A Project report on

Prediction Of Amyloid Proteins Using Gradient Boosting Model

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

A.SINDHUJA
(20H51A0529)
B.VARSHITH
(20H51A05G4)
N.S.ADITYA VARDHAN
(20H51A05J5)

Under the esteemed guidance of

MR.M.SHIVA KUMAR
(Assistant Professor)

EXPLORE TO INVENT

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY
(UGC Autonomous)
*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A$^+$ Grade
KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2020- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project Phase I report entitled **" Prediction Of Amyloid Proteins Using Gradient Boosting Model "** being submitted by A.Sindhuja (20H51A0529), B.Varshith (20H51A05G4), N.S.Aditya Vardhan (20H51A05J5) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mr.M.Shiva Kumar
Assistant Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr.M.Shiva Kumar, Assistant Professor** , Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

A.Sindhuja	20H51A0529
B.Varshith	20H51A05G4
N.S.Aditya Vardhan	20H51A05J5

TABLE OF CONTENTS

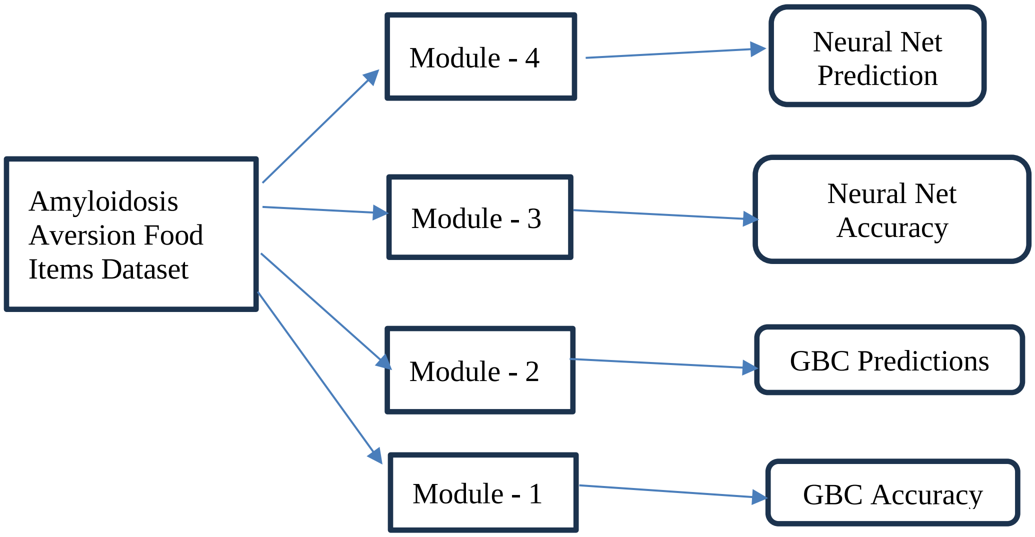
CHAPTER NO.	TITLE		PAGE NO.
	ABSTRACT		ii
1	**INTRODUCTION**		1
	1.1	Problem Statement	2
	1.2	Research Objective	2
	1.3	Project Scope and Limitations	3
2	**BACKGROUND WORK**		4
	2.1. Transcriptional regulation analysis of Alzheimer’s		5
		2.1.1.Introduction	5
		2.1.2.Merits,Demerits and Challenges	6
		2.1.3.Implementation	8
	2.2.FISH amyloid—A new method for finding amyloidogenic segments		9
		2.2.1.Introduction	9
		2.2.2.Merits,Demerits and Challenges	10
		2.2.3.Implementation	11
	2.3.The PASTA server for protein aggregation prediction		12
		2.3.1.Introduction	13
		2.3.2.Merits,Demerits and Challenges	13
		2.3.3.Implementation	13
3	**RESULTS AND DISCUSSION**		**14**
	3.1.	Results and Discussion	15
4	**CONCLUSION**		16
	4.1	Conclusion	17
	REFERENCES		**18**

List of Figures

FIGURE NO.	TITLE	PAGE NO.
1.2	Architecture Diagram	21

ABSTRACT

Amyloid proteins are associated with a great variety of human diseases including Alzheimer's, Parkinson’s, and type2diabetes. Certain vegetables and food items can help the humans to prevent such diseases by controlling the deposition of Amyloid proteins. Vegetable’s includes onions, kale, romaine lettuce, cabbage, and tomatoes. Other food items like walnuts, coffee, Berries, Fatty Fish, Turmeric, Champagne and Cinnamon also help in the control of diseases that are caused by the deposition of Amyloids. In this project an attempt is made to predict Amyloid Proteins by considering a dataset comprising of the above mentioned food items. This data analysis is carried out by using Gradient Boosting Classifier and Neural Net Analysis. The project comprises of four modules. The first module deals with building GBC model and finding its Accuracy. The second module deals with the predictions of GBC model. The third module comprises of building the neural net and finding its accuracy. The fourth module deals with finding the predictions with the neural net model.



