IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Early Stage Detection of Alzheimer Disease with Blood Plasma Proteins using Support Vector Machine

Shital P. Chattar[1], Snehal D. Ile[2], Sanika R. Jinde[3], Tanishka S. Kadam[4], Shreeya D. Kale[5]
Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2,3,4,5]
Pimpri Chinchwad Polytechnic, Pune, Maharashtra, India
ghanwat.shital@gmail.com,snehalile@gmail.com,sanuujinde@gmail.com,
tanishkadam510@gmail.com,siyakale17@gmail.com

Abstract: *This study aims to find an effective method for early detection of Alzheimer's disease (AD) by identifying blood-based non-amyloid bio-markers.[1] Current diagnostic methods focusing on amyloid-based markers have limitations in providing detailed information about the disease and detecting it in the early stages. The researchers used machine learning techniques, specifically support vector machines, to analyze complex data and identified five panels of non-amyloid proteins as potential bio-markers for early AD.[2]This approach using non-amyloid bio-markers demonstrates promise for early Alzheimer's detection compared to existing machine learning models.[3].*

Keywords: Alzheimer's disease, blood bio-marker, dementia, machine learning, support vector machine.

I. INTRODUCTION

The "Early Detection of Alzheimer's Disease with Blood Plasma Proteins using Support Vector Machine" project aims to develop a predictive model utilizing machine learning techniques, specifically the Support Vector Machine (SVM).[1] The goal is to identify potential indicators of Alzheimer's Disease (AD) in blood plasma protein data. AD is a prevalent neuro-degenerative disorder affecting millions worldwide, and early diagnosis is critical for effective medical care and management.

Alzheimer's disease (AD) is a major cause of dementia, presenting significant social and economic challenges.It reports for more than half of all dementia cases, influencing over 50 million individuals worldwide, a number projected to rise to 152 million by 2050. [2] Despite ongoing research, no cure for AD has been discovered. However, there is a concerted effort to develop clinical interventions targeting early stages of the disease, before extensive cell damage occurs.[3]

Studies suggest that AD is characterized by metabolic alterations that may precede amyloid pathology.[4]Early diagnosis is crucial for effective intervention. Current approaches approved established bio markers, like those based on amyloid-beta in cerebral spinal fluid (CSF) and brain amyloid imaging via positron emission tomography (PET).Yet, these methods have limitations. Blood, rich in metabolic information, is an attractive source for these bio-markers due to its non-invasiveness and potential cost-effectiveness compared to CSF and PET imaging.[5]

II. RELATED WORK

The project involves using blood plasma protein data obtained from individuals through blood tests. This data is then processed to identify patterns and features relevant to Alzheimer's disease risk.[2]

A machine learning model is trained on this pre-processed data to associate protein profiles with the risk of Alzheimer's. The model is subsequently applied to new data to identify individuals at risk, providing early detection.[4]

The input for the project involves blood plasma protein data obtained from individuals, typically through blood tests. The output is the identification of individuals at risk for Alzheimer's disease, providing early detection and potentially categorizing the risk level based on the blood plasma protein profile.[5]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

III. PROPOSED METHODOLOGY

3.1 Data Pre-Processing:

In order to reduce bias in machine learning (ML) procedures, the study formalized all data.[1] The standardization process was carried out by applying the formula

$$z = (x - \mu) / \sigma,$$

where μ and σ represent the mean and standard deviation respectively.

3.2 Replication and Evaluation of Existing Methods:

10-fold cross-validation was used to duplicate machine learning models for classifying people into two groups: Healthy Control (HC) and Alzheimer's Disease (ADD), with performance being equated over ten rounds.[2] The data set was divided into ten subgroups for this process, and the classifier underwent iterative training and testing. Robust performance estimation was ensured through repeated cross-validation.

