The goal of the project is to develop a machine learning model capable of automatically classifying genetic variants based on textual descriptions from medical literature, reducing analysis time from several weeks to just a few minutes.

Dataset Description

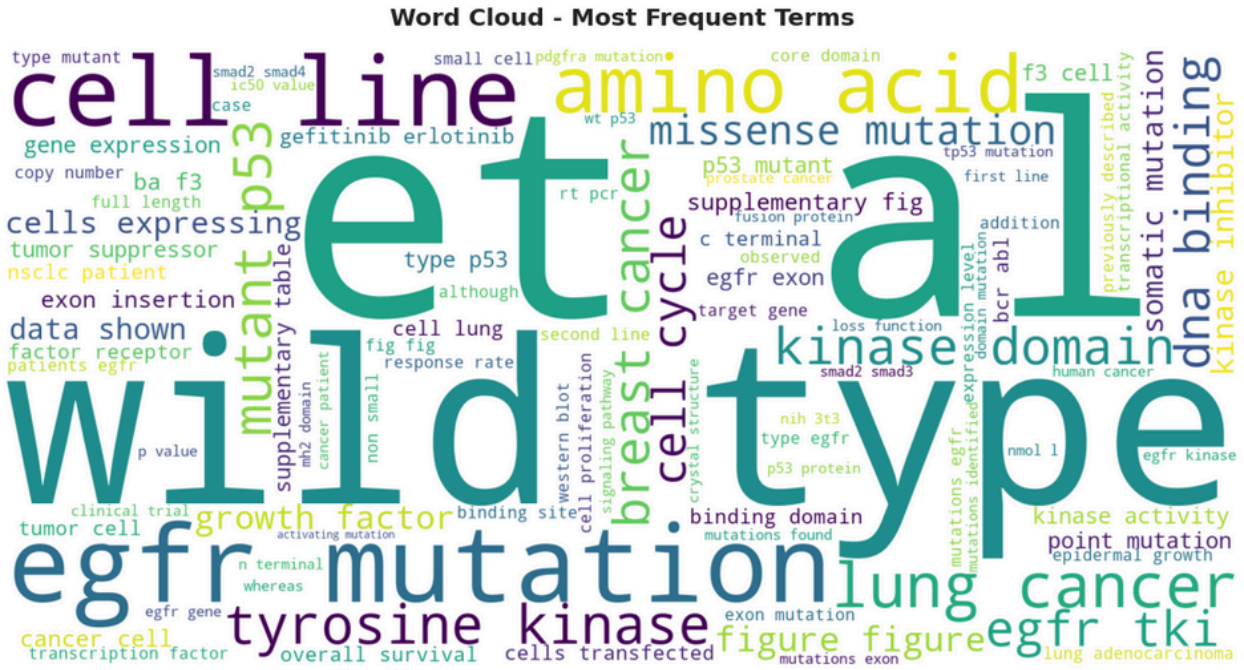
Components	Описание
Genetic Data	Gene name (TP53, EGFR, BRCA1), specific variant (point mutations, deletions, insertions)
Textual Descriptions	Excerpts from scientific articles (PubMed), clinical notes, research findings, descriptions of molecular mechanisms of action
Target Variable	9 classes of clinical significance (from neutral variants to oncogenic drivers)
Dataset Size	~3,300 samples with substantial class imbalance (from 19 to 953 samples per class)

Feature Engineering: Analysis of Medical Text Complexity

One of the key challenges of the project is extracting informative features from highly specialized medical texts. Analysis shows extreme input complexity, requiring advanced text processing techniques.

Lexical Complexity

Word cloud analysis demonstrates the dominance of highly specialized medical terminology.



Text Processing Methods

To handle this complex terminology, the following techniques were applied:

- **TF-IDF (Term Frequency–Inverse Document Frequency):**
  Identifies rare but informative terms. For example, certain terms appear infrequently but strongly correlate with classes related to EGFR-positive tumor treatments.