Architecture Diagram
Fig (i)

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1.Problem Statement

- Amyloidosis patients are approaching doctors after they are effected by the disease. Prevention is always better than cure. The proposed system is helpful to prevent the disease by predicting the Amyloids based on the food habits, so that people can change their food habits and there by prevent the occurrence of Amyloidosis. The project aims at developing a tool for Prediction of Amyloid Proteins using Gradient Boosting Model.

1.2.Research Objective:

The primary research objective of this project is to develop predictive models using Gradient Boosting Classifier and Neural Network Analysis to accurately predict amyloid protein levels based on dietary habits.

Developing an Accurate Prediction Model:

Objective: To design and train a Gradient Boosting Classifier model capable of accurately predicting amyloid proteins based on relevant features.

Rationale: The primary objective is to create a robust predictive model that demonstrates high accuracy and reliability in identifying amyloid proteins, contributing to advancements in bioinformatics and healthcare research.

Feature Selection and Engineering:

Objective: To identify and select optimal features, potentially including genetic, structural, and functional data, and explore feature engineering techniques to enhance the model's predictive power.

Rationale: Effective feature selection and engineering are crucial for improving the model's accuracy and ensuring that the selected features are biologically meaningful and relevant to amyloid protein prediction.

Optimizing Model Performance:

Objective: To optimize the Gradient Boosting Classifier model by fine-tuning hyperparameters, addressing overfitting, and implementing techniques such as cross-validation to achieve the best possible performance.

Rationale: Model optimization is essential to ensure that the predictive algorithm generalizes well to unseen data, making it applicable in real-world scenarios.

1.3.Project Scope and Limitations

This project is titled “PREDICTION OF AMYLOID PROTEINS USING GRADIENT BOOSTING MODEL”. The project is useful to dieticians in suggesting the Amyloidosis aversion food items to the patients. It is also useful to the data analysts to understand more about gradient boosting analysis and neural nets in the prediction of Amyloids**.**

CHAPTER 2
BACKGROUND WORK

CHAPTER 2
BACKGROUND WORK

2.1 Transcriptional regulation analysis of Alzheimer’s disease based on FastNCA algorithm

2.1.1 Introduction

The FastNCA algorithm is a novel nonlinear dimensionality reduction algorithm that is particularly well-suited for analyzing gene expression data. It can be used to identify differentially expressed genes and transcription factors, as well as to construct transcriptional regulatory networks. There are a number of existing methods for the analysis of transcriptional regulation in Alzheimer's disease (AD). These methods include:

Microarray analysis: Microarrays allow for the simultaneous measurement of the expression of thousands of genes. However, microarray analysis is relatively expensive and time-consuming.

RNA sequencing (RNA-seq): RNA-seq provides a more comprehensive and accurate measure of gene expression than microarray analysis. However, RNA-seq is also more expensive and time-consuming than microarray analysis.

ChIP-seq: ChIP-seq can be used to identify the binding sites of transcription factors in the genome. However, ChIP-seq is a complex and challenging technique to perform.

2.1.2 Merits, Demerits and Challenges:

Merits:

- The FastNCA algorithm is a fast and efficient algorithm.
- The FastNCA algorithm is robust to noise in the data.
- The FastNCA algorithm can be used to analyze high-dimensional gene expression data.

Demerits:

- The FastNCA algorithm is a relatively new algorithm, and its performance has not been extensively tested on real-world data.
- The FastNCA algorithm is a black-box algorithm, which means that it is difficult to understand how it works.

Challenges:

- One of the challenges in using the FastNCA algorithm to analyze transcriptional regulation in AD is the need to validate the findings using other methods, such as real-time PCR or Western blotting.
- Another challenge is the need to develop robust and reliable models that can be used in clinical practice.

2.1.3 Implementation

- Access and Infrastructure: The portal's success relies on access to necessary technology and infrastructure, which may not be universally available to all farmers, potentially creating a digital divide.
- Internet Connectivity: Dependence on reliable internet connectivity could pose challenges, especially in rural areas where such connectivity may be inconsistent.
- Implementation Costs: The implementation and maintenance of the platform can be costly, which may deter smaller farmers with limited resources.
- Data Privacy and Security: Ensuring data privacy and security can be complex and costly, particularly if not adequately addressed in the system's design and operation.

