-recall curve. Significance thresholds: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns $p \geq 0.05$

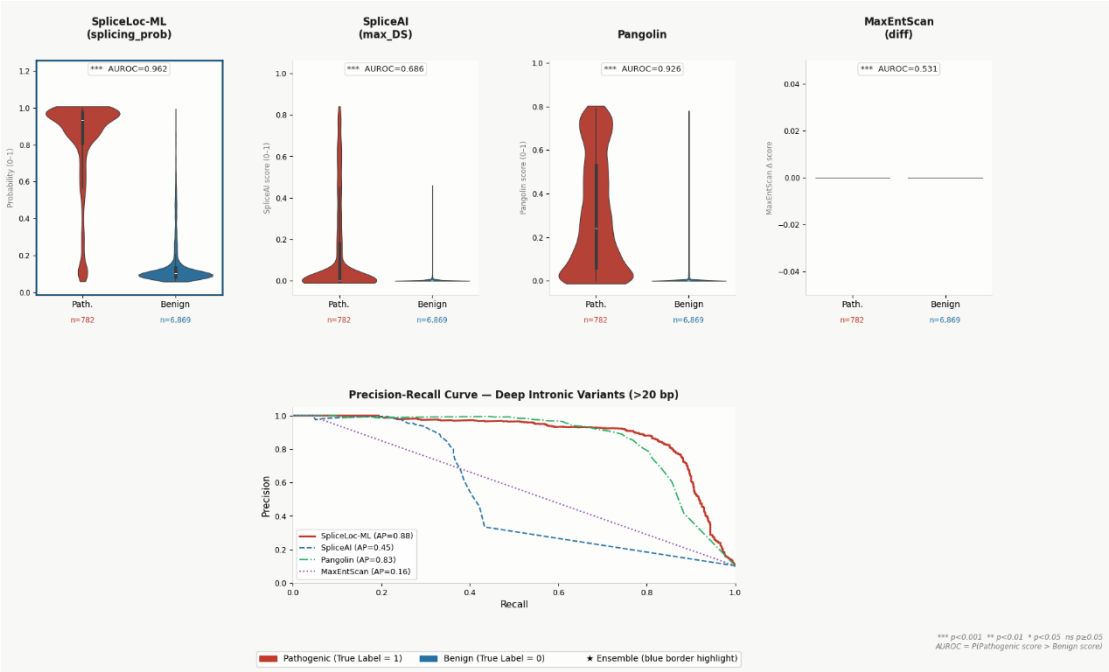


Figure 13| SpliceLoc-ML vs individual splicing tools in deep intronic variants (>20 bp). Violin plots showing pathogenicity score distributions between pathogenic (n=782) and benign (n=6,869) deep intronic variants, with corresponding precision-recall curve. Significance thresholds: *** p < 0.001, ** p < 0.01, * p < 0.05, ns p ≥ 0.05.

Table 8| Performance comparison of SpliceLoc-ML and individual splicing tools across genomic location categories. Higher values indicate better discriminative performance.

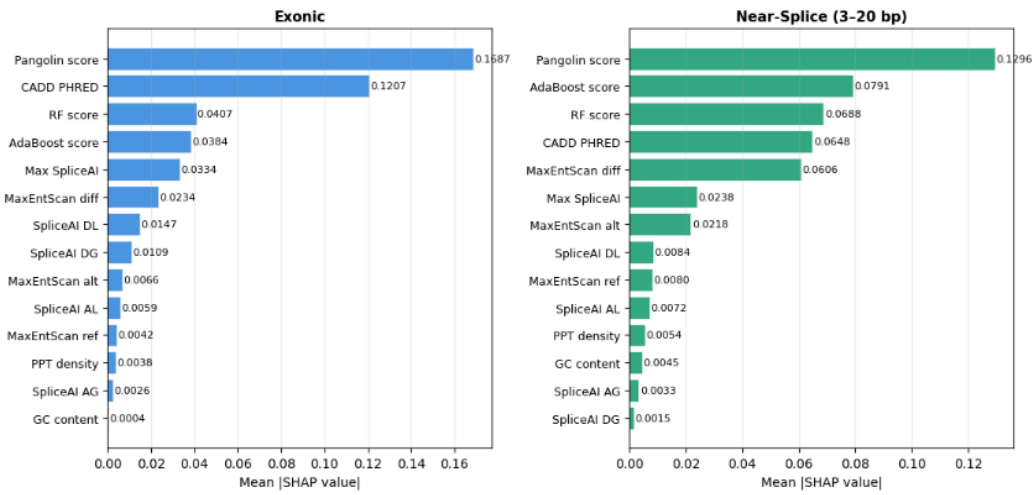
Category	Tool	AUROC	Avg. Prec	MCC
Exonic	SpliceLoc-ML	0.971	0.955	0.847
	SpliceAI	0.691	0.607	0.482
	Pangolin	0.932	0.900	0.757
	MaxEntScan	0.711	0.604	0.570
Near-splice	SpliceLoc-ML	0.980	0.964	0.856
	SpliceAI	0.692	0.607	0.499
	Pangolin	0.934	0.904	0.763
	MaxEntScan	0.888	0.853	0.741
Deep intronic	SpliceLoc-ML	0.962	0.884	0.771
	SpliceAI	0.686	0.445	0.377
	Pangolin	0.926	0.834	0.683
	MaxEntScan	0.531	0.157	0.235

4.5. Feature Importance

In Figure 14 SHAP values were computed for each variant-locus model on the held-out test set, it quantifies the marginal contribution of each feature to the model output for each variant, providing both the magnitude and direction of influence on pathogenicity prediction. Figure 14 shows the mean |SHAP| values representing overall feature importance per category.

Pangolin score was the most important feature across all three models, with mean |SHAP| values of 0.169 (exonic), 0.130 (near-splice), and 0.226 (deep intronic). This dominance reflects Pangolin's training on RNA-seq data from four tissues of four species, making it a biologically meaningful predictor of splice-disrupting variants in disease-relevant genes. CADD PHRED ranked second in the exonic (mean |SHAP| = 0.121) and deep intronic (0.092) models, indicating that evolutionary conservation serves as a complementary pathogenicity signal independent of splice site proximity. In the near-splice model, ada-score ranked second (0.079), followed by rf-score (0.069) and CADD PHRED (0.065), suggesting that ensemble meta-features capturing splice site boundary context contribute substantially in the 3–20 bp region.

MaxEntScan diff featured among the top five in the near-splice model with mean |SHAP| = 0.061, consistent with its established sensitivity to canonical splice site motif disruption in proximal intronic and exonic positions. In contrast, MaxEntScan features ranked among the lowest contributors in the deep intronic model with mean |SHAP| ≤ 0.001, confirms its known design limitations beyond canonical splice boundaries, as also reflected in its near-random discriminative AUROC of 0.531 in that category. SpliceAI delta scores showed low but nonzero SHAP contributions across categories, with greater relative importance in the deep intronic model where long-range splice site prediction becomes more relevant. GC content and PPT density consistently ranked lowest across all categories, yet their nonzero SHAP values confirm that local sequence composition features provide marginal discriminative signal in splicing tools.



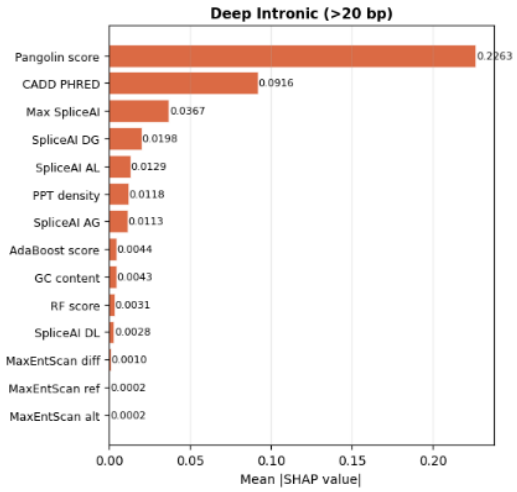


Figure 14| Mean |SHAP| feature importance rankings for SpliceLoc-ML across three genomic categories.

4.6. The five-Tier ACMG Classification Thresholds

Five-tier ACMG classification thresholds for each category-specific classifier were established by anchoring three boundaries to the published Bayesian posterior probability range. The fourth boundary (T_vus_lower) was derived from the training data using Youden's J index, it is calculated as (0.5275 for exonic), (0.4743 for near-splice), and (0.5125 for deep intronic).

Table 9| Five-tier ACMG classification thresholds for each category. Black = published constants from Tavtigian et al. (2018). **Red** = data-derived by Youden's J index on training set only.

Category	T_benign	T_vus_lower	T_likely_path	T_pathogenic
Exonic	0.10	0.5275	0.90	0.99
Near-Splice	0.10	0.4743	0.90	0.99
Deep Intronic	0.10	0.5125	0.90	0.99

4.7. Reclassification of the Full Dataset

The trained models were applied to the complete dataset of 27,994 variants to produce binary and five-tier ACMG classifications. Figure 15 shows the distribution of five-tier ACMG classification for the true benign and true pathogenic variants in the complete dataset. For true pathogenic variants, most of them were classified as pathogenic or likely pathogenic in all three categories, while for true benign variants, most of them were classified as benign or likely benign. The bar visualization in the figure confirms effective separation of the two classes under the five-tier system.

Of the 6,828 originally labeled pathogenic variants, 5,086 were reclassified as pathogenic or likely pathogenic by SpliceLoc-ML. The remaining 1,742 received lower classifications distributed between VUS and likely benign. Of the 21,166 originally labeled benign variants, 20,872 were confirmed as benign or likely benign, while 294 received higher classifications, reflecting the false positive rate.

For each category, the near-splice model confirmed 2,221 of 2,806 pathogenic variants as P/LP, the highest confirmation rate of the three categories. The exonic model confirmed 2,376 of 3,240, and the deep intronic model confirmed 489 of 782. The lower confirmation rate in the deep intronic category is not explained by the test-set sensitivity (0.872, similar to the exonic 0.866), and the high biological heterogeneity of deep intronic splice-altering mechanisms may contribute.

