Detection of Alzheimer's disease at Early Stage using Machine Learning

S.Pavalarajan,
Department of Computer Science and Business Systems
PSNA College of Engineering and Technology
Dindigul, India
e-mail:rajan@psnacet.edu.in

S.Shahul Hammed*
Department of Computer Science and Business Systems
PSNA College of Engineering and Technology
Dindigul, India
e-mail:shahul.y2s@gmail.com

C.Preethi
Department of Computer Science and Business Systems
PSNA College of Engineering and Technology
Dindigul, India
e-mail:preethic@psnacet.edu.in

B.Arun Kumar
Department of Computer Science and Engineering
Sri Krishna College of Engineering and Technology
Coimbatore, India
e-mail:arunkumar.oct06@gmail.com

K.Haripriya
Department of Computer Science and Business Systems
PSNA College of Engineering and Technology
Dindigul, India
e-mail:haripriyame@psnacet.edu.in

T.Mohanraj
Department of Computer Science and Engineering
Karpagam Academy of Higher Education
Coimbatore, India
e-mail:mohanrajt.me@gmail.com

2022 International Conference on Advanced Computing Technologies and Applications (ICACTA) | 978-1-6654-9515-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACTA54488.2022.9752827

Abstract—**Identification of dementia is an important concern in medical image processing. Alzheimer is a common kind of dementia. Four machine learning models were designed for identifying this disease. This is classified as a classification problem, and the classification algorithms tested include logistic regression, support vector classifier, decision tree, and random forest classifier. The models are fine tuned by choosing optimal values for parameters that influences the accuracy of the model. The optimal parameters are found using a K-fold cross validation score, and the models are generated using that. The dataset used in the model is longitudinal cross sectional data from OASIS. It has been inferred from the results that random forest classifier performs well than the other models.**

Keywords: logistic regression, support vector classifier, decision tree , random forest classifier, Alzheimer.

I. Introduction

The major form of dementia is the alzheimer disease. The result of this kind of disease the inability of the victim to perform his or her daily routines and considerable reduction in the thought process [1,2]. It has been observed from the studies that the number of patients currently affected with dementia all over the world is around 40 million. It is also expected that the number of victims would increase to around 150 million during the year 2050[3].

In this work the experimental analysis of the machine learning models with optimized parameter values to classify Alzheimer disease is made. Four machine learning models are created with ideal parameters that have a major impact on the model's accuracy. The optimized values of the parameters are identified with cross validation function. The results obtained with the four models are compared.

The dataset used in this paper is the longitudinal MRI data provided by OASIS. The dataset contains the following features. Mini-mental State Examination, Clinical Dementia Rating, Estimated Total Intracranial Volume, Normalized Whole Brain Volume, and Atlas Scaling Factor. In addition to this features two other general features are also added, years of education and socio economic status.

The paper is organised as follows: the first section describes state-of-the-art models, the third section describes the approach used, the fourth section provides a detailed analysis of the results obtained, and the fifth section provides the conclusion.

II. Related Work

Detection of Alzheimer disease is a critical necessity in the medical sphere, but not everyone can afford this model because its accuracy is dependent on the amount of the dataset. [4] have given a solution to this problem by means of transfer learning . two datasets were employed by the authors one from a local hospital and from the Alzheimer's Disease Neuro imaging Initiative database. The former is a small dataset and the later is large dataset. Three kinds of experimentation is done by the authors , one with the dataset from the local hospital and the second experiment, by combining the ADNI dataset for training along with the dataset of the local hospital. The last experiments is done by identifying the effective feature set from the large dataset and extracting the same set of features from the local hospital and implement the classification process. The classification process is done with singular value decomposition and the model proposed shows a comparatively higher accuracy than the other models.

The authors of [5] have examined various machine learning models. It includes ensemble models. Various slices are analysed and histograms are applied and the best slides are filtered and the further processing are done. Weighted voting and majority voting are used to find the best accuracy.

