IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF AMYLOID PROTEINS USING GRADIENT BOOSTING MODEL

[1] Mr. M. Shiva Kumar, [2] N. Sai Aditya Vardhan, [3] B. Varshith, [4] A. Sindhuja

[1] Assistant Professor, [2, 3, 4] UG Student

[1,2,3,4] Department of Computer Science and Engineering,

[1,2,3,4] CMR College of Engineering & Technology, Hyderabad, Telangana, India

Abstract: Numerous human disorders, such as Alzheimer's, Parkinson's, and type 2 diabetes, are linked to amyloid proteins. Through the regulation of Amyloid protein deposition, some foods and vegetables can assist in the prevention of certain disorders in humans. Onions, kale, tomatoes, cabbage, and romaine lettuce are examples of vegetables. Additional food items such as walnuts, coffee, berries, fatty fish, turmeric, champagne, and cinnamon also assist with the administration of illnesses brought on by the buildup of Amyloids. This experiment uses a dataset that includes the aforementioned dietary items in an attempt to predict amyloid proteins. Neural Net Analysis and Gradient Boosting Classifier are utilized in this data analysis. Under the current system, people with amyloidosis see doctors after developing the illness. Preventing is always better than curing. By forecasting the Amyloids based on dietary habits, the suggested system helps avoid the disease. By enabling people to modify their eating habits, it prevents the occurrence of amyloidosis. The project's goal is to develop a tool for Prediction of Amyloid Proteins using Gradient Boosting Model.

***Keywords* -** Amyloid Protein, Neural Net Analysis (NNA), Gradient Boosting Classifier (GBC), Prediction of Amyloid Protein, Dietary items, Alzheimer, Parkinson, Diabetes.

I. INTRODUCTION

The insoluble, inherently faulty proteins known as amyloid proteins (AMYs) are resistant to the actions of proteases. Additionally, they build up to create intracellular protein illusions or extracellular plaques in variety of tissues and organs, especially in diseased situations. The creation of biofilms, antimicrobial activity, binding and storage of peptide hormones, and contribute to the body's natural antiviral defenses are all significantly influenced by AMY proteins. The AMY proteins are not only valuable and the extracellular deposition of amyloid fibrils may result in several diseases. AMY proteins are nearly related with the cellular death process that leads to various complex diseases; similar to domestic Mediterranean fever, Alzheimer's, type II diabetes, Huntington's, and Parkinson's diseases. It's thus inferred that AMY proteins may be used as a novel therapeutic agent (breakthrough) for the diseases mentioned above.

II. RELATED WORK

These Modern initiatives often rely on pre-existing solutions as foundational elements to achieve efficiency and spark creativity. This method fosters a supportive atmosphere that fosters idea develop and meet new problems, in addition to acknowledging the knowledge and accomplishments of those of those who paved the way. We fully embrace this concept in our project, carefully incorporating components from pre-existing solutions to enhance our effort. These already-existing solutions act as beacons of light, providing frameworks and insights that influence the course of our endeavor.

A. Transcriptional regulation analysis of Alzheimer's disease based on FastNCA algorithm [1]:

Researchers investigated how genetic variations might influence gene activity in Alzheimer's disease (AD) using a technique called Fast Network Component Analysis (FastNCA). Their goal was to understand how these differences impact the underlying networks that control gene expression.

Data collection: They gathered two types of data from AD patients: gene expression information (microarray and RNA-seq data) and data on genetic variations (Single Nucleotide Polymorphisms or SNPs).

FastNCA analysis: They used FastNCA to analyze how SNPs affect the activity and influence of proteins called transcription factors (TFs) that control other genes (target genes).

Multi-data analysis: They combined the results from FastNCA with other data to determine which genes have changed expression levels and create networks displaying how these genes are regulated. They also compared the genes regulated by the same TFs across different datasets.

Findings: While the specific TFs controlling genes differed between datasets, highlighting the complexity of the control of genes in AD, these TFs often played a part in similar biological processes, such as immune response and cell communication. Even when involved in different processes, these pathways were still connected to AD.

In conclusion, the study used FastNCA to identify crucial TFs and pathways potentially participating in the evolution of AD, emphasizing the complex link between genetic variations and gene regulation in the disease. This method provides insightful information into the intricate mechanisms underlying AD.

B. Prediction of Amyloid Proteins Using Embedded Evolutionary & Ensemble Feature Selection Based Descriptors With eXtreme Gradient Boosting Model [2]:

This research explores a new approach for forecasting proteins that might create formations called amyloids, which are linked to various diseases. The technique involves several steps:

Extracting Key Features: Scientists examined the sequence of amino acids of proteins, searching for patterns typically found in amyloid proteins.

Selecting the Most Important Information: A machine learning technique called "Ensemble Feature Selection" helped identify the most critical features for accurate prediction.

