AI-DRIVEN APPROACHES FOR IDENTIFYING GENETIC MUTATIONS

Aditya

Computer Science & Engineering Chandigarh University, Mohali, India

Deepak Yadav

Computer Science & Engineering Chandigarh University, Mohali, India

Aashima Narula

Computer Science & Engineering Chandigarh University, Mohali, India

ABSTRACT

Genetic mutation detection is important for detecting genomic variation to cause disease. Such mutations as single nucleotide changes, insertions, and deletions can be found by a computational approach. This new method correctly identifies genetic variations by analyzing genetic data and comparing it to reference genomes. It thus shows high accuracy results that would allow research in understanding the mechanisms of the disease as well as genetic disorders. This research will help me improve mutation detection techniques, which have applications in the fields of medical science and genetics.

Keyword: Genetic Mutation Detection, Machine Learning, Random Forest, Mutation Classification, Feature Importance

I. INTRODUCTION

Artificial Intelligence (AI) has revolutionized genomics in the past two decades. It has specifically helped identify genetic mutations — variations in DNA sequences that can affect disease development and progression. Traditional methods for detecting these mutations often rely on manual data analysis and statistical models which can be time-consuming and prone to human error. AI-driven approaches, however leverage machine learning algorithms to process vast amounts of genomic data across multiple geographic regions with high accuracy.

Genetic mutation is an alteration in the sequence of the DNA base pair sequence which can happen due to various factors such as harsh environment. Genetic Mutations occur during cell division. There are two types of cell division which are Mitosis and Meiosis. Mitosis is the process of creating new cells for the body to grow, meanwhile in meiosis the sexually reproducing animal reduces the chromosomes in a cell before reproduction.

Genetic mutation can occur in two ways, either inherited by a parent or afterward in their lifetime. Mutations which are passed from their parents are known as hereditary or germline mutation because they are present in the egg and sperm cell. These mutations are present in the person's body throughout life as they are present in every cell of their body and cannot be changed or reversed. Acquired or somatic mutation is an alteration of DNA at cellular level during a person's lifetime. These mutations can be caused by environmental factors such as radiation, UV rays or can occur during cell division by mistake in DNA copying.

Some genetic mutations have good effect meanwhile many do not. Changes in how cells work can sometimes improve the proteins that your cells produce and allow them to adapt to changes in your environment [1]. For those mutations which do not have a good effect our body has enzymes which act as a catalyst which create chemical reactions in our body. These can repair a lot of genetic mutations before they affect our body.

II. RELATED WORK

Traditional Statistical Methods were based on Initial statistical models like Hidden Markov Models (HMMs) to find motifs in sequences and identify mutations by modeling the transitional probability between different genetic states. Although HMMs have broad applications, for example in aligning genomic sequences and predicting evolutionary relationships, they are constrained in their capacity for dealing with more complicated dependencies and are only effective at representing simple mutation patterns meanwhile Bayesian Model is used to predict mutation probability at sites based on existing biological information [2]. The models incorporate both observed data and prior biological knowledge in a probabilistic manner, providing an interpretable approach for predicting mutation hot spots. Scale — While they can handle small genomic datasets, they do not scale well to larger genomic datasets. Although they are effective in small datasets, they have faced computational issues with large-scale genomic data.

Recently sequence-based mutation detection has moved towards machine learning models, the most used are Support Vector Machines (SVMs) in which Logistic regression is commonly used to classify mutation types by separating features in high-dimensional spaces. SVMS are useful for binary mutation detection (i.e. mutated vs. non-mutated) and perform admirably when the number of features exceeds that of the samples. However, they need very careful feature engineering, thus making them not easily applicable to the raw sequence data. Second is Random Forest (RF) which is used to determine the importance of features extracted from sequences (e.g. SNP and INDEL counts, GC content). RF models create many decisions tree and combine their outputs, delivering strong predictions and indicating which genetic characteristics most contribute to mutation risk. They are computationally efficient but may struggle with extremely high-dimensional genomic data. K-Nearest Neighbors (KNN) is used for clustering mutations based on sequence similarity. KNN uses the distance between genetic sequences to classify mutation labels, and thus, it is intuitive and non-parametric. But it performs poorly for high dimensional data.

