Submitted By: Piyush Goel (goel.pi@husky.neu.edu)

a) **Project title** - Predicting the clinical impact of human mutation with deep neural networks

b) Names of the team members and their Northeastern University emails.
Piyush Goel - goel.pi@husky.neu.edu

c) **Objectives and significance**

  (a) The **goal** of the project is to reproduce and verify the paper - Sundaram, L., Gao, H., Padigepati, S.R. et al. Predicting the clinical impact of human mutation with deep neural networks. (Ref. a)

  (b) **Significance**: While most genetic mutations are benign, some have deleterious effects on health. If we could classify mutations as benign or harmful, then we could possibly use it to our advantage in making medicine, or preventing disease before it even starts to show any symptoms, the possibilities are be endless.

  (c) The **motivation** for doing this project is to reproduce the paper, verify the results, and learn in the process. Also, this problem seems like one which could have a huge positive impact on human health and lives in general, in the foreseeable future.

d) **Background**
Genetics (Ref. b):

  (a) DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism. A set of 3 of these bases is called a codon and each codon corresponds to one of twenty possible Amino Acids.

  (b) The human body has 23 pairs of chromosomes and each chromosome is made up of multiple genes. A gene is the basic physical and functional unit of heredity. Genes are made up of DNA.

  (c) A gene mutation is a permanent alteration in the DNA sequence that makes up a gene, such that the sequence differs from what is found in most people. Mutations range in size; they can affect anywhere from a single DNA building block (base pair) to a large segment of a chromosome that includes multiple genes.

  (d) Most disease-causing gene mutations are uncommon in the general population. However, other genetic changes occur more frequently. Genetic alterations that occur in more than 1 percent of the population are called polymorphisms. They are common enough to be considered a normal variation in the DNA. Polymorphisms are responsible for many of the normal differences between people such as eye color, hair color, and blood type. Although many polymorphisms have no negative effects on a person's health, some of these variations may influence the risk of developing certain disorders.

(e) Some genes act as instructions to make molecules called proteins. To function correctly, each cell depends on thousands of proteins to do their jobs in the right places at the right times. Sometimes, gene mutations prevent one or more of these proteins from working properly. By changing a gene's instructions for making a protein, a mutation can cause the protein to malfunction or to be missing entirely. When a mutation alters a protein that plays a critical role in the body, it can disrupt normal development or cause a medical condition. A condition caused by mutations in one or more genes is called a genetic disorder. In some cases, gene mutations are so severe that they prevent an embryo from surviving until birth.

(f) Only a small percentage of mutations cause genetic disorders—most have no impact on health or development. For example, some mutations alter a gene's DNA sequence but do not change the function of the protein made by the gene. A very small percentage of all mutations actually have a positive effect.

(g) Because a person's genetic code can have a large number of mutations with no effect on health, diagnosing genetic conditions can be difficult. Sometimes, genes thought to be related to a particular genetic condition have mutations, but whether these changes are involved in development of the condition has not been determined; these genetic changes are known as variants of unknown significance (VOUS) or (VUS). Sometimes, no mutations are found in suspected disease-related genes, but mutations are found in other genes whose relationship to a particular genetic condition is unknown. It is difficult to know whether these variants are involved in the disease.

Deep Learning:

(a) Deep Learning is a type of machine learning method which simply uses neural networks with multiple hidden layers to learn the input to output mapping function.

(b) Convolutional Neural Networks are a class of Deep Neural Networks which are primarily applied to vision based tasks. A Convolutional Layer consists of a kernel which when convoluted with the input produces the output of the layer.

(c) The Convolutional Layers, are generally applied to image type of data, but can also be applied to 1-D data like text or genetic sequence using a single dimensional kernel.

(d) Batch normalization is a technique for improving the speed, performance, and stability of artificial neural networks. It is used to normalize the input layer by adjusting and scaling the activations.

(e) Sometimes the output of a layer is sent to not just the next layer but also to a layer further down the path, i.e. skipping a few layers, this is called a skip connection.