Powerful Model for Prediction: The researchers used a machine learning model called "eXtreme Gradient Boosting" to forecast whether a protein is probably going to form amyloids.

To assess their method, they tested it on known amyloid and non-amyloid proteins. The outcomes were impressive, achieving 93.10% accuracy on the training data set and 89.67% on independent data, significantly exceeding existing methods. This promising new approach has the potential to aid in developing cures for illnesses linked to amyloid proteins.

C. Unveiling an Amyloid-Forming Segment within the Bap Protein of Staphylococcus epidermidis [3]:

Researchers investigated a specific protein from Staphylococcus epidermidis, a bacterium known for biofilm formation, to determine if it contained regions that could form amyloid fibers. These fibers are associated with various diseases and play a role in bacterial processes.

To achieve this, they employed a multi-step approach:

a. Software selection: They utilized three computer programs, AGGRESCAN, PASTA, and TANGO, to analyze the protein sequence and pinpoint areas with a high likelihood of forming amyloids.

b. Peptide synthesis: Based on the software predictions, they synthesized a specific seven-amino acid segment derived from the identified region of the protein.

c. Amyloidogenicity assays: To assess the peptide's ability to form amyloid fibers, they employed various techniques:

Thioflavin T fluorescence: This method measured an increase in fluorescence, indicating amyloid fiber formation.

Infrared spectroscopy: This technique identified the presence of beta-sheet structures, a characteristic of amyloid fibers.

Atomic force microscopy: This technique visualized the self-assembled structures at a nanoscale level, confirming the presence of amyloid fibers.

The combined analysis of software predictions and experimental techniques revealed that the synthesized peptide could indeed form amyloid fibers. This suggests that this specific region within the Bap protein might be involved in protein-protein interactions and contribute to biofilm formation in Staphylococcus epidermidis.

In essence, the study successfully identified a potential amyloidogenic domain within the protein using computational tools and validated its functionality through experimental methods.

D. Predicting diabetes mellitus with machine learning methods [4]:

Researchers investigated the possibility of machine learning to forecast diabetes mellitus. They explored various techniques and tools:

a. Gathering Data: They used a dataset with information on healthy individuals and diabetic patients, including factors potentially linked to diabetes.

b. Focusing on Key Factors: To simplify the data and determine what indicators were most important for prediction, they employed two methods:

Principal Component Analysis (PCA): This technique condenses the data by capturing the most important variations within the features.

Minimum Redundancy Maximum Relevance (mRMR): This approach selects features highly relevant to predicting diabetes while minimizing redundancy between them.

c. Testing Different Algorithms: They compared the performance of three machine learning techniques for forecasting diabetes:

Decision Tree: This algorithm creates a tree-like structure to classify data points according to series of decision rules.

Random Forest: This method combines multiple decision trees, leading to improved accuracy and robustness.

Neural Network: This algorithm mimics the human brain's structure and function, recognizing intricate patterns in data for prediction.

d. Evaluating Effectiveness: Assessing the effectiveness of each algorithm, they used a method known as "five-fold cross-validation." This method splits the data into five subsets, trains the model on four folds, and tests it on the remaining fold, iterating this process for all five folds to obtain a reliable estimation of the model's generalizability.

Their study revealed that the Random Forest algorithm achieved the maximum precision (80.84%) in predicting diabetes when using all available features. This implies that integrating several decision trees can effectively capture complex relationships within the data for better prediction compared to individual decision trees or neural networks in this specific case.

In conclusion, the study highlights the capability of machine learning, particularly Random Forest, for predicting diabetes using relevant patient data.

III. Methods And Experimental Details

A. Dataset

The first crucial step in developing an automated predictor is choosing a suitable training dataset. This data serves two key purposes: it allows us to train machine learning models and provides a benchmark for evaluating their accuracy. Our system utilizes the Amyloidosis Dataset as input, accessible through user interfaces. The system then calculates several key outputs, including Gradient Boosting Accuracy & Predictions, Neural Net Accuracy & Predictions. Similar to previous methods, the training patterns are divided into training data set and testing data set. The training data set is utilized to train the machine learning algorithms (machine learning and deep learning), while the testing set is utilized to estimate the model's performance. Notably, the training data set and testing data set are distinct to avoid bias. Our specific dataset incorporates information about a person's dietary intake, suggesting a potential connection between diet and the evolution of amyloidosis.

