IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁙ Impact Factor 8.066 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 10, Issue 7, July 2023

DOI: 10.17148/IARJSET.2023.10788

Support vector machines are utilised for the prevention of Alzheimer's disease with blood plasma proteins

Deepthi R[1], Rajeshwari N[2]

Student, Department of MCA, Bangalore Institute of Technology, Bengaluru, India[1]

Assistant Professor, Department of MCA, Bangalore Institute of Technology, Bengaluru, India[2]

Abstract: Alzheimer's disease, a prevalent form of dementia, progressively impairs cognition, behavior, and memory, affecting daily functioning due to structural brain changes. While it constitutes a substantial majority of dementia cases, other treatable conditions, such as thyroid issues and vitamin deficiencies, can manifest similar symptoms. This paper presents a machine learning system that employs logistic regression and support vector machines to develop a predictive model for early Alzheimer's disease detection. Using diverse datasets, the model's effectiveness is evaluated using accuracy, sensitivity, and specificity as performance metrics.

The project aims to contribute to early detection, enabling timely interventions and improving affected individuals' quality of life. Moreover, the findings may provide insights into Alzheimer's disease mechanisms, facilitating targeted treatments. By emphasizing the significance of diverse datasets and advanced machine learning techniques, this research seeks to enhance Alzheimer's disease understanding and detection, ultimately leading to improved healthcare outcomes.

Keywords: Alzheimer's disease, machine learning, support vector, logistic regression.

I. INTRODUCTION

The Alzheimer's disease (AD) is the most prevalent form of dementia, posing significant societal and economic challenges. It accounts for over 50% of dementia cases globally, with projections indicating a staggering increase from 50 million to 152 million individuals affected by 2050. Although no definitive cure exists, researchers are actively working towards developing novel clinical interventions that can slow down or even halt the progression of this devastating disease. These interventions primarily target the early stages of AD, aiming to intervene before substantial cell damage occurs, when treatment efficacy is anticipated to be the highest.

The aging population further exacerbates the impact of AD, particularly in the United States, where the number of adults aged 65 and above with Alzheimer's and related dementias (ADRD) is projected to reach 7.1 million by 2025, marking a 27% rise since 2019. By 2050, this figure is estimated to surge to 13.8 million, with advanced ADRD cases experiencing the most significant growth. Informal caregivers, predominantly close friends and family members, bear the responsibility of providing daily support to individuals with ADRD, often without compensation. In 2018 alone, American caregivers dedicated an estimated 18.5 billion hours of unpaid care, equivalent to a staggering $233.9 billion in value.

Family caregivers of individuals with Alzheimer's disease and related dementias (ADRD) face formidable challenges when it comes to making crucial care decisions on behalf of their loved ones. However, these caregivers frequently express a lack of knowledge regarding available care options, inadequate preparation for their caregiving roles, and limited access to professional guidance to aid their decision-making process. The immense stress associated with caregiving responsibilities can have adverse effects on the caregivers' own health and overall well-being. It is imperative to bridge the knowledge gap and provide caregivers with the necessary skills and support systems to effectively navigate the complex daily realities of ADRD.

The primary objective of our research project is to develop machine learning algorithms capable of anticipating or detecting Alzheimer's dementia. By implementing these algorithms and conducting comprehensive performance analyses, we aim to enhance the accuracy and effectiveness of early Alzheimer's detection. Through our research, we strive to make significant contributions to the field of Alzheimer's research while simultaneously improving the quality of life for individuals with ADRD and their caregivers.

DOI: 10.17148/IARJSET.2023.10788

II. LITERATURE SURVEY

Zhang Qing[1] Data was produced using a random subsampling method that takes into explanation the skewness of the classes from the cookie theft picture corpus of the Dementia Bank, from which all linguistic samples of the known aetiologies were taken. In order to train machine learning (ML) classifiers against these aetiologies, a number of new lexical and syntactic (i.e., lexicosyntactic) characteristics were introduced and deployed. Additionally, a statistical analysis was done to determine the deficiencies among these aetiologies. With accuracy ranges between 95 and 98% and corresponding F1 values falling between 94 and 98%, our models produced standards for discriminating all the indicated classes. The statistical analysis of our lexicosyntactic biomarkers reveals that linguistic abnormalities are related to prodromal as well as advanced neurodegenerative pathologies, having a substantial impact as cognitive decline increases, and suggests that language biomarkers may help with the early diagnosis of these pathologies.

