

Cairo University
Faculty of Science
Department of Mathematics



Machine Learning Algorithms to Predict Alzheimer's Disease

**Research submitted to Faculty of Science Department of Mathematics which is
part of the requirements of obtaining a bachelor's degree in computer science**

➤ **Presented by:**

Asmaa Abdelfattah Mohamed

Hayam Tarek Fathy

Shrouk Saeed Ahmed

Nada Shaban Sayed.

➤ **Supervisor:**

Dr. Sameh Basha

2022

Machine Learning Algorithms to Predict Alzheimer's Disease

Faculty of Science, Cairo University, Egypt

Abstract. Alzheimer's disease (AD) is Serious mental illness affecting the brain. the reason for dementia in the elderly people. Around 45 million people are suffering from this disease. Alzheimer's disease spreads rapidly over the years. This disease has a significant impact on the social, financial, and economic aspects. Early treatment of Alzheimer's disease is more effective and causes less brain damage. Since Alzheimer's is a chronic disease, we should have taken care of early detection using machine learning. This proposed model represents the analysis and the result regarding detecting Dementia from various machine learning models. Longitudinal Magnetic Resonance Imaging (MRI) data from OASIS has been used for the development of the system. Several techniques such as Decision Tree, Random Forest, Support Vector Machine, Logistic Regression and keras neural network have been employed to identify the best parameters for Alzheimer's disease prediction. we obtained 0.97, 0.1, 0.944 and 0.971 for the Accuracy, precision, Recall, F1_score respectively.

Keywords: Alzheimer, Machine Learning classification, neural network.

1 Introduction

In 1906, Dr. Alois Alzheimer noticed changes in the brain tissue of a woman who had died of an unusual mental illness. Her symptoms included memory loss, language problems, and unpredictable behavior. After she died, he examined her brain and found many abnormal clumps and tangled bundles of fibers. After that many organizations form around the world to raise funds for Alzheimer's research to find causes, cure and define Alzheimer. Alzheimer is a type of dementia which is a neurological disorder that occurs through proteins in the brain that do not perform their function normally affecting the work of brain cells releasing a series of toxic substances. But unfortunately, they could not be assured about it causes they determine some causes such as genetic factor, age, family history and synaptic injury. And they could not find cure for Alzheimer however, doctors prescribe medications to reduce symptoms such as anxiety and depression.

Alzheimer is currently ranked as the seventh leading cause of death in the United States and is the most common cause of dementia among older adults.[1] But early detection of the disease helps to reduce symptoms and not reach advanced stages, and here comes the role of machine learning which defined as the study of computer programs that leverage algorithms and statistical models to learn through inference and patterns without being explicitly programmed, the model's algorithm uses a collection of data for training and builds a way to predict the output and saves that procedure for future purposes. Model such as Support Vector Machine, Logistic Regression, Decision Tree and Random Forest a good application to detect whether the patient is demented by Alzheimer or not with use of balanced data such as OASIS data set. [2]

The remainder of the paper is organized as follows:

Section 2 discusses Related Work, Section 3 Description of dataset, Section 4 methods and methodology, Section 4 Proposed Model and Finally Section 5, the conclusion of the presented work is provided.

• **Related Work**

This section discusses previously authored works in the same domain for Alzheimer Classification using Machine learning.

At VIT university 2022 team works on Alzheimer detection and they define Alzheimer as diseases is progressive neurological condition that leads to short term memory loss which lasts for years or the rest of life Detection by applying on data some algorithms Such as SVM, Random Forest classifier, decision tree, voting classifier and XGBoost and they has 85.91 accuracy using XGBoost. They also display another results for detecting Alzheimer using ML and deep learning.

And they conclude causes of Alzheimer as

1. It is more likely for men to have demented, or Alzheimer's Disease, than for women.
2. In terms of years of education, demented patients were less educated.