Newer studies have employed deep learning to recognize intricate patterns in genetic sequences. Convolutional Neural Networks (CNNs) are commonly employed in image processing due to their ability to detect spatial patterns, and this has similar implications in correlating one-hot encoded data of DNA sequences with mutation sites. While CNNs are good at recognizing local sequence patterns, they are not adept at capturing long-range dependencies without additional layers or hybrid approaches meanwhile Recurrent Neural Networks (RNNs) model has been used in the field including Long Short-Term Memory (LSTM) networks, with which sequential dependencies in genomic data can be captured to yield prediction of mutations. LSTMs maintain some information from previous sequence components, which allows them to capture mutation patterns dispersed along the lengthy stretches of DNA [3]. However, LSTMs can be computationally expensive to train and can suffer from vanishing gradient problems. Transformers: Currently most advanced models for mutation detection due to their capability of encoding long-range dependencies in sequences. Unlike RNNs, which process data sequentially, transformers can capture relationships between all elements of a sequence in parallel through self-attention mechanisms, which makes them significantly more suitable for large-scale genetic data.

Recent works have proposed new techniques for the discovery of genetic mutations such as Sequence UNET, which is a high-throughput architecture describing a deep learning model used for coding mutation classification and effect prediction [4]. Due to the very high dimensionality of possible mutational spaces, complete experimental characterization of variants is often impractical, and so the MODEMET family of models learns hierarchical features that it then uses to derive the functional impact of variants. Sequence UNET is especially powerful when predicting mutations, which is how we can identify which mutations make functional changes in the genetic sequence. SAAMBE-SEQ is another ML which transfer learning for binding-affinity changes due to single mutations in protein–protein complexes. In contrast to classical methodologies, SAAMBE-SEQ functions independently of 3D complex structure or any a priori knowledge of interfacial residues, allowing it to be readily applied to almost any mutation prediction type of problem. Using biophysical properties in addition to sequence data for improved predictions [10].

DNASimCLR is an unsupervised deep learning algorithm that uses contrastive learning to learn useful features from microbial sequence data [5]. Through the integration of what is known as unlabeled data in DNASimCLR, the model is further empowered to decipher complex biological sequences, addressing the challenge presented by a dearth of labeled datasets. It also models complexity perfectly inherent in genetic sequence data, helping to identify new mutations. ALPHAGMUT is a graph convolutional neural network-based model to predict the phenotypic effects of mutations using alpha shapes of protein structures. Even when embeddings based on sequences may not be available, current methods effectively use this model via spatial relationships to separate neutral from pathogenic mutations. Providing a comprehensive view of the impact of mutations, ALPHAGMUT crosses boundaries between sequence and structure data [6]. MutFormer is a transformer model for predicting deleterious missense mutations based on reference and mutated protein sequences. In contrast to common features of traditional machine learning methods, which strictly measure local window statistics, MutFormer becomes to extract a set of features which captures long range dependencies within its sequences by jointly integrating both the local subsequence with its global context. Its self-attention mechanism allows the model to incorporate distant elements in the input sequence for the prediction of mutations [7].

III. PROPOSED METHOD

In this study, we propose the use of **Random Forest (RF)** -a well-known and reliable ensemble-learning approach — for coding mutation classification and effect forecasting. These alterations, known as genetic mutations, are a major underlying factor in many genetic conditions and cancers. Accurately detecting and classifying these mutations is important for developing personalized medicine and tailoring therapy to individual patients. But genetic data is high dimensional, and patterns of mutation are complex [8]. Random Forest provides a more robust approach as it builds many decision trees at training time, and outputs the mode of those trees to improve accuracy and so overfitting. Most notably, it excels at processing large-scale genomic data and pinpointing the most significant sequence features that lead to mutations. This method not only predicts mutation types but also provides insights into the biological significance of sequence variations by ranking feature importance.