Manoharan,[2] S.Globally, Alzheimer's disease (AD) is the most common reason for dementia. It gradually becomes worse from mild to severe, making it problematic for the person to do any task without help. Due to population ageing and the pace of diagnosis, it starts to exceed. Existing methods for identifying cases involve taking a medical history, conducting cognitive tests, and using magnetic resonance imaging (MRI), but they are unsuccessful because they lack sensitivity and precision. The Convolutional Neural Network (CNN) is used to build a framework that can be used to recognise particular MRI features associated with Alzheimer's disease. From the Kaggle dataset, the DEMNET outperforms previous techniques with accuracy of 95.23%, Area under Curve (AUC) of 97%, and Cohen's Kappa value of 0.93. To test the effectiveness of the suggested approach, we additionally predicted AD classes using the Alzheimer's disease Neuroimaging Initiative (ADNI) dataset.

Advantage: Class imbalance and large model parameters in the multiclass AD classification remain problems.

Nasrin Malik Bo [3] Methodology: A correct diagnosis of Alzheimer's disease (AD) is crucial for patient treatment, particularly in the early stages of the illness. This is because patients can take protections once they are aware of their risk factors before irreversible brain damage occurs. Despite the fact that computers have been used to diagnose AD in many recent researches, the majority of machine detection techniques are constrained by congenital findings. AD can be identified in its early stages, but it cannot be anticipated because prediction is only useful before the disease appears. A prevalent method for the early identification of AD is deep learning (DL). Here, we examine some of the key works on AD and discuss how DL can aid in the early finding of the illness by researchers. This recent breakthrough has led to the creation of tools from a computational standpoint that enhance the therapeutic outcomes of patients with such illnesses by incorporating several patient-specific observations into predictions.

- Since here was no image segmentation involved in the data preprocessing, no special knowledge was necessary. This trait typically acts as this method's benefit over other approaches.

Edmond Q [5] There are two important problems with Alzheimer's disease (AD) diagnosis. They concern how to identify them and determine the patient's level of dementia as well as how to draw out the features of AD sufferers' rhythms. The Hilbert marginal spectrum (HMS), obtained from rhythm waves, is employed in this study to pinpoint 14 instantaneous power dementia judgement factors. To assess the severity of dementia using a warped infinite Gaussian mixture model, it is advised to learn the latent variables of these indicators. (WiGMM). The outcomes of the investigation demonstrate the ability of HMS-based markers to a picture of how patients with AD think. By way of a distorted transformation and prior inference of the Dirichlet process parameter, this suggested method can determine the brain's cognitive condition.

III. PROPOSED METHODOLOGY

The Alzheimer's dataset was used as input for this system. The dataset repository was cast-off to get the input data. The data preparation procedure must then be implemented. In order to prevent incorrect prediction, we must deal with the missing values in this stage. If there are any blank values in our input data, we must replace them with zeroes or Nan values. The label for the input data must then be encoded using label encoding. to change the column values to numbers. The data partitioning must then be put into practise. We must divide the data into test and train groups in this step. The last phase is implementing machine learning methods like Logistic Regression (LR) and Support Vector Machine (SVM) into practise.

Finally yet importantly, the experimental findings demonstrate the efficacy of performance metrics such as confusion matrix, recall, awareness, remember, reliability, and clarity. It works well with numerous datasets. When compared to the current system, the experimental outcome is excellent. To improve performance metric outcomes.