(f) Activation functions are applied to the outputs of the neural network layer to introduce non-linearity in the network which helps it learn complex functions. ReLU (Rectified Linear Unit) is one such very widely used activation function. Its equation is $f(x) = max\{0, x\}$. The sigmod function is another activation function which is generally applied to the last layer of the neural network to make binary classifications. Its equation is $f(x) = \frac{1}{1+e^{-x}}$

Uniqueness:

(a) This paper tries to classifies mutations by training on gene mutation data not only from humans but also from the primates genetically close to humans since most of the mutations are extremely rare in humans and hence are rarely observed, this provides a way to collect large enough data to train a deep neural network. It also has a good accuracy.

e) **Proposed approach**

Method and Implementation:

(a) Deep Neural Networks will be used, constituting mainly 1-D Convolutional Layers with Batch Normalization and the ReLU activation function. The output layer would be using the Sigmoid activation function instead of ReLU, and then one Global Max-pooling Layer (which simply outputs the maximum) as well. The network contains some skip connections as well.

(b) The output of the network is the pathogenicity score for the given mutation.

(c) the model takes multiple matrices as input for one mutation:

    i. A 51 length reference amino acid sequence, i.e. the sequence before the mutation, centered at the point where the mutation will occur.

    ii. A 51 length alternate amino acid sequence, i.e. the sequence after the mutation, centered at the point where the mutation occurred.

    iii. A 51 length position weight matrix calculated using the data from 11 primates.

    iv. A 51 length position weight matrix calculated using the data from 50 mammals excluding primates.

    v. A 51 length position weight matrix calculated using the data from 38 vertebrates.

    vi. Output of a pretrained solvent accessibility prediction network which takes as input the position weight matrix calculated using the data from all 99 vertebrates.

    vii. Output of a pretrained secondary structure prediction network which takes as input the position weight matrix calculated using the data from all 99 vertebrates.

(d) Since there are 20 different possible Amino Acids, therefore each of the 51 length matrix would be 51x20 instead, since either one-hot encoding or the position weight matrix are used.

(e) Python 3 will be used to do most of the coding, and some minor shell scripting to manipulate or look through the data (since the data files are extremely large and can not be loaded into the ram whole at once). The major third-party python libraries required are Tensorflow, Keras, Pandas and Numpy, more may be included if necessary.

(f) The model architecture to be implemented is on page 5.

(g) The architecture for the solvent accessibility and secondary structure prediction models is the same and is on page 6.

Data:

(a) The genetic sequences for GRCh37/hg19 (homo-sapiens) are obtained from the Genome Aggregation Database (gnomAD) (Ref. e), and through the REST API provided by UCSC Genome Browser (Ref. c) to access the data as required. The authors of the paper also provided some data which contains all of the mutations which they used for training and testing.

(b) The actual neural network which predicts the probability of a mutation being pathogenic requires the outputs of two previously trained neural networks which determine a protein's (a sequence of amino acids) secondary structure and solvent accessibility. The plan is to use these weights for now. But if time permits and retraining of these networks is tried then data from the Protein Data Bank (Ref. f) would be required.

Evaluation Strategy:

(a) The classification accuracy would be used to determine if the model works as intended.

Expected Outcomes:

(a) The accuracy of model on both withheld common primate variants and clinical variants increases with the number of benign variants used to train the network.

(b) Training on variants from each of the six non-human primate species independently contributes to an increase in the performance of the network, whereas training on variants from more distant mammals negatively impacts the performance of the network.

Fallback Option:

(a) In case the initial idea fails, the input to the model could be modified to use only the reference and alternate amino acids. This would significantly simplify the network, and even eliminating the need for the secondary structure and solvent accessibility prediction networks. But this would also very likely result in reduced accuracy, since now the network will have lesser information about the amino acid sequence, for example it would now have no idea about its secondary structure.

f) **Individual tasks**

(a) There is only one team member so all of the work will be done by one person.

g) **References**

(a) https://www.nature.com/articles/s41588-018-0167-z

(b) https://ghr.nlm.nih.gov/primer

(c) http://genome.ucsc.edu/

(d) https://en.wikipedia.org/

(e) https://gnomad.broadinstitute.org/

(f) https://www.rcsb.org/

Benign Pathogenic

Frequency

Pathogenicity score for
Chr12: 103246653 C>T
(PAH gene R>Q)