978-1-6654-9515-8/$31.00©2022IEEE

In [6] the authors suggested a strategy for analysing and classifying brain pictures in order to detect Alzheimer's disease. The author's key contribution is to analyse the various feature extraction mechanisms and to select the optimum technique to use in the classification process. The authors have compared eight feature extraction models. The steps that follow feature extraction are T-test and the implementation of classifier. The classifier model that is used by the authors is K-Nearest Neighbour. It has been inferred that shearlet transform performs well than the other models. Multistage classifier is proposed by the authors of [7]. Three different machine learning models are employed by the authors and particle swam optimization is used for optimizing feature extraction. The results are found to be comparatively better than the single classifiers.

The authors of [8] have used deep learning model and they considered voxels'sgray matter as the important parameter. Once this is extracted, the voxels are improved with the help of the gausian filter and from this enhanced voxels irrelevant tissues are removed with the use of the stripping algorithm. This process is followed by the segmentation of the voxel which is given as input to the convolutional neural network model. This model provides better accuracy than the other base line models.With the advancement in artificial neural networks and the related technologies there are various pre trained models that are trained with the general dataset. Various researches are being in process for utilizing the pre trained model in the medical image processing. The authors of [9] have used three such pretrained models for building a convolutional neural network and the results are compared.In [10] the authors have considered the identification of slides that would have high impact on the results rather than employing all the slides. The model used is the convolutional neural network layer, In the middle layer for extraction of features that have high impact, models such as DenseNet121 and variants of it are used.After many experimentation, the authors of [11] have argued that the slices in the central have high impact on the classification process with more number of tissues in them. The authors have modified the classical Alexnet classifier and for convolution they have used 5 layers and for pooling they have used three layers. Totally there are eight layers. The results of this model are comparatively good.

From the MRI dataset, the authors of [12] [15]have separated and extracted the white and gray matter and given as inputs to the convolutional neural network. Convolutional base forms the first two layers of the convolutional neural network. The fully connected layer forms the third part of the model. Different weights were initialized and the optimized weights are identified and the model is trained. Bernoulli function is used for optimizing the weights and to avoid overfit of data.

The method imposed in [13] [14]for choosing the features that are optimal is grey wolf optimization. Six different features are extracted by the authors and two machine learning model and a deep learning model is implemented and tested.

III. METHODOLOGY

3.1 Data pre-processing

The following stages are included in the dataset pre-processing.

- Handling missing data
- Feature scaling
- Train test split

When evaluating the dataset, it was discovered that there are some missing values. The median value is used to fill in the blanks. Feature scaling is used to convert all of the data in a dataset to values that fall within a specific range. MinMxScaler is the scaler used in this project. This type of scaler's default range is 0 to 1. The formula is

$x\_new = (x-min(x)) / (max(x)–min(x))$

Where x is the original data and x_new refers to the scaled value of x. min(x) and max(x) refers to the minimum value and the maximum value. For measuring the performance of the model the dataset should be split into train set and test set. The default test size of 25% is used.

3.2 Machine Learning Model

The following machine learning models are built for detecting the azhemier disease and the results are compared.

- Logistic Regression
- Support Vector Machine
- Decision Tree
- Random Forest

The uniqueness of the approach is that efforts were taken to find the optimal parameters for each model and the model is built as shown in the following figure.

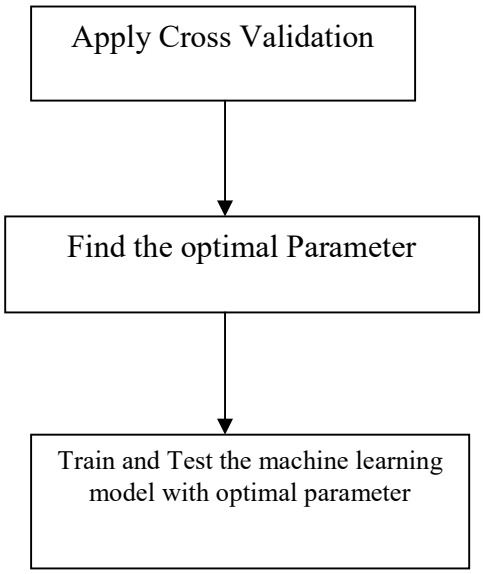


Fig. 1 :overview of the system

K-Fold cross validation mechanism is employed and the optimal parameters are identified and the best scores are used in the model. The steps are as follows