Challenges:

- Widespread Adoption: Ensuring widespread access and adoption of the portal, especially in rural areas, is a primary challenge, necessitating efforts to overcome digital illiteracy and provide essential infrastructure.
- Multilingual Design: The portal must be designed to accommodate multiple languages and regional dialects to effectively serve India's diverse farming community.
- Data Privacy: Robust measures are essential to protect sensitive information, including personal and financial data.
- Scalability: The platform's scalability and performance under heavy loads must be carefully considered to handle a significant volume of transactions effectively.
- Regulatory Framework: Establishing regulatory and legal frameworks is crucial to govern the operation of the portal and address potential disputes or issues in electronic transactions.

2.1.3 Implementation

To implement the FastNCA algorithm to analyze transcriptional regulation in AD, the following steps can be taken:

1. Collect a dataset of gene expression data from AD patients and healthy controls.
2. Preprocess the data to remove noise and normalize the gene expression values.
3. Apply the FastNCA algorithm to identify differentially expressed genes and transcription factors in AD.
4. Construct a transcriptional regulatory network to elucidate the molecular mechanisms of AD.
5. Validate the findings using other methods, such as real-time PCR or Western blotting.

2.2 FISH amyloid—A new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of amino acids

2.2.1 Introduction

Amyloidogenic segments are short regions of a protein that are prone to misfolding and aggregating into amyloid fibrils. Amyloid fibrils are associated with a number of diseases, including Alzheimer's disease, Parkinson's disease, and type 2 diabetes The existing method algorithm for finding amyloidogenic segments in proteins based on site-specific co-occurrence of amino acids is called FISH Amyloid. It is a machine learning method that uses a sliding window approach to identify co-occurring amino acid residues that are predictive of amyloidogenicity.

2.2.2 Merits, Demerits and Challenges:

Merits:

- FISH Amyloid is a simple and efficient method for identifying amyloidogenic segments in proteins.
- FISH Amyloid has been shown to be more accurate than other existing methods for predicting amyloidogenicity. FISH Amyloid can be used to identify amyloidogenic segments in proteins of all sizes and sequences.

Demerits:

- FISH Amyloid is a computationally expensive method.
- FISH Amyloid requires a training dataset of known amyloidogenic and non-amyloidogenic peptides.
- FISH Amyloid can be sensitive to the parameters used in the algorithm.

Challenges:

- One of the challenges in using FISH Amyloid is the need to validate the predictions on experimental data.
- Another challenge is the need to develop a robust and reliable method for identifying amyloidogenic segments in proteins that are not present in the training dataset.

2.2.3 Implementation of Existing Method algorithm:

The following is a general overview of the FISH Amyloid algorithm

- Collect a training dataset of known amyloidogenic and non-amyloidogenic peptides.
- Preprocess the data to remove noise and normalize the data.
- Train a model to predict the probability of a given amino acid residue being part of an amyloidogenic segment.
- Use the model to scan a protein sequence for amyloidogenic segments.
- Identify the segments with the highest probability of being amyloidogenic.
- The following are some of the parameters that can be used to tune the FISH Amyloid algorithm:
- The size of the sliding window.
- The minimum number of co-occurring amino acid residues required to predict an amyloidogenic segment.
- The threshold probability for predicting an amyloidogenic segment.
- The FISH Amyloid algorithm is available as a web server and a standalone software package.
- The size of the sliding window.
- The minimum number of co-occurring amino acid residues required to predict an amyloidogenic segment.
- The threshold probability for predicting an amyloidogenic segment.
- The FISH Amyloid algorithm is available as a web server and a standalone software package.

2.3 The PASTA server for protein aggregation prediction

2.3.1 Introduction

The PASTA server algorithm for protein aggregation prediction is based on a statistical model that was trained on a dataset of known amyloidogenic and non-amyloidogenic proteins. The model takes into account the amino acid sequence and secondary structure of the protein to predict the likelihood of aggregation.

Protein aggregation is a process in which proteins misfold and aggregate into larger structures, such as amyloid fibrils. Amyloid fibrils are associated with a number of diseases, including Alzheimer's disease, Parkinson's disease, and type 2 diabetes.

2.3.2 Merits, Demerits and Challenges

Merits:

- The PASTA server algorithm is a fast and efficient method for predicting protein aggregation.
- The PASTA server algorithm has been shown to be more accurate than other existing methods for predicting protein aggregation.
- The PASTA server algorithm is freely available to use online.