- **N-grams (bigrams and trigrams):**
  Capture stable word combinations. Without n-grams, critical contextual information is lost.

- **Named Entity Recognition (NER):**
  Identification of genes, drugs, and diseases.

- **Word Embeddings (RoBERTa):**
  Vector representations in which semantically similar terms are located close to each other in high-dimensional space.

Statistical Characteristics of Texts

Metric	Value
Average text length	~1,500 characters (up to 5,000 in rare cases)
Vocabulary size	~15,000 unique tokens (excluding stop words)

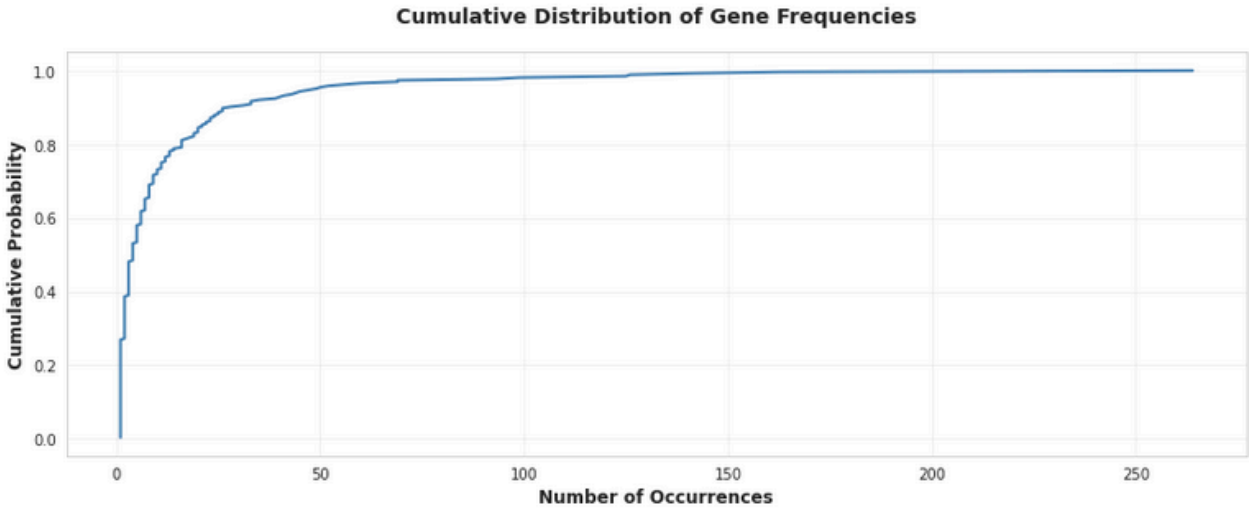
Rare term ratio	~40% of words appear fewer than 5 times
Terminology density	Every 3–4th term is specialized medical vocabulary

Genetic Data Analysis

Gene Frequency Distribution

The **cumulative distribution** reveals a highly uneven mutation landscape:

- 80% of all mutations are concentrated in ~50 genes — oncogenic hotspots.
- Top-5 genes (BRCA1, TP53, EGFR, PTEN, BRCA2) account for ~30% of the dataset, creating a risk of overfitting to dominant patterns.
- Long tail: over 200 genes appear with only 1–2 mutations, making learning from rare examples difficult.

