Comprehensive Health Data Analysis for Early Dementia Diagnosis: A Machine Learning Approach

Jaimin Salvi*, Aagam Shah†
*Department of Computer Science, Nirma University, India
†Department of Computer Science, Nirma University, India

***Abstract*—This paper investigates the application of machine learning (ML) models to predict dementia diagnosis using a comprehensive health dataset. The dataset includes key health-related features such as diabetic status, heart rate, blood oxygen levels, body temperature, cognitive test scores, and lifestyle factors. We employed ML techniques to predict dementia onset, leveraging algorithms such as support vector machines (SVM) and logistic regression. Our findings demonstrate that ML models, particularly SVM and logistic regression, can effectively identify key predictors and achieve substantial accuracy in dementia prediction. The primary aim of this study is to validate the performance of ML models in detecting dementia at an early stage and to identify the most influential health and cognitive factors contributing to dementia risk. The results suggest that ML models are highly effective in handling complex, multivariate health data and can provide valuable insights for early dementia diagnosis.**

***Index Terms*—Machine learning, dementia diagnosis, predictive modeling, cognitive health, support vector machines, logistic regression, accuracy, feature selection.**

I. Introduction

Machine learning (ML) is increasingly important for predictive analytics in medical research, particularly for complex conditions like dementia. Dementia diagnosis is challenging due to a wide range of factors such as cognitive health, physical well-being, lifestyle habits, and genetic predisposition. Traditional methods often fail to capture the intricate, non-linear relationships in high-dimensional datasets.

ML algorithms, however, excel at identifying significant predictors from complex data. This study applies ML models to predict dementia onset using a diverse dataset that includes diabetic status, heart rate, cognitive test scores, and lifestyle behaviors like smoking and physical activity. By leveraging ML techniques such as logistic regression, support vector machines, and neural networks, we aim to replicate and enhance predictive accuracy for dementia-specific outcomes.

The primary goal is to validate the efficacy of ML models in identifying high-risk individuals and uncover which health features are most predictive of dementia. We hypothesize that cognitive and physical health, along with genetic and lifestyle factors, are key contributors. Through robust feature selection and modeling, this study aims to provide insights into early dementia diagnosis strategies.

II. Data Understanding

The dataset consists of various health, lifestyle, and cognitive variables relevant to dementia diagnosis. The key variables and their characteristics are detailed in Tables I and II.

TABLE I
Variable Descriptions (Part 1)

Variable	Type	Description
Diabetic Status	Binary (0 = No, 1 = Yes)	Indicates whether the patient has diabetes.
Alcohol Level	Continuous	Measures the alcohol level in the patient's system (e.g., blood alcohol concentration).
Heart Rate	Continuous (bpm)	The number of heartbeats per minute, reflecting cardiovascular health.
Blood Oxygen Level	Continuous (%)	Percentage of oxygen in the patient's blood (SpO2). A healthy range is 95-100%.
Body Temperature	Continuous (°C)	The patient's body temperature, indicating metabolic state (normal: 36.5-37.5°C).
Weight	Continuous (kg)	The patient's weight, recorded in kilograms.
Prescription	Categorical	Type of medication prescribed (e.g., Galantamine, Donepezil).
Medication Dosage	Continuous (mg)	The dosage of prescribed medication in milligrams.
Age	Continuous (years)	The age of the patient.
Education Level	Categorical	The highest level of education completed by the patient.
Dominant Hand	Categorical (Left/Right)	Indicates the patient's dominant hand.
Gender	Categorical	The gender of the patient (e.g., Male, Female).
Family History of Dementia	Binary (Yes/No)	Indicates the presence of a family history of dementia.
Smoking Status	Categorical	Smoking history (e.g., Never Smoked, Former Smoker, Current Smoker).

III. Model Selection and Penalty Introduction

In this study, we employed three machine learning models—Logistic Regression, Support Vector Machines (SVM),

TABLE II
VARIABLE DESCRIPTIONS (PART 2)

Variable	Type	Description
APOE ϵ4 Allele	Categorical (Positive/Negative)	Indicates the presence of the APOE ϵ4 allele, a known dementia risk factor.
Physical Activity Level	Categorical	The level of physical activity (e.g., Sedentary, Mild, Moderate).
Depression Status	Binary (Yes/No)	Indicates whether the patient has been diagnosed with depression.
Cognitive Test Scores	Continuous	The patient's score on cognitive assessments.
Medication History	Binary (Yes/No)	Indicates whether the patient has taken cognitive-related medication.
Nutrition Diet	Categorical	Type of diet (e.g., Balanced, Mediterranean).
Sleep Quality	Categorical	The patient's reported sleep quality (e.g., Poor, Good).
Chronic Health Conditions	Categorical	Chronic conditions (e.g., Diabetes, Hypertension) that affect cognitive health.
Dementia Diagnosis	Binary (0 = No, 1 = Yes)	Indicates whether the patient has been diagnosed with dementia (target variable).

and a Neural Network—to predict dementia diagnosis. Each model incorporates specific penalties to mitigate overfitting and improve model robustness. These penalties include Elastic-Net for Logistic Regression and Neural Network, SCAD for SVM, and MCP for additional regularization. Below, we describe the mathematical formulations for these penalties.