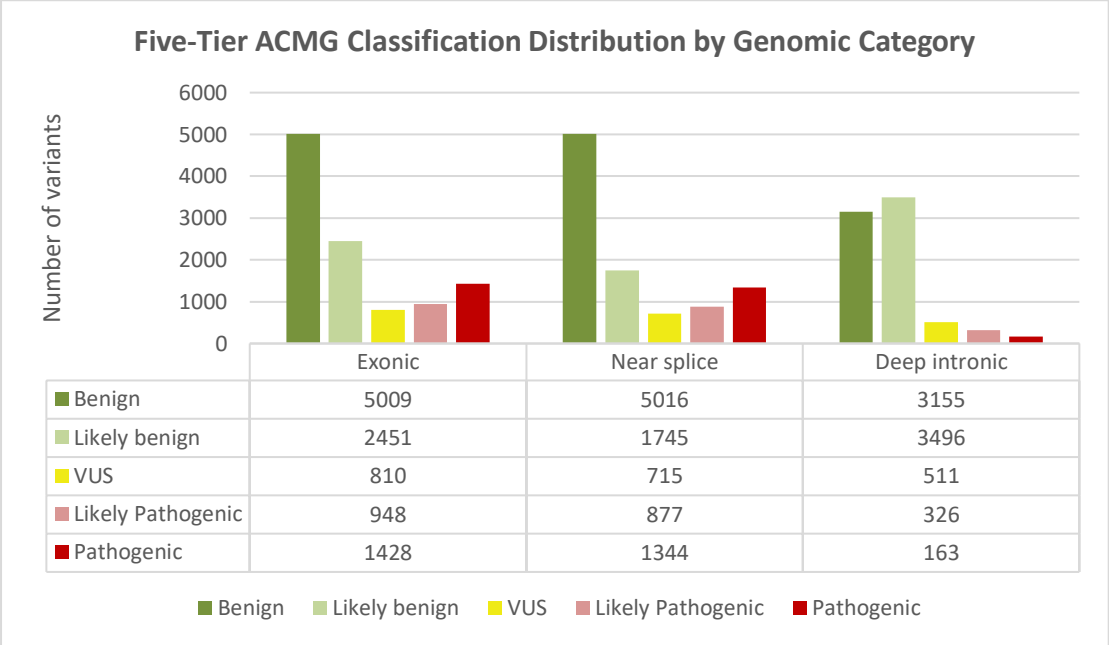


Figure 15| Distribution of the five-tier ACMG classification for all variants (n = 27,994 variants) per each category. Each graph shows the number of variants classified at each tier by using the SpliceLoc-ML model. There is a significant prevalence of benign or likely benign variants within all three genomic categories.

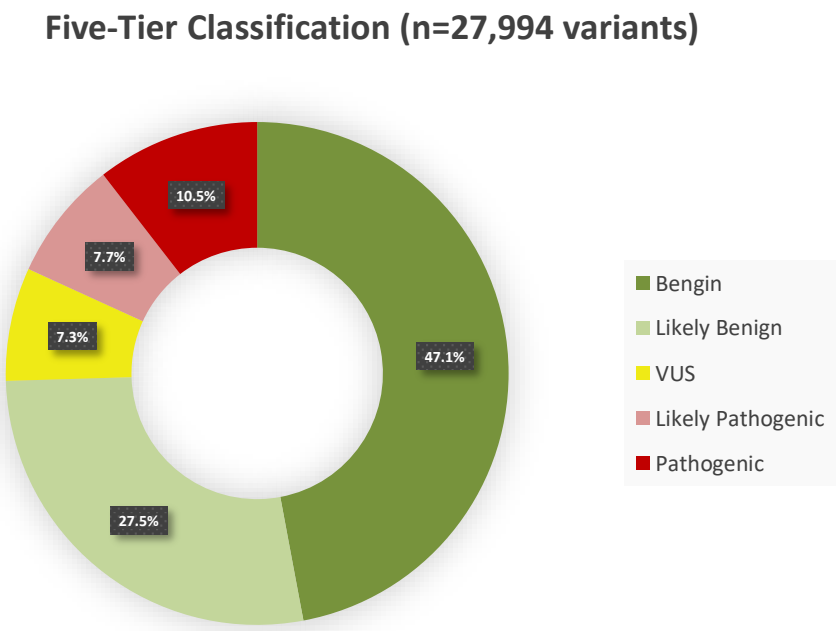


Figure 16|Distribution of the Five-tier ACMG classification in the total 27,994 variants in the full dataset. The Benign and Likely Benign variants include 74.6% of all classifications. The VUS rate of 7.3% indicates the need for further functional analysis to assign proper clinical significance.

4.8. Comparison with SpliceVarDB Experimentally Validated Variants

In order to analyze the relationship between model prediction and experimental ground truth, the detection rate of the SpliceLoc-ML model was compared between four laboratory validation methods of the pathogenic data from SpliceVarDB, RNA-sequencing, minigene assay, MaPSy, and MFASS. The detection rates are the ratio of pathogenic variations detected by the test that the model predicts as pathogenic or likely pathogenic. It was found a systematic decrease in the detection rate as the validation method moved progressively further from native biology. The highest detection rate was observed for RNA-sequencing validated variants 81.6% (4,212 of 5,163), which detect splicing changes using the RNA obtained from patient samples or from RNA sample that shows relevance to the disease under study, which represent the closest available to true in vivo splicing. The detection rate of the Minigene assay validated variants was 72.1% (240 of 333), corresponding to its near-native biological nature in which a genomic fragment sequence is expressed in a cellular system. The detection rate of MaPSy validated variants was found to be 60.6% (123 of 203), corresponding to the more artificial nature of this high-throughput method. The lowest detection rate was observed for MFASS validated variants 32.1% (340 of 1,060). The lower detection rate for these assays fits with the model reflecting native splicing better than reporter-based assays. It also fits with label noise in the

reporter-validated variants, and our data cannot separate the two explanations. MFASS uses synthetic reporter systems that express the protein in non-native cell lines and was shown to have the ability to detect minor splicing effects which may lead to no visible transcript changes in native human tissue. The features used in the current deep learning models (SpliceAI, Pangolin, and MaxEntScan) are trained to make predictions on native splicing changes, and therefore expected to show lower consistence with assays that measure a partially different biological mechanism.

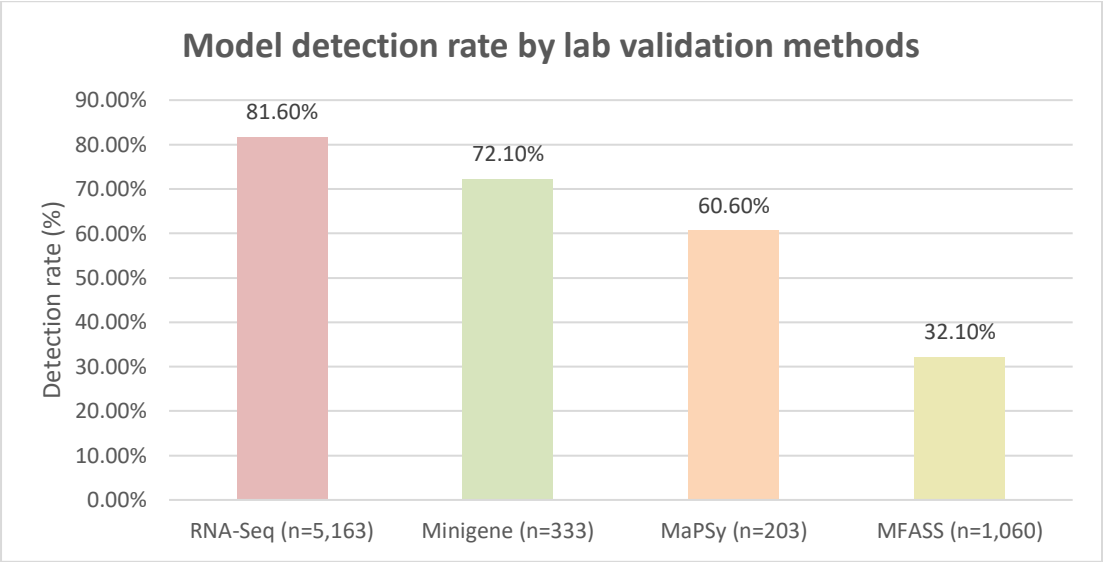


Figure 17|Model detection rate (ratio of laboratory-confirmed pathogenic variants classified as P/LP) by laboratory validation method.

Table 10|Model detection rate by laboratory validation method. **Green** = high detection, **amber** = intermediate, **red** = low detection. Assay context reflects proximity to native human splicing biology.

Validation method	Total lab-confirmed pathogenic	Detected by model (P/LP)	Detection rate	Assay
RNA-Sequencing	5,163	4,212	**81.6%**	Native tissue, in vivo
Minigene Assay	333	240	**72.1%**	Synthetic, near-native
MaPSy	203	123	**60.6%**	Synthetic, high-throughput
MFASS	1,060	340	**32.1%**	Synthetic reporter, non-native

Part B| Case study – Long QT Syndrome

The total number of ClinVar SNV for the 8 genes equal 7,892. The remaining 4,349 non-canonical splice variants were kept for reclassification. The near-splice category is the most dominant in all eight genes with total variants 4,133 representing the highest number of variants in each one. This reflects the expectation from biology since the extended splicing region is always more populated compared to the deep intronic region.

4.9. Reclassification of LQTS Variants

After the trained SpliceLoc-ML model was applied to the 4,349 LQTS non-canonical splice variants, the classification changed significantly when compared with the original ClinVar classification. Table 11 shows the full before and after comparison for all five-tier ACMG classifications.

Table 11| Summary of before and after reclassification of 4,349 LQTS non-canonical splice variants. ClinVar = original database classification. SpliceLoc-ML model = five-tier ACMG classification made by the ensemble RF model. Signs: ▲ = increase; ▼ = decrease.