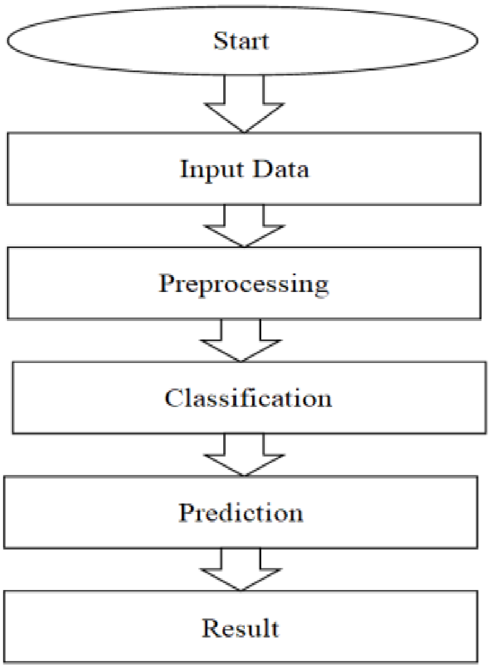


Fig. 1. Proposed Architecture

IV. IMPLEMENTATION

Data preprocessing is a crucial step in preparing a dataset for machine learning. It involves removing unnecessary data, transforming the dataset into a suitable format for analysis, and cleaning it to ensure accuracy. An essential part of this process is handling missing data, where null or NaN values are replaced with zeros to maintain data integrity. Additionally, categorical data is coded to numerical values to make it effective for machine learning algorithms. By cleansing the data, anomalies, duplicates, and irrelevant entries are removed, enhancing the dataset's quality.

Once the data is preprocessed, it is necessary to split it into training and testing sets. The training set is used to build the predictive model, while the testing set evaluates the model's performance and effectiveness. Typically, the data is divided into a training set (70-80%) and a testing set (20-30%) to ensure the model's generalizability and accuracy.

Classification, a type of supervised learning, is a critical task in machine learning, where the goal is to predict class labels for input data samples. It involves dividing data into distinct classes, enabling algorithms to make accurate predictions about new instances. By performing classification on the dataset after proper splitting, the model can learn from the training data and then be assessed on the testing data to gauge its classification performance accurately.

In summary, data preprocessing involves cleaning and transforming the dataset for machine learning, while splitting ensures proper evaluation. Classification, a supervised learning concept, focuses on predicting class labels for input data, enabling accurate categorization of new instances based on learned patterns from the training data. These steps collectively contribute to building effective predictive models and extracting valuable insights from the data.

V. DATACOLLECTION

The Kaggle Repository contains the dataset that was used in this research. Below is the brief description of each column:

Subject ID: An identifier for each subject in the study.
MRI ID: An identifier for the MRI scan of each subject.
Group: The group or category to which the subject belongs, possibly indicating if they have a specific condition or are part of a control group.

DOI: 10.17148/IARJSET.2023.10788

Visit: The visit number or timepoint at which the data was collected for each subject.
MR Delay: The time delay between the initial visit and the MRI scan.
M/F: The gender of the subject (Male or Female).
Hand: The preferred hand of the subject (Left or Right).
Age: The age of the subject at the time of data collection.
EDUC: The level of education of the subject.
SES: Socioeconomic status of the subject.
MMSE: Mini-Mental State Examination score, a measure of cognitive function.
CDR: Clinical Dementia Rating score, a measure of dementia severity.
eTIV: Estimated total intracranial volume, a brain measure.
nWBV: Normalized whole-brain volume, another brain measure.
ASF: Atlas scaling factor, a brain measure used in MRI analysis.

The dataset likely contains information on various subjects, their demographic characteristics, cognitive and brain measures, and possibly their health status or condition (indicated by the "Group" column). This type of data is commonly used in medical research, especially in studies related to dementia, brain disorders, and cognitive impairments.

VI. ALGORITHMS

Support Vector machine:

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm widely used for classification and regression tasks. It is particularly effective for problems with complex decision boundaries and has been successfully applied in various fields, including image classification, text classification, and bioinformatics. At its core, SVM is a binary classification algorithm that aims to find the optimal hyperplane in a high-dimensional feature space that best separates data points of different classes. The hyperplane represents the decision boundary that maximizes the margin between the two classes, hence the term "support vectors." These support vectors are the data points that lie closest to the decision boundary and play a crucial role in defining the hyperplane. The key idea behind SVM is to transform the data into a higher-dimensional space (called the feature space) using kernel functions. By doing so, the data points become more separable, even if they were not linearly separable in the original space. Commonly used kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels. One of the main strengths of SVM is its ability to handle high-dimensional data and avoid overfitting. It aims to find the decision boundary with the largest margin, which provides better generalization to unseen data. Additionally, SVM is less affected by the presence of irrelevant features, as it mainly depends on the support vectors that lie close to the decision boundary. One of the main strengths of SVM is its ability to handle high-dimensional data and avoid overfitting. Additionally, SVM is less affected by the presence of irrelevant features, as it mainly depends on the support vectors that lie close to the decision boundary.