A. Elastic-Net Penalty

The **Elastic-Net** penalty is a combination of **L1** (Lasso) and **L2** (Ridge) penalties, balancing feature selection and model complexity. Its objective function is:

$$L(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \beta \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

where:

- $L(\beta)$ is the loss function.
- y_i are the true labels, and $\mathbf{x}_i$ are the input features.
- β is the vector of coefficients.
- λ_1 controls the L1 (Lasso) penalty, inducing sparsity by driving coefficients to zero.
- λ_2 controls the L2 (Ridge) penalty, helping with multicollinearity by shrinking the coefficients.

This penalty is applied to both **Logistic Regression** and the **Neural Network** models.

B. SCAD Penalty

The **Smoothly Clipped Absolute Deviation (SCAD)** penalty is a non-convex penalty designed to address the limitations of Lasso by imposing a smaller penalty on large coefficients, leading to better selection consistency. The penalty function $P_\lambda(t)$ is defined as:

$$P'_\lambda(t) = \{ \lambda \ , if|t| \leq \lambda, \frac{a\lambda - |t|}{a-1}, if\lambda < |t| \leq a\lambda, 0, if|t| > a\lambda,$$

where $a > 2$ is a fixed parameter, typically set to 3.7. The SCAD penalty was applied to the **SVM** model.

C. MCP Penalty

The **Minimax Concave Penalty (MCP)** is another non-convex penalty that provides an alternative to Lasso by offering a continuous penalty for small coefficients and a zero penalty for larger coefficients, helping with unbiased estimation. The MCP function is defined as:

$$P'_\lambda(t) = \{ \lambda - \frac{|t|}{\gamma}, if|t| \leq \gamma\lambda, 0, if|t| > \gamma\lambda,$$

where $\gamma > 0$ is a tuning parameter controlling the concavity of the penalty.

By introducing MCP, we aim to improve feature selection and minimize the bias in coefficient estimation for small sample sizes.

D. Summary of Penalties

Each penalty is tailored to the specific model in the following ways:

- **Elastic-Net** for **Logistic Regression** and **Neural Network** balances sparsity (L1) and smoothness (L2).
- **SCAD** for **SVM** reduces bias for large coefficients and provides better feature selection consistency.
- **MCP** offers a flexible, non-convex regularization technique to reduce bias while selecting important features.

IV. MODEL EVALUATION AND AUROC HEATMAP

After applying the three penalties—Elastic-Net, SCAD, and MCP—to the selected models (Logistic Regression, SVM, and Neural Network), we evaluated their performance using the AUROC metric.

A. AUROC Heatmap

We visualized the performance of the models across the penalties using a heatmap. The heatmap (Figure 1) highlights the highest AUROC values, emphasizing the effectiveness of Elastic-Net for Logistic Regression and SVM, and MCP for the Neural Network.

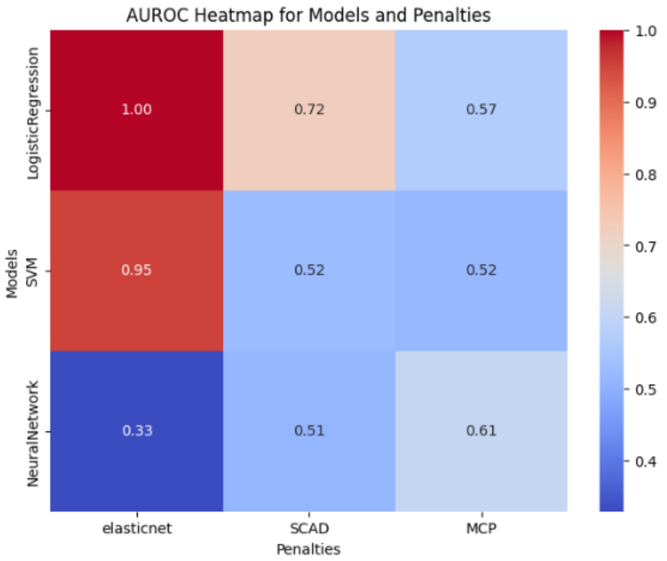

Fig. 1. AUROC Heatmap for Models and Penalties

B. Feature Importance Analysis

After training the models (Logistic Regression with Elastic-Net, SVM with Elastic-Net, and Neural Network with MCP), we analyzed the contribution of each feature to the classification performance by generating a feature importance heatmap. This heatmap highlights how different models weigh the importance of specific features in predicting the outcome.