0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

Global Maxpooling
(1)

(51,1)

BatchNormalization +
Sigmoid+
1D Convolution
(1,1,1)

(51,1)

BatchNormalization +
Relu+
1D Convolution
(1,40,40)

(51,40)

+

(51,40)

6 Layers

+ (51,40)

BatchNormalization +
Relu +
1D Convolution
(5,40,40)

(51,40)

BatchNormalization +
Relu +
1D Convolution
(5,40,40)

(51,40)

(51,40)

BatchNormalization +
Relu +
1D Convolution
(5,80,40)

(51,80)

BatchNormalization +
Relu +
1D Convolution
(5,40,40)

(51,40)   Concat   (51,40)

BatchNormalization +
Relu +
1D Convolution
(5,40,40)

(51,40)   (51,40)

(51,40)   Pretrained Solvent
Accessibility Layers   (51,40)

(51,20)

99 Vertebrate
Conservation

(51,20)

(51,40)   Pretrained Secondary
Structure Layers   (51,40)

(51,20)

99 Vertebrate
Conservation

(51,20)

(51,40)   1D Convolution
(1,20,40)   (51,40)

(51,20)

1D Convolution
(1,20,40)

(51,40)

Only Primate
Conservation

(51,40)   +   +   (51,40)

(51,20)   (51,40)   1D Convolution
(1,20,40)   (51,40)   1D Convolution
(1,20,40)

Reference Sequence   (51,20)   Alternative Sequence

Only Mammal
Conservation

(51,20)

(51,40)   1D Convolution
(1,20,40)   (51,40)

(51,20)

Only Vertebrate
Conservation

...CCCCCBBBCCCCCCCCCCCHCCHHHBBBCCCBBCCCCCCCCCCCCCCCCC.... Secondary Structure Prediction

...EEEEBEBIEBIEIBIIIIBBBBBBBBBBBBBBBIBBIEEEEBIEIEIEEBBEE... Solvent Accessibility Prediction

```
                    ┌──────────────────┐
                    │     Softmax       │
                    │    ( 1,40,3)      │
                    └──────────────────┘
                            ↑ ( 51,40)        ┌─────────────────────┐
                    ┌──────────────────┐─────→│ To Pathogenicity Model │
                    │ BatchNormalization + │   └─────────────────────┘
                    │      Relu +       │
                    │   1D Convolution  │
                    │    ( 1,40,40)     │
                    └──────────────────┘
                            ↑ ( 51,40)
                    ┌──────────────────┐
                    │ BatchNormalization + │
                    │      Relu +       │
                    │   1D Convolution  │
                    │    ( 5,40,40)     │
                    └──────────────────┘
                            ↑
             ┌─────────────────────────────┐
             │              ( 51,40)        │
             │              ⊕               │
   6 Layers  │                              │
             │      ┌──────────────┐        │
             │      │     ⊕ ( 51,40)│       │
             │      │ BatchNormalization + │ │
             │      │      Relu +   │        │
             │      │ 1D Convolution│        │
             │      │   ( 5,40,40)  │        │
             │      └──────────────┘        │
             │              ↑ ( 51,40)       │
             │      ┌──────────────┐        │
             │      │ BatchNormalization + │ │
             │      │      Relu +   │        │
             │      │ 1D Convolution│        │
             │      │   ( 5,40,40)  │        │
             │      └──────────────┘        │
             │              ○ ( 51,40)       │
             └─────────────────────────────┘
                            ↑ ( 51,40)
                    ┌──────────────────┐
                    │  1D Convolution   │
                    │    ( 1,40,40)     │
                    └──────────────────┘
                            ↑ ( 51,40)
                    ┌──────────────────┐
                    │  1D Convolution   │
                    │    ( 1,20,40)     │
                    └──────────────────┘
                            ↑ ( 51,20)
                    ┌──────────────────┐
                    │ PWM conservation  │
                    │(Generated using RAPTORX)│
                    └──────────────────┘
```