B. System Design

1. Amyloidosis Dataset:

We meticulously compiled a dataset that captured a person's daily and weekly dietary intake, with a particular emphasis on twelve food items: onions, kale, romaine lettuce, cabbage, tomatoes, walnuts, coffee, berries, fatty fish, turmeric, champagne, and cinnamon. The selection of these specific food items stemmed from their potential role in combating amyloidosis, a condition marked by the abnormal accumulation of proteins in the body. To delve deeper into the potential connections, we harnessed the power of machine learning. Two distinct models, Gradient Boosting and Neural Networks, were employed to analyze the intricate dietary data. By meticulously training these models, we aimed to identify patterns in food choices that might signal a person's susceptibility to amyloidosis. Following the training phase, the system rigorously evaluated the accuracy of each model. Armed with this knowledge, the models were then used to generate predictions about the potential link between specific dietary choices and a person's aversion to amyloidosis. This innovative approach paves the way for exploring the possibility of using personalized dietary data to assess an individual's risk of amyloidosis, potentially resulting to the development of more targeted dietary interventions in the future.

2. Training and Test Data:

Training Data: For machine learning models, the training data's quality is crucial. In this project, the training data resembled a well-stocked recipe database. Each entry detailed a food item's nutritional content, composition, and other relevant factors alongside a label indicating the presence of amyloid proteins. A diverse dataset, encompassing a wide variety of food items, was essential to prevent biased predictions and ensure the models could generalize their knowledge to unseen food combinations. This richness allowed the models to grasp the underlying relationships between food characteristics and amyloid proteins, transforming them from memorizers to sophisticated food analysts.

Testing Data: Within the machine learning workflow, the testing data serves as a critical benchmark for evaluating model generalizability, a cornerstone for real-world applicability. This independent dataset mirrors the structure of the training data set, providing details regarding a range of food items such as nutritional content, composition, and other relevant features. However, the key distinction lies without labels indicating whether amyloid proteins is present or not. This deliberate omission fosters an unbiased evaluation scenario. By presenting the models with unseen data, we can objectively assess their capacity for transfer the knowledge gleaned from the training phase. In essence, the testing stage poses a crucial question: can the models leverage their learned patterns to accurately predict the amyloid protein content in entirely novel food combinations? A high degree of success in this evaluation signifies that the models have transcended rote memorization and achieved a genuine understanding of the intricate relationships between food characteristics and amyloid proteins. This capacity for generalizability is paramount for practical applications, where the models are likely to encounter food combinations beyond the scope of the training data.

3. Amyloid Proteins Prediction Tool:

This user-friendly tool is built using PyUIC, a Python library specifically designed to convert user interface (UI) descriptions written in Qt Designer into Python code.

The tool itself functions as an entry point, presenting users with various options upon launch. These options revolve around the two machine learning models employed by the system: Gradient Boosting and Neural Networks. Users can access the accurateness of each model, giving valuable perceptivity into their overall effectiveness in predicting the existence of amyloid proteins in food items. Additionally, they can view the specific predictions made by each model, allowing for a deeper understanding of how the models interpret the provided dietary data and arrive at their conclusions.

4. Gradient Boosting Accuracy:

The system puts the trained model to the test, using it to predict amyloid protein levels in unseen data (testing data). This assesses the model's ability to generalize its knowledge. Accuracy is then computed utilizing the accuracy_score function, reflecting the model's success in making correct forecasts regarding the new data.

5. Neural Net Accuracy:

To assess the model's effectiveness, the system employs accuracy as the main performance metric. During training, the model iterates through the training data set 100 times (epochs) while adjusting its internal parameters to minimize loss and improve accuracy. Finally, the trained model is examined on data to analyze its capacity for generalization and precise forecasting.

6. Prediction on GBC and NN:

Takes the input from the user and gives the output whether the user is in Low, Needed, or More Amyloids.

C. Architecture:

Input Data: The project utilizes a dataset of food items, potentially related to amyloidosis development.

Output: The project delivers four main results:
a. GBC Model Establishment and Accuracy: This module focuses on building and assessing the accuracy of the GBC model in predicting amyloidosis-linked foods.
b. GBC Model Predictions: This module analyzes the specific predictions generated by the trained GBC model.
c. Neural Network Development and Accuracy: This module involves constructing and evaluating the accuracy of a neural network model for the same prediction task.
d. Neural Network Predictions: This module analyzes the predictions generated by the trained neural network model.

