National College of Ireland

Protein Sequence Classification using Machine Learning and Deep Learning

MSc Research Project
Data Analytics

Shravanee Shekhar Siddha

Student ID: x18180949

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shravanee Shekhar Siddha
Student ID:	x18180949
Programme:	Data Analytics
Year:	2019-2020
Module:	M.Sc. Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	17/08/2020
Project Title:	Configuration manual: Protein Sequence Classification using machine learning and deep learning
Word Count:	7151
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Shravanee Shekhar Siddha
Date:	17th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	Q
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	Q
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	Q

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Protein Sequence Classification using Machine Learning and Deep Learning

Abstract

A number of protein sequences are found and added to the database but its functional properties are unknown. The experiments carried out in the laboratory consume a considerable amount of time for predicting the functions of a protein. Thus, this gives rise to the need of using computational methods for the classification of protein sequences into the respective family. Protein family classification can significantly contribute in the prediction of protein function based on sequence motifs. These factors promote proteomics as a very important area in the field of modern computational biology. This project provides an approach for protein sequence classification using Natural Language Processing (TF-IDF and Word Embedding). Different machine learning models like Decision Tree, Random Forest and deep learning models like Convolutional Neural Network, Long Short-Term Memory were developed and compared for generating efficient protein classification system. The results showcased that Decision Tree showed the highest accuracy of 78.71%, followed by Random Forest and were much faster.

Keywords: Protein Sequence Classification, Proteomics, NLP, TF-IDF, Word Embedding Decision Tree, Random Forest, Convolutional Neural Network, Long Short-Term Memory.

1 Introduction

Proteins play a crucial role by carrying out a number of functions within an organism to sustain its life. It is responsible for performing fundamental functions such as replication of DNA, accelerating necessary metabolic reactions, transporting molecules from cell to cell, reproduction, etc. In addition to this, a protein sequence is formed by twenty different amino acids arranged in a particular order (Carter B et al. (2019)). Both the features, length and sequence of the amnio acid prove to be equally beneficial in the process of gene encoding and also, for a protein to function efficiently. In short, even a single mutation in the gene may result in introducing a wrong amino acid in the sequence.

The fact that it performs all the tasks within an organism based on a very common principle – twenty amino acids that can form a protein, makes it remarkable than the other biomolecules. This is the reason why researchers have focused more on studying proteins by considering its factors such as structure, function, composition, etc. The answer to the questions such as how cancer is caused, why people age, how drugs could be designed for various diseases, how life has evolved on this planet is found by understanding how these proteins fold, how do they function, how they are transformed into complexes, etc (Majhi V. et al. (2019)).

1.1 Motivation and Background

As mentioned earlier, protein sequence is formed by twenty different amino acids. Thus, the structure and function of each protein is decided on the basis of how the amino acids are arranged and what kind of amino acids are used in the formation. As the functional properties of protein are encoded in its sequence, decoding this connection between the protein sequence and its function has continued to be a subject of research in the field of molecular biology due

to its profound implications (Lee E. and Wong A. (2012)).
Determining functions of a protein often includes traditional techniques such as crystallography i.e. structural studies or biochemical studies, which prove to be time consuming. In the recent years, various computational methods have been developed for predicting the function of a protein. A more general and easily accessible approach for prediction of protein function is through inheritance. This implies the idea that proteins with similarity in sequences, carry out similar functions. Hence, protein families are defined to club together, the proteins which have similar functions.
Protein sequence classification into its respective family would give a deeper understanding in the structure, function and metabolic activities of the particular protein. It would help in recognizing the proteins which are unknown or difficult to characterize with help of pairwise alignments. Generally, proteins are categorised with respect to their similarity in structure or sequence. Hence, representing well characterized proteins with known functions. In the cases where a novel protein is identified, its functional properties could be found by considering the group to which it is categorised to belong (Rentzsch, R. and Orengo, C. (2011)). On the contrary, it is observed that proteins have the tendency of transforming its structure or sequence while retaining its functions. Protein sequence classification presents an effective means for retrieving accurate biological information from huge number of data. Additionally, over the years, there has been an immense rise in the generation of biological data, particularly protein data. Therefore, creating a need for developing a new method which could be performed by using advanced tools and techniques for the classification of protein.

1.2 Research Question

Protein function prediction represents a major side of proteomics as it acts as an important factor in biological processes. An accurate classification of protein sequences into their respective families makes the process of molecular analysis much easier and efficient, as it can only be performed on a family of protein sequence and not an individual component of protein. A common approach towards recognizing functional characteristics of an unknown protein is by investigating sequence similarity with annotated protein sequence. This process is time consuming and complicated and hence, classification of protein sequence is preferred over this process. Classification of proteins sequences proves to beneficial in many ways since, it gives idea related to the structure, metabolic activity, etc. of the protein. Moreover, it also enhances the process of identifying protein which are difficult to be characterized.
Hence, there is a need to design a model which can classify protein sequences into protein families efficiently.
Thus, the research question comes -

RQ: "To what extent does the classification of protein sequences by using different feature extraction techniques TF-IDF with Machine learning models (Decision Tree and Random Forest) and Word Embedding with Deep learning models (Convolutional Neural Network and Long Short-Term Memory) help in improving the efficiency of classification, for predicting the function of proteins in the field of proteomics?"

A huge amount of newly discovered protein sequences data is accumulated in the field a bioinformatics. A classification system which classifies protein sequences into their respective family would be useful for determining the structure or function of an unidentified protein. The current classification results are not satisfactory due to improper features in the data (Iqbal M. et al. (2014)). Thus, implementing a new classification system with accurate features would solve this problem.

The research question stated above is solved by following and implementing the research objectives mentioned in section 1.3 (Table 1).

1.3 Research Objectives

Research objectives gives a brief idea on the major activities involved in any research. The research objectives stated below have been designed by studying and analysing different works done in the field of Protein Classification.

Table 1: Research Objectives

Objectives	**Description**	**Evaluation**
Objective - 1	Pre-processing of protein sequence classification dataset	
Objective - 2	Implementing Natural Language Processing (NLP)	
Objective – 2(a)	Implementing TF-IDF for feature extraction	
Objective - 2(b)	Implementing Word Embedding using Keras for feature extraction	
Objective – 3	Using Chi square test for feature selection	
Objective - 4	Implementation, Evaluation and Results of Machine Learning Algorithms	
Objective - 4(a)	Implementation, Evaluation and Results of Decision Tree	Accuracy and Classification Report (Precision, Recall, F1-Score)
Objective - 4(b)	Implementation, Evaluation and Results of Random Forest	
Objective - 5	Implementation, Evaluation and Results of Deep Learning Algorithms	
Objective - 5(a)	Implementation, Evaluation and Results of Convolutional Neural Network	
Objective - 5(b)	Implementation, Evaluation and Results of Long-Short Term Memory	
Objective - 6	Comparison of Machine Learning Techniques with Deep Learning Techniques	Accuracy

Contributions

The major contribution resulting from this project was the implementation of machine learning models such as Decision Tree, Random Forest and deep learning models like Convolutional Neural Network and Long Short-Term Memory that are capable to detect protein sequence classification with good efficiency.
The minor contribution from this project was comparison of the developed models.
The rest of the technical report consists the following chapters: Chapter 2 represents Related Work, Chapter 3 consists of Scientific Methodology and Design Specifications, Chapter 4 talks about Data Pre-processing, Implementation, Evaluation, and Results of Protein Sequence Classification Models, Chapter 5 contains Discussion and Comparison of Results and eventually, Chapter 6 consists of Conclusion and Future Work.