3.3 Novel Panel Identification and Model Development:

Feature Subset Preselection: To find relevant proteins for categorizing AD and Healthy Control (HC) individuals, correlation-grounded point subset selection was utilized (Data-set As a result, the data's dimensionality was decreased.
Novel Panel Formation and SVM-Based Evaluation: Using a brute-force method, protein panels were constructed using the preselected attributes. Additionally, utilizing the Data-set, Support Vector Machine (SVM) classifiers were trained and cross-validated using several kernels. 1.
Kernelized SVM Classification: SVM was selected because to its resilience. The SVM classifiers used kernel functions, such as polynomial kernels, for non-linear data.

3.4 Implementation and Performance Evaluation:

The Correlation-based Feature Subset Selection (CFS) method was used to choose points. The emphasis was on Sensitivity (SN) and Specificity (SP), in accordance with global guidelines for clinically applicable biomarkers of Alzheimer's disease (AD).[3] When developing the model, a performance criterion of 70% was established for both SN and SP, taking into account the unique delicacy marks.

$$\text{Logit}(p)=\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n$$

Here,
p is the probability of the positive class(at risk for Alzheimer's).
β_0 is the intercept.
$\beta_1, \beta_2, \ldots \beta_n$ are the co-efficients for the features $x_1, x_2, \ldots x_n$.

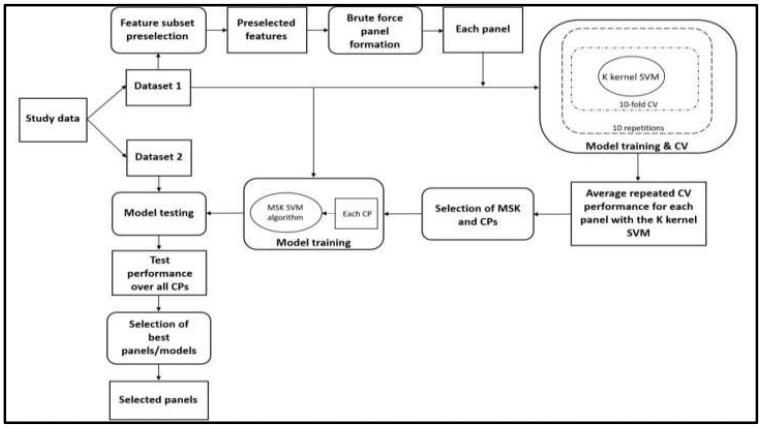

Fig. 1. Overall framework for identification of novel putative bio-marker panels and model development for early AD

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

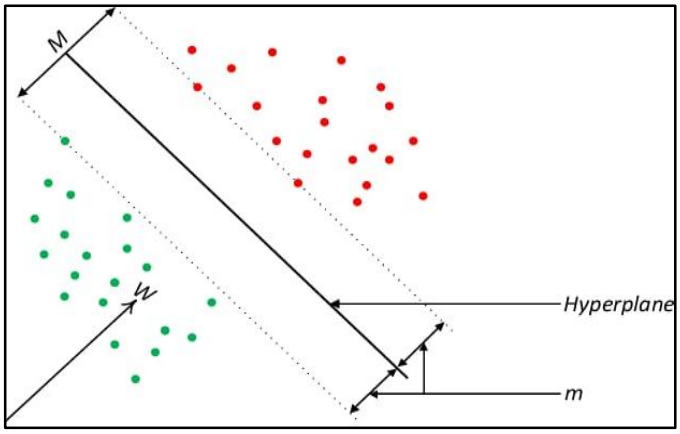

Fig. 2. Mechanism of SVM classification

IV. PROPOSED ALGORITHM

Step 1: First, load and preprocess the data.
Considering that you have a data collection with labels (AD or not) and features (blood plasma rates)
Step 2: Choosing Features
Make use of a feature selection technique (Correlation-based Feature Subset Selection, for example).
Step 3: Divide the data into training and testing sets
Divide the data set in half, using 80% for training and 20% for testing.
Step 4: SVM Training Train_svm_model(training_data, training_labels) is the model.
Step 5: Assess the model's sensitivity, specificity, and accuracy by using the formula
evaluate_model(model, testing_data, testing_labels).
Step 6: Prompt Identification
On fresh data, apply the learned model for early detection.
load_new_data() = new_data
For early detection prediction, assuming you have new data,
predict_with_svm(model, new_data)
Step 7: Show Results display_results(prediction, sensitivity, accuracy, and specificity)
Step 8: Conclude.