3. Brain volume in non-demented groups is greater than in demented groups.

4. Among the demented group there is a higher concentration of 70-80-year-olds than in the non-demented patients. [2]

Model	Accuracy	Precision	Recall	F1-score
Decision tree classifier	80.46%	0.80	0.79	0.78
Random forest classifier	86.92%	0.85	0.81	0.80
Support vector machine	81.67%	0.77	0.70	0.79
XGBoost	85.92%	0.85	0.83	0.85
Voting classifier	85.12%	0.83	0.83	0.85

At Bennett University 2021 team works on comparative analysis of Machine Learning Algorithms to Predict Alzheimer's Disease they could define Alzheimer such as an inherited, irreversible brain condition that steadily affects the ability to perform the necessary things, memory, and reasoning skills. they also work on oasis dataset and explain how to preprocessing the dataset by converting categorical data to numerical zero and one, and missing values filled by taking the mean of the feature and finally split data to train and test but this creates an imbalance between training and testing split. So, stratified sampling has been applied with a training-validation size of 80% and a testing size of 20%. After that, standardization has been applied to do the scaling of the features

support vector machine (SVM)which provides the best results among the models, logistic regression, decision tree, and random forest have been used for prediction. [3]

Model	Accuracy (%)	Recall (%)	Precision (%)	AUC (%)	F1 score
SVM	92.0	91.9	91.9	91.9	91.9%
Logistic regression	74.7	70.3	76.5	74.6	73%.3
Decision tree	80.0	59.4	100	79.7	74.5%
Random forest	81.3	70.3	84.4	81.2	76.7%

At Suez Canal University 2022 team works on Machine Learning Framework for the Prediction of Alzheimer’s Disease Using Gene Expression Data Based on Efficient Gene Selection thy intend to fill the gap of data shortage by introducing a symmetric framework to predict AD from GE data, with the aim to produce the most accurate prediction using the smallest number of genes using Integration of Datasets ,Preprocessing ,Gene Selection (GS) such as Chi squared (χ^2) , Analysis of variance (ANOVA) , Mutual Information (MI) and Classification using following algorithms SVM , RF , LR . And their result found that The best model is the one that achieves the highest performance with the smallest number of genes. In our experiments on an integrated dataset of 39,280 genes, this model has turned out to be SVM. It reached an accuracy of 97.5% using the 1058 hand-crafted genes, obtained by intersections and unions of genes with high filter values. [4]

Validation results of the SVM classifier

Metric	Value
Precision	0.980
Sensitivity (Recall)	0.970
Specificity	0.970
Kappa	0.945
Acc	0.975
AUC	0.972

2 Description of the dataset

The main goal is to predict Alzheimer disease on different patients based on different attributes. The longitudinal Magnetic Resonance Imaging (MRI) data from OASIS [5] has been used for the prediction. OASIS dataset has a dimension of 373 rows x 15 columns. Table 1 show eight different attribute which used in the proposed model.

Table1: Dataset description of proposed machine learning system

attribute	Description
M/F	Gender
Age	Person's age
EDUC	Years of education
SES	Socioeconomic status
MMSE	Mini-mental state examination
eTIV	Estimated total intracranial volume
nWBV	Normalized whole brain volume
ASF	Atlas scaling factor

3 Theories and Method

3.1 Logistic regression

is a supervised learning algorithm used in machine learning to predict the probability of a binary outcome. A binary outcome is limited to one of two possible outcomes. Examples include yes/no, 0/1 and true/false. Figure 1 shows the flow diagram of the whole logistic regression model.

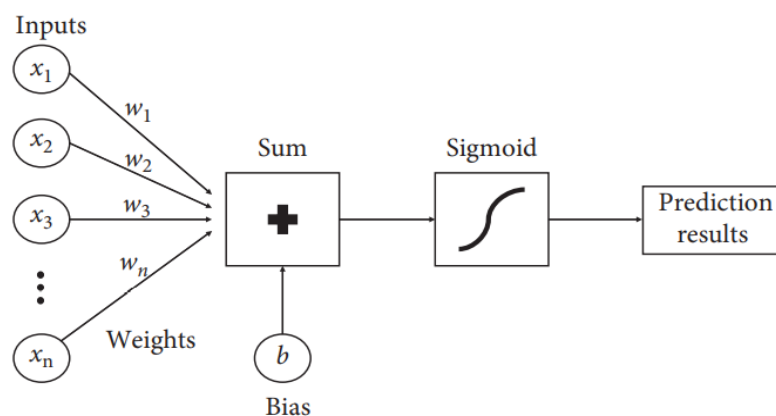


Figure 1:Logistic regression

3.2 Random Forest

a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Figure 4 shows the flow diagram of the whole random forest model.