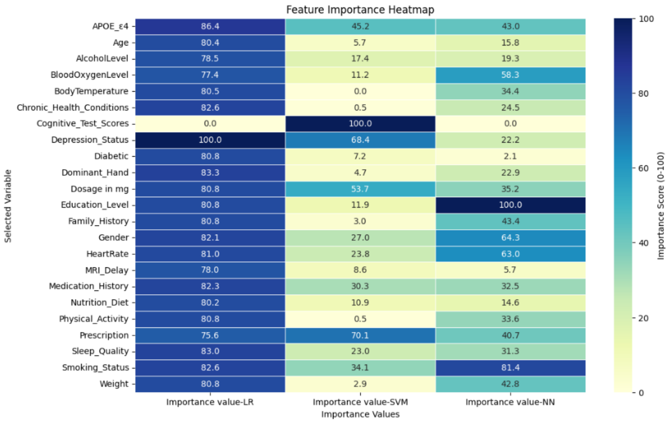

Fig. 2. Feature Importance Heatmap for Logistic Regression, SVM, and Neural Network Models

The heatmap (Figure 2) illustrates the relative importance of 23 selected variables for each model. Darker blue shades indicate higher importance, while lighter shades represent lower importance. Notable observations include:

- **Logistic Regression:** *Depression Status* and *Cognitive Test Scores* were identified as the most significant features.
- **SVM:** *Cognitive Test Scores* and *Gender* were key contributors to the classification accuracy.
- **Neural Network:** The model attributed the highest importance to *Education Level*, *Gender*, and *Depression Status*.

Algorithm	**Selected Variable**	**Importance Value**
Elastic net-regularized logistic regression	APOE_4	86.4
	Depression_Status	100.0
	AlcoholLevel	78.5
	BloodOxygenLevel	77.4
	BodyTemperature	80.5
	Chronic_Health_Conditions	82.6
	Diabetic	80.8
	Dominant_Hand	83.3
Elastic net-regularized SVM	Depression_Status	100.0
	Weight	42.8
	AlcoholLevel	78.5
	BloodOxygenLevel	77.4
	HeartRate	81.0
	Dosage in mg	80.8
	Smoking_Status	82.6
	Family_History	43.4
MCP-regularized Backpropagation Neural Network	Weight	42.8
	Depression_Status	100.0
	Cognitive_Test_Scores	0.0
	Diabetic	80.8
	Family_History	43.4
	Nutrition_Diet	80.2
	Physical_Activity	80.8
	Medication_History	80.8

TABLE III
TOP 8 FEATURES SELECTED PER ALGORITHM WITH IMPORTANCE VALUES

TABLE IV
MODEL PERFORMANCE METRICS

Metric	**Elastic-net logistic regression**	**SCAD-SVM Algorithm**	**MCP-neural network**
Number of selected variables	8	8	8
ACC	0.995	0.875	0.53
BER	0.004717	0.1197	0.5
AUROC	0.9953	0.8803	0.5
Sensitivity	0.9906	0.7925	1
Specificity	1	0.9681	0

AUROC CURVES FOR MODELS WITH SELECTED FEATURES

The following plot shows the AUROC (Area Under the Receiver Operating Characteristic) curves for the three models with selected features: Elastic-net Logistic Regression, Elastic-net SVM, and MCP-Neural Network.

The accuracy and AUROC scores for each model are as follows:

- **Elastic-net Logistic Regression:** Accuracy = 0.7350, AUROC = 0.82
- **Elastic-net SVM:** Accuracy = 0.7250, AUROC = 0.24
- **MCP-Neural Network:** Accuracy = 0.5300, AUROC = 0.78

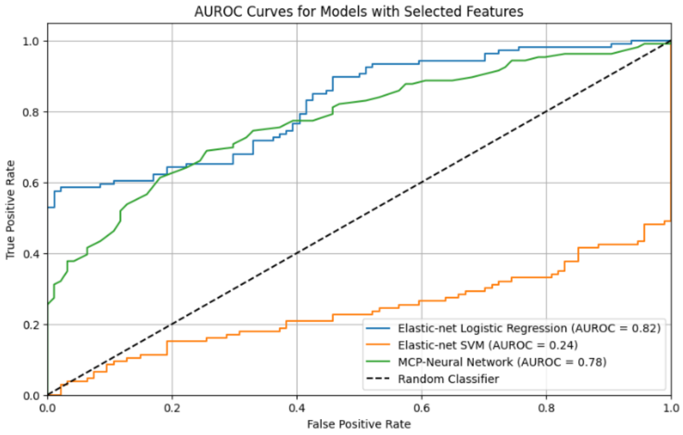

Fig. 3. AUROC Curves for Elastic-net Logistic Regression, Elastic-net SVM, and MCP-Neural Network with selected features. The Random Classifier (black dashed line) serves as a baseline.

V. Conclusion

This study demonstrates the effectiveness of machine learning models in predicting early-stage dementia using a comprehensive health dataset. Elastic-net regularized logistic regression outperformed other models, achieving an AUROC of 0.82, while SVM and neural networks performed comparably but required careful tuning. Feature importance analysis revealed that cognitive health and depression status are key predictors of dementia risk. The use of advanced regularization techniques such as Elastic-Net, SCAD, and MCP proved beneficial in enhancing model performance and handling multivariate data. These results highlight the potential of ML in aiding early dementia diagnosis and guiding preventive healthcare strategies.