V. RESULTS

5.1 Replication and evaluation of existing models:

In replicating 7 existing models for Alzheimer's Disease (ADD) and Healthy Control (HC) classification, one model from a prior work couldn't be repeated due to unavailability of the original dataset.[1] Table II shows the models' average cross-validated performance over ten runs, demonstrating high area under the curve (AUC), specificity (SP), and sensitivity (SN).[2]This implies that there may be a significant rate of false positives, which means the underlying protein panels could not be useful biomarkers for the early diagnosis of illness.

5.2 Feature subset preselection:

From the initial 146 proteins, 16 proteins with a merit of 0.36 were preselected utilizing the Correlation-based Feature Subset Selection (CFS) strategy using the new method.[3] These proteins are shown in Table III together with the z-test-calculated statistical significance.

5.3 Novel panel formation and SVM-based evaluation:

Two to sixteen distinct protein panels were created from the sixteen proteins that CFS had preselected. SVM with a 2-degree polynomial kernel was used. Among the 10,699 models that satisfied the performance standard (SN and SP ≥

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

70%) for ADD in comparison to HC, two models featuring six and eight protein panels demonstrated exceptional functioning.[4] When compared to the HC categorization, these models fared poorly in the evaluation of mild cognitive impairment (MCI). For MCI vs. HC classification, five models with certain protein panels performed best, identifying AD participants with SN and SP above 80% and 70%, respectively, at dementia and MCI phases.[5]

TABLE II
PERFORMANCE OF EXISTING BLOOD BIOMARKER PANELS FOR AD DETECTION

Panel size	Panel	ML model	ADD vs. HC (Dataset 1)			MCI vs. HC (Dataset 2)		
			SN	SP	AUC	SN	SP	AUC
11	Adip, B2M, CRP, FABP, FVII, IL18, MCP1, PPP, TLSP, TNC, VCAM	Random forest	85.2	25.9	0.62	81.6	46.3	0.72
5	A1M, ApoE, BNP, IL16, SGOT	Logistic regression	85.2	74.1	0.90	79.0	50.0	0.70
8	A1M, ApoA2, ApoE, BNP, Eot3, IGM, PLGF, SGOT	Random forest	88.0	72.4	0.87	80.9	46.3	0.69
5	A1M, ApoA2, ApoE, BNP, SGOT		87.0	62.1	0.83	83.1	38.9	0.67
13	ApoA2, ApoE, BNP, Eot3, HBEGF, IGM, IL16, PLGF, PYY, SGOT, TNC, TTR, Vit		92.6	60.3	0.87	85.3	42.6	0.72
14	A1M, A2M, ApoA2 ApoE, BNP, BTC, CRP, Eot3, IGM, IL16, MPO, PLGF, RAGE, SGOT		92.6	67.2	0.91	83.1	44.4	0.70
6	A1M, A2M, AAT, ApoE, CC3, PPP	Naive Bayes	86.1	63.8	0.82	78.3	37.0	0.62
5	A1M, A2M, CC3, IGM, TNC	SVM	81.1	60.5	0.77	75.7	35.2	0.65