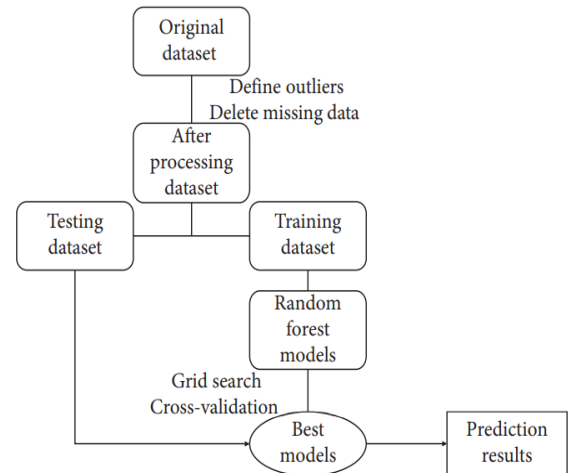


Figure 2:Random Forest

3.3 Support vector machine

(SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. we perform classification by finding the hyper-plane that differentiates the two classes very well. Figure 3 shows the flow diagram of the whole SVM model

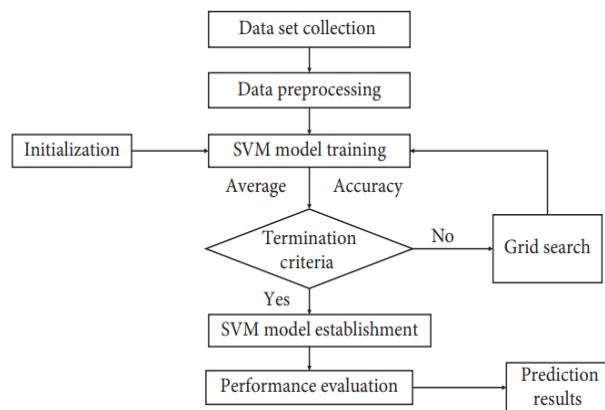


Figure 3: SVM

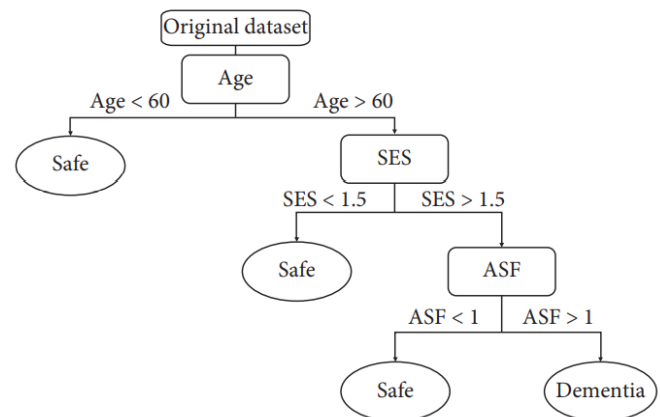


Figure 4:Decision tree

3.4 Decision Trees

are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. Figure 4 shows the flow diagram of the whole decision tree model.

3.5 Confusion matrix

performance measurement technique for Machine learning classification. It is a kind of table which helps you to know the performance of the classification model on a set of test data for that the true values are known. The term confusion matrix itself is very simple.

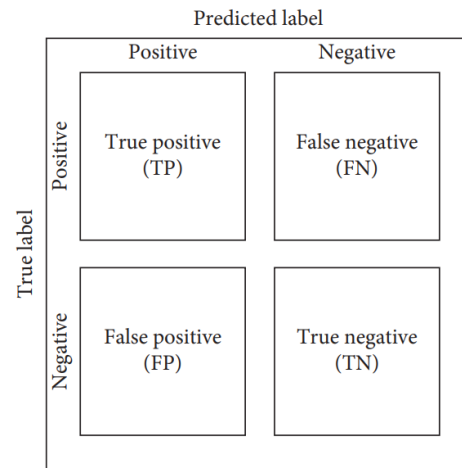


Diagram of confusion matrix.

