Prediction of Progeria using Support Vector Machine
D. Deva Hema¹, Vikrant², Anish Joshi³, Vishnu Sharma⁴
Assistant Professor¹, Student²,³,⁴
Department of Computer Science and Engineering
SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Abstract:
Progeria is an extremely rare disease that causes the body of children to age faster. Children with progeria do not live past 13 years. Ageing plays a very important role in today’s world. It has huge impact on human life and health and also the economy. There are more than 300-400 genes which are related to human ageing and many of them function as a type of gene which are called as protein coding genes. These genes encode protein which consists of a series of three nucleotide sequences which are called as codons in genetic language. The molecular basis of ageing is poorly understood. The genes which are related to Ageing are not studied in a strenuously manner and also their analysis is not done in an efficient manner. In our project we are using protein samples from biological databases and then the data is extracted and classified using Machine Learning algorithms. The protein samples are classified as ageing or non-ageing on the basis of a number of protein features. On the basis of these features the values are predicted and these values help us in classifying as ageing or non-ageing. After these classifications, we check for the protein stability in the ageing-related protein samples. For this process, we again use support vector machine algorithm. The support vector algorithm uses the maximum margin plane which is called as the “Hyperplane”. On the basis of these margins, we are able to predict if the given protein sample has progeria genes.

Keywords: Progeria, Protein, Machine Learning.

I. INTRODUCTION
In this project, a simple model is constructed which will help in distinguishing ageing related proteins and non-ageing related proteins. Then these proteins will be used in predicting whether the person has progeria or not. Our project mainly aims at classifying proteins as ageing and non-ageing proteins. The further process of detecting whether the protein is related to progeria or not is a scientific process which requires heavy machinery. Machine learning is a subset of AI. It focuses on the development of programs with the help of which the computer can learn automatically from its past experiences and then improve it without being programmed again and again. ML have algorithms that can access data, classify it and the use it for themselves. Progeria is an extremely rare disease that causes the body of children to age faster. Children with progeria do not live past 13 years. Progeria occurs when there is a mistake in a certain gene which causes it to become an abnormal gene. Progeria affects all the sexes equally. It is found in 1 in 4 million births and hence is said as an extremely rare genetic condition. Progeria is not inherited or passed down in families. Symptoms of progeria- Head grows in size, Enlarged eyes, small lower jaw, Ears sticking out, A high-pitched voice, loss of body fat and muscles, Hair loss, etc. These are some of the symptoms of progeria.

II. RELATED WORKS-:
The Existing System for protein analysis uses three machine learning algorithms, namely,

1] XG Boost Algorithm-:
XG Boost is an advanced machine learning algorithm. It is used for classification process and also to increase the computational power. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

XGBoost is used in the project for two reason-;
1-Execution Speed
2-Model Performance

2] Logical Regression-:
Logical Regression was mainly used in the biological science in early 20th century. It is used when the dependent variable target is categorical.

3] The Support Vector Machine-:
The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. This system utilises four databases and creates a dataset out of the databases. After ID mapping and GO ancestor determination, several feature sets are extracted. Firstly the most important feature sets are selected in various steps which reduces the most dimensionality of the feature space.
III. PROPOSED SYSTEM:

The existing system uses three algorithms for the analysis of protein. In the proposed approach, we are going to use these three algorithms and also the support vector machine algorithm. The Support Vector Machine algorithm can be used for both linear and non-linear data. It uses the graph method for classifying the data. In our approach, SVM will help us in analysing the change in protein stability. We use support vector machines to predict protein stability changes for single amino acid mutations leveraging both sequence and structural information. Our approach was evaluated using cross-validation methods on a large dataset of single amino acid mutations.

Logistic Regression
In this project, XGBoost algorithm is used as a classification algorithm. The XGBoost does two jobs in this project
1] Maintains the computational power,
In the XGBoost algorithm, Gradient boosted tree algorithms are used. Gradient boosted tree algorithms are capable of selecting the most important uncorrelated features by building small decision trees of a few of the most important and gradually refining the small models by adding new trees.

XGBoost Algorithm
In the XGBoost algorithm, some feature sets are selected by the algorithm. These feature sets are processed by Logical Regression and Support Vector Machine algorithm. On the basis of processing, predictions are made on the basis of the results of the LR and SVM. After these steps are completed, the prediction is made whether the protein sample is ageing related or non-ageing related. The ageing related samples are completely distinguished. These ageing related proteins are then checked for stability. The protein samples consists of lamin which is a main thing required for body growth and ageing. The protein stability is used to detect progeria. If the person is having progeria then the protein sample of the person should be unstable. We are going to use support vector machine for this work. We are going to observe the protein stability in different protein samples and compare them with the stability values of proteins of patients who are suffering from progeria.

---

**Figure 1. Architecture Diagram**
IV. RESULTS AND DISCUSSION:

The main result of the project is the creation of a model which will help to predict progeria by using machine learning. The entire process is described in the proposed system section. We start by extracting important features from the proteins and then use ML to distinguish it as ageing and non-ageing. The table below gives the predicted value of the proteins classified as ageing and non-ageing related.

<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Feature Description</th>
<th>Ageing/Non-Ageing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Damaged DNA Binding</td>
<td>8.8/0.4</td>
</tr>
<tr>
<td>2</td>
<td>Cellular process</td>
<td>9.6/7.2</td>
</tr>
<tr>
<td>3</td>
<td>Growth factor binding</td>
<td>5.3/0.1</td>
</tr>
<tr>
<td>4</td>
<td>Regulation of growth</td>
<td>20.0/2.1</td>
</tr>
<tr>
<td>5</td>
<td>Protein binding</td>
<td>75.7/24.4</td>
</tr>
</tbody>
</table>

The diagram shown below is the predicted graph which will be used to compare protein stability using Support Vector Machine.

V. CONCLUSION

Progeria is predicted using Machine Learning. Ageing related proteins are classified and then are checked for stability and compared with stability of the progeria gene proteins. These method gives predicted results and does not involve the implementation of the actual process. All the values given in the results section are predicted values. The graph obtained is also predicted.

VI. FUTURE WORKS

The project aims in the Prediction Progeria using Support Vector Machine. This project can be further extended by implementing the entire process. By this project, we are able to give a predicted result of the process. The future work includes exact implementation of this process.

VII. REFERENCES


[7]. A Gentle Introduction to XGBoost for Applied Machine Learning by Jason Brownlee on August 17, 2018 in XGBoost.

[8]. Logistic Regression- Detailed Overview, Saishruthi Swaminathan