ACMG Tier	ClinVar n	ClinVar %	Model n	Model %	Change
Benign	476	10.9 %	2,496	57.4 %	▲ 2,020
Likely Benign	2,571	59.1 %	1,587	36.5 %	▼ 984
VUS	489	11.2 %	212	4.9%	▼ 277
Likely Pathogenic	0	0	27	0.6 %	▲ 27
Pathogenic	0	0	27	0.6 %	▲ 27
Conflicting	70	1.6 %	0	0	▼ 70
No ClinVar data	743	17.1 %	0	0	▼ 743

The benign variants originally were only 476 variants (10.9%) with a ClinVar benign label, while the SpliceLoc-ML model classified 2,496 variants (57.4%) to this tier, an increase of 2,020 variants. The likely benign tier number decreased from 2,571 variants (59.1%) by ClinVar to 1,587 variants (36.5%) by the model (a reduction of 984 variants). Together, the two benign tiers now account for 93.9% of all classified variants, compared with 70.0% in ClinVar, reflecting the model's ability to resolve a large proportion of previously uncertain or unclassified entries as non-pathogenic with high confidence.

The VUS category was reduced by more than half, from 489 variants (11.2%) in ClinVar to 212 variants (4.9%) after SpliceLoc-ML model reclassification. ClinVar contained no likely pathogenic or pathogenic classifications among these non-canonical splice variants, the SpliceLoc-ML model identified 27 variants in each of these high-risk tiers (0.6%, 27 variants for each). The 70 ClinVar conflicting variants (1.6%) and 743 variants with no ClinVar data (17.1%) were each resolved into definitive five-tier classifications by the model, contributing to the reductions that were seen in the uncertain categories.

4.10. The five-Tier Classification Results per Gene

Table 12 and Figure 18 show the five-tier reclassification results according to each gene. The KCNQ1 and KCNH2 together have the largest numbers (1,548 and 1,450 variants, respectively), reflecting their central roles in LQTS genetics and the greater efforts of clinical sequencing that are directed at these genes. The VUS rates varied across genes, from 1.1% in KCNJ2 to 9.3% in TRDN. TRDN also showed the highest pathogenic count (n = 15) and the highest combined LP/Pathogenic rate (3.2%), which indicate that this gene may carry a disproportionately large number of variants that affect splicing and require further clinical investigation.

Table 12| Reclassified results with the five-tier ACMG reclassification by gene. Total = all variants that were filtered. VUS = variants requiring additional functional evidence. LP = Likely Pathogenic. SCN5A is not included as no variants passed the canonical transcript filtering for that gene.

Gene	Benign	Likely Benign	VUS	Likely Path.	Pathogenic	Total	% Path.
KCNQ1	1,075	412	48	8	5	1,548	0.8 %
KCNH2	703	661	74	6	6	1,450	0.8 %
TRDN	326	283	65	7	15	696	3.2 %
KCNJ2	200	69	3	1	0	273	0.4 %
CALM3	67	67	7	3	0	144	2.1 %
CALM2	69	57	7	2	0	135	1.5 %
CALM1	56	38	8	0	1	103	1.0 %
Total	**2,496**	**1,587**	**212**	**27**	**27**	**4,349**	**1.2 %**

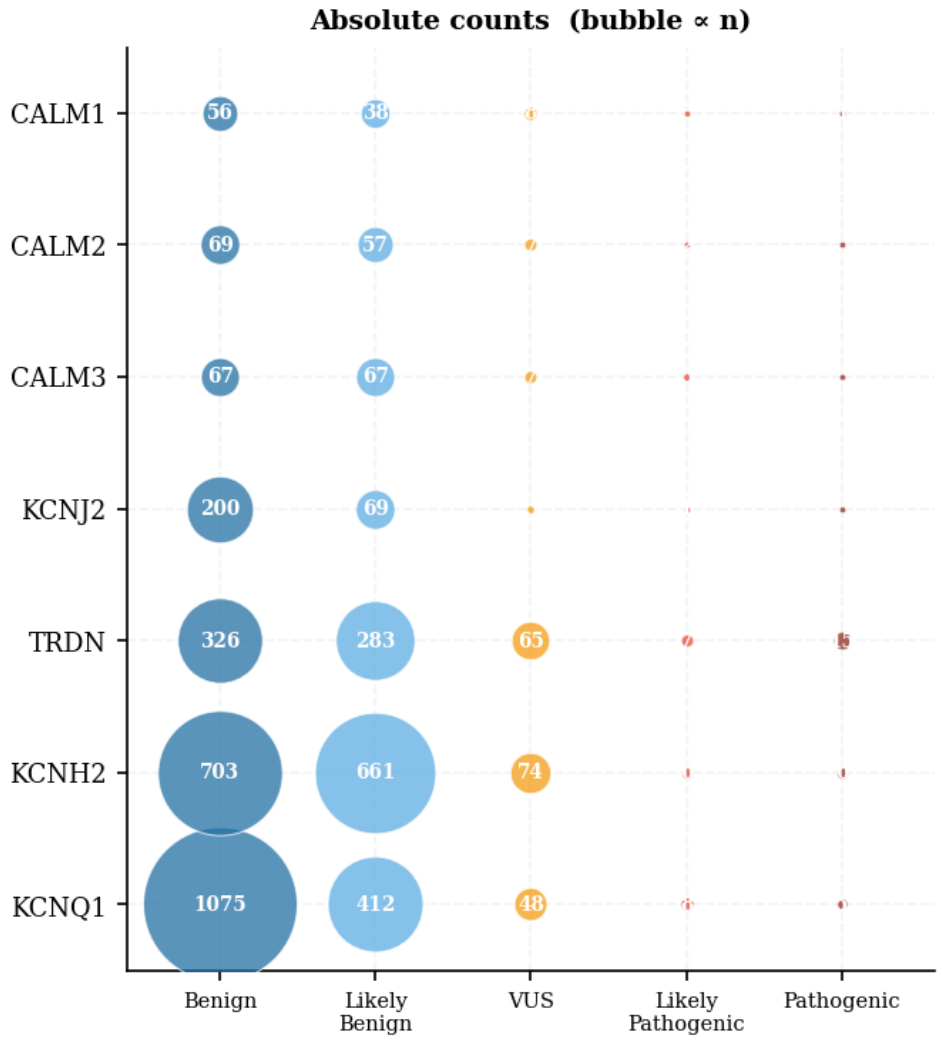


Figure 18| Numbers of five-tier ACMG classification per LQTS gene shown as a bubble matrix. KCNQ1 and KCNH2 have the highest number of variants. TRDN shows the highest rate of Pathogenic/Likely Pathogenic classifications rate (3.2%) and the highest VUS rate (9.3%).

4.11. Variant of Uncertain Significance Analysis

The 212 variants classified as VUS after SpliceLoc-ML model reclassification represent the point of uncertainty for the model, VUS where the posterior probability calculated by the model falls between the (T_vus_lower and T_likely_path) thresholds (0.47–0.53 to 0.90, depending on the category), where computational evidence alone is insufficient for confident classification.

Figure 19 presents four panels for analysis of these variants. Panel A shows the VUS count per gene. KCNH2 had the largest number of variants (n = 74), followed by TRDN (n = 65) and KCNQ1 (n = 48). The three calmodulin genes (CALM1, CALM2, CALM3) each contribute 7 to 8 VUS, and KCNJ2 had the fewest number (n = 3). Panel B shows that the majority of VUS are from the near-splice category (76.4%, n =

162) and 23.6% (n = 50) are from the deep intronic category. This distribution reflects the larger overall size of the near-splice variants category and confirms that deep intronic variants are disproportionately enriched for uncertain or pathogenic classifications relative to their total count.

Panel C shows the distribution of model splicing probabilities within the VUS group and categorized by original ClinVar classification. Across all ClinVar origin groups (ClinVar VUS, ClinVar B/LB, No ClinVar, and Conflicting), the overall median splicing probability is 0.631. This uniformly high position within the VUS probability zone (0.512–0.900) demonstrates that these 212 variants are not a residual noise group, they carry real and intermediate-strength splicing signals that could not be resolved further by the model's available computational features alone. Panel D shows the histogram of splicing probabilities across all 212 VUS. The distribution is broadly spread across the 0.512–0.900 range, indicating continuous variation in splice-disruption probability among the unclassified variants.

Only a small number of variants exceed the likely pathogenic threshold of 0.900, and very few exceed the pathogenic threshold of 0.990. These cases represent the highest-priority candidates for follow-up experimental work, specifically RNA sequencing from patient-derived tissue or minigene splicing assays.