4 Proposed Model

Figure 6 show the flow chart of the machine learning proposed model.

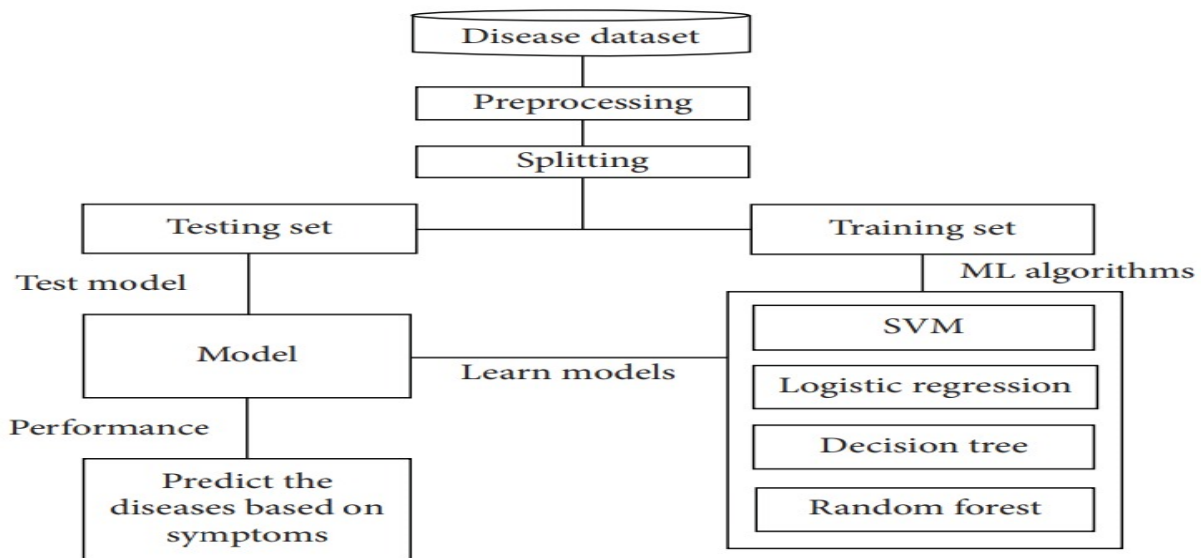



Figure 5:Proposed model

4.1 Data Preprocessing

4.1.1 converting categorical attribute columns like group and gender to numerical values 0 and 1. Figure 7 show the result of conversion.



```


<class 'pandas.core.frame.DataFrame'>
RangeIndex: 373 entries, 0 to 372
Data columns (total 15 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Subject ID   373 non-null    object
1   MRI ID       373 non-null    object
2   Group        373 non-null    int64
3   Visit        373 non-null    int64
4   MR Delay     373 non-null    int64
5   Gender       373 non-null    object
6   Age          373 non-null    int64
7   EDUC         373 non-null    int64
8   SES          354 non-null    float64
9   MMSE         371 non-null    float64
10  CDR           373 non-null    float64
11  eTIV          373 non-null    int64
12  nWBV         373 non-null    float64
13  ASF          373 non-null    float64
dtypes: float64(5), int64(7), object(3)
memory usage: 43.8+ KB

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 373 entries, 0 to 372
Data columns (total 15 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Subject ID   373 non-null    object
1   MRI ID       373 non-null    object
2   Group        373 non-null    int64
3   Visit        373 non-null    int64
4   MR Delay     373 non-null    int64
5   Gender       373 non-null    object
6   Age          373 non-null    int64
7   EDUC         373 non-null    int64
8   SES          354 non-null    float64
9   MMSE         371 non-null    float64
10  CDR           373 non-null    float64
11  eTIV          373 non-null    int64
12  nWBV         373 non-null    float64
13  nWBV         373 non-null    float64
14  ASF          373 non-null    float64
dtypes: float64(5), int64(7), object(3)
memory usage: 43.8+ KB

```

Figure 6:converting of grouping attributes to numerical values

4.1.2 checking for the missing values and fill these values by median value. Figure 8 show the result of filling.



```

Subject ID      0
MRI ID          0
Group           0
Visit           0
MR Delay        0
Gender          0
Hand            0
Age             0
EDUC            0
SES             0
MMSE            0
CDR             0
eTIV            0
nWBV            0
ASF             0
dtype: int64

Subject ID      0
MRI ID          0
Group           0
Visit           0
MR Delay        0
Gender          0
Hand            0
Age             0
EDUC            0
SES             19
MMSE            2
CDR             0
eTIV            0
nWBV            0
ASF             0
dtype: int64

```

Figure 7:filling missing values with median value

4.1.3 drop unnecessary columns like subject ID, MRI ID, hand columns because they do not effect on the output (The Target Value “Group”). Figure 9 show the final dataset.

