**Retea neuronala(LSTM)**

De ce LSTM:

Retelele ce utilizeaza layere de acest tip clasifica mai bine instantele in care este relevant ordinea componentelor imputului, in cazul nostru ordinea aparitiilor aminoacizilor este relevanta in determinarea a ce tip de proteina se regaseste.

Procesul de Clasificare:

Reteaua primeste ca input o secventa de aproximativ 4000 de vectori( vectorul este format din 19 de 0 si un 1 pe pozitia ce indica tipul de aminoacid ) fiecare vector reprezantand un tip de aminoacid( sunt 20 de tipuri diferite de aminoacizi in datele oferite). Prelucreaza Inputul prin intermediul a 2 layere de tip LSTM( un layer de acest tip spre deosebire de un layer de tip dens preaia valoriile de la toate nodurile de la layerul precedent cat si de la nodul precedent de pe acest layer), rezultatul este prelucrat de 2 layere de tip dens ultimul avand aproximativ 9000 de noduri (nr de proteine cunoscute din date). Rezultatul acestei retele este unul categoric si va fi de forma unui vector cu o valore de 1 si restul de 0, cifra de 1 va fi pe indicele ce reprezinta proteina identificata. Dupa identificarea proteinei algoritmul va utiliza un api pentru a extrage efectele proteinei respective, dupa extragerea tuturor proteinelor putem determina efectele genomului.

Rezultate Clasificare:

Reteaua are o acuratete de sub 6% pe datele de validare.

Aceste rezultat se datoreaza unor serii de factori multipli:

1. Numarul scazut de neuroni.( fiecare layer intermediar ar trebui sa aiba minim de 10 ori mai multe noduri de cat layer ul de output pentru a putea crea o functie ce sa determine in mod corect rezultatul, nu am avut la dispozitie un server pentru a crea o astfel de retea)
2. Variatia ridicata a datelor( din cele 9000 de proteine gasite majoritatea au intre 1-3 aparitii ceea ce face mai dificil asocierea unei secvente de output cu una de input si astfel procesul de antrenare va necesita mai mult timp)