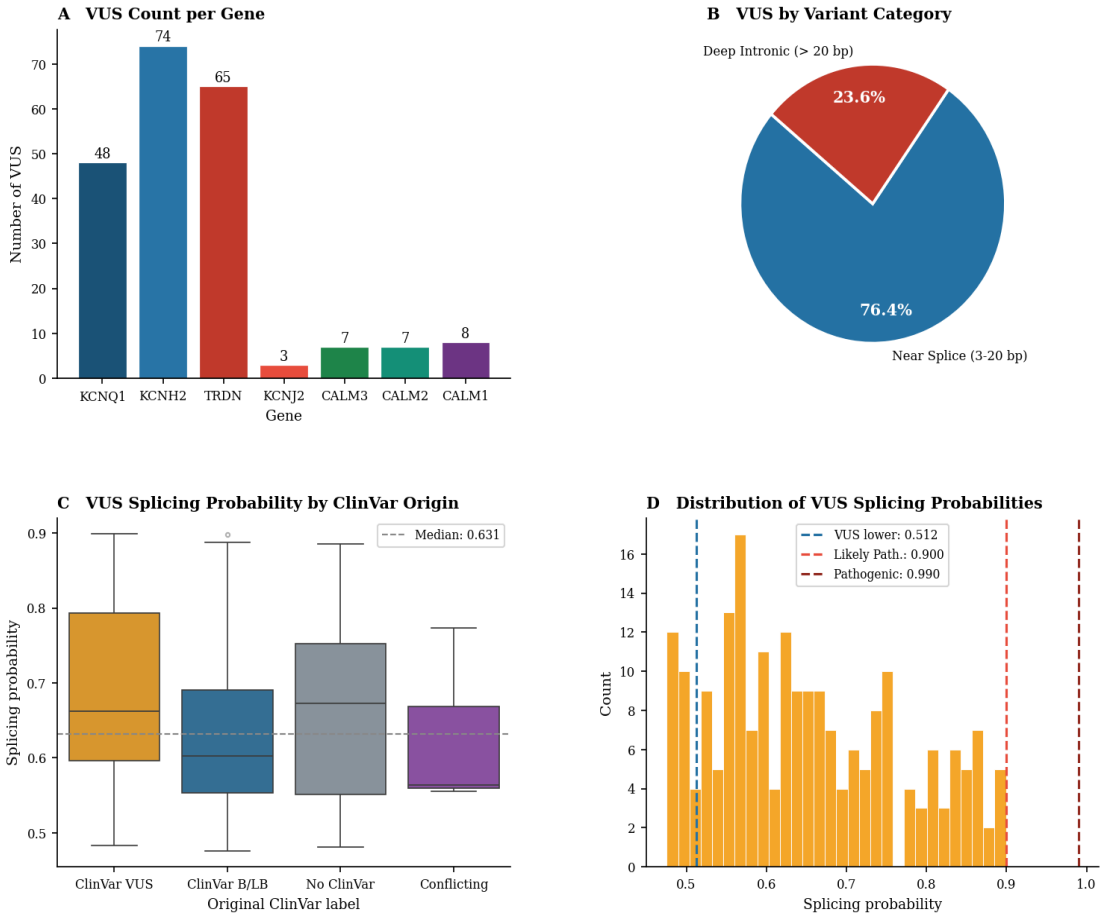


Figure 19|In-depth analysis of the 212 VUS variants. A: VUS count per gene, KCNH2 has the highest count (n = 74) and KCNJ2 the lowest (n = 3). B: VUS by variant category, 76.4% from the Near-Splice category. C: Model splicing probability stratified by original ClinVar label, overall median = 0.631. D: Histogram of VUS splicing probabilities with threshold lines at 0.512 (VUS lower), 0.900 (Likely Pathogenic), and 0.990 (Pathogenic).

4.12. Feature Score Distributions by ACMG classification

Figure 20 displays violin plots of four main model features (SpliceLoc-ML model splicing probability, SpliceAI delta score, CADD Phred, and Pangolin), separated by the five-tier ACMG classification assigned to the 4,349 LQTS variants. The sample sizes are identical for all four models (benign 2,496, likely benign 1,587, VUS 212, likely pathogenic 27, and pathogenic 27). These distributions confirm that the model is working properly, if the model is performing correctly by capturing real biological signal, then the feature scores should increase gradually from benign to pathogenic categories, reflecting real biological signals.

For model splicing probability, the expected gradually increase is clearly observed, median probabilities rise from approximately (0.05 benign), to (0.15 likely benign)

and (0.63 VUS) to above (0.90 in likely pathogenic and pathogenic), it shows the clearest tier separation.

The SpliceAI score shows near-zero median values in the benign (~ 0.01) and likely benign (~ 0.03), with only minimal outliers in each group. This confirms that most low-risk variants do not lie within the positional windows that SpliceAI is most sensitive to. The VUS shows a broader distribution, a teardrop shape extending from near zero up to approximately 0.65 with a median of around 0.14, reflecting the intermediate and variable splice-site proximity of this group. The likely pathogenic shows a large, wide violin spanning the full 0 –1.0 range with a median near 0.30, indicating a heterogeneity in the SpliceAI signal in the variants that are confidently classified as high-risk by the ensemble model. The pathogenic has a similarly wide distribution but with a considerably higher median of approximately 0.67, showing that the most confident pathogenic tend to carry stronger SpliceAI signals.

CADD Phred scores show the most gradual increase across tiers. The benign has a broad bottom-heavy distribution with a median of approximately 3 and a long upper tail reaching ~15, reflecting the wide range of possible CADD values even at low-probability positions. The likely benign is taller and slightly narrower with a median near 8. The VUS has a tall, tapered violin with a median of approximately 13. The likely pathogenic is compact and narrow with a median near 17, and the pathogenic is similarly compact with a median of approximately 21. This consistent stepwise progression across all five tiers driven by a tool that operates entirely independently of splicing mechanics, provides particularly strong evidence that the five-tier classification reflects real differences in variant deleteriousness.

The benign and likely benign Pangolin scores both show distributions that are tightly concentrated at or near zero, with medians of approximately 0.01 and 0.02 respectively. This absence of Pangolin signal in the low-risk classes reflects Pangolin's high specificity for real splice-disrupting variants. The VUS begins to show a modest signal with a narrow teardrop distribution and a median of approximately 0.10. The likely pathogenic shows a large, broad violin spanning from near zero to approximately 0.85, with a median near 0.16, reflecting heterogeneity within this class in terms of Pangolin-predicted splice disruption. The pathogenic is markedly different, a compact violin with a median near 0.64 and a distribution concentrated above 0.40. This sharp jump between the likely pathogenic and

pathogenic in the Pangolin score is consistent with Pangolin being the most important feature in the trained ensemble models and explains why crossing the 0.990 pathogenic threshold requires a strong multi-tool.

In all four features, the directional trend from benign to pathogenic is consistent and statistically expected, confirming that the model's five-tier classification aligns with independent biological evidence. The differing shapes of the violins from the extremely tight benign splicing-probability violin to the broad, heterogeneous SpliceAI and Pangolin violins in the high-risk classes, also explain why a single-tool threshold approach would be insufficient, variants in the LP and pathogenic carry strong ensemble-level signals even when individual tool scores vary widely.

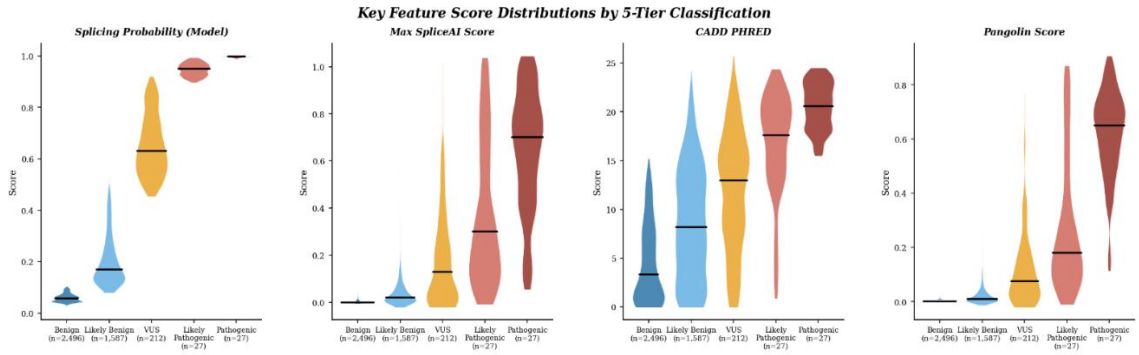


Figure 20|Violin plots of four key feature scores stratified by ACMG five-tier classification (Benign n = 2,496, Likely Benign n = 1,587, VUS n = 212, Likely Pathogenic n = 27, and Pathogenic n = 27). The models are: SpliceLoc-ML model splicing probability, maximum SpliceAI delta score, CADD PHRED score, and Pangolin score. Horizontal black lines indicate tier medians. All four features show a monotonic increase from benign toward pathogenic, providing convergent independent validation of the model classification.

Chapter 5| Discussion

5.1.Overview

This research developed, validated, and clinically applied SpliceLoc-ML model for the pathogenicity classification of non-canonical splice variants, a clinically significant but computationally neglected class of genomic variants. By training three category-specific Random Forest (RF) models on experimental dataset of 27,994 variants collected from SpliceVarDB and gnomAD v4.1 and integrating biological features covering splice-site disruption, evolutionary conservation, and local sequence context, the model achieved high discriminative performance in all three genomic location categories. Then 4,349 non-canonical splice variants from 8 Long QT Syndrome (LQTS) genes were reclassified by this model, confirming its clinical applicability.

5.2.Dataset Design and Class Imbalance Management

The construction of a high-quality, experimentally validated dataset was foundational to the accuracy of the current model. The use of SpliceVarDB as the positive control source ensured that all 6,828 pathogenic variants were supported by functional splicing evidence rather than using computational prediction data alone. This methodology is important because training a classifier on variants with pathogenicity is computationally predicted would introduce circularity, especially when the same tools that were used for predicting pathogenicity serve as features. So, using experimentally confirmed splicing disruption in this study avoids such problems and measures actual performance. The use of common gnomAD v4.1 variants with allele frequency more than 0.001 as negative controls is well-established in the literature of clinical variant interpretation. According to the principle of Mendelian disease genetics, if the frequency of a variant is significant in the population, then it can be considered benign since it cannot have high penetrance because of its pathogenicity.

The class imbalance issue, specifically in the deep intronic category reaching 8.8:1 benign to pathogenic ratio, it was solved by two-stage solution of random downsampling for exonic and near-splice categories and random downsampling combined with cost-sensitive class weighting for the deep intronic category. If pure downsampling were applied until the class ratios equal 1:1 for deep intronic, more than 6,000 benign variants would be lost, which could lead to biased results. Instead, downsampling until a 3.84:1 ratio with inverse frequency weighting is preferred for an imbalanced class with biological heterogeneity. On the test set the deep intronic model had a sensitivity of 0.872 and a specificity of 0.953, so the class imbalance did not lead to low sensitivity in this category. The tier-level rate uses a stricter probability threshold (0.90 for Likely Pathogenic) than the binary cutoff (0.512), and we did not investigate it further. Heterogeneity of the assays (Section 5.10) may also contribute.