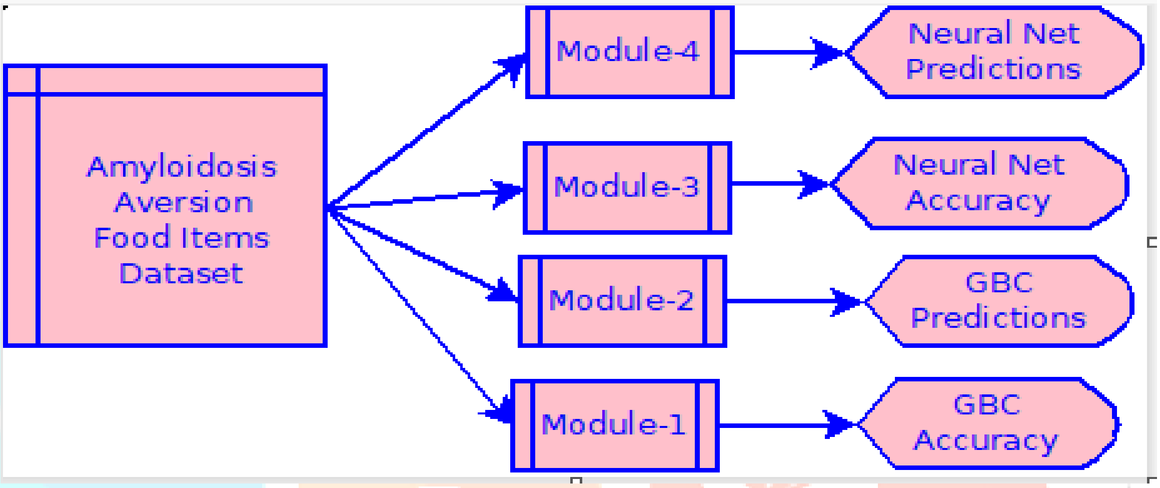


Fig : 1 Architecture

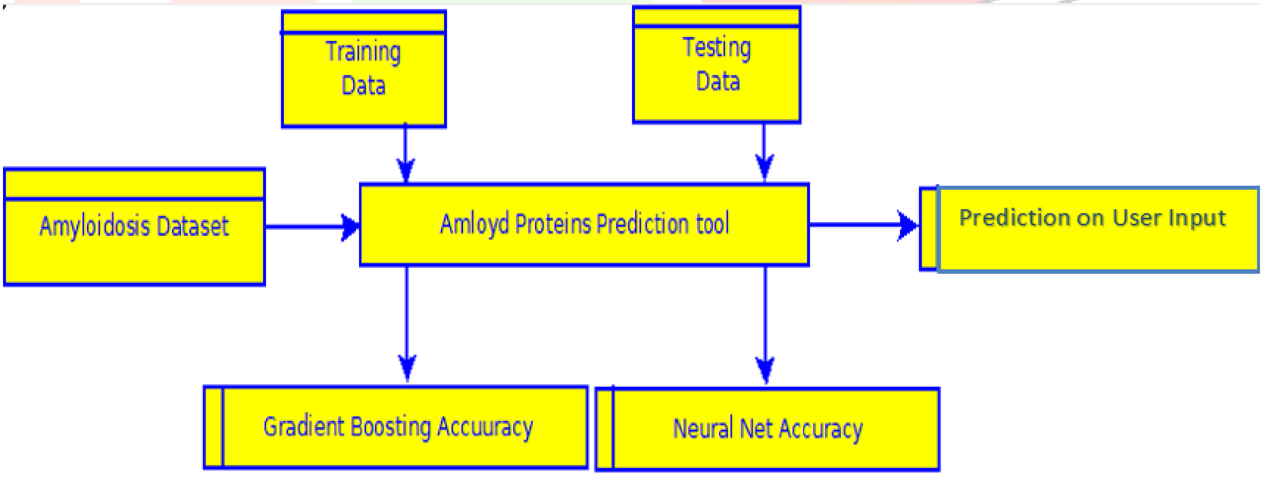


Fig : 2 Data Flow

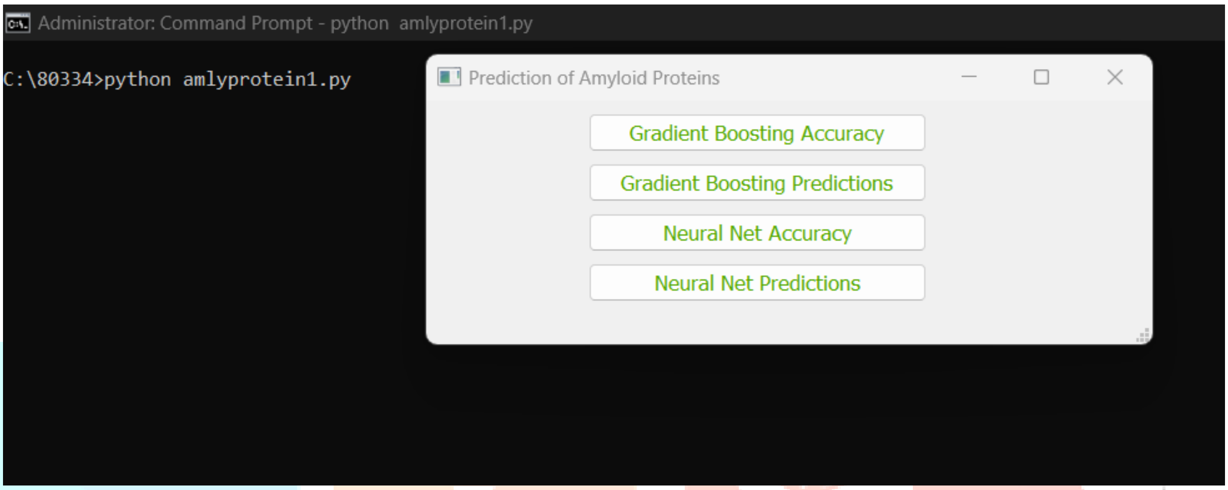
IV. RESULTS AND DISCUSSION

This section showcases our project's functionality through a series of steps and accompanying visuals.