	Group	Visit	Gender	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	0	1	1	87	14	2.0	27.0	0.0	1987	0.696	0.883
1	0	2	1	88	14	2.0	30.0	0.0	2004	0.681	0.876
2	1	1	1	75	12	2.0	23.0	0.5	1678	0.736	1.046
3	1	2	1	76	12	2.0	28.0	0.5	1738	0.713	1.010
4	1	3	1	80	12	2.0	22.0	0.5	1698	0.701	1.034
...

4.2 Data splitting

stratified sampling has been applied with a training-validation size of 80% and a testing size of 20%. To evaluate the performance of the machine using unknown data.

5 Results and Analysis

5.1 models

5.1.1 Logistic regression: Figure 9 show the model prediction's prediction accuracy, test precision, test recall, f1 score and the confusion matrix, the number of correct predictions is 71 and the wrong predictions is 4.

Accuracy: 94.67%
Test precision: 97.06%
Test recall: 91.67%
Test F1: 94.29%

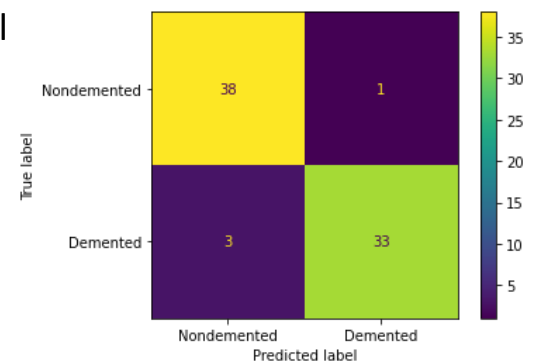


Figure 8: result of logistic regression

5.1.2 SVM: Figure 10 show the model prediction's accuracy, test precision, test recall, f1 score and the confusion matrix, the number of correct predictions is 72 and the wrong predictions is 3.

Accuracy: 96.00%
Test precision: 100.00%
Test recall: 91.67%
Test F1: 95.65%

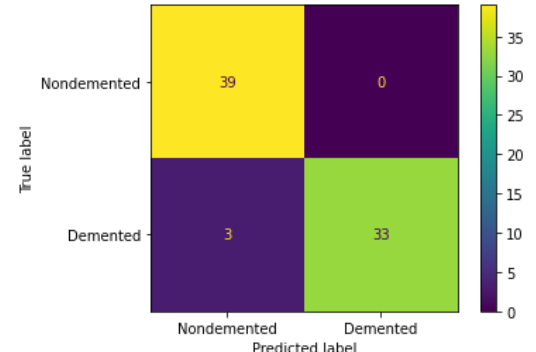


Figure 10: result of SVM

5.1.3 Decision tree: Figure 11 show the model prediction's accuracy, test precision, test recall, f1 score and the confusion matrix, the number of correct predictions is 72 and the wrong predictions is 3.

Accuracy: 96.00%
Test precision: 97.14%
Test recall: 94.44%
Test F1: 95.77%

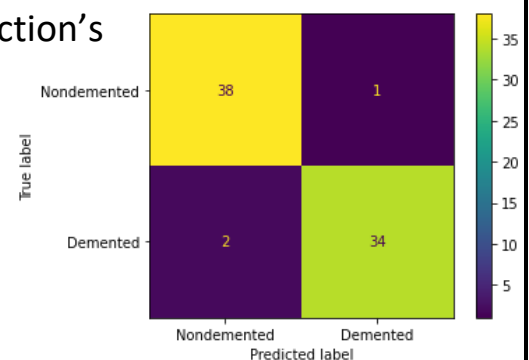


Figure 11: result of DS

5.1.4 Random Forest: Figure 12 show the model prediction's accuracy, test precision, test recall, f1 score and the confusion matrix, the number of correct predictions is 73 and the wrong predictions is 2.