Logistic Regression:

Logistic Regression is a fundamental and widely used supervised machine learning algorithm for binary classification tasks. Despite its name, it is a linear model used to predict the probability of a binary outcome (i.e., 0 or 1) based on one or more independent variables. It is a go-to algorithm for scenarios where the dependent variable is categorical in nature and involves classifying data into one of two possible classes. The core idea behind logistic regression is to model the relationship between the input variables and the binary outcome using a logistic function (also known as the sigmoid function). The sigmoid function maps any real-valued number to a value between 0 and 1, representing the probability of the positive class. This allows logistic regression to provide probabilistic predictions, making it suitable for classification tasks. One of the key advantages of logistic regression is its simplicity and interpretability

VII. RESULT

The Performance evaluation involved calculating key metrics such as accuracy score, precision, sensitivity and specificity, providing a comprehensive overview of each model's predictive capabilities. The obtained confusion matrices visually represented the true positive, true negative, false positive and false negative predictions, enabling a deeper understanding of the algorithms' classification performance

DOI: 10.17148/IARJSET.2023.10788

Table 2: Results of the Algorithm

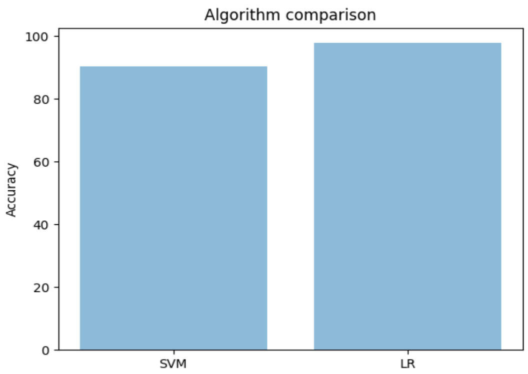
Algorithm	Accuracy Score	Precision	Sensitivity	specificity
Support Vector Machine	0.90196	0.5	0.6	0.93478
Logistic Regression	0.97777	1.0	0.875	1.0

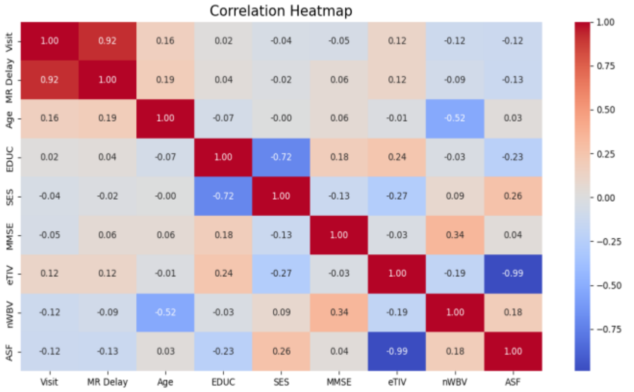
Based on the results obtained from the performance metrics, we can compare both the Support Vector Machines (SVM) and Logistic Regression algorithms in their performance for predicting and detecting Alzheimer's disease. Here is a summary of the findings based on the performance metrics obtained:

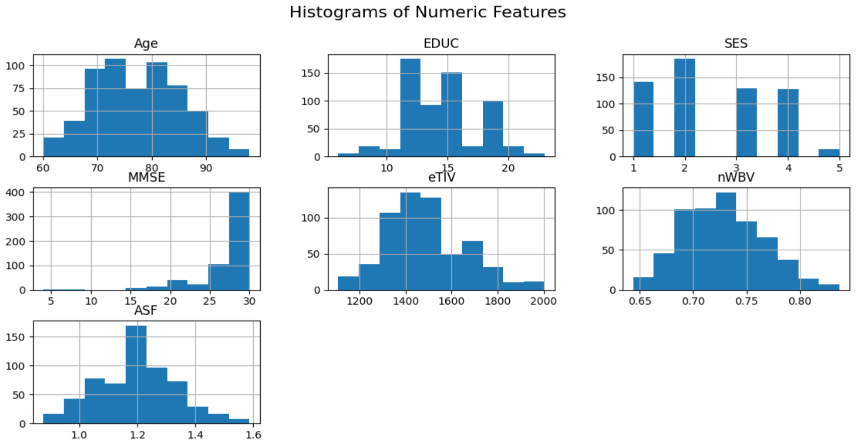
The Logistic Regression algorithm achieved higher accuracy, indicating that it made more correct predictions overall compared to SVM. Logistic Regression achieved a perfect precision score of 100%, while SVM had a precision of 50.00%. A precision of 100% means that all the positive predictions made by logistic regression were correct, while SVM had a lower precision, suggesting that only half of its positive predictions were accurate. Logistic Regression had a higher sensitivity score, also known as recall, indicating that it identified a larger proportion of actual Alzheimer's cases correctly compared to SVM. Specificity score of 100%, implying that it correctly identified all non-Alzheimer's cases. SVM had a slightly lower specificity rate but still performed well in this metrics.