Executing the Application:
1. Launch the Command Prompt as Administrative.
2. Navigate to the Project Directory: This step instructs the program on location of the necessary files.
3. Run the File amlyproteinl1.py: Executing this file initiates the application.

Upon program execution, the "Amyloid Proteins Prediction Tool" window launches as shown in Fig: 3. This interface comprises four distinct modules:

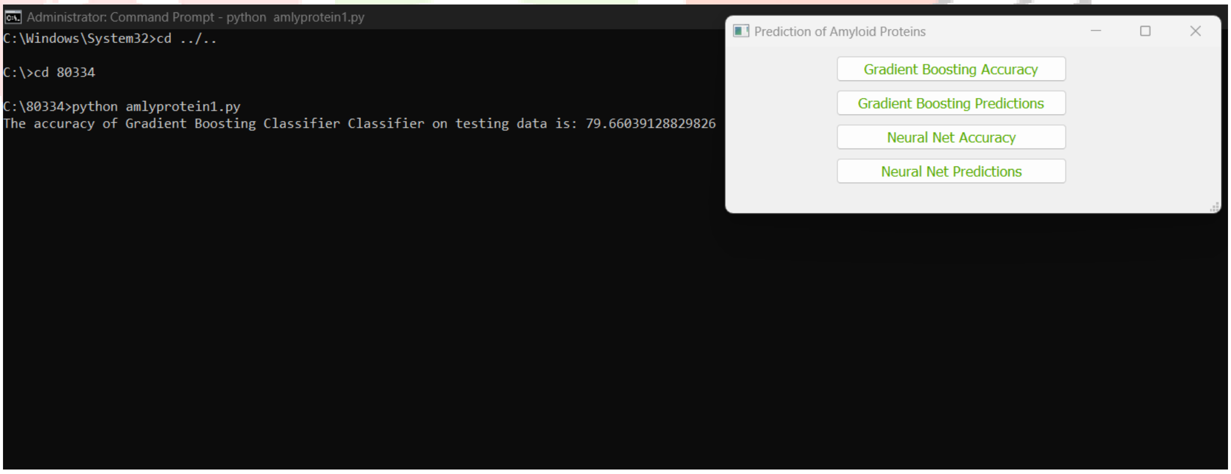
- Gradient Boosting Accuracy: This section displays the model's overall accuracy in predicting amyloid protein level as shown in Fig: 4.
- Gradient Boosting Prediction: This module showcases the model's predicted amyloid protein levels of user as shown in Fig: 5.
- Neural Network Accuracy: This section presents the accuracy of the neural network model in predicting amyloid protein levels as shown in Fig: 6.
- Neural Network Prediction: This module displays the neural network model's predicted amyloid protein levels of user as shown in Fig: 7.



```
C:\80334>python amlyprotein1.py
```

Administrator: Command Prompt - python amlyprotein1.py
Prediction of Amyloid Proteins
Gradient Boosting Accuracy
Gradient Boosting Predictions
Neural Net Accuracy
Neural Net Predictions