2 Related Work

2.1 Introduction

In the recent years, there has been a remarkable growth in biological data which is complex in nature. This data is available in different types in volume and nature. The data is represented in the form of DNA, RNA and Protein Sequences. Hence, there is a need to improve the existing computer-based techniques in order to extract maximum knowledge; which is required to design the storage, preserve the data, etc. (Kabli F. et al., (2017)).

This section investigates the work done in protein sequence classification and highlights new and significant findings in it. The review is categorised into the following subsections – (i) A Review on Datasets and Features used in Protein Sequence Classification (ii) A critique of Existing Methods, Techniques, Algorithms used in Protein Sequence Classification (iii) An Investigation of Protein Sequence Classification and the Identified Gaps.

2.2 A Review on Datasets and Features used in Protein Sequence Classification

Protein sequence classification has become a backbone for the recent biological information science which involves study of proteins and other molecules, in order to determine the function of a number of new proteins. The research conducted by Li M. et al. (2017) involved classification of GCPRs (G-protein Coupled Receptor) protein sequences. The dataset was divided into three categories. First category contained GCPRs superfamily with total 1019 protein sequences. Each family was further classified into level I subfamily with 991 protein sequences and further into level II subfamily with 872 protein sequences. All these sequences were reviewed in UniProtKB. The data was pre-processed and Term Frequency - Inverse Document Frequency (TF-IDF) and N-gram were used as feature selection techniques.

A study conducted by Lee T and Nguyen T (2019) involved learning of dense vector representation by investigating raw protein sequences. The data was collected from Universal Protein Resource (UniProt) database incorporating 3,17,460 protein sequences with 589 families. A distributed representation was created by representing each sequence as a series of trigram (overlapping), using Global Vectors for Word Representation (GloVe).

Vazhayil A et al., 2019 considered protein family as a set of proteins which share similar

structure at both - sequence as well as molecular level and represent same functions. Even though a huge number of sequences are known, it is observed that there exists a lack of knowledge for the functional properties of the protein sequences. For this research, a data consisting of 40433 protein sequences had been collected from Swiss-Prot Protein family database (Pfam), incorporating 30 different families. The dataset was checked for data redundancy, which showed that no redundancy of protein sequences was found. Initially, Keras word embedding and N-gram was applied on the text data to represent discrete characters as vectors of continuous number.

For this project, TF-IDF and Keras Word embedding have been chosen as feature extraction techniques, so that Machine Learning and Deep Learning algorithms could capture optimal information form the protein sequences. Term Frequency - Inverse Document Frequency (TF-IDF) as it calculates the importance of a protein sequence from a document. It is used to highlight the relevant features and suppress the less important features by managing the feature weight. Additionally, Keras Word Embedding is chosen due to its ability to restore the sequential information in the protein sequences.

2.2 A Comparison of the Reviewed techniques

Table 2: Comparison of reviewed datasets and techniques

Datasets	Applied techniques	Result	Authors
Protein Data Bank	Decision Tree, Random Forest and Extra Trees	Decision Trees with 91.6, Random Forest with 92.7%	Parikh Y. et al. (2019)
PROSITE database	Support Vector Machine, Decision Tree and Naive Bayes	Decision Trees with 77.3%	Yang Y. et al. (2007)
SCOP	Random Forest, K star, Nearest Neighbour, Logistic Regression, Support Vector Machine and Naïve Bayes	Random Forest with 73.7%	Li J. et al. (2013)
Structural Protein Sequences	Convolutional Neural Network (CNN)	Convolutional Neural Network with 90%	Shinde A and D'Silva M (2019)
4prot and CB513 CullPDB	Convolutional Neural Network (CNN)	Convolutional Neural Network with 90.93%	Jalal S. et al. (2019)
Swiss-Prot	Deep Neural Network (DNN), Recurrent Neural Network (RNN), Long-Short Term Memory (LSTM) and Convolutional Neural Network (CNN)	Long-Short Term Memory with 91.24%	Naveenkumar K S et al. (2019)

From the comparison between works done by various researchers, it is clear that Machine Learning Techniques like Support Vector Machine (SVM) and Decision Tree performs well in Protein Sequence Classification whereas considering Deep Learning Techniques – Convolutional Neural Network and Long-Short Term Memory gives better results as compared to the other techniques.

2.4 An Investigation of Protein Sequence Classification and the Identified Gaps

In the recent years, bioinformatics field has witnessed many challenges including prediction of enzyme class from a newly discovered, unidentified protein. This prediction of enzyme class creates a good chance for the researchers, due to addition of new proteins (Bihari A et al. (2019)). The main objective of this research was to implement machine learning techniques for feature selection and prediction and select the best suitable technique. Seven different types of classification techniques were implemented, examined and compared for choosing the technique with the best performance. The dataset consisted of 4,368 protein sequences which belonged to six distinct enzyme classes. After evaluation, it was observed that enzyme classes four, five and six are imbalanced in comparison with the other classes. Moreover, the precision and recall values showed comparatively low values than the other classes. Hence, the research concluded that the factors like small size of the data and imbalanced data majorly affected the performance of the classification techniques.

To tackle the same problem, Suvarna Vani K and T.D.Sravani (2014) used SMOTE (Synthetic Minority Over-Sampling Technique) algorithm for rebalancing the data and improving the score of accuracy. In addition to this, SMOTE has come across very useful when it comes to classification problems related to proteins. But then again, it also has some limitations like over-fitting and over-generalization. Thus, to overcome this problem the authors introduced a sampling method designed by them. Application of this new sampling methods could not balance the dataset but it minimized the problem of imbalanced data to some extent and also, improved the accuracy in comparison to the previous method. Another limitation of this research was the size of the dataset. The dataset used for this research consisted of only 659 protein sequences. Thus, due to insufficient data, the model could show a better accuracy.

A study conducted by G. Mirceva (2019) on different feature selection techniques for improving the performance of protein classification put-forth that since feature selection helps to choose most important features in a data and also, decreases the dimensionality of a dataset, it is looked upon as a crucial step in classification of proteins. Selecting important features makes the process of classifying proteins much easier. Therefore, this research followed the process of feature extraction by using voxel-based descriptor wherein each protein was represented as point in feature vector space. Then, important features from the descriptor were selected by implementing a few traditional feature selection techniques –Gain Ratio, Pearson's Correlation Coefficient, Relief and Greedy Stepwise. Finally, by using the selected features a classification models were executed. The estimated results showed less accuracy for the models which used Gain Ratio and Relief as feature selection techniques. Thus, this research gives attention to the fact that how important it is to select an appropriate feature selection technique in order to generate better accuracy and improve the classification of proteins.
The above section represents gaps observed in previous works. Thus, considering these gaps, this project would aim to bridge the gaps identified and generate a better performance.

2.6 Conclusion

From the above reviewed work, it can be seen that there is a need for developing efficient classification system by using Natural Language Processing (NLP) which can obtain accurate results and answer the research question and objectives mentioned in the previous sections. Also, traditional machine learning algorithms derive information directly from amino acid sequences where as Deep learning algorithms use abstraction layers to capture complex representations from the raw data. Thus, implementing machine learning and deep learning with different feature extraction techniques together would provide deeper insights and showcase variations in the results. Hence, by considering all the significant information and identified gaps obtained from various research, a good classification system can be developed by performing required data pre-processing, using appropriate feature selection techniques and implementing best possible models.
The next chapter discusses the scientific methodology and design specifications selected to develop the Protein Sequence Classification model which will be useful in the field of bioinformatics.