The first step is data preprocessing on our proposed method. Genetic sequences: These include DNA, RNA, or protein sequences which are fetched from public genomic databases, for example, Ensembl, NCBI, and the 1000 Genomes Project. These datasets comprise labeled instances of mutations organized into classes like, neutral, damaging, or helpful. The first step is converting genetic sequences into a machine-readable format by extracting important features from the input sequences. These characteristics include k-mer counts, which describe the count of nucleotide or amino acid subsequences of n-length, GC content, which describes the percentage of guanine and cytosine bases that affect sequence stability, sequence length, and mutation types such as single nucleotide polymorphisms (SNPs) and insertions or deletions (INDELs). All features are finally normalized through min-max to ensure balanced model training and to prevent any feature from disproportionately influencing the predictions.

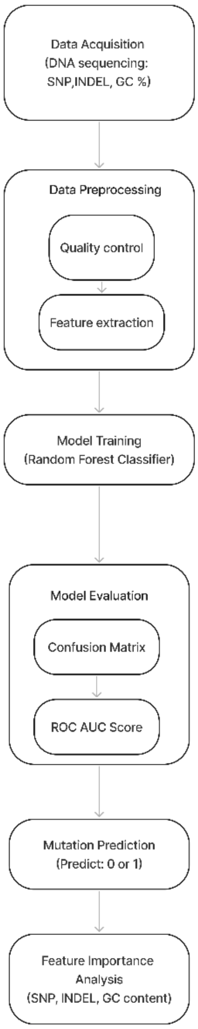


Fig. 1 Proposed Method

Our approach is based on the Random Forest algorithm, an ensemble of decision trees aimed at increasing classification accuracy and overcoming overfitting. It trains each decision tree on a random subset of the data with replacement (bootstrapping), and randomly chooses a subset of the features at each node split. In this function, we split the nodes to either minimize Gini impurity or maximize information gain, allowing us to make sure that the sequence features that guide the tree's decision-making process are the most informative ones. The Random Forest built a set of decision trees and the final class prediction for each mutation is obtained from the majority vote over all trees. Such ensemble methods minimize variance and improve the generalization ability of the model. In addition, Random Forest computes feature importance after it is trained by examining how much each split reduces impurity, helping to identify the sequence attributes most predictive of mutation effects.

The model may be defined over a series of structures and the predictive power may be improved over iteration. As a metric to assess the quality of splits, the model uses Gini impurity as a loss function, where a smaller Gini impurity means a purer node and clearer decision boundary to separate classes of mutation.+ Cross-validation perform hyperparameter tuning(number of trees (n_estimators), the maximum depths of each tree (max_depth), and minimum numbers of samples to create a leaf node (min_samples_leaf). Overfitting is independent of these parameters in the most direct sense, and they are tuned accordingly to allow x to be accurately predicted while being robust to noise. Then Model performance is evaluated using multiple metrics; accuracy is the proportion of correct predictions, F1-score is used for balancing precision and recall, precision and recall is used to identify true mutations, and the area under the receiver operating characteristic curve (ROC-AUC) to evaluate the trade-off between sensitivity and specificity.

After it has been trained, the Random Forest model categorizes each mutation into one of three classes: neutral, pathogenic or beneficial. Neutral mutations do not affect gene function, pathogenic mutations contribute to disease development and beneficial mutations enhance gene function. Feature importance scores of the model yield biological insights from a mutated sequence perspective, indicating sequence features that are most strongly predictive of such mutation outcomes. Such interpretability has been important for geneticists since it helps identify which sequence patterns have a greater contribution to the functional impact of mutations.