Accuracy: 97.33%
 Test precision: 100.00%
 Test recall: 94.44%
 Test F1: 97.14%

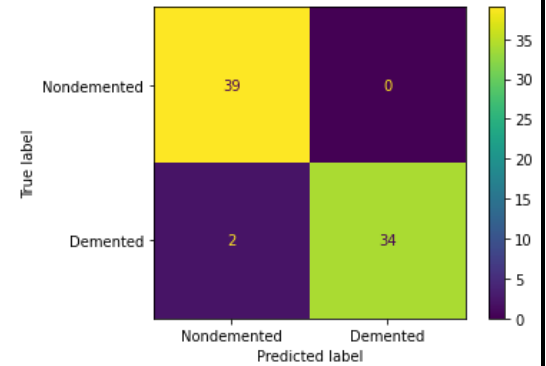


Figure 12: result of random forest

- Accuracy can be defined as the ratio of the number of correctly classified cases to the total of cases under evaluation. The best value of accuracy is 1 and the worst value is 0.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{\text{Number of correctly classified cases}}{\text{Total number of cases under evaluation}}$$

- Precision of positive class is intuitively the ability of the classifier not to label as negative a sample that is positive. The best value of precision is 1 and the worst value is 0.

$$Precision \text{ (positive class)} = \frac{TP}{TP+FP} = \frac{\text{True Positive}}{\text{Number of cases predicted as positive}}$$

- Recall of positive class is also termed sensitivity and is defined as the ratio of the True Positive to the number of actual positive cases.

$$Sensitivity \text{ (or Recall of positive class)} = \frac{TP}{TP+FN} = \frac{\text{True Positive}}{\text{Number of actual positive cases}}$$

- F1-score is the weighted average of recall and precision of the respective class. Its best value is 1 and the worst value is 0.

$$F1 - score = \frac{TN}{TN+FP} = \frac{2 * Precision * Recall}{Precision + Recall}$$

5.2 cross validation

Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. It is mainly used in settings where the goal is prediction. The goal of cross-validation is to test

the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting.

- **K-Fold Cross-Validation**

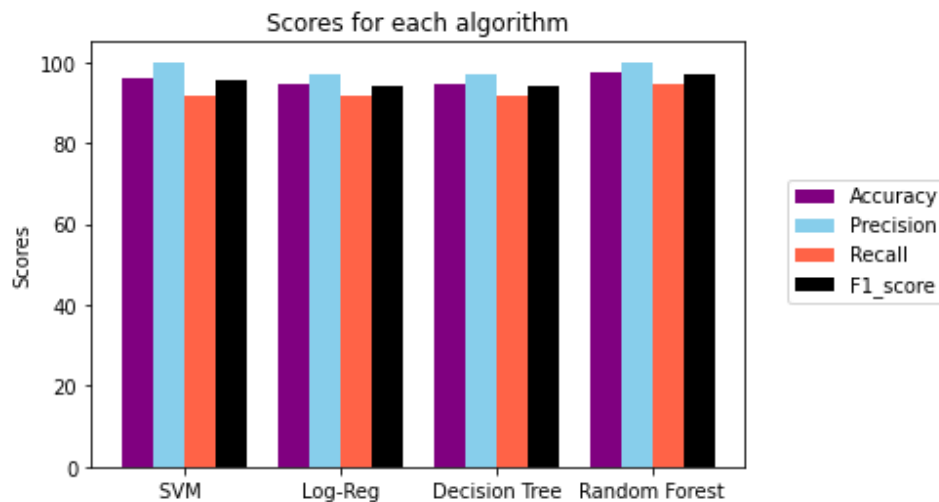
The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

- **The stratified k fold cross-validation**

Stratified k-fold cross-validation is the same as just k-fold cross-validation, But Stratified k-fold cross-validation, it does stratified sampling instead of random sampling. 5-fold cross-validation has been applied to evaluate all possible combinations, The following results were obtained for each model.