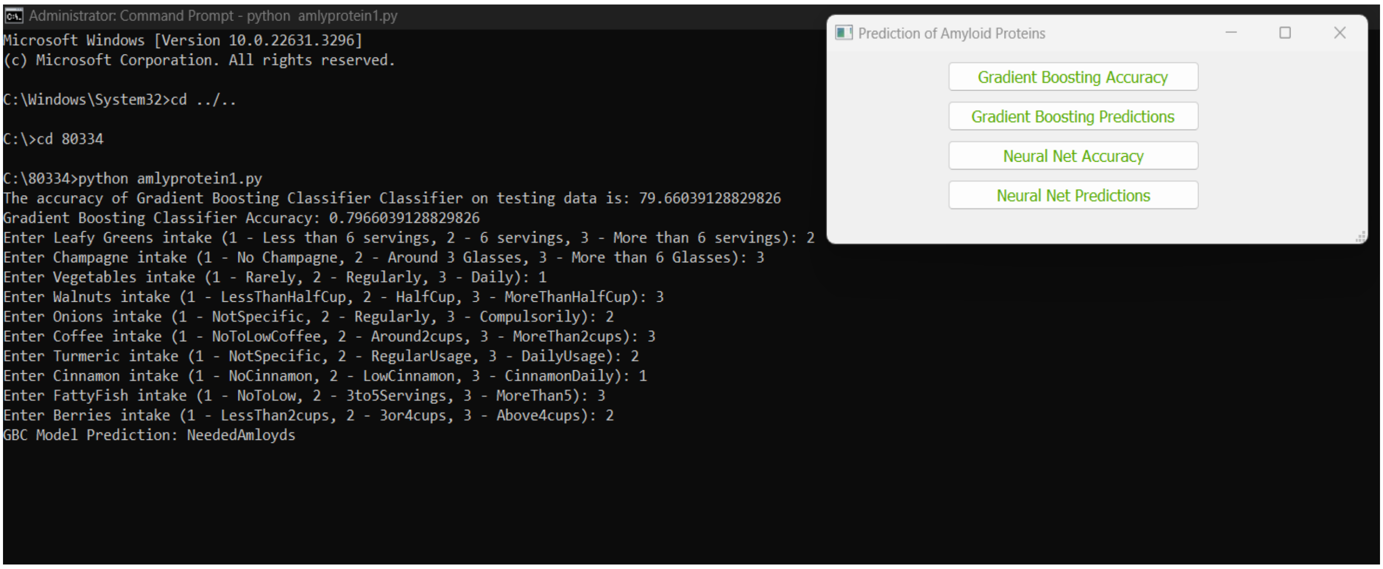
Fig: 3 Amyloid Protein Prediction Tool



```
C:\Windows\System32>cd ../..

C:\>cd 80334

C:\80334>python amlyprotein1.py
The accuracy of Gradient Boosting Classifier Classifier on testing data is: 79.66039128829826
```

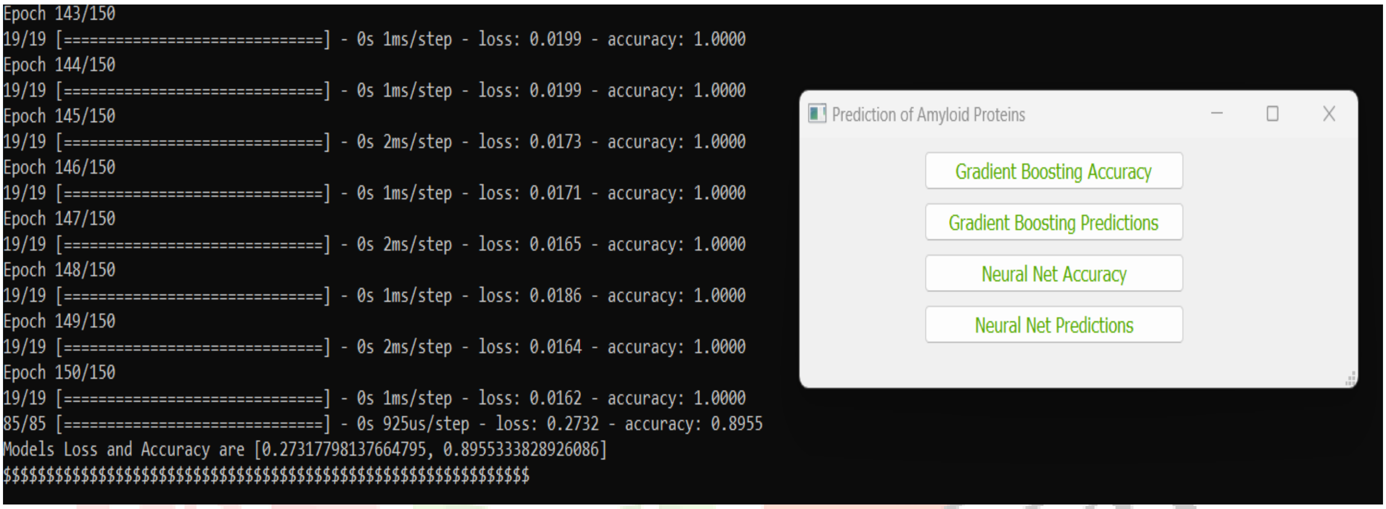
Administrator: Command Prompt - python amlyprotein1.py
Prediction of Amyloid Proteins
Gradient Boosting Accuracy
Gradient Boosting Predictions
Neural Net Accuracy
Neural Net Predictions

Fig: 4 GBC Accuracy


Administrator: Command Prompt - python amlyprotein1.py
Prediction of Amyloid Proteins
Gradient Boosting Accuracy
Gradient Boosting Predictions
Neural Net Accuracy
Neural Net Predictions