For Each ModelIdentify the parameters(P1,P2…P3) that has high impact on the accuracy.

For Each parameter PiList out the possible values v1,v2…vn of the parameter

For each value vi Perform K-fold cross validation and find the best mean score

Save the value of the parameter that produces the best result.

The value of the parameter that produces better mean score in terms of accuracy is considered as the best parameter and then the model is built it with. The parameter that is considered for each model is explained below.

978-1-6654-9515-8/$31.00©2022IEEE

In the logistic regression model the parameter considered for tuning is Inverse of Regularization

Inverse of regularization: Inverse of regularization is the parameter that is used for avoiding the over fitting problem that arises in the dataset with small amount of data and in the dataset considered we have only 150 rows. A good regularisation value would lower the error value by reducing the disparity between the original and forecasted values.

In the case of the support vector machine, the parameters that are taken into account include

- Kernel Type
- Kernel coefficient
- Penalty parameter

The model of the support vector machine employed in the work is support vector classification called as C-support vector classification. Penalty parameter the objective of any support machine is to identify the hyperplane that has the minimum margin and to identify a hyperplane that is capable of separating maximum number of instances. Both of these criteria cannot be met in all dataset. the penalty parameter C is used for identifying the plane which can best separate the maximum number of instances. The model that is proposed should not have any overfitting values and hence a best penalty parameter should be chosen. Kernel Type : Kernel type plays a key role in designing the support vector machine and there are different kernel functions. The work of the kernel function is to map the observations that are non linear in nature to a higher dimension space.

Kernel coefficient gamma : it helps in fitting the hyperplane exactly. But when the value of the gamma increases there are chances of the data being overfitted so an optimal value should be found

In case of the decision tree classifier the parameter considered is the max_depth which obviously specifies the maximum depth that the tree can grow. In general higher the depth of the tree , the training error value will be low but if a larger value is assigned there is a chance of the model to get overfit.

In case of the random classifier the parameters considered are

- The maximum number of trees a forest can have(n_estimator)
- The maximum number of features that could be considered in case of best split(max_features)
- Tree's maximum depth(max_depth)

n_estimators and max_depth features are obvious with their names and max_features represents the number of subset in the features that has to be considered when a node is split. All these three parameters have a high influence on the accuracy and so optimal values of the parameters should be found.

Figure 2 Comparison of accuracy of the four classifier models

Once the optimal values for the parameters are found, the model is built. The results obtained are compared and analyzed in the next section.

IV. RESULTS

The parameters that are considered for evaluation are as follows.

- Precision(P)
- Recall(R)
- F1-score(F1)

Precision, recall and F1-Score are calculated with the following parameters

True Positive rate

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP+FP} \quad (1)$$

Where,

TP =True Positive, FP =False Positive

False Positive rate

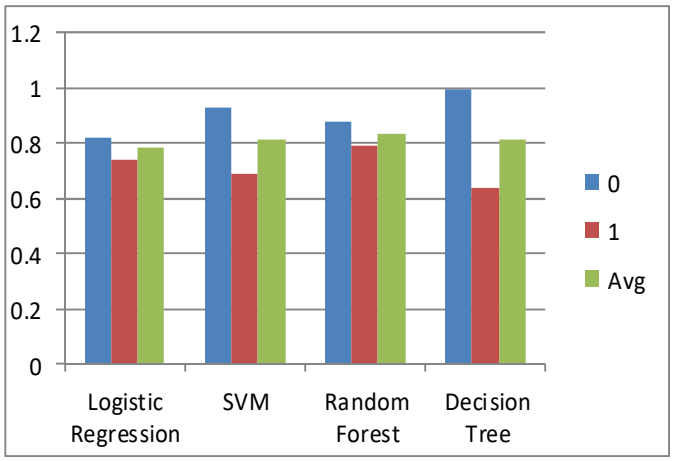
$$\text{False Positive Rate (FPR)} = \frac{FP}{FP+TN} \quad (2)$$