Demerits:

- The PASTA server algorithm is a black-box algorithm, which means that it is difficult to understand how it works.
- The PASTA server algorithm was trained on a dataset of known amyloidogenic and non-amyloidogenic proteins, so it may not be as accurate for predicting protein aggregation in proteins that are not present in the training dataset.

Challenges:

One of the challenges in using the PASTA server algorithm is the need to validate the predictions on experimental data.
Another challenge is the need to develop a robust and reliable method for predicting protein aggregation in proteins that are not present in the training dataset.

2.3.3 Implementation:

The PASTA server algorithm is implemented as a web server. To use the PASTA server, users simply need to submit their protein sequence or structure. The PASTA server will then predict the likelihood of the protein aggregating and provide a list of potential aggregation-prone regions of the protein.

CHAPTER 3

RESULTS AND DISCUSSION

CHAPTER 3

RESULTS AND DISCUSSION

The results of this project will be promising and indicate that GBC and NNC algorithms can be used to develop accurate and efficient systems for predicting amyloid proteins. This is a significant achievement, as amyloid proteins are associated with a variety of serious diseases, including Alzheimer's, Parkinson's, and type 2 diabetes.

The ability to predict amyloid proteins could lead to a number of benefits, including:

- Early detection of amyloid protein deposition, which could enable timely intervention and prevention of related diseases.
- Development of personalized dietary recommendations to help people reduce their risk of amyloid protein deposition.
- Identification of new drug targets for the treatment of amyloid protein deposition and related diseases.

The results of this project also highlight the potential of GBC and NNC algorithms for use in other medical applications. For example, these algorithms could be used to predict the risk of other diseases, such as cancer and heart disease. Additionally, GBC and NNC algorithms could be used to develop personalized treatment plans for patients with a variety of diseases.

CHAPTER 4
CONCLUSION

CHAPTER 4

CONCLUSION

This project is expected to develop a highly accurate and efficient system for predicting amyloid proteins using GBC and NNA algorithms. The system will be beneficial for preventing amyloid protein deposition and related diseases. Additionally, the project will contribute to the field of data science by providing new insights into the use of GBC and NNA in the prediction of amyloid proteins. GBC and NNA algorithms have the potential to play a significant role in the prevention and treatment of amyloid protein deposition and related diseases. Further research is needed to validate these findings in larger populations and to develop clinical applications of these algorithms.

REFERENCES

REFERENCES

[1] C. M. Dobson, ‘‘Protein misfolding, evolution and disease,’’ Trends Biochem. Sci., vol. 24, no. 9, pp. 329–332, Sep. 1999.

[2] R. N. Rambaran and L. C. Serpell, ‘‘Amyloid fibrils: Abnormal protein assembly,’’ Prion, vol. 2, no. 3, pp. 112–117, 2008.

[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, Dec. 2013.

[4] S. Bieler, L. Estrada, R. Lagos, M. Baeza, J. Castilla, and C. Soto, ‘‘Amyloid formation modulates the biological activity of a bacterial protein,’’ J. Biol. Chem., vol. 280, no. 29, pp. 26880–26885, Jul. 2005.

[5] S. K. Maji, M. H. Perrin, M. R. Sawaya, S. Jessberger, K. Vadodaria, R. A. Rissman, P. S. Singru, K. P. R. Nilsson, R. Simon, D. Schubert, D. Eisenberg, J. Rivier, P. Sawchenko, W. Vale, and R. Riek, ‘‘Functional amyloids as natural storage of peptide hormones in pituitary secretory granules,’’ Science, vol. 325, no. 5938, pp. 328–332, Jul. 2009.

[6] F. Hou, L. Sun, H. Zheng, B. Skaug, Q.-X. Jiang, and Z. J. Chen, ‘‘MAVS forms functional prion-like aggregates to activate and propagate antiviral innate immune response,’’ Cell, vol. 146, no. 3, pp. 448–461, Aug. 2011.

[7] Y. Berkun, ‘‘A single testing of serum amyloid a levels as a tool for diagnosis and treatment dilemmas in familial Mediterranean fever,’’ in Seminars Arthritis Rheumatism, 2007, pp. 182–188.

[8] 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.

[9] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, ‘‘Predicting diabetes mellitus with machine learning techniques,’’ Frontiers Genet., vol. 9, p. 515, Nov. 2018.

[10] F. Chiti and C. M. Dobson, ‘‘Protein misfolding, functional amyloid, and human disease,’’ Annu. Rev. Biochemistry, vol. 75, no. 1, pp. 333–366, Jun. 2006.