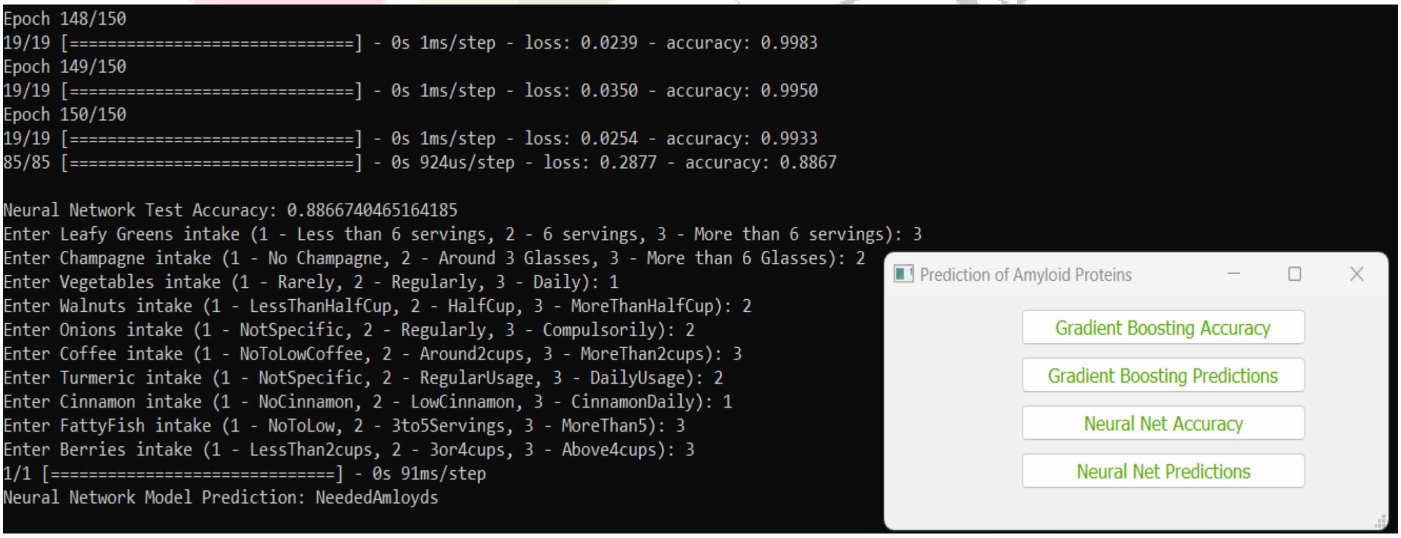
```
Microsoft Windows [Version 10.0.22631.3296]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>cd ../..

C:\>cd 80334

C:\80334>python amlyprotein1.py
The accuracy of Gradient Boosting Classifier Classifier on testing data is: 79.66039128829826
Gradient Boosting Classifier Accuracy: 0.7966039128829826
Enter Leafy Greens intake (1 - Less than 6 servings, 2 - 6 servings, 3 - More than 6 servings): 2
Enter Champagne intake (1 - No Champagne, 2 - Around 3 Glasses, 3 - More than 6 Glasses): 3
Enter Vegetables intake (1 - Rarely, 2 - Regularly, 3 - Daily): 1
Enter Walnuts intake (1 - LessThanHalfCup, 2 - HalfCup, 3 - MoreThanHalfCup): 3
Enter Onions intake (1 - NotSpecific, 2 - Regularly, 3 - Compulsorily): 2
Enter Coffee intake (1 - NoToLowCoffee, 2 - Around2cups, 3 - MoreThan2cups): 3
Enter Turmeric intake (1 - NotSpecific, 2 - RegularUsage, 3 - DailyUsage): 2
Enter Cinnamon intake (1 - NoCinnamon, 2 - LowCinnamon, 3 - CinnamonDaily): 1
Enter FattyFish intake (1 - NoToLow, 2 - 3to5Servings, 3 - MoreThan5): 3
Enter Berries intake (1 - LessThan2cups, 2 - 3or4cups, 3 - Above4cups): 2
GBC Model Prediction: NeededAmloyds
```