Where TN=True Negative

Precision

$$\text{Precision} = \frac{TP}{TP+FN} \quad (3)$$

Where FN is False Negative



Recall

$$\text{Recall} = \frac{TP}{TP+FP} \quad (4)$$

F-Measure

$$\text{F-Measure} = (5)2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Fig. 3 Comparison of AUC of the four classifier models

The system's accuracy is calculated using the formula below.

Accuracy= (TP+TN)/ (TN+FP+FN+TP)

Area under the curve is calculated with false positive rate and true positive rate.

978-1-6654-9515-8/$31.00©2022IEEE

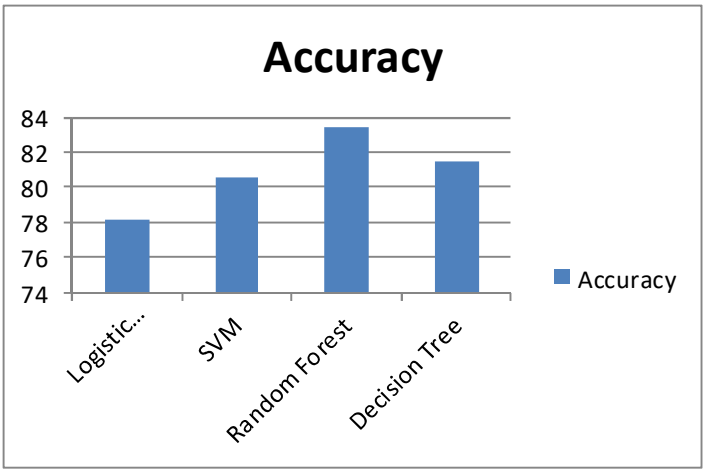


Fig. 4 Comparison of precision of the four classifier models

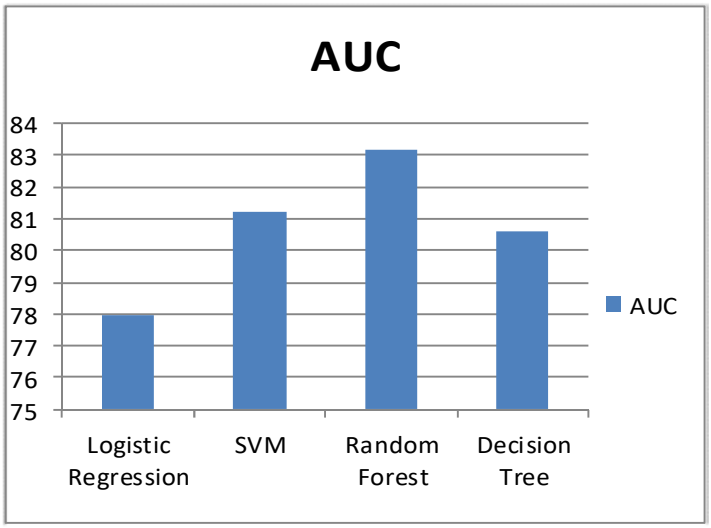


Fig. 5 Comparison of recall for the four classifiers

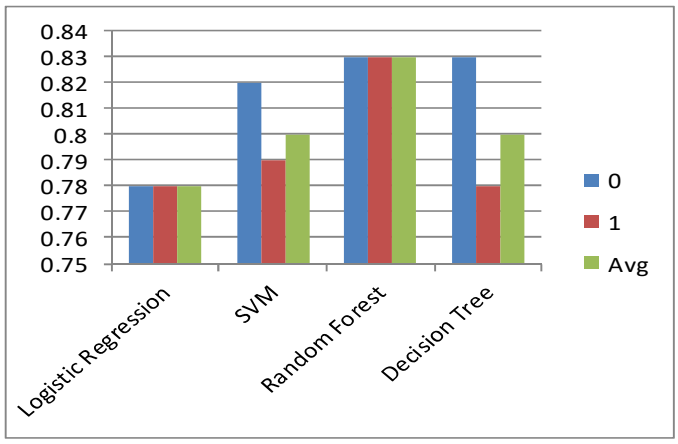


Fig. 6 Comparison of F1-Score for the four classifiers

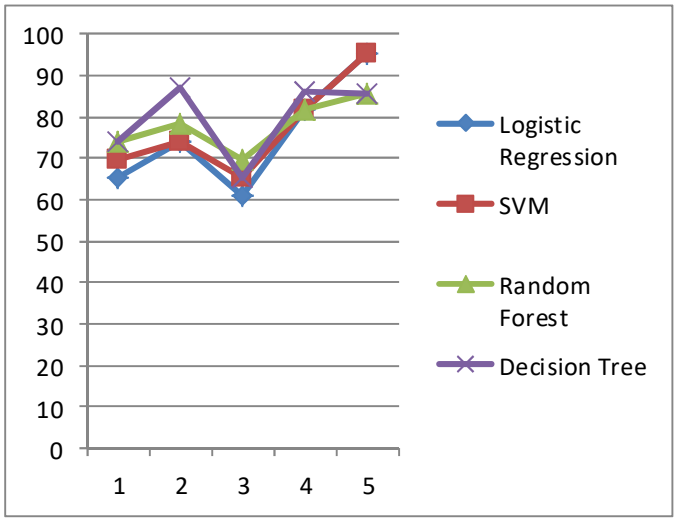


Fig. 7 Cross Validation Score of the Four Classifiers

V. Conclusion and Future Work

Four machine learning models were designed for classifying the alzheimer disease. The models were designed with optimal values for the performance parameters that affect the accuracy of the model. The optimal values are identified with cross validation mechanism. Among the four models, random forest classifier performs well than the other models. A deep learning model for detecting Alzheimer's disease will be developed in the future.