5.3.Model Performance and Generalization

5.3.1. Absolute Performance

All three category-specific RF models exceeded the pre-defined minimum performance thresholds across all 6 metrics, with AUROC values of 0.969, 0.975, and

0.936 for the exonic, near-splice, and deep intronic models respectively. These values are higher than those for any individual splicing tool on non-canonical variants, establishing the SpliceLoc-ML model as the most accurate system evaluated in this study for this variant class. The MCC values (0.837, 0.853, 0.811) have additional relevance because it provides a more robust assessment of model performance than standard metrics, and by integrating all four parameters of the confusion matrix, the MCC remains unaffected by class imbalance, ensuring that the high predictive accuracy is not an artifact of the majority class distribution, unlike accuracy or F1 score (Chicco & Jurman, 2020). MCC values above 0.80 indicate that the models are truly and significantly capturing biologically meaningful signals across both classes.

The near-splice category yielded the best overall performance, which makes biological sense. These variants are within the extended splice site sequence motifs that serve as targets for several splicing control mechanisms (polypyrimidine tract, branch point sequences, and splicing enhancers). This region contains the most information regarding SpliceAI's delta scores, MaxEntScan's position weight matrices, and Pangolin's prediction results, providing the highest multi-featured signal for the classification model. As a result, the near-splice model benefits from the most complementary and redundant features compared to the other two categories, which is a well-established driver of ensemble classifier performance (Breiman, 2001).

The deep intronic model has a lower AUROC 0.936 that reflects the complexity associated with the classification of distant variants located more than 20 bp from annotated splice sites, this is due to the heterogeneous nature of the disease mechanisms. However, this score remains significantly superior to any of the individual scores in the individual tools, with the next one being Pangolin AUROC = 0.926 in this analysis. The high NPV of the deep intronic model 0.966 is clinically relevant, it indicates that variants with a low predicted probability by the model are highly unlikely to be pathogenic, so the chance of that variant causing disease is low.

5.3.2. Train-Test Methodology and Absence of Overfitting.

The separation of the test set from all model development steps, including feature imputation, threshold selection, and class weight calculation, all of which were derived exclusively on the training dataset strengthens the model. The Youden's J binary thresholds 0.5275 exonic, 0.4743 near-splice, and 0.5125 deep intronic were

therefore truly predicted to the test set. In contrast, several published variant classification tools that derive optimal thresholds from the same data used to evaluate them, introducing optimistic bias in reported sensitivity and specificity values. The reproducibility of the present results is further supported by the fixed random seed (RANDOM_STATE = 42) applied to all probability steps, enabling exact replication from the same data.

5.4.Comparison with Individual Splicing Tools

5.4.1. SpliceAI

SpliceAI achieved AUROC values of 0.691, 0.692, and 0.686 for exonic, near-splice, and deep intronic models respectively, which are significantly lower than the ensemble model and Pangolin. This finding is not in conflict with the original SpliceAI publication, which reported a precision-recall AUC of 0.98 based on canonical splice site variants. This study analyzed SpliceAI on non-canonical variants only, a type of variants that was specifically left out during the original training and evaluating of SpliceAI. The lower performance of SpliceAI on non-canonical exonic variants is particularly notable because its delta scores were specifically designed to capture changes in acceptor and donor probability at canonical positions. When the variants are positioned within the exon's boundaries or farther from the canonical ±1/±2 boundaries, the receptive field of SpliceAI's dilated convolutional network (10 kb) is not the critical factor but the biological significance of the delta score calculation is diminished. This interpretation is consistent with the violin plot analysis, where both pathogenic and benign SpliceAI scores cluster near zero across all three categories, indicating that the majority of non-canonical variants in the dataset are not associated with significant delta score differences regardless of their true pathogenicity.

These findings have direct practical implications for clinical genomic laboratories that use SpliceAI as a primary filter for splice variant classification. The present results suggest that SpliceAI alone would produce a high false negative rate for non-canonical variants and should be supplemented with complementary tools and ensemble-based approaches.

5.4.2. Pangolin

Pangolin had the highest performance score in all the three categories when compared with the individual tools, achieving AUROC values of 0.932, 0.934, and 0.926. Pangolin was trained on splice site usage measured by RNA-seq in four tissues from four species, so it can model tissue-specific regulation of splicing beyond the motifs captured by MaxEntScan (Zeng & Li, 2022). The high performance of Pangolin in the intronic region may be due to its capability to detect mutations causing disruptions in the regulatory sequences within introns for splice site selection across large regions. So, any increase in performance of the SpliceLoc-ML model over Pangolin for exon and near-splice classes likely arise from the complementary nature of information from MaxEntScan, CADD, and sequence features that provide additive discriminative information beyond what Pangolin alone captures.

The consistency of the present Pangolin AUROC values (0.932–0.934) with those reported by Zeng et al. (2022) in their independent standard serves as important external validation that the present dataset and pipeline are correctly implemented and that the results are generalizable beyond this study's specific sample.

5.4.3. MaxEntScan

MaxEntScan demonstrated the most strongly location-dependent performance of any tool evaluated, with AUROC values of 0.711 exonic, 0.888 near-splice, and 0.531 deep intronic. The near-splice performance is consistent with MaxEntScan's design, its maximum entropy position weight matrices were derived from experimentally characterized 5' and 3' splice site sequences and are most informative for variants that alter the thermodynamic quality of canonical or near-canonical splice site motifs. The lack of discriminative ability in deep intronic variants (AUROC = 0.531, $p < 0.001$) is not a failure of the analysis but rather it represents a successful verification of the mechanistic properties of the algorithm. MaxEntScan's scoring window does not extend to deep intronic positions, and the scores that were observed to be near-zero for both classes in that panel correctly reflect that this tool has no informative signal for variants far from annotated splice boundaries. In this context, we can see that choosing to use the MaxEntScan features among others in the classification ensemble was the right choice, because MaxEntScan provides maximum information in the near-splice category as the second most important feature according to SHAP-based feature importance, while other features cover the gaps in information for MaxEntScan.

5.5.Feature Importance and Biological Interpretability

The consistent dominance of Pangolin score as the top-ranked feature across all three models, with mean |SHAP| values of 0.169 exonic, 0.130 near-splice, and 0.226 deep intronic, is biologically sense and strengthens the external validation provided by the tool comparison. Pangolin's multi-tissue training enables it to encode splicing regulatory information across diverse biological contexts, making it the single most information-rich feature available in the feature set. The higher importance in deep intronic variants (0.226) compared to exonic variants (0.169), in the deep intronic, the other tools contribute less discriminative signal, so the Random Forest algorithm correctly places greater weight on the one feature that retains predictive power. SHAP values quantify the contribution of each feature to individual predictions, so when one feature dominates, the mean |SHAP| value reflects its consistent influence across all test variants.

CADD PHRED ranking second in exonic and deep intronic models reflects the importance of evolutionary conservation as a complementary pathogenicity signal. Variants that disrupt splicing often also fall in regions of high conservation across vertebrate genomes because the functional consequences of splice disruption, like loss of protein domains or nonsense-mediated decay, are strongly selected against. CADD combines signals of conservation (PhyloP, PhastCons), regulatory annotations, and protein-level features, providing a multi-dimensional conservation signal. This is precisely the characteristic that makes it a valuable ensemble feature, it captures biological signals that are not accounted for by splicing-specific tools. Also, the current comparison does not include CADD PHRED score in the violin plot across tools specifically to avoid categorizing a general pathogenicity model as a splicing tool while retaining it as an ensemble feature.

The contribution of GC content and PPT density, while not large but it confirms that local sequence context features provide discriminative signals beyond what existing tools capture. High PPT density near the 3' splice site is a well-established predictor of strong splice acceptor function, and GC content plays a role in the formation of secondary structures in the RNA and the kinetics of spliceosome assembly. Their contribution to the predictive model is biologically sound, and their non-zero mean |SHAP| values confirm that the Random Forest identified true biological signal rather than noise in these features.

5.6.ACMG Five-Tier Classification Framework

5.6.1. Threshold Derivation

The combination of the Tavtigian Bayesian prior probability model with data-derived Youden's J thresholds represents a clinically principled approach to translating continuous model probabilities into the five-tier ACMG classification system. By setting the thresholds for T_benign 0.10, T_likely_pathogenic 0.90, and T_pathogenic 0.99 based on the published Bayesian posterior probability ranges, the outputs from the classification model are easily interpretable. The T_vus_lower threshold values obtained from the data 0.5275, 0.4743, and 0.5125 represent the boundary at which the model's discrimination is maximized, as determined by Youden's J criterion on the training dataset. The small differences between categories are due to the different class distributions and feature information content in each genomic category.

5.6.2. Full Dataset Reclassification

The reclassification of the complete 27,994-variant dataset showed that the model predictions agreed with the ground-truth labels: 74.5% of the originally pathogenic variants (5,086 of 6,828) received a P/LP classification, and 98.6% of the benign variants (20,872 of 21,166) received a B/LB classification. The remaining 25.5% of the pathogenic variants (1,742) received intermediate or lower classifications. Their pathogenicity in the original SpliceVarDB annotation was supported by one or more of the four methods (RNA-seq, minigene assay, MaPSy or MFASS). Because this dataset includes the 70% of variants used for training, these figures describe agreement with the labels, and the held-out test results in Table 7 remain the unbiased estimate of performance.

The VUS rates on the test set (10.3% exonic, 9.6% near-splice, and 9.9% deep intronic) are relatively low due to the strong discrimination of the SpliceLoc-ML model, and it classifies most variants confidently into the pathogenic or benign probability zones.

5.7.The SpliceVarDB Laboratory Validation methods and Model Calibration

The rate of models detection using all four in vitro tests, RNA-sequencing 81.6%, minigene assay 72.1%, MaPSy 60.6%, and MFASS 32.1%. This gradient can be

explained in two ways: either the model is calibrated to native splicing, or some of the reporter-validated labels do not reflect splicing disruption in native tissue. The exclusion analysis in 5.10 is needed to tell them apart.