3 Scientific Methodology Approach and Design Specification

Data Mining process constitutes of various techniques like CRISP-DM, SEMMA and KDD. After understanding and examining these techniques, KDD i.e. Knowledge Discovery in Database was found to be systematic and hence, was selected as it best fits the requirement of this project.

3.1 An approach for Protein Sequence Classification

The methodology of Protein Sequence Classification is an iterative process and involved conversion of raw data into high-level information. Following are the steps implemented in this methodology –
(a) Data selection - Data selection is the fundamental step where the dataset was collected from Kaggle. Since, the data consisted of 4,67,305 amino acid sequences, it was found to be appropriate for this project due to its large data size. The dataset was available in two csv files.
(b) Data cleaning and pre-processing - Then, cleaning and pre-processing of the data was done as the data after dataset selection is not always ready to use. This involved combining the two .csv format files, removing unwanted rows and columns, missing values, etc.
(c) Data transformation - Transformation of the data was done by using TF-IDF and Keras Word Embedding and chi-square as feature selection technique for generating optimal feature representation. For the problem of imbalanced data, random under-sampling was used.
(d) Data mining - This phase consisted of implementing Machine Learning (Decision Trees and Random Forest) and Deep Learning models (Convolutional Neural Network and Long Short-Term Memory) on the transformed data.
(e) Evaluation - Finally, all the models were evaluated by using Accuracy, Precision, Recall and F1-score as evaluation metrics and the performances of the models were compared by using accuracy score.
Figure 1 shows the step by step methodology for Protein Sequence Classification.

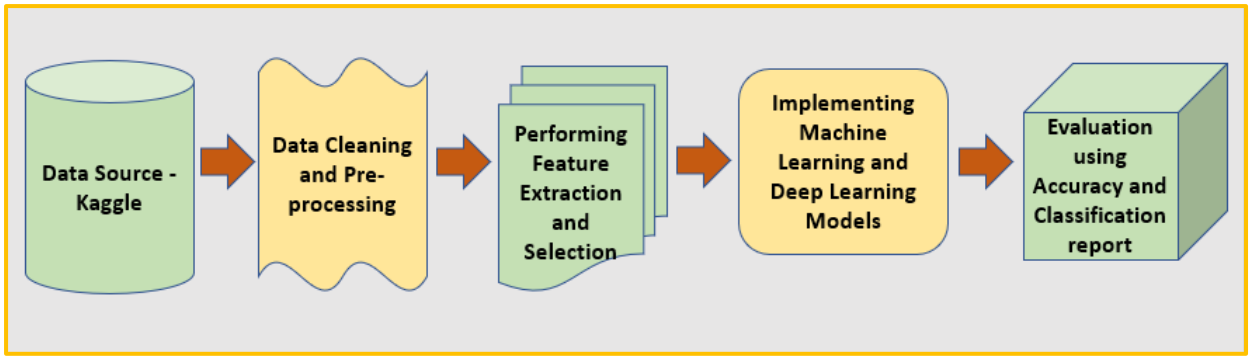

Figure 1: Approach for Protein Sequence Classification

3.2 Project Design Specification

The design process for the Protein Sequence Classification involves two layers – Presentation Layer and Business Logic Layer. Both the layers represent its own significance. The dataset was available in a well-structured format and thus, required only data processing. Hence, the 2-tier architecture was preferred. These two layers together depict the entire process followed in this project. The Presentation Layer represents the evaluation of results by creating data visualizations in Python. The other layer which is the Business Logic Layer shows the process of data collection, data pre-processing and transformation and the implementation of different models.
The Figure 2 represents the design process for Protein Sequence Classification.

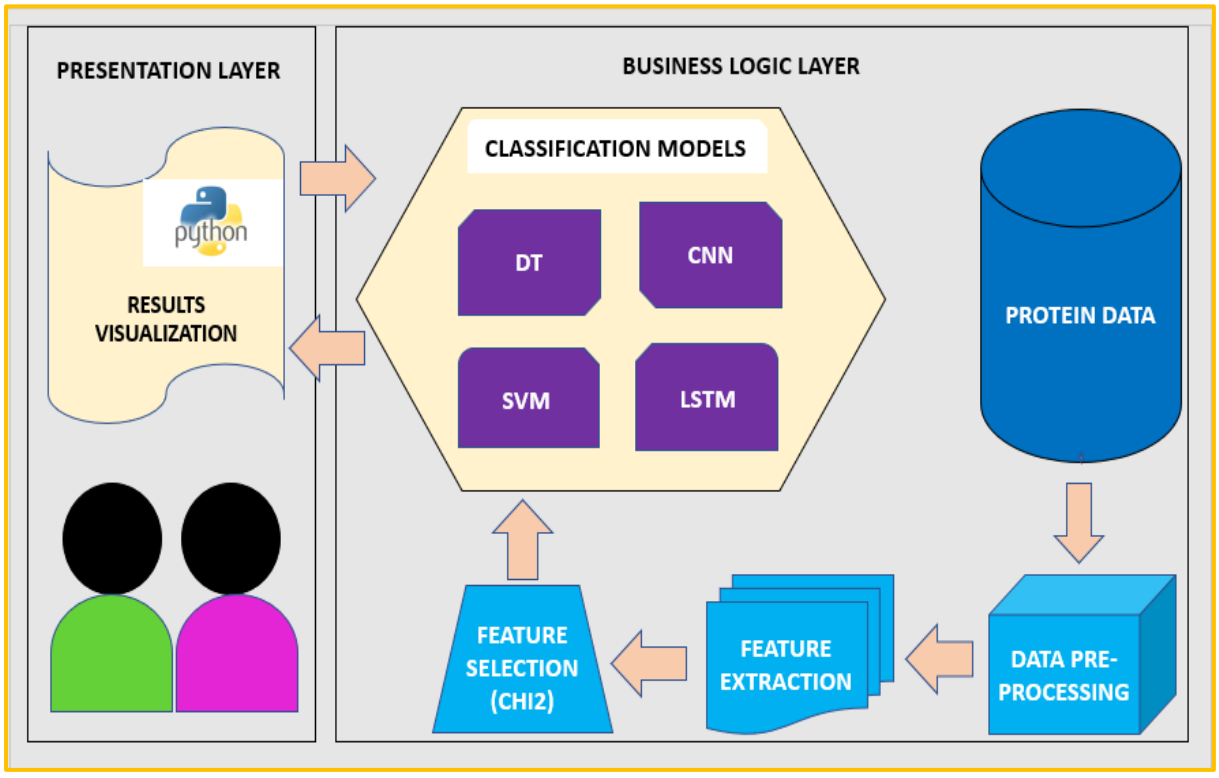

Figure 2: Design Process for Protein Sequence Classification

The implementation, evaluation and results of Machine Learning and Deep Learning Techniques have been discussed in detail in the next chapter.

4 Implementation, Evaluation and Results of Protein Sequence Classification Techniques

4.1 Introduction

This chapter gives an overview of steps included in the implementation and evaluation of results. The below figure 3 depicts the workflow for Protein Sequence Classification. This chapter also, elaborates the workflow of this project.