References

[1] Beheshti I., Demirelb H., Matsudaaf H. Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. Comput. Biol. Med. 2017;83:109–119.

[2] Klöppel S., Stonnington C.M., Chu C., Draganski B., Scahill R.I., Rohrer J.D., Fox N.C., Jack C.R., Jr., Ashburner J., Frackowiak R.S. Automatic classification of MR scans in Alzheimer's disease. Brain. 2008;131:681–689.

[3] M. Prince, A. Wimo, M. Guerchet, A. Gemma-Claire, Y.-T. Wu, and M. Prina, "World Alzheimer Report 2015: The Global Impact of Dementia - An analysis of prevalence, incidence, cost and trends," Alzheimer's Dis. Int., p. 84, 2015.

[4] Elshatoury, Heba et al. 'Volumetric Histogram-Based Alzheimer's Disease Detection Using Support Vector Machine'. 1 Jan. 2019 : 515 – 524.

[5] Acharya, U.R., Fernandes, S.L., WeiKoh, J.E. et al. Automated Detection of Alzheimer's Disease Using Brain MRI Images– A Study with Various Feature Extraction Techniques. J Med Syst 43, 302 (2019).

[6] K.R. Kruthika, Rajeswari, H.D. Maheshappa, "Multistage classifier-based approach for Alzheimer's disease prediction and retrieval",Informatics in Medicine Unlocked,Volume 14,2019,Pages 34-42.