SVM-Model	Log-Reg	Decision Tree	Random Forest
Accuracy:0.944	Accuracy:0.941	Accuracy:0.898	Accuracy:0.941

5.3 Model Comparison

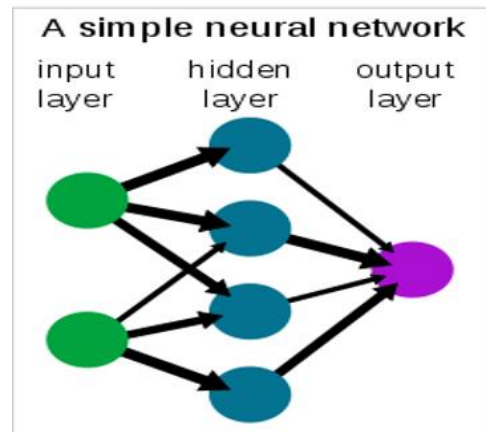


Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score
SVM	96	100	91.67	95.65
log-reg	94.67	97.06	91.67	94.29
decision tree	96	97.14	94.44	95.77
Random forest	97.33	100	94.44	97.14

6 Neural Network

A neural network is a series of algorithms that seeking to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.

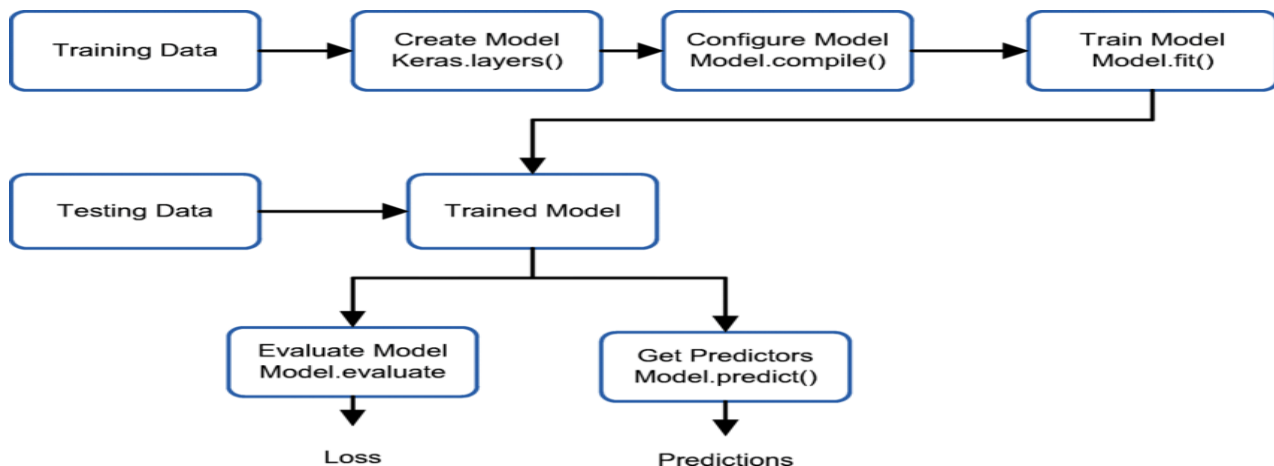
As application of neural network in Alzheimer predication we used Keras which is a high-level, deep learning API developed by Google for implementing neural networks. It is written in Python and is used to make the implementation of neural networks easy, we used two model sequential model and Functional model to predict.



Keras proposed model

After stage of preprocessing to OASIS dataset and splitting data

We follow proposed model for keras library



Compilation stage

Data goes to input layer and then made on it some mathematical calculations in hidden layer we used dense layer it is a type of fully connected layer receives output from each neuron in hidden layer and made some functions on it like activation function and optimization.

Activation function is used to transform mathematical calculations in hidden layer to output or activation the node, we used rectified linear activation function(relu)

which is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero.

Model could made false prediction out of these calculations so we have to use loss function to reduce the difference between real results and the predicted valued in classification problem which we had it best to use binary cross-entropy.