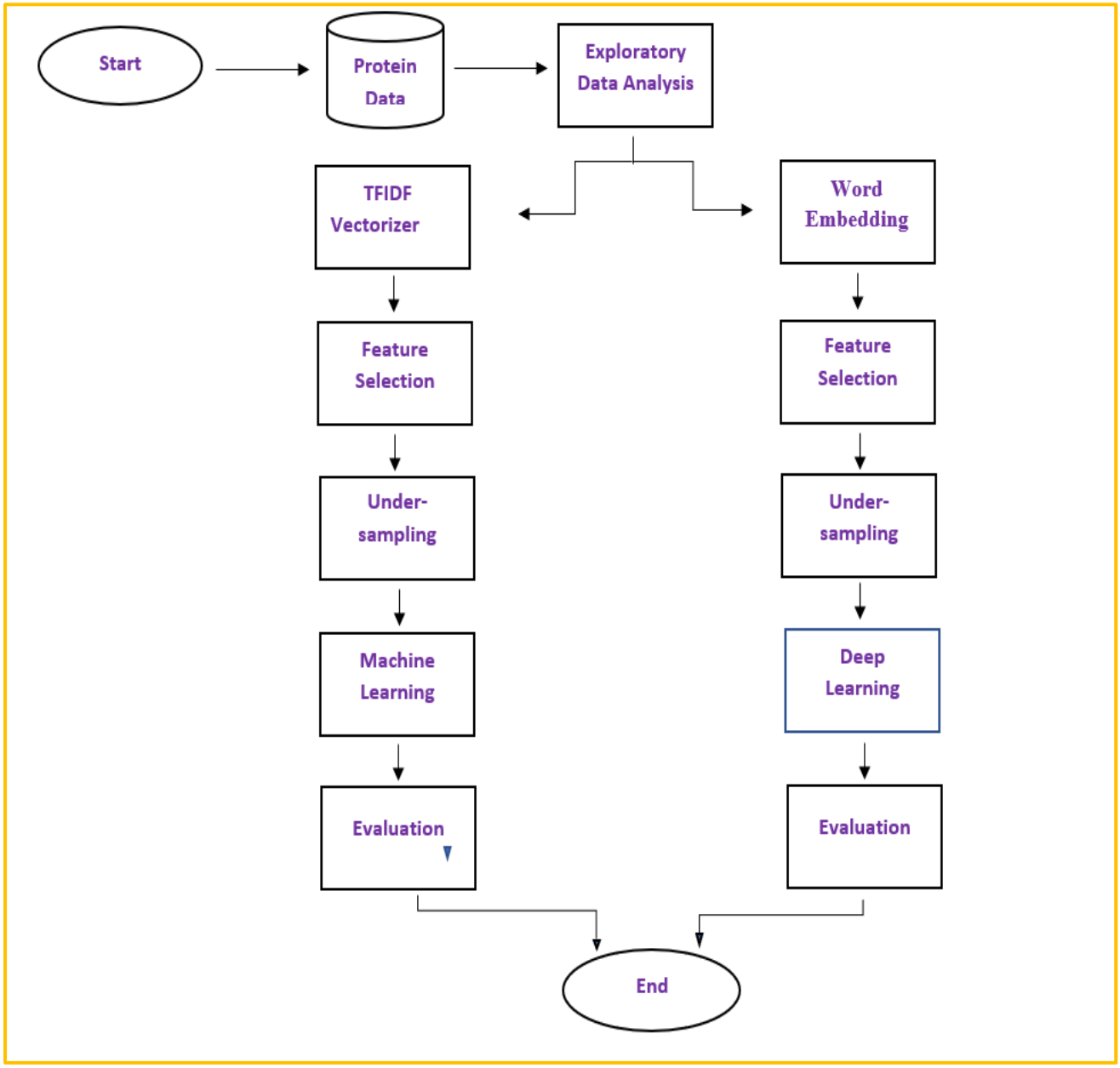

Figure 3: Workflow diagram for Protein Sequence Classification

4.2 Data Pre-processing

The raw data is not often ready to be used for running different models after the selection process. Hence, the data had to undergo a cleaning process which ensured removal of missing values and incorrect data. It involved discarding the errors, fluctuations or variance in the data, also referred to as noisy data.
The dataset was available in two separate csv files. Hence, the two datasets were merged by using a common attribute “StructureID”. The merging of datasets was followed searching for the rows which showed protein as their macromolecule type because this dataset offered a variety of macromolecules like RNA, DNA, Protein, etc. After selecting the protein macromolecule, it was observed that 4,67,305 consisted of protein sequences. Then, more cleaning of the data was done by removing all the missing values i.e. eliminating the rows which had no sequences or labels and columns which were not important. Some of the sequences contained the letter ‘X’ which represented unknown protein sequence. Hence, those sequences were also dropped. By implementing Exploratory Data Analysis, the dataset was reduced to top 20 most common protein classes. The labels of those classes were transformed from string to numeric represenation by using LabelEncoder() function form the sklearn library.

4.3 Exploratory Data Analysis

Exploratory Data Analysis is one of the best process which helps in understanding the raw data and suggests what processing needs to be done on the data. It is a process which involves investigating the raw data in order to detect the underlying patterns, recognize abnormality. An additional attribute “Sequence length” was considered while performing exploratory data analysis. The cleaned features were stored in a new file and exploratory data analysis was performed on the new data. As this project involved classification of protein sequence, this analysis showed the trends related to protein family i.e. classes such as frequency distribution of the sequence length and top 20 classes which have longest sequence length (Refer figure 5 and 6).