[7] Basheera, Shaik & Ram, M. (2019). "Convolution neural network–based Alzheimer's disease classification using hybrid enhanced independent component analysis based segmented gray matter of T2 weighted magnetic resonance imaging with clinical valuation".Alzheimer's & Dementia: Translational Research & Clinical Interventions. Volume 5.pp. 974-986.

[8] B. Khagi, B. Lee, J. Pyun and G. Kwon, "CNN Models Performance Analysis on MRI images of OASIS dataset for distinction between Healthy and Alzheimer's patient," *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, Auckland, New Zealand, 2019, pp. 1-4.

978-1-6654-9515-8/$31.00©2022IEEE

[9] E. Jabason, M. O. Ahmad and M. N. S. Swamy, "Classification of Alzheimer's Disease from MRI Data Using an Ensemble of Hybrid Deep Convolutional Neural Networks," *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*, Dallas, TX, USA, pp. 481-484.

[10] [M. Raza, M. Awais, W. Ellahi, N. Aslam, H.X. Nguyen, H. Le-Minh,"Diagnosis and monitoring of Alzheimer's patients using classical and deep learning techniques",Expert Systems with Applications,Volume 136,2019,Pages 353-364

[11] Huanhuan Ji, Zhenbing Liu, Wei Qi Yan, and Reinhard Klette. 2019. Early Diagnosis of Alzheimer's Disease Using Deep Learning. In Proceedings of the 2nd International Conference on Control and Computer Vision (ICCCV 2019). Association for Computing Machinery, New York, NY, USA, 87–91.

[12] [Shankar K, Lakshmanaprabu S.K., Ashish Khanna, Sudeep Tanwar, Joel J.P.C. Rodrigues, Nihar Ranjan Roy,

[13] [Alzheimer detection using Group Grey Wolf Optimization based features with convolutional classifier,Computers & Electrical Engineering,Volume 77,

[14] Pavalarajan.S and Vincent Antony Kumar.A,'Prediction based object recovery using sequential Monte Carlo method', International Journal of Engineering and Technology 5 (5) PP. 4206 - 4213,2013.

[15] B. A. Kumar, T. Mohanraj, S. ShahulHammed and R. Santhosh, "A Study of BlockChain Technologies and health Care Systems," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, pp. 265-267, doi: 10.1109/I-SMAC49090.2020.9243529. R. Lokeshkumar, E. Maruthavani, and A. Bharathi , A new perspective for decision makers to improve efficiency in social business intelligence systems for sustainable development, International Journal of Environment and Sustainable Development 2018 17:4, 404-416.

[16] S. Balakrishnan, V. K. and M. S. S. Hameed, "An Embarking User Friendly Palmprint Biometric Recognition System with Topnotch Security," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1028-1032, doi: 10.1109/ICICCS51141.2021.9432230.

[17] V. Arulkumar A Survey on Multimedia Analytics in Security Systems of Cyber Physical Systems and IoT, 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC) pp 1556-1561, Nov 2021. 10.1109/ICOSEC51865.2021.9591754

[18] Ravi Kumar Poluru & Lokesh Kumar R (2021) An Improved Fruit Fly Optimization (IFFOA) based Cluster Head Selection Algorithm for Internet of Things, International Journal of Computers and Applications, 43:7, 623-631, DOI: 10.1080/1206212X.2019.1600831

[19] Selvan.C, S.R.Balasundaram, (2020), Data Analysis in Context Based Statistical Modeling in Predictive Analytics, IGI Global, Handbook of Research on Engineering, Business, and Healthcare Applications of Data Science and Analytics, pages 98-114.

[20] V. Arulkumar An Intelligent Face Detection by Corner Detection using Special Morphological Masking System and Fast Algorithm, 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC) pp 1556-1561, Nov 2021.

[21] N. C. Brintha, C. Preethi and J. T. Winowlin Jappes, "Exploring Malicious Webpages Using Machine Learning Concept," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1-5, doi: 10.1109/INCET51464.2021.9456222.

978-1-6654-9515-8/$31.00©2022IEEE