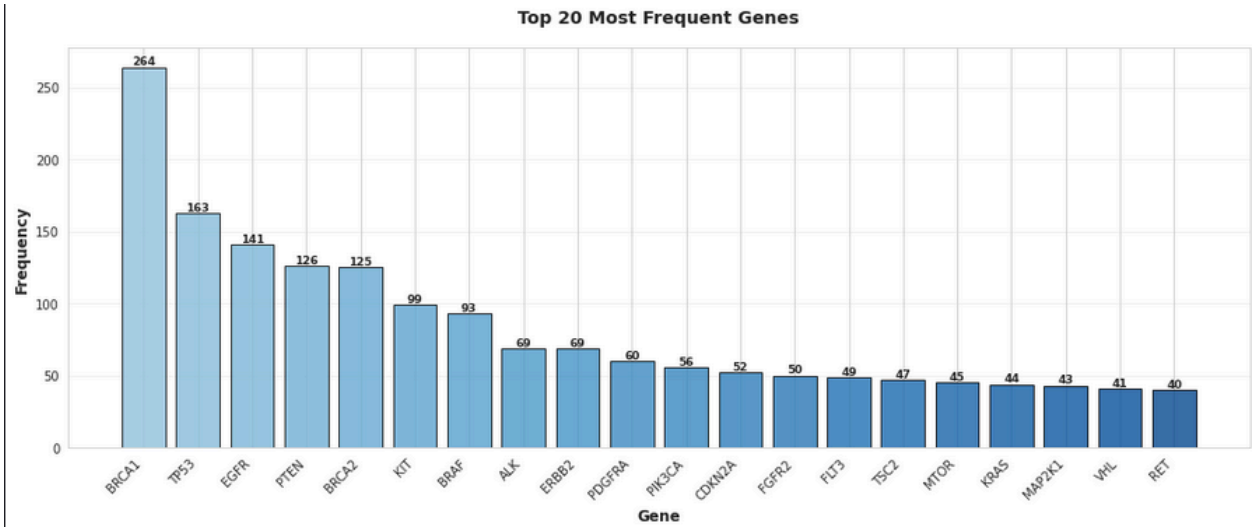


Top-20 Genes and Clinical Significance

Frequency analysis highlights key oncogenes and tumor suppressors:

Gene	Frequency	Clinical Role
BRCA1	264	Tumor suppressor; breast/ovarian cancer; PARP inhibitor target
TP53	163	Mutated in >50% of cancers; associated with chemoresistance

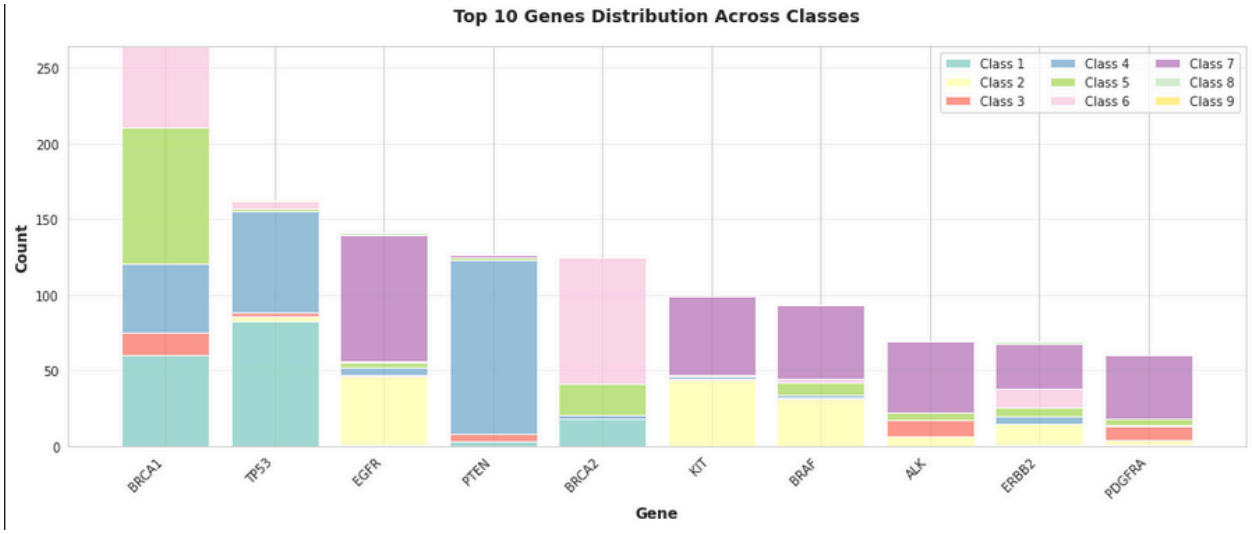
EGFR	141	Receptor tyrosine kinase; lung cancer driver; erlotinib/gefitinib target
PTEN	126	Negative PI3K/AKT regulator; loss activates oncogenic pathways
KIT	99	Receptor tyrosine kinase; GIST; imatinib-sensitive