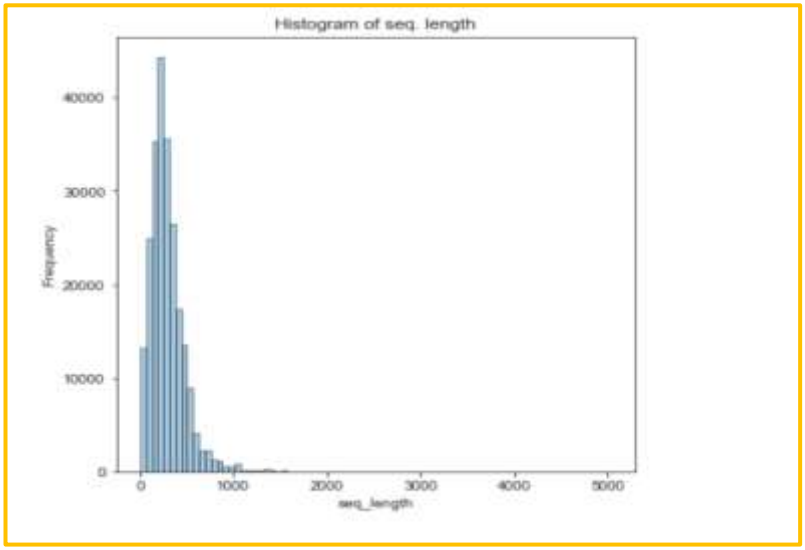

Figure 4: Length of protein sequences

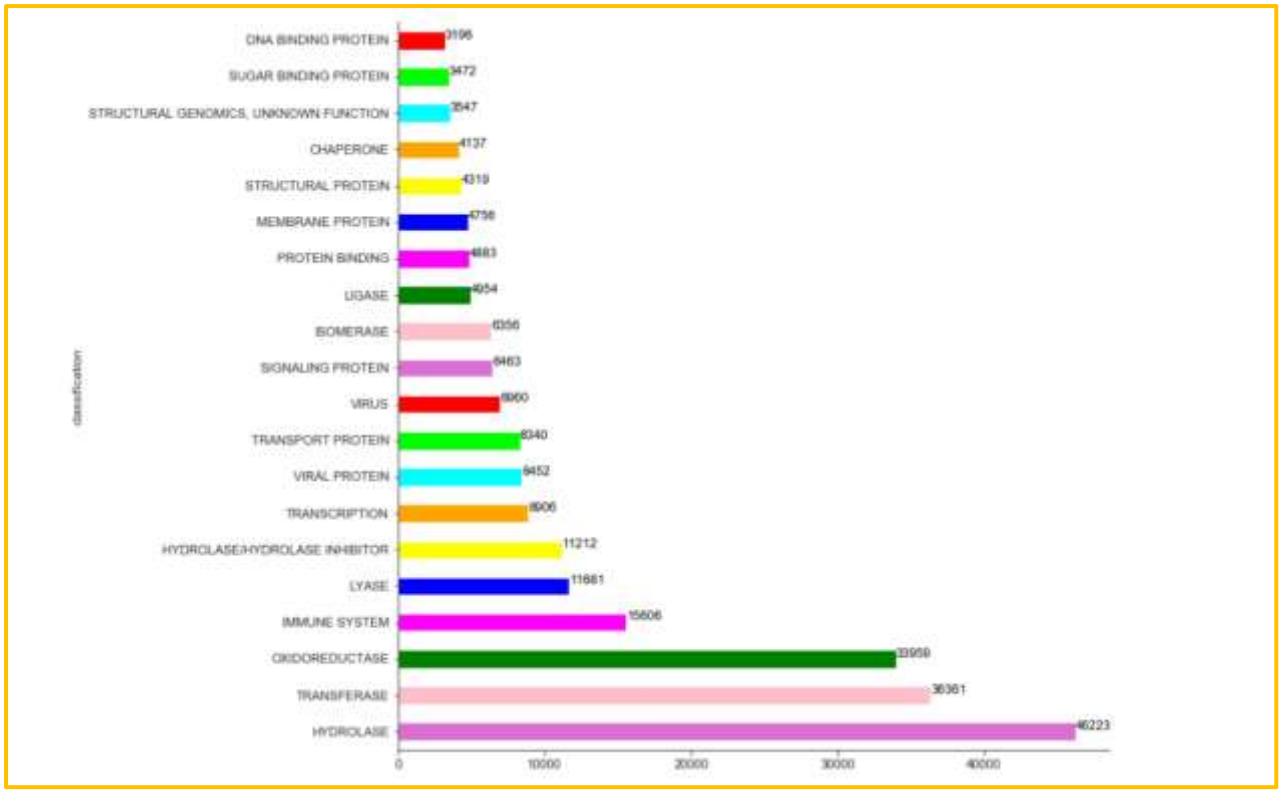

Figure 5: Bar plot with top 20 frequently occurring classes

4.4 NLP Techniques

Natural Language Processing i.e. NLP is considered to be an area of study in the field of Artificial Intelligence. NLP involves understanding, analysing and processing of human language. With the advancement in Machine Learning techniques, NLP is used primarily in different organizations for chatbots, sentiment analysis, market intelligence, etc. Natural Language Processing offers a variety of techniques for extracting significant features from the raw data, so that it can be used for training a classification model. This project studies the traditional TF-IDF (used with machine learning) and the well-known Word Embedding using Keras (used with deep learning) and compares the results developed. Thus, this project demonstrates 2 different feature extraction techniques on multiclass dataset.

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF or Term Frequency-Inverse Document Frequency is a statistical technique which examines how important a word is for a document. TF-IDF focuses on relevant words and supress the words which are irrelevant. Hence, this technique is used for extracting significant features from the protein sequences. In this project, a feature was considered as a highly distinctive feature, if it showed a greater frequency of occurring in a protein sequence and that of lower in the other protein sequences. For implementing TF-IDF, TF-IDFVectorier() function was used along with n-gram.

Word Embedding using Keras

Word embedding is a technique used for denoting a word as an n-dimensional vector. It uses Keras library that includes a layer– Embedding(). This layer serves a class in Keras and is generally treated as the first layer in a sequential model, while performing Natural Language

Processing. Also, for converting every character in a sequence into number, tokenizer() function has been used. Additionally, pad_sequences is used to maintain the length of the sequences.

4.5 Feature Selection

Feature selection is a critical part which involves selecting features that are really important for a model. This in turn enhances the speed of any machine learning algorithm and also, minimizes the complexity. Moreover, it provides a good accuracy as well as also, reduces the chances of overfitting. In short, feature selection is responsible for the performance of a model. There are different techniques available for feature selection. For this project, Chi Square technique was implemented as a feature selection technique.

4.5.1 Chi-square Test

Before implementing machine learning algorithms, it is necessary to filter the original features. Chi-square test is a statistical test that plays a crucial role in feature selection. It investigates the relationship between the dependent and independent features of a data and eventually, discards the features which are less important. By using chi-square test between the feature and classification category, the correlation degree between the same was examined and then, the features were chosen. As the number of features extracted by the feature extraction techniques were very large, chi-square test was chosen as a feature selection technique. The work done by Cheng B. et al. (2005) shows that chi-square techniques has been useful as it reduced the number of features along with discarding the noisy features for generating an accurate classification. Thus, chi2 and SelectKBest packages were used from the sklearn library for implementing feature selection.

4.6 Random under-sampling

Imbalanced data is one of the most crucial issue in the field of machine learning. It is often observed that sizes of classes in a particular data vary from one another. The classes which contain greater number instances are defined to be majority whereas the classes with lesser number of instances are defined to be minority. To handle this issue of imbalanced data, Random Under-sampling is used. As, the name suggests, it involves removal of classes which are over-represented. In simpler words, it discards samples of the majority classes. This technique has been chosen for balancing, as the data is huge.
For this project, imblearn package was imported for implementing this technique. This package provided a function RandomUnderSampler() which fostered under-sampling of the majority classes. This technique reduced the number of samples from the classes which were over-represented.
This has achieved the research objectives from objective-1 to objective-3 stated in Section 1.3. The next section discusses the implementation and evaluation of machine learning and deep learning techniques in detail.

4.7 Implementation, Evaluation, and Results of Protein Sequence Classification Models

After feature extraction and feature selection, machine learning models like Decision Tree and

Support Vector Machine and deep learning models like Convolutional Neural Network and Long-Short Term Memory were implemented. The performance of these models was estimated by using Accuracy, Precision, Recall, F1-score as evaluation metrics. A classification report was generated for measuring the value of predictions on per-class basis. The metrics were explained in terms of true positives, false positives, true negatives and false negatives, which helped in checking whether the predictions are correct or incorrect –

True Positive (TP) - Occurred when a class was positive and predicted positive.
False Positive (FP) - Occurred when a class was negative but predicted positive.
True Negative (TN) - Occurred when a class was negative and predicted negative.
False Negative (FN) - Occurred when a class was negative but predicted positive.

In this case, positive and negatives were the generic names for classification problem.