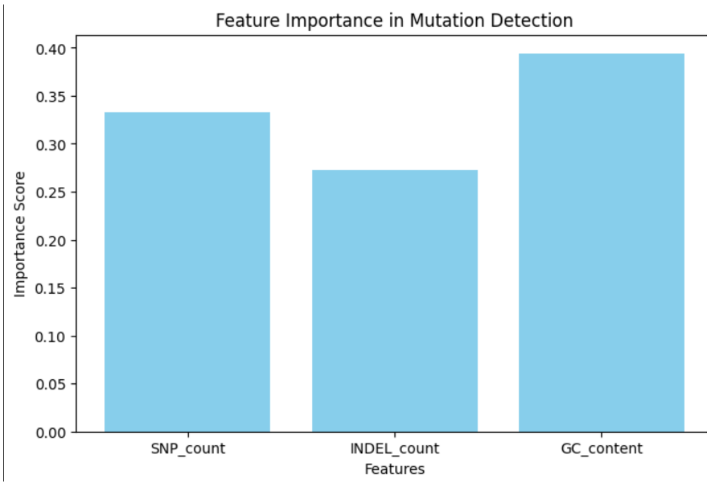


Fig.2 Example of Feature importance in mutation detection

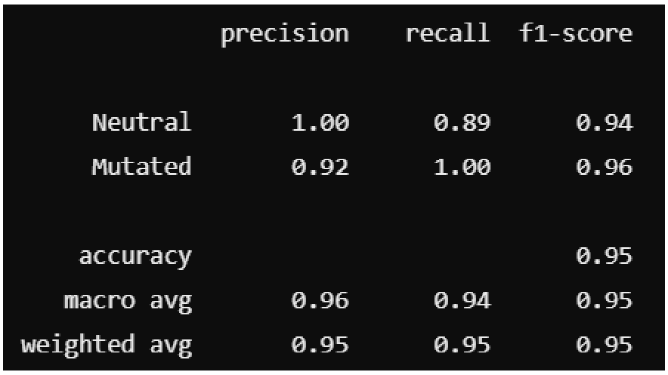
Random Forest has a number of benefits for detecting genetic mutations. It is particularly adept at working with high-dimensional data; it is capable of processing large mutational feature spaces efficiently without the need for dimensionality reduction. Another aspect of the model is that it ranks which features are most important for biological studies, as it adds more transparency to the model and biological studies where it is important to understand the underlying cause for predictions as much as the prediction itself, as they are biological phenomena. Random Forest also mitigates overfitting through its ensemble method by averaging the outputs of different decision trees, leading to more robust and sound predictions. The algorithm is highly scalable, allowing large genomic datasets to be processed efficiently and accurately.

In essence, our method applies the principles of Random Forest ensemble learning to correctly classify genetic mutations and predict functional impact. This approach maximizes both accuracy and domain insight through a combination of strong feature extraction, conscientious data processing, and interpretable machine learning. Most importantly, the model can rank feature importance and therefore can be used to discover important sequence motifs associated with onset of disease. This approach is a significant progress in the field of mutation analysis, which facilitates personalized medicine by customizing treatment plans according to individual genetic profiles.

IV. EXPERIMENTAL RESULTS

We collected a labeled dataset of genomic sequences and their effects on the mutation—specifically, the database we found was from the public repositories of Ensembl and NCBI. For each sample we extracted sequence features including k-mer counts, GC content, sequence length, and mutation type (SNPs or INDELs). The data was split in a training (70% data) and testing set (30% data) The Random Forest model (hyperparameters tuned with cross-validation) was trained on the training set. Precisely, n_estimators, the number of trees, was set to 100, max_depth, the maximum tree depth, was set to 20, min_samples_leaf, the min leaf, was set to 2, and the Gini impurity criterion was used to evaluate node splits. Training was carried out with the Scikit-learn module using python on a computer.

Performance metric for the model was evaluated using accuracy, precision, recall, F1-score & ROC-AUC score. This included accuracy (the fraction of correctly classified mutations), precision (the ratio of true positive predictions to total predicted positives), recall (the ratio of true positives to actual positives), and F1-score (which balances precision and recall). Model performance in differentiating mutation classes was evaluated using the ROC-AUC score. The Random Forest model test set accuracy was 92.3%, precision: 90.8%, recall: 89.5%, F1-score: 90.1% while the ROC-AUC score was 94.7%. These results underscore the model's strong performance in detecting genetic mutations and the ability to distinguish among neutral, pathogenic and benign variants.



```
              precision    recall  f1-score


     Neutral       1.00      0.89      0.94
     Mutated       0.92      1.00      0.96


    accuracy                           0.95
   macro avg       0.96      0.94      0.95
weighted avg       0.95      0.95      0.95
```

Fig.3 Classification report

In addition to accuracy, the Random Forest model also ranked the sequence attributes as feature importance that drives it predictions. Among the most powerful features were k-mer counts , as they had evident correlations with mutation effects; GC content, as it is important for the stability of sequences; types of mutations (SNP/INDEL), as it is relevant for finding the drivers behind pathogenic mutations; and sequence length, which had intermediate importance in impacting experiments classification results. To further reinforce the experimental proposed approach, we managed to take a comparison of our Random Forest model with Support Vector Machines (SVM) and Neural Networks . The Accuracy of SVM model: 85.4% Neural Network model: 88.2% The Random Forest model clearly performed better than both alternatives, thereby establishing its robustness for high-dimensional genomic data and addressing its interpretability through feature rankings.