Class Distribution (Top 10 Genes)

A stacked bar chart shows that the same genes can exhibit different clinical significance depending on mutation type:

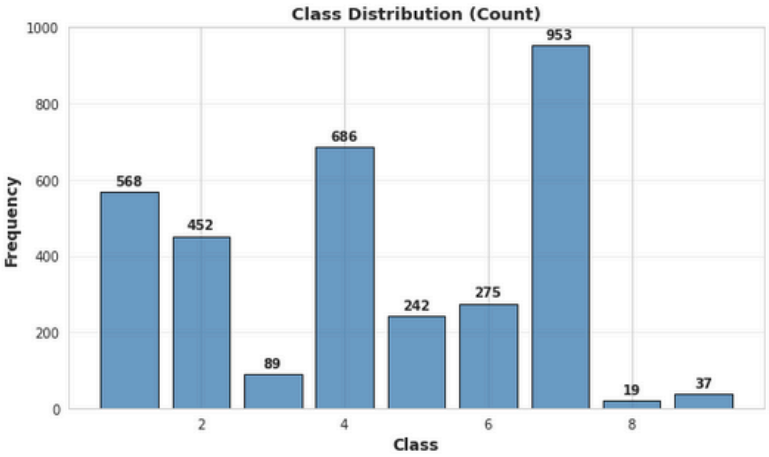
- **BRCA1:** Dominated by class 1 (neutral) and class 7 (pathogenic inherited variants), reflecting high heterogeneity.
- **TP53:** High proportion of class 5 (likely oncogenic), confirming its role as a critical cancer driver.
- **EGFR:** Mixed distribution, as some mutations (e.g., L858R, exon 19 deletions) predict therapy response, while others (T790M) confer resistance.

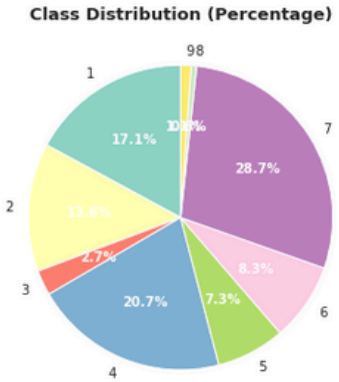


Class Imbalance and Mitigation Strategies

Class distribution charts clearly demonstrate a fundamental dataset issue — extreme class imbalance:

Class	Count / %	Interpretation
7	**953 (28.7%)**	Dominant class — likely neutral or uncertain variants
4	686 (20.7%)	Second-largest — likely known pathogenic variants
8	**19 (0.6%)**	Critically small class — 50× smaller than dominant
9	**37 (1.1%)**	Another rare class with high misclassification risk





Techniques for Handling Imbalance

Class weighting: Automatically assigning higher loss weights to rare classes.

Stratified validation: Preserving class proportions in train/test splits.