Accuracy - Accuracy of a model shows how a model a may perform and whether it is being trained accurately or not. The greater value for accuracy represents better performance of the model.

$$\textbf{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

Precision - Precision is the ratio of correctly classified positive instances to the overall classified positive instances. This evaluation metric represents the exactness. It can also be referred to as accuracy generated due to positive predictions.

$$\textbf{Precision} = (TP)/(TP+FP)$$

Recall - Recall often known as Sensitivity is the ratio of instances which are correctly class as positive to the overall instances. This denotes the measure of completeness. In simpler terms, it is the fraction of positives that were correctly detected.

$$\textbf{Recall} = (TP)/(TP+FN)$$

F1-score - F1-score represents the harmonic mean of precision and recall value. The F1-score highlights how accurate and robust the classifier is. F1-scores tend to be lower than the accuracies as they compute mean of precision and recall.

$$\textbf{F1-Score} = 2*((Precision*Recall)/(Precision+Recall))$$

4.8 Experiment 1: Implementation, Evaluation and Results of Decision Tree Model

4.8.1 Implementation

Decision Trees are considered as one of the strongest predictors that provide a clear explanation of concept for particular data (Sikandar A. et al. (2017)). The fact that they are potentially fast and generate models which perform good on different features, makes decision tree a well-known model. The Decision Tree Classifier while performing, makes use of predictive tree type of model i.e. the Decision Tree. The model thus, determines the value of the class in the

leaves by observing the classes in branches.
The Decision Tree Classifier was implemented by dividing the data into train and test. The function DecisionTreeClassifier() provided by the sklearn library was used to implement this classifier. The model was implemented on the data extracted using TF-IDF.

4.8.2 Evaluation and Results

The overall accuracy gained by Decision Tress is 78.71%, which is quite good. A classification report (refer figure 6) was generated for checking the values of precision, recall, F1-score for twenty different classes. The last class i.e. the Virus showed the highest values for precision - 97%, recall - 90% and F1-score - 93% respectively.

	precision	recall	f1-score
CHAPERONE	0.93	0.82	0.87
DNA BINDING PROTEIN	0.84	0.67	0.75
HYDROLASE	0.66	0.32	0.44
HYDROLASE/HYDROLASE INHIBITOR	0.92	0.81	0.86
IMMUNE SYSTEM	0.94	0.77	0.84
ISOMERASE	0.92	0.82	0.87
LIGASE	0.89	0.78	0.83
LYASE	0.94	0.90	0.92
MEMBRANE PROTEIN	0.85	0.71	0.78
OXIDOREDUCTASE	0.95	0.88	0.91
PROTEIN BINDING	0.80	0.52	0.63
SIGNALING PROTEIN	0.86	0.62	0.72
STRUCTURAL GENOMICS, UNKNOWN FUNCTION	0.80	0.68	0.73
STRUCTURAL PROTEIN	0.88	0.66	0.76
SUGAR BINDING PROTEIN	0.93	0.70	0.80
TRANSCRIPTION	0.85	0.57	0.68
TRANSFERASE	0.92	0.86	0.89
TRANSPORT PROTEIN	0.91	0.76	0.83
VIRAL PROTEIN	0.91	0.78	0.84
VIRUS	0.97	0.90	0.93
micro avg	0.92	0.79	0.85
macro avg	0.88	0.73	0.79
weighted avg	0.91	0.79	0.84
samples avg	0.79	0.79	0.79

Accuracy: 0.7871615422924694

Figure 6: Decision Tree Results

4.9 Experiment 2: Implementation, Evaluation and Results of Random Forest Model

4.9.1 Implementation

Random Forest is one of the popular techniques preferred for executing classification problems (Hakala K. et al. (2019)). It is a type of prediction algorithm which is based on the ensemble of decision tree. Advantage of using Random Forest for protein sequence classification is that

it generates good classification performance and also, support the multi-label classification on a huge number of labels.
The Random Forest Classifier was implemented by dividing the data into train and test. The function RandomForest() provided by the sklearn. library was used to implement this classifier. Eventually, the model was implemented on the extracted data using TF-IDF.

4.9.2 Evaluation and Results

The overall accuracy gained by Random Forest Classifier is 77.24 %. The class Oxidoreductase showed the maximum values for precision - 98%, whereas Virus showed highest value for recall - 90% and f1-score - 93%. The classification report evaluated below (refer figure 7) gives a summary of the precision, recall and f1-score values for all the twenty classes.

	precision	recall	f1-score
CHAPERONE	0.96	0.80	0.87
DNA BINDING PROTEIN	0.88	0.64	0.74
HYDROLASE	0.77	0.30	0.43
HYDROLASE/HYDROLASE INHIBITOR	0.93	0.81	0.87
IMMUNE SYSTEM	0.95	0.76	0.85
ISOMERASE	0.95	0.81	0.87
LIGASE	0.93	0.75	0.83
LYASE	0.96	0.89	0.92
MEMBRANE PROTEIN	0.85	0.71	0.77
OXIDOREDUCTASE	0.98	0.86	0.92
PROTEIN BINDING	0.85	0.49	0.62
SIGNALING PROTEIN	0.91	0.59	0.72
STRUCTURAL GENOMICS, UNKNOWN FUNCTION	0.89	0.62	0.73
STRUCTURAL PROTEIN	0.90	0.65	0.76
SUGAR BINDING PROTEIN	0.96	0.68	0.80
TRANSCRIPTION	0.87	0.56	0.68
TRANSFERASE	0.97	0.84	0.90
TRANSPORT PROTEIN	0.94	0.74	0.83
VIRAL PROTEIN	0.93	0.76	0.84
VIRUS	0.97	0.90	0.93
micro avg	0.94	0.77	0.85
macro avg	0.92	0.71	0.79
weighted avg	0.94	0.77	0.84
samples avg	0.77	0.77	0.77

Accuracy : 0.7724568163350896

Figure 7: Random Forest Results

4.10 Experiment 3: Implementation, Evaluation and Results of Convolutional Neural Network Model

4.10.1 Implementation

According to the authors, Zhang D. and Kabuka M. (2015), Convolutional Neural Networks or

CNN has been adopted from the concept of neurobiology (visual cortex) incorporating different convolutional layers with fully connected layers. Between these layers, pooling functions are embedded so as to lessen the features. CNN has received a lot appreciation and popularity in the computer vision area, because of the convolutional filters which were used to find the image patterns. Thus, this asset can also be shared with protein sequencing data.

By using multiple combination of layers, the final model for CNN had two convolutional layers. Each layer was followed by max pooling. The flat array passed the extracted feature ti dense network with relu as an activation function. Eventually, a softmax classifier was used which classified the sequence into respective families/classes. The model was compiled by using categorical cross entropy and adam as an optimizer.

4.10.2 Evaluation and Results

The overall accuracy obtained by Convolutional Neural Networks was 75%. A classification report for all the twenty classes has also been generated below (refer figure 8). It was observed that Virus class showed the maximum value for precision - 94%, recall - 90% and f1-score - 92%.

	precision	recall	f1-score
CHAPERONE	0.79	0.67	0.73
DNA BINDING PROTEIN	0.83	0.39	0.53
HYDROLASE	0.49	0.22	0.31
HYDROLASE/HYDROLASE INHIBITOR	0.83	0.80	0.81
IMMUNE SYSTEM	0.68	0.80	0.74
ISOMERASE	0.85	0.81	0.83
LIGASE	0.77	0.69	0.73
LYASE	0.83	0.87	0.85
MEMBRANE PROTEIN	0.60	0.64	0.62
OXIDOREDUCTASE	0.81	0.91	0.86
PROTEIN BINDING	0.52	0.22	0.31
SIGNALING PROTEIN	0.56	0.44	0.50
STRUCTURAL GENOMICS, UNKNOWN FUNCTION	0.53	0.23	0.32
STRUCTURAL PROTEIN	0.69	0.57	0.62
SUGAR BINDING PROTEIN	0.76	0.71	0.73
TRANSCRIPTION	0.38	0.64	0.48
TRANSFERASE	0.86	0.83	0.85
TRANSPORT PROTEIN	0.76	0.71	0.73
VIRAL PROTEIN	0.71	0.73	0.72
VIRUS	0.94	0.90	0.92
accuracy			0.75
macro avg	0.71	0.64	0.66
weighted avg	0.75	0.75	0.75

Figure 8: Convolutional Neural Network Results

4.11 Experiment 4: Implementation, Evaluation and Results of Long Short-Term Memory Model

4.11.1 Implementation

Long Short-Term Memory or LSTMs prevent the problem of passing irrelevant information by using input gate which is responsible for storing information. This is an efficient unit as it passes the relevant information, during sequence processing. The input gate and the output gate together form the memory cell. Thus, LSTM network consist of many such memory cells or sub-architectures (Hochreiter S. and Obermayer K. (2005)).
In this project, an embedding layer for mapping input layer is used. Then, a layer of LSTM is used. The dense layer with relu activation is used that adds representational capacity to the model. Eventually, a dense layer is used as output layer with softmax activation. The model is trained using categorical_crossentropy and compiled using adam optimizer.