RNA sequences from patient or diseased tissue allow us to analyze the actual splicing environment in human cells, including the regulatory environments of specific tissues, the entire assembly of the spliceosome, and the complete set of splicing regulatory proteins expressed in the relevant tissue. The high level of detection 81.6% of variants that have been confirmed by RNA-seq indicates that there is a high degree of similarity between the output of the model and the gold standard of splice-disruption evidence. The lower detection rate for minigene assays 72.1% is consistent with the known limitations of minigene systems, specifically that the construct only includes genomic sequences from several hundred base pairs to a few kilobases and this may lead to missing distal regulatory sites affecting splice variation when present in the native genomic sequence. On the other hand, the lower MaPSy detection rate 60.6% reflects the high-throughput but lower-resolution nature of this method, which uses competitive splicing of synthetic constructs and may detect precise isoform shifts that do not represent clinically significant splice disruption in native tissue. The MFASS detection rate 32.1% is very low because it uses synthetic reporter constructs in non-native cell lines to measure exon inclusion rates and is capable of detecting very small splicing changes, including minor isoform shifts that may have negligible functional consequences in the disease-relevant tissue.

All the features used in the current ensemble, SpliceAI, Pangolin, and MaxEntScan were all trained to predict biologically consequential splicing changes, not minor disruptions in a reporter assay. The expected agreement between these features and MFASS only validated variants is low, and the 32.1% detection rate is consistent with published observations that MFASS has a higher sensitivity but lower specificity for clinically relevant splice disruption compared to RNA-sequencing. Whether this gradient reflects calibration to native splicing or label noise cannot be resolved with the present data (Section 5.10).

5.8.Clinical Application: Long QT Syndrome

5.8.1. VUS Resolution

We queried eight LQTS-associated genes. After filtering to MANE Select transcripts and single nucleotide variants, 4,349 non-canonical splice variants from seven genes remained, because no SCN5A variant passed the filter. SpliceLoc-ML placed 212 variants (4.9%) in the VUS tier, while 489 (11.2%) were VUS in ClinVar, a net decrease of 277 variants (56.6%). The two VUS sets come from different classifiers and are not nested, so this is a change in the number of variants per tier and not the resolution of individual variants. The model placed 27 variants in the Likely Pathogenic tier and 27 in the Pathogenic tier. None of the 4,349 variants was Likely Pathogenic or Pathogenic in ClinVar, because these were removed before the analysis (Section 3.12). The tiers come from applying the posterior probability thresholds of Tavtigian et al. (2018) to the model output, so they are based on computational evidence only. In the ACMG/AMP framework, computational prediction counts as supporting evidence (PP3), and with calibrated thresholds it can reach strong evidence (Pejaver et al., 2022; Walker et al., 2023). It cannot establish a Likely Pathogenic or Pathogenic classification alone. We had no functional assays, segregation data or case-level evidence for any of these variants, so the tiers are not clinical classifications, and the 54 variants should be seen as candidates for further testing. The move of variants to the benign tiers is also computational and was not validated experimentally. For a disease with a risk of sudden cardiac death, this can help to decide which variants to test first.

5.8.2. Gene-Specific Findings

There was quite an inconsistency in the LP/Pathogenic classification ratios across different genes, with the lowest being 0.4% for KCNJ2 to the highest 3.2% for TRDN. The protein encoded by the TRDN (Triadin) gene is part of the cardiac ryanodine receptor complex, and pathogenic variants in this gene are associated with Triadin Knockout Syndrome, a rare but severe form of LQTS with a high risk of exercise-induced sudden cardiac arrest. The fact that TRDN has a disproportionately high proportion of the pathogenic load (15 pathogenic, 7 likely pathogenic out of 696 variants, representing 3.2% at the pathogenic tier alone) is consistent with the known clinical severity of TRDN-related disease and the biological expectation that a gene with severe disease consequence is under stronger selection against functional variants, including those affecting splicing. The high VUS rate in TRDN (9.3%) also

indicates that this gene needs more functional validation studies, particularly given the high clinical risks of missed pathogenic variants.

KCNJ2 showed the lowest VUS rate (1%) and the lowest LP/Pathogenic rate (0.4%), consistent with the known relatively mild phenotype (Andersen-Tawil Syndrome) and the enrichment of loss-of-function coding variants rather than splice-disrupting variants as the primary pathogenic mechanism in this gene. SCN5A was excluded from the analysis due to the absence of variants passing the canonical transcript filtering, which likely reflects the unique splicing complexity of this gene.

5.8.3. Features and Model Validation

The violin plots of feature score distributions stratified by classification tier for LQTS variants provide an independent internal validation of the quality of the model's classification. The observation that Pangolin score, SpliceAI delta score, CADD PHRED, and ensemble probability all show monotonically increasing median values from benign through pathogenic tiers across independently annotated features confirms that the model's tier classification results are consistent across different biological aspects. These observations are not necessarily mathematically proven because the model may give very high probability estimates for variants based on one dominant feature regardless of other features that do not support this classification. The consistency of the gradient detected during LQTS feature analysis indicates a true biological signal and substantially reduces the concern that model performance is driven by overfitting to a single feature or by the training data itself.

5.9.Comparison with Published Frameworks

The current method can be compared with several recently published approaches for non-canonical splice variant classification. PDIVAS (Kurosawa et al., 2023) is the most methodologically similar published tool, employing a Random Forest trained on experimentally validated splice variants with Pangolin as the primary feature. PDIVAS reported an average precision of 0.92 and MCC of 0.88 on deep intronic variants. The current SpliceLoc-ML model achieved average precision values of 0.955 exonic, 0.964 near-splice, and 0.884 deep intronic, with MCC values of 0.847, 0.856, and 0.771 respectively. We cannot make a comparison between numbers due to

the varying nature of datasets used, variant filtering criteria, and the use of category-specific versus unified models.

PDIVAS only evaluated the deep-intronic variants (≥ 50 bp from splice sites) and reported a SpliceAI AUROC of 0.87 on that dataset, while the lower SpliceAI AUROC achieved by this study's ensemble model is 0.686–0.692, so three factors may lead to this lower value, a less benign allele frequency threshold (AF > 0.1% in the ensemble model vs ≥ 5% in the PDIVAS), leading to a more challenging to distinguish benign population, a pathogenic dataset with greater mechanistic diversity due to the incorporation of variants from SpliceVarDB that include exonic, near-splice, and deep-intronic validated variants not just deep-intronic pseudoexons alone, and the inclusion of exonic variants for which SpliceAI was not trained to predict. However, the higher performance of Pangolin and MaxEntScan in this study compared to PDIVAS is due to the near-splice category, which belongs to the optimal scoring range of both tools and was entirely excluded from PDIVAS evaluation. These differences highlight a major limitation of single-category comparison, a tool optimized for deep-intronic variants will outperform other tools in this specific category but cannot predict all non-canonical splice-altering variants. This work solves this gap by training category-specific RF models for all three non-canonical variant classes and showing that the SpliceLoc-ML model achieves an AUROC of 0.958–0.980 across all categories together, which is a result that is not achievable by any individual tool and not addressed by PDIVAS. Then the application to LQTS variant reclassification provides additional evidence on its translational potential.

The unique strength of the current model that makes it different from other models is its stratification into three location categories, each with its own model and threshold. PDIVAS and most comparable tools apply a single unified model across all variants, which may result in a standard decision boundary for all variants with different feature distributions and class ratios. This study provides proof that category-specific performance improvements are achievable, particularly in deep intronic variants where specialized features (PPT density, GC content) and adjusted class weighting are essential. The deep intronic AUROC of 0.936 represents a significant achievement over the performance of any single tool in that category and reflects the design advantage of category-stratified modeling.

5.10. Limitations

There are some limitations in this study that should be acknowledged. First, the models were trained and evaluated on single-nucleotide variants only. Indel mutations affecting splice regulatory sequences represent a clinically significant variant class that is not covered by this model.

SpliceAI and Pangolin scores were used as input features, and both tools are themselves trained models. SpliceAI was trained on GENCODE-annotated splice sites from genes on chromosomes 2, 4, 6, 8, 10–22, X and Y, and tested on genes from chromosomes 1, 3, 5, 7 and 9 (Jaganathan et al., 2019). Pangolin was trained on splice site usage measured by RNA-seq in four tissues from human, rhesus macaque, rat and mouse, with the same chromosome split for human genes (Zeng & Li, 2022). Neither tool was trained on labelled pathogenic or benign variants, and SpliceAI saw only the reference sequence. Still, we did not select our variants by chromosome, so part of them lie in regions on which these tools were trained. Jaganathan et al. (2019) found no significant difference in validation rate or sensitivity between variants on training and test chromosomes, but we did not repeat this comparison for our ensemble model. For this reason, we cannot exclude that part of the performance of SpliceLoc-ML comes from leakage. Evaluating the model only on variants from chromosomes 1, 3, 5, 7 and 9, or retraining on the sequence features behind these tools instead of their final scores, would show how large this effect is.

A further limitation is the composition of the pathogenic variant set used for training and evaluation. The SpliceVarDB source includes variants validated by MFASS and MaPSy, two high-throughput synthetic reporter assays that operate outside the native cellular environment and are known to detect minor isoform shifts that may not produce clinically significant splicing changes in native human tissue. As shown in the section 4.8, the SpliceLoc-ML model's detection rate for MFASS-validated variants was only 32.1%, compared to 81.6% for RNA-sequencing-validated variants, which may reflect a mismatch between what these assays measure and what the model is calibrated to predict. The inclusion of these variants in the pathogenic training set may therefore have introduced label noise, so some variants annotated as pathogenic based on synthetic assay evidence which do not represent true in vivo splice-disrupting events. Label noise of this kind would lower the measured sensitivity and could explain part of the lower confirmation rate in the deep intronic category, but we did not measure its effect. MFASS and MaPSy validated variants

make up as many as 1,263 of the 6,828 pathogenic variants (18.5%). Retraining without them and comparing performance on the remaining variants would test this directly.