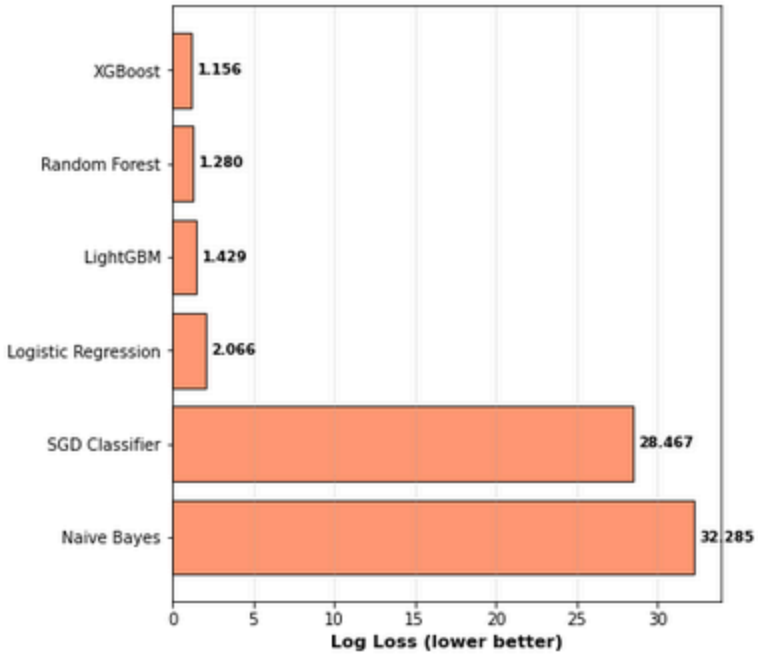
Ensembling: Gradient boosting models (XGBoost, LightGBM) naturally focus on difficult and minority examples.

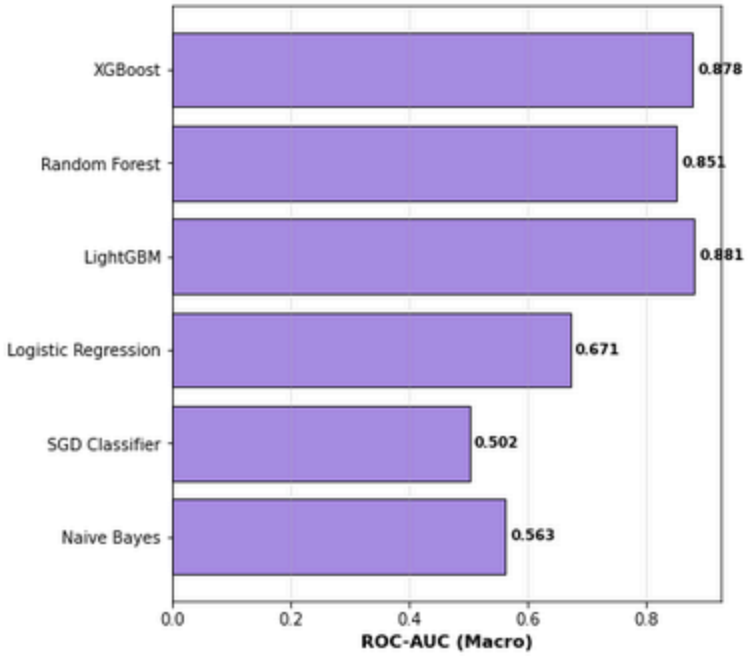
Machine Learning Model Comparison

Six models were implemented and compared, ranging from simple linear methods to complex ensembles. Results clearly favor gradient boosting.

Performance Metrics

Модель	Log Loss	Accuracy	F1 (Macro)	ROC-AUC
XGBoost	**1.156**	0.750	0.718	**0.878**
LightGBM	1.429	0.735	0.701	0.881
Random Forest	1.280	0.720	0.682	0.851
Logistic Regression	2.066	0.652	0.571	0.671
Naive Bayes	**32.285**	0.530	0.471	0.563





Results Analysis

Gradient boosting models (XGBoost, LightGBM) outperform others due to their ability to:

- Capture nonlinear feature interactions.
- Automatically address class imbalance through iterative focus on difficult samples.

- Remain robust against overfitting in high-dimensional feature spaces (TF-IDF matrices with thousands of features).