4.11.2 Evaluation and Results

Long Short-Term Memory showed the lowest overall accuracy i.e. 51%. The classification report for all the twenty protein classes is given below (refer figure 9). It can be seen that Virus class showed the highest value for precision - 83% and f1-score - 79% and class Immune System showed highest value for recall - 79%.

	precision	recall	f1-score
CHAPERONE	0.63	0.48	0.55
DNA BINDING PROTEIN	0.51	0.14	0.22
HYDROLASE	0.50	0.01	0.03
HYDROLASE/HYDROLASE INHIBITOR	0.76	0.67	0.71
IMMUNE SYSTEM	0.61	0.79	0.69
ISOMERASE	0.59	0.18	0.28
LIGASE	0.64	0.10	0.17
LYASE	0.54	0.35	0.42
MEMBRANE PROTEIN	0.56	0.23	0.33
OXIDOREDUCTASE	0.50	0.72	0.59
PROTEIN BINDING	0.18	0.05	0.07
SIGNALING PROTEIN	0.28	0.10	0.15
STRUCTURAL GENOMICS, UNKNOWN FUNCTION	0.20	0.02	0.03
STRUCTURAL PROTEIN	0.61	0.20	0.30
SUGAR BINDING PROTEIN	0.52	0.52	0.52
TRANSCRIPTION	0.27	0.36	0.31
TRANSFERASE	0.46	0.70	0.55
TRANSPORT PROTEIN	0.49	0.27	0.35
VIRAL PROTEIN	0.57	0.52	0.54
VIRUS	0.83	0.75	0.79
accuracy			0.51
macro avg	0.51	0.36	0.38
weighted avg	0.52	0.51	0.48

Figure 9: Long Short-Term Memory

4.12 Conclusion

On the basis of implementation and estimated results, the research objectives till Objective - 5 (section 1.3) have been achieved. The models developed - Decision Tree, Random Forest with TF-IDF performed better than Convolutional Neural Network and Long Short-Term Memory with Word Embedding.
The next section shows the comparison between the developed models with the help of visualization.

5 Discussion and Comparison of Results

5.1 Comparison of Developed Models

This section comprises of the comparison between the developed models i.e. comparison between machine learning and deep learning models. The comparison between these models is done by using accuracy as an evaluation metrics. The Decision Tree Classifier shows the highest accuracy i.e. 78.71% and hence, performs better than the other models. In addition to this, Random Forest Classifier also performed well and showed an accuracy of 77.24%. Both the machine learning models gave a better performance with TF-IDF as feature extraction technique. Additionally, the Convolutional Neural Network performed well and gave an accuracy of 75.16%, which was slightly lower than the machine learning models. On the other hand, Long Short-Term Memory gives a satisfactory accuracy of 51.44% and lowest of all the models. Thus, when compared the machine learning models performed better than deep learning models. This proves that the traditional feature extraction technique TF-IDF with machine learning performed better than Word Embedding technique with deep learning.

The table below (refer figure 10) highlights the accuracy scores for the machine learning models - Decision Tree, Random Forest and deep learning models - Convolutional Neural Network and Long Short-Term Memory (Test accuracies have been considered.)
The Figure 11 represents comparison between the accuracy scores for machine learning and deep learning models.

	Accuracy
Decision Tree	0.787162
Random Forest	0.772457
Convolutional Neural Network	0.751645
Long Short-Term Memory	0.514403

Figure 10: Accuracy of all the models

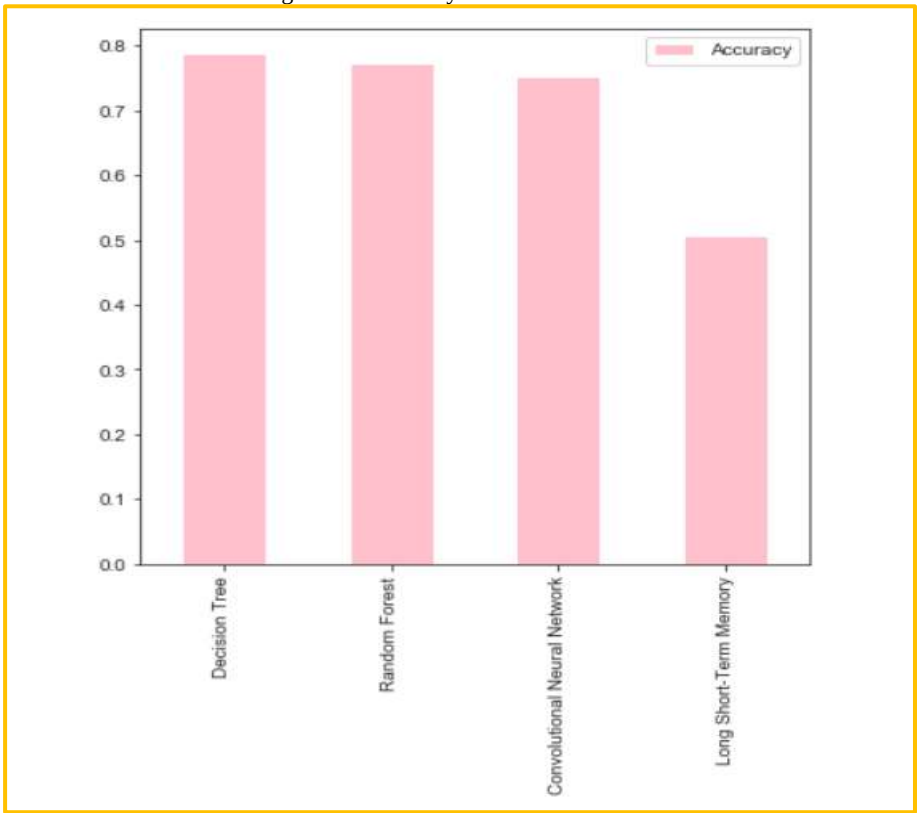

Figure 11: Comparison between the accuracies of machine learning and deep learning models

Thus, comparison of accuracies between all the developed models suggest that Decision Tree, Random Forest i.e. machine learning models show better performance than deep learning models and can be used for protein sequence classification to predict the function of protein.
This has achieved the last research objective (section 1.3) i.e. objective – 6 along with the research question (section 1.2).
The next chapter discusses the conclusion and future work for protein sequence classification.