At the end we apply Adam optimization function which is algorithm reduce loss and improve accuracy of the model. Next, we test model accuracy and loss in the following we will define used models

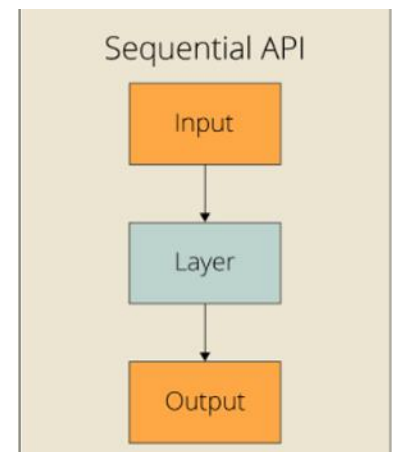
and their test results.

Sequential Model

The sequential model is a linear stack of layers, allows you to create models layer-by-layer for most problems. It is limited in that it does not allow you to create models that share layers or have multiple inputs or outputs.

Results

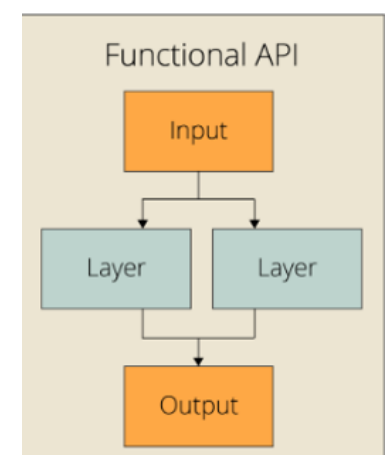
```
3/3 [=====] - 0s 3ms/step - loss: 0.4473 - accuracy: 0.9600  
Test loss: 0.4473148584365845  
Test accuracy: 0.9599999785423279
```



Functional Model

The Keras functional API is a way to create models that are more flexible than the Sequential API it can handle models with non-linear topology, shared layers, and even multiple inputs or outputs.

```
3/3 [=====] - 0s 4ms/step - loss: 0.3525 - accuracy: 0.9600  
Test loss: 0.35251584649086  
Test accuracy: 0.9599999785423279
```



7 CONCLUSION

AD is a serious disease which need a lot of attention, research and development so that we can predict it early and have a chance to reduce the symptoms.

In this paper we have presented proposed model for predicting AD by Machine Learning. the Open Access Series of Imaging Studies (OASIS) dataset, which is resulting from the analysis of a number of MRI scans, is used in our proposed model.

First, we converted categorical data to numerical data in columns. we checked for null and missing values. the median values are used to fill in those missing values and dropped unnecessary columns. the values were standardized to make sure that the data easily fit in the Machine learning models. The dataset was used in training and testing SVM, Random Forest, Decision Tree, Logistic Regression and Keras Neural Network. these models used for evaluating Accuracy, Precision, Recall, F1_score and confusion matrix . Cross-Validation was used to improve the system results.

The system got the best result by Random Forest.

Future Work :

The proposed model can be improved by using larger dataset and more Machine learning models.

References

[1] the NIH National Institute on Aging (NIA). NIA scientists and other experts review this content to ensure it is accurate and up to date. July 08, 2021

Alzheimer diseases fact sheet

<https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet#symptoms>

[2] C. Kavitha, Vinodhini Mani, S. R. Srividhya¹, Osamah Ibrahim Khalaf and Carlos Andrés Tavera Romero. January 12, 2022

Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models

<https://www.frontiersin.org/articles/10.3389/fpubh.2022.853294/full>

[3] Morshedul Bari Antor , A. H. M. Shafayet Jamil , Maliha Mamtaz , Mohammad Monirujjaman Khan , Sultan Aljahdali , Manjit Kaur , Parminder Singh , and Mehedi Masud . March 17, 2021

A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease

<https://www.hindawi.com/journals/jhe/2021/9917919/>

[4] Aliaa El-Gawady, Mohamed A. Makhlouf, BenBella S. Tawfik and Hamed Nassar. January 24, 2022

Machine Learning Framework for the Prediction of Alzheimer's Disease Using Gene Expression Data Based on Efficient Gene Selection

<https://www.mdpi.com/2073-8994/14/3/491>

[5] "The OASIS brains datasets," 2016, https://www.kaggle.com/jboisen/mri-and-alzheimers?select=oasis_longitudinal.csv