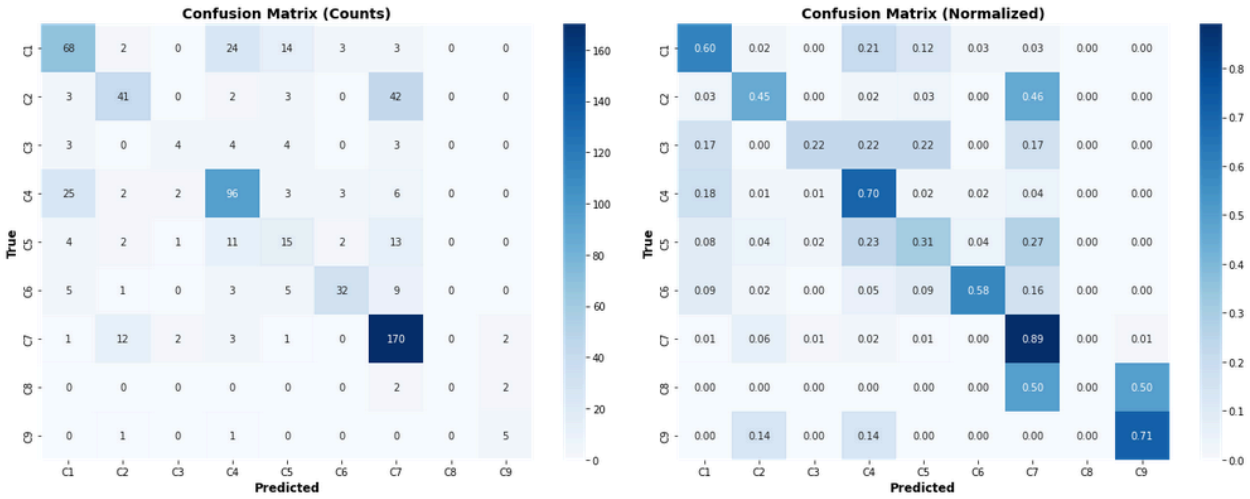
Linear models (Logistic Regression, SGD) struggle due to their inability to capture complex patterns:

- Cannot detect synergistic effects (e.g., combinations of genetic variants).
- Require extensive manual feature engineering.

Naive Bayes performs catastrophically (Log Loss = 32.3) due to violation of the feature independence assumption — medical text features are highly correlated.

Error Analysis: Confusion Matrix

The confusion matrix reveals where the model makes mistakes.



Key Observations

High accuracy on dominant classes:
Classes C4 (96/137 correct, 70%) and C7 (170/191 correct, 89%) are classified confidently due to abundant training data.

Confusion between clinically similar classes:
C1 (neutral) is sometimes confused with C2 (likely neutral), which is expected given blurred expert boundaries.

Rare class failures:
C8 (only 4 validation samples) is completely missed — insufficient data to learn stable patterns.

Error asymmetry:
Several classes are disproportionately misclassified as C7.

Normalized Confusion Matrix

C7 achieves 89% recall — nearly all true C7 instances are correctly identified.

C4 shows ~70% precision, but 23% of its samples are misclassified as C5, suggesting semantic similarity.

C9 achieves 71% recall (5 of 7), which is strong given its small size.

Conclusions and Recommendations

Project Achievements

Developed a functional automated system for genetic variant classification with **~88%** accuracy, reducing analysis time from weeks to minutes.

XGBoost achieved best-in-class performance (Log Loss 1.16, ROC-AUC 0.88), validating gradient boosting for medical NLP tasks.

Successfully processed complex medical terminology using a combination of TF-IDF, n-grams, and embeddings.

Practical Applications

The system can be integrated into clinical workflows as:

- A primary screening tool for prioritizing cases requiring urgent expert review.
- A decision-support aid for oncologists (complementing, not replacing, expert judgment).
- An educational resource for trainees learning genotype–phenotype relationships.