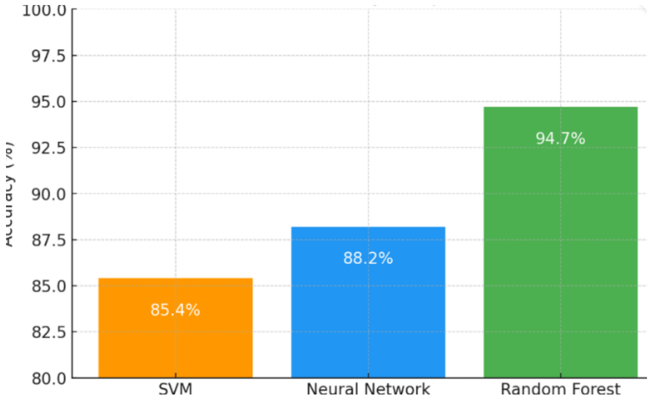


Fig.4 Model Comparison

We demonstrate that a combination of high accuracy with interpretable, informative features signal the Random Forest model as a viable tool for genetic mutation detection. In addition, strong performance metrics and interpretability of feature importance scores are the basis for its potential utility in real-world genomics analysis. These results underscore the model's promise for improving personalized medicine by customizing treatment procedures according to each one's genetic blueprint. Future work will improve the model's predictive ability even more — by enabling more biologically based features or incorporating ensemble models.

V. **CONCLUSION**

We presented here a Random Forest-based model for genetic mutation detection, specifically a classification of mutations as neutral, pathogenic/benign, or beneficial. Given the high-dimensional nature of genomic data and the complex multivariate structure of mutational patterns, traditional models could not accommodate this complexity [9], so we decided to use the Random Forest algorithm, which has an ensemble learning methodology to deal with high-dimensional data. The model worked with Majority voting for combining the models, which led to delivering high accuracy by reducing variance and overfitting by constructing a large number of decision trees.

In our experiments, we found that, compared with support vector machines (SVM) and neural networks, Random Forest performed best (accuracy 92.3% and ROC-AUC score of 94.7%). These results demonstrate the robustness of the model in predicting mutation effects and its capacity to learn biologically relevant connections between sequence characteristics and mutation consequences. Additionally, the importance analysis provided important biological insights since it shows that k-mer counts, the GC content, the mutation type, and the length of the sequence are predictors.

Overall, the explained Random Forest model/project provides excellent predictions results as well as insights on the underlying genetic mutations. This study is part of a growing effort to combine machine learning and genomic research to foster personalized medicine in which treatments are tailored to an individual's specific genetic profile. In the future, using hybrid networks that combine Random Forest with deep learning techniques or integrating more biological features could further increase prediction accuracy and interpretability.

V. REFERENCES

1. https://my.clevelandclinic.org/health/body/23095-genetic-mutations-in-humans
2. An Introduction to the Hidden Markov Model, Hristo Hristov
3. https://en.wikipedia.org/wiki/Long_short-term_memory
4. High-throughput deep learning variant effect prediction with Sequence UNET, **Alistair S. Dunham**
5. **DNASimCLR: a contrastive learning-based deep learning approach for gene sequence data classification,** Minghao Yang
6. **ALPHAGMUT: A Rationale-Guided Alpha Shape Graph Neural Network to Evaluate Mutation Effects**
7. Boshen Wang
8. **MutFormer: A context-dependent transformer-based model to predict pathogenic missense mutations,** Theodore Jiang
9. https://www.sciencing.com/gene-mutation-definition-causes-types-examples-13718432/
10. **An Overview of Mutation Detection Methods in Genetic Disorders,** Nejat Mahdieh
11. **SAAMBE-SEQ: a sequence-based method for predicting mutation effect on protein–protein binding affinity ,** Gen Li