Also, the negative control set was constructed using common SNVs from gnomAD v4.1 with allele frequency greater than 0.001. This approach follows the frequency-based logic of the ACMG/AMP benign criteria. It is based on this strategy, a variant causing severe Mendelian disease cannot persist at high frequency in the general population due to natural selection. The gnomAD v4.1 dataset itself, despite it is the largest population variant database currently available, still underrepresents certain populations, which could create population-specific annotation gaps. In addition, some genes are associated with variable expression and incomplete penetrance, meaning that low-penetrance pathogenic variants might reach allele frequencies above 0.001 in the general population without being selected against. If any such variants were included in the gnomAD control set, they would be incorrectly labeled as benign, introducing false negatives into the negative class.

5.11. Future work

There are several areas that can be considered for future research that could enhance this current study. First of all, the use of deep learning methods for sequence representation, such as those based on RNA foundation models or transformers trained on large splice sequences datasets, could provide more informative features to be used as inputs to the model than the current feature set obtained through scoring systems. Recent studies show that machine learning-based embedding techniques can capture some regulation not covered by position weight matrices and delta scoring methods (Chen et al., 2022; Dalla-Torre et al., 2025), so their integration as new features to complement the Random Forest model is a logical extension of this study.

The 54 variants placed in the Likely Pathogenic or Pathogenic tiers have not been tested functionally. A first step is to review the ClinVar entries, published case reports, inheritance and population frequency of each variant. Where samples are available, RNA analysis or minigene assays would test the predicted splicing effect directly. A retrospective comparison of the model output with ClinVar variants whose classification was later changed would give an independent estimate of accuracy.

The model currently handles single-nucleotide variants only, extending the pipeline to insertions and deletions, and retraining would close this gap. Also, retraining on RNA-sequencing validated variants only, and comparing the result to the current model, this would clarify the label noise that affects performance and may improve deep intronic sensitivity specifically.

The dependence of the model on the final scores of SpliceAI and Pangolin should be tested by evaluating it on variants from the chromosomes held out during training of these tools, and by retraining on the underlying sequence features.

Chapter 6| Conclusion

This study develops and validates a SpliceLoc-ML model with five-tier ACMG pathogenicity classification of non-canonical splice variants, that is shown to have practical clinical application in Long QT Syndrome genes. The category-specific Random Forest models showed highly discriminative performance (AUROC 0.936–0.975) in all genomic location categories, outperforming each individual splicing tools evaluated. The systematic comparison of the three splicing tools SpliceAI, Pangolin, and MaxEntScan confirmed expected performance hierarchy in the results derived from previous literature but was expanded to include their performance on non-canonical splice variants. Applied to 4,349 non-canonical splice variants from LQTS genes, the model placed 212 variants (4.9%) in the VUS tier compared with 489 (11.2%) in ClinVar, a net decrease of 56.6%, and placed 54 variants in the Likely Pathogenic or Pathogenic tiers. These assignments are computational. Without functional or segregation evidence they are not clinical classifications, and the 54 variants are candidates for further testing.

Chapter 7| Appendices

Appendix A| Software Packages and Libraries

Package / Library	Version	Purpose
Python Environment		
Python	3.10	Core programming language for all analyses
Google Colaboratory	—	Cloud-based computational environment for all Python analyses
Ubuntu	—	Local operating system environment for running Pangolin and BCFtools
Machine Learning & Statistics		
scikit-learn	1.0.2	Random Forest model training, hyperparameter optimization, 5-fold cross-validation, performance metrics calculation, class weight computation, and missing value imputation
imbalanced-learn	0.11	Random downsampling of majority class to address class imbalance in training data
SHAP	0.42	For feature importance analysis and model interpretability
joblib	—	Saving of trained models, binary cutoff values, and five-tier classification thresholds
Data Processing & Visualization		
pandas	2.0	Data loading, manipulation, filtering, and preprocessing of variant datasets
NumPy	1.25	Numerical computations and array operations
matplotlib	3.7	Generation of ROC curves, bar charts, performance metric figures, and distribution plots
seaborn	0.12	Generation of violin plots and score distribution visualizations
Variant Annotation Tools		
Ensembl VEP	110	Variant Effect Predictor, primary annotation pipeline using GRCh38 reference assembly
SpliceAI plugin (VEP)	—	Extraction of four delta scores per variant (acceptor gain, acceptor loss, donor gain, donor loss)
MaxEntScan plugin (VEP)	—	Extraction of reference allele score, alternate allele score, and difference score for short-range splice site motif strength evaluation
CADD plugin (VEP)	—	Extraction of CADD-PHRED deleteriousness scores (0–99) reflecting evolutionary constraint across species
Pangolin	1.1	Tissue-specific splicing probability prediction across brain, heart, liver, and testis, run locally on Ubuntu with --mask True and -d 300 parameters
Data Download & Processing		
gsutil	—	Downloading gnomAD v4.1 genome sequencing VCF files from Google Cloud Storage
BCFtools	—	Filtering of gnomAD VCF files by variant type (SNVs only) and allele frequency (AF > 0.001)

Appendix B| Model Input Features

Feature Name	Source Tool	Scale / Range	Missing Value Strategy
Δ Acceptor Gain	SpliceAI	0 – 1	Zero (null splice effect)
Δ Acceptor Loss	SpliceAI	0 – 1	Zero (null splice effect)
Δ Donor Gain	SpliceAI	0 – 1	Zero (null splice effect)
Δ Donor Loss	SpliceAI	0 – 1	Zero (null splice effect)
max_splice_val	SpliceAI	0 – 1	Zero (null splice effect)
Pangolin score	Pangolin	0 – 1	Zero (null splice effect)
MaxEntScan_ref	MaxEntScan	−20 to +12	Median per category
MaxEntScan_alt	MaxEntScan	−20 to +12	Median per category
MaxEntScan_diff	MaxEntScan	Variable	Zero (no change)
CADD_PHRED	CADD	0 – 99	Median per category
GC content	Custom Python	0 – 1	Zero
PPT density	Custom Python	0 – 1	Zero

Appendix C| Hyperparameters

Hyperparameter	Selected Value	Explanation
Number of estimators (trees)	500	More trees produce more stable probability estimates, 500 was sufficient without being computationally expensive
Maximum tree depth	10	Deep enough to capture real patterns in the data, shallow enough to avoid memorizing the training set
Minimum samples per leaf	5	Prevents the model from making decisions based on too few variants, which reduces overfitting to rare or noisy examples
Features considered per split	√(n_features)	Each tree sees a different random subset of features at each split, which is what makes the ensemble more robust than any single tree
Optimization method	RandomizedSearchCV	Using 5-fold cross-validation on the training set only, the test set was never involved in this process
Random seed	42 (fixed throughout)	Fixed throughout all steps so that the entire analysis is exactly reproducible

Chapter 8| References

Ackerman, M. J., Priori, S. G., Willems, S., Berul, C., Brugada, R., Calkins, H., Camm, A. J., Ellinor, P. T., Gollob, M., & Hamilton, R. (2011). HRS/EHRA expert consensus statement on the state of genetic testing for the channelopathies and cardiomyopathies: this document was developed as a partnership between the Heart Rhythm Society (HRS) and the European Heart Rhythm Association (EHRA). *Europace*, *13*(8), 1077–1109.

Adler, A., Novelli, V., Amin, A. S., Abiusi, E., Care, M., Nannenberg, E. A., Feilotter, H., Amenta, S., Mazza, D., & Bikker, H. (2020). An international, multicentered, evidence-based reappraisal of genes reported to cause congenital long QT syndrome. *Circulation*, *141*(6), 418–428.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., & Postolsky, B. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell reports*, *1*(5), 543–556.

Balmaña, J., Digiovanni, L., Gaddam, P., Walsh, M. F., Joseph, V., Stadler, Z. K., Nathanson, K. L., Garber, J. E., Couch, F. J., & Offit, K. (2016). Conflicting interpretation of genetic variants and cancer risk by commercial laboratories as assessed by the prospective registry of multiplex testing. *Journal of Clinical Oncology*, *34*(34), 4071–4078.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J., & Frey, B. J. (2010). Deciphering the splicing code. *Nature*, *465*(7294), 53–59.

Barbosa, P., Savisaar, R., Carmo-Fonseca, M., & Fonseca, A. (2023). Computational prediction of human deep intronic variation. *GigaScience*, *12*, giad085.

Blakes, A. J., Wai, H. A., Davies, I., Moledina, H. E., Ruiz, A., Thomas, T., Bunyan, D., Thomas, N. S., Burren, C. P., & Greenhalgh, L. (2022). A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. *Genome medicine*, *14*(1), 79.

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

Brnich, S. E., Abou Tayoun, A. N., Couch, F. J., Cutting, G. R., Greenblatt, M. S., Heinen, C. D., Kanavy, D. M., Luo, X., McNulty, S. M., & Starita, L. M. (2020). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome medicine*, *12*(1), 3.

Canavati, C., Sherill-Rofe, D., Kamal, L., Bloch, I., Zahdeh, F., Sharon, E., Terespolsky, B., Allan, I. A., Rabie, G., & Kawas, M. (2024). Using multi-scale genomics to associate poorly annotated genes with rare diseases. *Genome medicine*, *16*(1), 4.

Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature reviews genetics*, *3*(4), 285–298.

Chen, K. M., Wong, A. K., Troyanskaya, O. G., & Zhou, J. (2022). A sequence-based global map of regulatory activity for deciphering human genetics. *Nature genetics*, *54*(7), 940–949.

Chiba-Falek, O., Kerem, E., Shoshani, T., Aviram, M., Augarten, A., Bentur, L., Tal, A., Tullis, E., Rahat, A., & Kerem, B. (1998). The molecular basis of disease variability among cystic fibrosis patients carrying the 3849+ 10 kb C→ T mutation. *Genomics*, *53*(3), 276–283.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 6.

Clarke, A. (2024). Plenary sessions Pl1 ELPAG award lecture. *Eur. J. Hum. Genet*, *32*, 799–903.

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M., & MacArthur, D. G. (2020). A brief history of human disease genetics. *Nature*, *577*(7789), 179–189.