6 Conclusion and Future Work

The need for precise, efficient and automated protein sequence classification methods continues to grow as advances in bioinformatics disclose new proteins. A basic task in protein sequence analysis is to discover the family/ class of a protein sequence as it reflects the function of a protein. Most of the work mentioned above (section 2), involved dealing with datasets of comparatively small size. Experimenting with different feature extraction technique with machine learning and deep learning models was the main motivation of this project.
A larger dataset was used to generate accurate results and different feature extraction techniques like TF-IDF and Word Embedding using Keras was implemented for representing protein sequence into numeric vectors. To pass the relevant features to the models, chi-square technique was applied for feature selection. Before passing the data to the models, the data was resampled to handle problem of imbalanced data. The random under-sampling was used due to large the data size. It removed majority classes. Lastly, the models Decision Tree, Random Forest, Convolutional Neural Network and Long Short-Term Memory were implemented and evaluated by using accuracy and classification report for generating results.

It can be concluded that both TF-IDF and Word embedding worked well for representation of protein sequence. The models Decision Tree, Random Forest and Convolutional Neural Networks gave good performance whereas Long Short-Term Memory gave poor performance and showed the lowest accuracy out of all the models. The execution time for deep learning models was found to be time-consuming specially for LSTM. On the other hand, machine learning models were fast enough to execute and evaluate the results. Also, it was observed that combination of TF-IDF and machine learning models generated better results. It is clear that machine learning models with TF-IDF as feature extraction technique developed more efficient and fast results than deep learning models with Word Embedding. Therefore, the research objectives were accomplished along with good results.

6.1 Future Work

Further improvements can be done in this project by training the models on all the classes. For this project, top 20 protein classes have been used whereas this project could be extended further by considering all the classes i.e. 4416 classes. This would increase the size of the data and make the models more reliable. The larger data, the better performance. Also, a different technique to handle the imbalanced data could be used for generating good results.

Acknowledgement

I would like to express my heartfelt gratitude towards my mentor Dr.Catherine Mulwa. This project would not have been possible without her guidance, motivation and constant support. She always went the extra mile to help me whenever I was in need. I would also like to thank my family for keeping faith in me.

References

Bihari, A., Gupta, C. and Tripathi, S. (2019). Protein Classification using Machine Learning and Statistical Techniques: A Comparative Analysis, *Recent Advances in Computer Science and Communications,* pp. 1-19

Carter, B., Bileschi, M., Smith, J., Sanderson, T., Bryant, D., Belanger, D. and Colwell, L. (2019). Critiquing Protein Family Classification Models Using Sufficient Input Subsets, *Journal of Computational Biology,* pp. 1-11, DOI: 10.1101/674119

Cheng, B., Carbonell, J. and Seetharaman, J. (2005). Protein Classification Based on Text Document Classification Techniques, *PROTEINS: Structure, Function, and Bioinformatics*, pp. 955-970, DOI: 10.1002/prot.20373

G, Mirceva., I, Ivanoska., A, Naumoski. and A, Kulakov. (2019). Feature Selection for Improved Classification of Protein Structures, *42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* pp. 1013-1018, DOI:10.23919/MIPRO.2019.8757005

Georgara, D., Kermanidis, K. and Mariolis, I. (2012). Support Vector Machine Classification of Protein Sequences to Functional Families Based on Motif Selection, *IFIP International Federation for Information Processing,* pp. 28-36, DOI: 10.1007/978-3-642-33409-2_4

Hakala, K., Kaewphan, S., Bjorne, J, Mehryary, F., Moen, H., Tolvanen M., Salakoski, T. and Ginter,

F. (2019). Neural network and random forest models in protein function prediction, *Creative commons*, pp. 1-15, **doi:** https://doi.org/10.1101/690271

Hochreiter, S. and Obermayer, K. (2005). Sequence Classification For Protein Analysis, pp. 1-2

Iqbal, M., Faye, I., Belhaouari, S. and Said, A. (2012). Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics, *The Scientific World Journal*, pp. 1-12, doi: 10.1155/2014/173869

Jalal, S., Zhong, J. and Kumar, S. (2019). Protein Secondary Structure Prediction using Multi input Convolutional Neural Network, *SoutheastCon*, pp. 1-5, **doi:** 10.1109/SoutheastCon42311.2019.9020333

Kabli, F., Hamou, R. and Amine, A. (2017). New Classification System for Protein Sequences , *First International Conference on Embedded & Distributed Systems (EDiS)*, pp. 1-6, doi**:** 10.1109/EDIS.2017.8284029

Lee, E. and Wong, A. (2012). Identifying Protein Binding Functionality of Protein Family Sequences by Aligned Pattern Clusters, *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 276-281, doi: 10.1109/BIBM.2012.6392682

Lee, T. and Nguyen, T. (2019). Protein Family Classification with Neural Network, *Stanford University*, pp. 1-9

Li, J., Wu, J. and Chen, K. (2013). PFP-RFSM: Protein fold prediction by using random forests and sequence motifs, *Biomedical Science and Engineering*, pp. 1161 – 1170, doi: 10.4236/jbise.2013.612145

Li, M., Ling, C, and Gao, J. (2017). An Efficient CNN-based Classification on G-protein Coupled Receptors Using TF-IDF and N-gram, *IEEE Symposium on Computers and Communications (ISCC*), pp. 1-8, doi: 10.1109/ISCC.2017.8024644

Majhi, V., Paul, S. and Jain, R. (2019). Bioinformatics for Healthcare Application, *Amity International Conference on Artificial Intelligence (AICAI)*, pp. 204 – 207, DOI:10.1109/AICAI.2019.8701277

Naveenkumar K S, Mohammed Harun Babu R, Vinayakumar R and Soman KP (2019). Protein Family Classification using Deep Learning, *Center for Computational Engineering and Networking (CEN)*, pp. 1-9, DOI:10.1101/414128

Parikh, Y. and Abdelfattah, N. (2019). Machine Learning Models to Predict Multiclass Protein Classifications, *IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference*, pp. 0300-0304, doi: 10.1109/UEMCON47517.2019.8993049

Rentzsch, R. and Orengo, C. (2011). Protein function prediction using domain families, *BMC Bioinformatics*, pp. 1-14, doi: 10.1186/1471-2105-14-S3-S5

Satpute, B. and Yadav, R. (2018). Machine Intelligence Techniques for Protein Classification, *3rd International Conference for Convergence in Technology (I2CT)*, pp. 1-4, DOI:10.1109/I2CT.2018.8529495

Shinde, A. and D'Silva, M. (2019). Protein Sequence Classification using Natural Language Processing, *International Journal of Engineering Development and Research*, pp. 169-175

Sikandar, A., Waqas, A., Usama, B., Wang, X., Sikandar, M., Yao, L., Jiang, Z. and Chunkai, Z. (2017). Decision Tree Based Approaches for Detecting Protein Complex in Protein-Protein Interaction Network

(PPI) via Link and Sequence Analysis, *IEEE Access (Volume: 6)*, pp. 22018-22120, doi: 10.1109/ACCESS.2018.2807811

Suvarna, Vani K. and T, D, Sravani. (2014). Multiclass unbalanced protein data classification using sequence features, *IEEE Conference on Computational Intelligence in Bioinformatics and Computational*, pp. 1-8, doi: 10.1109/CIBCB.2014.6845517

Vazhayil, A., Vinayakumar, R. and Soman, KP. (2019). DeepProteomics: Protein family classification using Shallow and Deep Networks, *Center for Computational Engineering and Networking (CEN)*, pp. 1-17, DOI:10.1101/414631

Yang, Y., Liang, Lu, B. and Yun, Yang, W. (2007). CLASSIFICATION OF PROTEIN SEQUENCES BASED ON WORD SEGMENTATION METHODS, *Proceedings of the 6th Asia-Pacific Bioinformatics Conference*, pp. 1-10, doi: 10.1142/9781848161092_0020