TABLE III
CFS-BASED PRESELECTED PROTEINS

Protein	*P*	
	ADD vs. HC (Dataset 1)	MCI vs. HC (Dataset 2)
A1M	2.9E-6	3.3E-1
A2M	2.5E-3	3.2E-1
ApoA2	3.2E-8	1.1E-1
ApoE	1.1E-7	3.8E-4
BNP	7.7E-7	5.2E-2
BTC	4.4E-2	2.4E-1
CD5L	1.0 E-1	8.6E-1
Eot3	5.5E-5	6.2E-3
IGM	9.7E-7	3.9E-5
IL3	8.1E-3	6.9E-15
MCSF1	4.0E-1	8.4E-2
PAPPA	7.7E-4	1.6E-1
PLGF	1.3E-5	3.2E-1
PYY	2.7E-6	5.9E-1
RAGE	6.5E-3	6.3E-1
SGOT	9.2E-6	2.2E-6

TABLE IV
PERFORMANCE OF NOVEL CANDIDATE BLOOD BIOMARKER PANELS FOR EARLY DETECTION OF AD

Panel size	Panel	ADD vs. HC (Dataset 1)			MCI vs. HC (Dataset 2)		
		SN	SP	AUC	SN	SP	AUC
7	A2M, ApoE, BNP, Eot3, PLGF, RAGE, SGOT	88.5	70.4	0.87	80.1	70.4	0.80
7	A2M, ApoE, BNP, Eot3, PYY, RAGE, SGOT	88.9	73.8	0.89	77.9	74.1	0.80
8	A2M, ApoE, Eot3, IGM, MCSF1, PYY, RAGE, SGOT	85.3	71.6	0.86	83.8	70.4	0.83
9	A2M, ApoA2, ApoE, BNP, BTC, Eot3, PYY, RAGE, SGOT	85.0	75.0	0.89	80.1	72.2	0.80
10	A1M, A2M, ApoE, BNP, BTC, Eot3, IGM, MCSF1, PAPPA, SGOT	88.1	72.9	0.89	83.1	70.4	0.80

VII. APPLICATIONS

There are several real-world uses for the project utilizing support vector machines to detect Alzheimer's disease early using blood plasma proteins:

1. Diagnosis Clincial
2. Programs for Screening
3. Medical Studies
4. Investigation and Biomarker Finding
5. Evaluation of Risk
6. Interventions Preventive

VII. CONCLUSION AND FUTURE WORK

Using a novel approach, we developed potential models and identified five candidate non-amyloid biomarker panels for early detection of Alzheimer's Disease (AD).[1]

Based on the detected panels, these models showed that they could classify AD dementia, prodromal AD, and normal controls with an AUC of at least 0.80, a sensitivity of more than 80%, and a specificity of more than 70%. This raises

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

the possibility of identifying non-amyloid proteins that represent metabolic processes related to or independent of AD in the early stages of the illness.[2]

Finding the condition early on could help with early intervention and provide new understanding of amyloid pathology[3]. Improved clinical trial interventions could result from a better understanding of protein interactions.[4]

Talk about possible directions for further study or model enhancements. Think about extending the research to encompass more varied data sets or applying further sophisticated machine learning methods.[5]

REFERENCES

[1] Association, "2018 Alzheimer's disease facts and figures," Alzheimer's & Dementia, vol. 14, no. 3, pp. 367-429, 2018.

[2] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World Alzheimer report 2016: improving health-care for people living with dementia: coverage, quality and costs now and in the future," 2016.

[3] B. Dubois et al., "Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria," Alzheimer's & Dementia, vol. 12, no. 3, pp. 292-323, 2016.

[4] C. S. Eke, E. Jammeh, X. Li, C. Carroll, S. Pearson, and E. Ifeachor, "Identification of Optimum Panel of Blood-based Bio-markers for Alzheimer's Disease Diagnosis Using Machine Learning," presented at the Proc. IEEE Eng Med Biol Soc, Jul, 2018.

[5] G. M. McKhann et al., "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on AgingAlzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," Alzheimer's & Dementia, vol. 7, no. 3, pp. 263- 269, 2011.