Coolidge, C. J., Seely, R. J., & Patton, J. G. (1997). Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic acids research*, *25*(4), 888–896.

Cooper, D. N., Ball, E. V., & Krawczak, M. (1998). The human gene mutation database. *Nucleic acids research*, *26*(1), 285–287.

Coppieters, F., Casteels, I., Meire, F., De Jaegere, S., Hooghe, S., Van Regemorter, N., Van Esch, H., Matulevičienė, A., Nunes, L., & Meersschaut, V. (2010). Genetic screening of LCA in Belgium: predominance of CEP290 and identification of potential modifier alleles in AHI1 of CEP290-related phenotypes. *Human mutation*, *31*(10), E1709–E1766.

Cormier, M. J., Pedersen, B. S., Bayrak-Toydemir, P., & Quinlan, A. R. (2022). Combining genetic constraint with predictions of alternative splicing to prioritize deleterious splicing in rare disease studies. *BMC bioinformatics*, *23*(1), 482.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., De Almeida, B. P., & Sirelkhatim, H. (2025). Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, *22*(2), 287–297.

Danis, D., Jacobsen, J. O., Carmody, L. C., Gargano, M. A., McMurry, J. A., Hegde, A., Haendel, M. A., Valentini, G., Smedley, D., & Robinson, P. N. (2021). Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *The American Journal of Human Genetics*, *108*(9), 1564–1577.

Ellingford, J. M., Ahn, J. W., Bagnall, R. D., Baralle, D., Barton, S., Campbell, C., Downes, K., Ellard, S., Duff-Farrier, C., & FitzPatrick, D. R. (2022). Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome medicine*, *14*(1), 73.

Fraile-Bethencourt, E., Valenzuela-Palomo, A., Díez-Gómez, B., Caloca, M. J., Gómez-Barrero, S., & Velasco, E. A. (2019). Minigene splicing assays identify 12 spliceogenic variants of BRCA2 exons 14 and 15. *Frontiers in genetics*, *10*, 503.

Ghosh, R., Harrison, S. M., Rehm, H. L., Plon, S. E., Biesecker, L. G., & Group, C. S. V. I. W. (2018). Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Human mutation*, *39*(11), 1525–1530.

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., & Smoller, J. W. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human mutation*, *36*(5), 513–523.

Guez, J., Goodrich, J. K., Moldovan, M. A., Chao, K. R., Kar, P., Panchal, R., Wilson, M. W., Laricchia, K. M., Rohlicek, G., & Biba, D. (2026). Integrating 730,947 exome sequences with clinical literature improves gene discovery. *medRxiv*.

Harrison, S. M., & Rehm, H. L. (2019). Is 'likely pathogenic'really 90% likely? Reclassification data in ClinVar. *Genome medicine*, *11*(1), 72.

Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., & Schwartz, G. B. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, *176*(3), 535–548. e524.

Jung, H., Lee, K. S., & Choi, J. K. (2021). Comprehensive characterisation of intronic mis-splicing mutations in human cancers. *Oncogene*, *40*(7), 1347–1361.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., & Birnbaum, D. P. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443.

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*(3), 310–315.

Kurosawa, R., Iida, K., Ajiro, M., Awaya, T., Yamada, M., Kosaki, K., & Hagiwara, M. (2023). PDIVAS: pathogenicity predictor for deep-intronic variants causing aberrant splicing. *BMC genomics*, *24*(1), 601.

Leman, R., Parfait, B., Vidaud, D., Girodon, E., Pacot, L., Le Gac, G., Ka, C., Ferec, C., Fichou, Y., & Quesnelle, C. (2022). SPiP: Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effects on mRNA splicing. *Human mutation*, *43*(12), 2308–2323.

Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J., & Fairbrother, W. G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences*, *108*(27), 11093–11098.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., & Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *FEBS letters*, *579*(9), 1900–1903.

Lord, J., Gallone, G., Short, P. J., McRae, J. F., Ironfield, H., Wynn, E. H., Gerety, S. S., He, L., Kerr, B., & Johnson, D. S. (2019). Pathogenicity and selective constraint on variation near splice sites. *Genome research*, *29*(2), 159.

Makhnoon, S., Shirts, B. H., & Bowen, D. J. (2019). Patients' perspectives of variants of uncertain significance and strategies for uncertainty management. *Journal of Genetic Counseling*, *28*(2), 313–325.

Marwaha, S., Knowles, J. W., & Ashley, E. A. (2022). A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome medicine*, *14*(1), 23.

Moles-Fernández, A., Domènech-Vivó, J., Tenés, A., Balmaña, J., Diez, O., & Gutiérrez-Enríquez, S. (2021). Role of splicing regulatory elements and in silico tools usage in the identification of deep intronic splicing variants in hereditary breast/ovarian cancer genes. *Cancers*, *13*(13), 3341.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., & Nickerson, D. A. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, *42*(1), 30–35.

Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., & Rath, A. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European journal of human genetics*, *28*(2), 165–173.

Nicora, G., Zucca, S., Limongelli, I., Bellazzi, R., & Magni, P. (2022). A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. *Scientific reports*, *12*(1), 2517.

Okubo, M., Noguchi, S., Awaya, T., Hosokawa, M., Tsukui, N., Ogawa, M., Hayashi, S., Komaki, H., Mori-Yoshimura, M., & Oya, Y. (2022). RNA-seq analysis, targeted long-read sequencing, and in silico prediction to unravel pathogenic intronic events and predict the regulatory mechanisms underlying complex splicing abnormalities in patients with dystrophinopathy.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825–2830.

Pejaver, V., Byrne, A. B., Feng, B.-J., Pagel, K. A., Mooney, S. D., Karchin, R., O'Donnell-Luria, A., Harrison, S. M., Tavtigian, S. V., & Greenblatt, M. S. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *The American Journal of Human Genetics*, *109*(12), 2163–2177.

Petrazzini, B. O., Naya, H., Lopez-Bello, F., Vazquez, G., & Spangenberg, L. (2021). Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData mining*, *14*(1), 44.

Pros, E., Gómez, C., Martín, T., Fábregas, P., Serra, E., & Lázaro, C. (2008). Nature and mRNA effect of 282 different NF1 point mutations: focus on splicing alterations. *Human mutation*, *29*(9), E173–E193.

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*(1), 31.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., & Spector, E. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine*, *17*(5), 405–423.

Sangermano, R., Garanto, A., Khan, M., Runhart, E. H., Bauwens, M., Bax, N. M., van den Born, L. I., Khan, M. I., Cornelis, S. S., & Verheij, J. B. (2019). Deep-intronic ABCA4 variants explain missing heritability in Stargardt disease and allow correction of splice defects by antisense oligonucleotides. *Genetics in medicine*, *21*(8), 1751–1760.

Scotti, M. M., & Swanson, M. S. (2016). RNA mis-splicing in disease. *Nature reviews genetics*, *17*(1), 19–32.

Smith, C., & Kitzman, J. O. (2023). Benchmarking splice variant prediction algorithms using massively parallel splicing assays. *Genome biology*, *24*(1), 294.

Soemedi, R., Cygan, K. J., Rhine, C. L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., & Fairbrother, W. G. (2017). Pathogenic variants that alter protein code often disrupt splicing. *Nature genetics*, *49*(6), 848–855.

Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D. N., & Sanford, J. R. (2011). Loss of exon identity is a common mechanism of human inherited disease. *Genome research*, *21*(10), 1563.

Sullivan, P. J., Gayevskiy, V., Davis, R. L., Wong, M., Mayoh, C., Mallawaarachchi, A., Hort, Y., McCabe, M. J., Beecroft, S., & Jackson, M. R. (2023). Introme accurately predicts the impact of coding and noncoding variants on gene splicing, with clinical applications. *Genome biology*, *24*(1), 118.

Sullivan, P. J., Quinn, J. M., Wu, W., Pinese, M., & Cowley, M. J. (2024). SpliceVarDB: A comprehensive database of experimentally validated human

splicing variants. *The American Journal of Human Genetics*, *111*(10), 2164–2175.

Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., Biesecker, L. G., & Group, C. S. V. I. W. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in medicine*, *20*(9), 1054–1060.

Vaz-Drago, R., Custódio, N., & Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. *Human genetics*, *136*(9), 1093–1111.

Vincent, G. M. (2000). Long QT syndrome. *Cardiology clinics*, *18*(2), 309–325.

Walker, L. C., de la Hoya, M., Wiggins, G. A., Lindy, A., Vincent, L. M., Parsons, M. T., Canson, D. M., Bis-Brewer, D., Cass, A., & Tchourbanov, A. (2023). Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. *The American Journal of Human Genetics*, *110*(7), 1046–1067.

Wang, R., Helbig, I., Edmondson, A. C., Lin, L., & Xing, Y. (2023). Splicing defects in rare diseases: transcriptomics and machine learning strategies towards genetic diagnosis. *Briefings in bioinformatics*, *24*(5), bbad284.

Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A. H., Karczewski, K., Alexander, Y., Barton, P. J., Funke, B., Cook, S. A., & MacArthur, D. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in medicine*, *19*(10), 1151–1158.

Wojcik, M. H., Reuter, C. M., Marwaha, S., Mahmoud, M., Duyzend, M. H., Barseghyan, H., Yuan, B., Boone, P. M., Groopman, E. E., & Délot, E. C. (2023). Beyond the exome: what's next in diagnostic testing for Mendelian conditions. *The American Journal of Human Genetics*, *110*(8), 1229–1248.

Yeo, G., & Burge, C. B. (2003). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Proceedings of the seventh annual international conference on Research in computational molecular biology,

Zeng, T., & Li, Y. I. (2022). Predicting RNA splicing from DNA sequence using Pangolin. *Genome biology*, *23*(1), 103.

Zurynski, Y., Deverell, M., Dalkeith, T., Johnson, S., Christodoulou, J., Leonard, H., Elliott, E. J., & group, A. R. D. I. o. F. S. (2017). Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays. *Orphanet journal of rare diseases*, *12*(1), 68.