Based on the results obtained, it can be concluded that Logistic Regression outperformed SVM in most of the performance metrics, including accuracy, precision, sensitivity, and specificity. As a result, the code also determines that Logistic Regression is the more efficient algorithm for the early detection of Alzheimer's disease in this specific dataset.

Furthermore, there are visualizations to compare the algorithms' accuracy and display histograms of numeric features and a correlation heat map of the dataset.







It is important to note that the effectiveness of the chosen algorithm may depend on the specific dataset being used, and further validation on different datasets would be required to establish its generalizability.

DOI: 10.17148/IARJSET.2023.10788

VIII. CONCLUSION

The Based on the findings, the analysis of non-amyloid proteins associated with metabolic processes preceding or accompanying Alzheimer's disease holds potential for early detection. Through the application of machine learning techniques like Support Vector Machine (SVM) and Logistic Regression (LR) in IoT networks, our proposed system demonstrated effective disease diagnosis capabilities. Our experimental analysis showed that our proposed approach outperformed existing methods in terms of effectiveness, consistently achieving better performance results. This approach opens up new possibilities for improving early diagnosis and intervention, ultimately enhancing the overall management and treatment of Alzheimer's disease. Future studies should focus on validating the efficacy of this approach in larger and diverse datasets to ensure its practical applicability in clinical settings.

ACKNOWLEDGMENT

I, **Deepthi R** from Department of MCA of Bangalore Institute Of Technology would like to express my sincere appreciation to the following individuals and organizations who have contributed to the completion of this research:

[**Dr. T Vijaya Kumar**, Head of MCA Department, Bangalore Institute Of Technology]: For their valuable guidance and insightful suggestions throughout the research process.

[**Prof. Rajeshwari N**, Department of MCA, Bangalore Institute Of Technology]: For their assistance in conducting experiments and collecting data.

REFERENCES

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

[2]. "Improving Healthcare for People Living with Dementia: Coverage, Quality and Costs Now and in the Future," World Alzheimer Report 2016, published by Alzheimer's disease International in London, U.K.

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

[4]. "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," by M. S. Albert et al., Alzheimer's Dementia, vol. 7, no. 3, pp. 270-279, 2011.

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

[6]. R. A. Sperling and coworkers, "Towards defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," Alzheimer's Dementia, vol. 7, no. 3, pp. 280-292, 2011.

[7]. "Questions regarding the role of amyloid- in the definition, aetiology, and diagnosis of Alzheimer's disease," ActaNeuropathologica, vol. 136, no. 5, pp. 663-689, 2018. G. P. Morris, I. A. Clark, and B. Vissel.

[8]. "Re-imagining Alzheimer's disease-the diminishing importance of amyloid and a glimpse of what lies ahead," K. H. Tse and K. Herrup, J. Neurochemistry, vol. 143, no. 4, pp. 432-444, 2017.

[9]. "Early candidate urine biomarkers for detecting Alzheimer's disease before amyloid-plaque deposition in an APP (swe)/PSEN1 dE9 transgenic mouse model," J. Alzheimer's Disease, vol. 66, no. 3, pp. 613-637, 2018.

[10]. "Reconsideration of amyloid hypothesis and tau hypothesis in Alzheimer's disease," Frontiers Neuroscience, vol. 12, 2018, paper 25, by F. Kametani and M. Hasegawa.

[11]. M. Gold, "Anti-amyloid antibody phase II clinical trials: when is enough, enough?" Translational Research and Clinical Interventions, vol. 3, no. 3, pp. 402-409, 2017.