Fig: 5 GBC Prediction


Prediction of Amyloid Proteins
Gradient Boosting Accuracy
Gradient Boosting Predictions
Neural Net Accuracy
Neural Net Predictions

```
Epoch 143/150
19/19 [==============================] - 0s 1ms/step - loss: 0.0199 - accuracy: 1.0000
Epoch 144/150
19/19 [==============================] - 0s 1ms/step - loss: 0.0199 - accuracy: 1.0000
Epoch 145/150
19/19 [==============================] - 0s 2ms/step - loss: 0.0173 - accuracy: 1.0000
Epoch 146/150
19/19 [==============================] - 0s 1ms/step - loss: 0.0171 - accuracy: 1.0000
Epoch 147/150
19/19 [==============================] - 0s 2ms/step - loss: 0.0165 - accuracy: 1.0000
Epoch 148/150
19/19 [==============================] - 0s 1ms/step - loss: 0.0186 - accuracy: 1.0000
Epoch 149/150
19/19 [==============================] - 0s 2ms/step - loss: 0.0164 - accuracy: 1.0000
Epoch 150/150
19/19 [==============================] - 0s 1ms/step - loss: 0.0162 - accuracy: 1.0000
85/85 [==============================] - 0s 925us/step - loss: 0.2732 - accuracy: 0.8955
Models Loss and Accuracy are [0.27317798137664795, 0.8955333828926086]
$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```

Fig: 6 Neural Net Accuracy


Prediction of Amyloid Proteins
Gradient Boosting Accuracy
Gradient Boosting Predictions
Neural Net Accuracy
Neural Net Predictions

```
Epoch 148/150
19/19 [==============================] - 0s 1ms/step - loss: 0.0239 - accuracy: 0.9983
Epoch 149/150
19/19 [==============================] - 0s 1ms/step - loss: 0.0350 - accuracy: 0.9950
Epoch 150/150
19/19 [==============================] - 0s 1ms/step - loss: 0.0254 - accuracy: 0.9933
85/85 [==============================] - 0s 924us/step - loss: 0.2877 - accuracy: 0.8867

Neural Network Test Accuracy: 0.8866740465164185
Enter Leafy Greens intake (1 - Less than 6 servings, 2 - 6 servings, 3 - More than 6 servings): 3
Enter Champagne intake (1 - No Champagne, 2 - Around 3 Glasses, 3 - More than 6 Glasses): 2
Enter Vegetables intake (1 - Rarely, 2 - Regularly, 3 - Daily): 1
Enter Walnuts intake (1 - LessThanHalfCup, 2 - HalfCup, 3 - MoreThanHalfCup): 2
Enter Onions intake (1 - NotSpecific, 2 - Regularly, 3 - Compulsorily): 2
Enter Coffee intake (1 - NoToLowCoffee, 2 - Around2cups, 3 - MoreThan2cups): 3
Enter Turmeric intake (1 - NotSpecific, 2 - RegularUsage, 3 - DailyUsage): 2
Enter Cinnamon intake (1 - NoCinnamon, 2 - LowCinnamon, 3 - CinnamonDaily): 1
Enter FattyFish intake (1 - NoToLow, 2 - 3to5Servings, 3 - MoreThan5): 3
Enter Berries intake (1 - LessThan2cups, 2 - 3or4cups, 3 - Above4cups): 3
1/1 [==============================] - 0s 91ms/step
Neural Network Model Prediction: NeededAmloyds
```

Fig: 7 Neural Net Prediction

V. Conclusion

Our study successfully developed and evaluated two machine learning models: Gradient Boosting Classifier (GBC) and a Neural Network. These models aim to predict the quantity of amyloid protein in various foods, including common ones like onions, kale, and berries. Both models performed well on entirely new data, suggesting they might possess the ability to identify foods linked to various levels of amyloid proteins.

The GBC model attained a precision of 79.6%, while the Neural Network reached 89%. Notably, both models could predict the amyloid protein levels in unseen food items.

Overall, this project contributes to the field of using ML (machine learning) to analyze dietary patterns. It makes room for more research into the connection between specific foods and the accumulation of amyloid proteins. This research possesses the potential to educate dietary strategies for individuals worried about health problems associated with amyloid buildup.

References

[1] Q.Sun, W. Kong, X. Mou, and S. Wang, ‘‘Transcriptional regulation analysis of Alzheimer’s disease based on FastNCA algorithm,’’ Current Bioinf., vol. (14), no. 8, pp. 771–782, Dec. 2019.

[2] Ishahid Akbar, Hashim Ali, Ashfaq Ahmad, Mahidur R. Sarker,Aamir Saeed, Ely Salwana, Sarah Gul, Ahmad Khan1 And Farman AL: Prediction of Amyloid Proteins Using Embedded Evolutionary & Ensemble Feature Selection Based Descriptors With eXtreme Gradient Boosting Model, IEEE Access

[3] P. Lembré, C. Vendrely, and P. Martino, ‘‘Identification of an amyloidogenic peptide from the Bap protein of Staphylococcus epidermidis,’’ Protein Peptide Lett., vol. 21, no. (1), pp. 75–79

[4] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, ‘‘Predicting diabetes mellitus (dm) with ML(machine learning) methods,’’ Frontiers Genet., vol. 9,p. 515..

[5] https://www.python.org/

[6] https://github.com/baoboa/pyqt5/blob/master/pyuic/uic/pyuic.py

[7] https://www.numpy.org/

[8] https://riverbankcomputing.com/software/pyqt/intro