International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

# Prediction Of Diabetes During Covid – 19 Using Cloud Based App

*Dr Sakthivel V[1], Jayavignesh P[2], Akshaya Rohith KB[3], Avinash P[4]*

*[1]Assistant Professor(Sr.G), School Of Cse, VIT- Chennai Campus,Chennai, India, [2]Student, School Of Cse, VIT- Chennai Campus,Chennai, India, [3]Student, School Of Cse, VIT- Chennai Campus,Chennai, India, [4]Student, School Of Cse, VIT- Chennai Campus,Chennai, India*

*Email: [1]sakthivel.V@Vit.Ac.In, [2]jayavignesh.P2019@Vitstudent.Ac.In, [3]akshayarohith2019@Vitstudent.Ac.In, [4]avinash.A2019@Vitstudent.Ac.In*

***Abstract: Covid – 19 Is A Dreadful Disease Which Cost Us A Lot Of Lives. It Is Similar To The Flu Which Attacked Us In 1914.We Suffered A Very Huge Loss Of Human Population. But What Is The Solution For This? Should We Be Afraid Of This Disease Our Entire Lives Or Should We Face This Disease And Live With It Until A Proper Cure Is Found? , Is COVID – 19 The Only Disease That Affects Humanity? No. There Are Lots Of Diseases That Still Affect Human Lifestyle Apart From COVID-19. People Are Still Not Ready To Go To Hospitals For Their Regular Checkups Example Diabetes Patients. Because They Think They Could Get Affected By This Dreadful Disease If They Go To Hospitals, Well They Are Not Wrong. What If They Can Test Their Diseases Online, What If They Don't Need To Go To Hospitals For Their Checkups But Instead Stay At Their Homes And Take A Test At Their Comfort. So This Project Aims At Creating A Web App Which Will Be Useful For Patients To Test Any Kind Of Disease Like Diabetes, Flu Etc. Using Powerful And Persistent Datasets Available At Kaggle And With Use Of Machine Learning And Python To Train A Model Deploy It Using A Cloud Based App Development Tool With The Use Of HTML,CSS***

***Keywords: Diabetes, Machine Learning, COVID-19.***

## 1. INTRODUCTION

The Main Use Of This Project Is That Patients Can Check Their Diseases At Their Home[1]. The Other Main Reason Is That We May Be In Need Of More Doctors And Hospitals To Treat And Cure COVID Patients[2]. We Can't Waste The Knowledge Of Our Doctors In These Diseases Which Are Comparatively Less Effective Than COVID-19[3]. As A Startup This Project Aims On Predicting Diabetes As The Main Disease Using The Famous "PIMA Dataset" Available In UCI Repository[4]. Which Consists Of Over 769 Experiences Of Patients Suffered From Diabetes And Their Symptoms[5]. The Classifiers That We Use Here Are Random Forest, KNN, Linear Regression. We Will Be Using Method Selection To Select The Best Suitable Classifier To Train And Test The Model[6]. The Missing Values Will Be Filled With Mean Or Median Of The Specific Column So That We

Don’t Affect The Continuity Of The Data Values[7]. Graphical Representation Between Various Columns And A Confusion Matrix Will Be Provided In Order To Select The Necessary Attributes And Ignore The Ones With Low. Forms Will Be Designed For As The Front End To Get The Inputs From The User. Initially The App Shall Be Tested In Local Server And Then HEROKU A Cloud Based Tool Will Be Used To Deploy The Web App On The Global Server.

### *1 Literature Review*

Karmul Hasan Et Al Uses Ensembling Methodology To Predict Diabetes. The Classifiers That Are Ensembled Are LR, RF, NB, KNN And Decision Tree. The Advantage Of Ensembling Technique Combines The Power Of All The Above-Mentioned Classifiers, And Reduces The Complexity. The Issues In These Methods Are Different Ensembling Methods Must Be Used With Different Datasets. Due To This The Accuracy Rate Will Be Severely Damaged. No UI Is Included Therefore Cannot Be Used By Normal People To Predict. Quin Wang Et Al Uses NB To Counter The Missing Values. ADASYN To Counter Imbalance In The Classification Of The Dataset. RF To Train And Test The Dataset. The Advantages Of This Method Is One Of The Best And Popular Method For Data Prediction And Gives The Utmost Accuracy For Most Of The Datasets. Using NB For Missing Values Increases The Accuracy Rate And The Prediction Rate. The Issues Related To RF May Be Suitable To Many Cases, But In Case Of Dataset Having A Large Amount Of Data, We Must Use Cross Validation For Different Classifiers And Select The Best Model. ADASYN When Used In A Dataset With Average Number Of Data's Increases The BIAS. Dhiraj Dahiwade Et Al Uses Ensembling Mechanism. The Classifiers That Are Ensembled Are KNN And CNN. The Advantages Of CNN Is Used For More Complex Diseases Like Heart Diseases Cancer Etc. The Issues In These Method Is In General Diseases CNN May Not Be Powerful, At Least Not More Than RF. Use Of CNN May Over Fit The Data. Ensembling Mechanism Is Only Powerful When We Combine More Than At Least Three Classifiers. We Can Use Cross Validation Instead Of Ensembling. Mrunmayi Patil Et Al Uses SVM For General Disease Prediction. The Advantage Of SVM Works Relatively Well When There Is Clear Margin Of Separation Between Classes. SVM Is Relatively Memory Efficient. The Issues In SVM Algorithm Is Not Suitable For Large Data Sets, Such As PIMA Dataset. For The Implementation Of SVM There Must Be A Clear Boundary Between The Classes, There Must Not Be Overlapping Of Classes. In Our Case There Will Be Overlapping Since The Values Of Certain Attributes Range From 1-4 Etc. It Cannot Be Effectively Used In A Large Dataset. Rohit Binu Mathew Et Al Developed An App Which Will Be Easily Handled By The Users. This Model Uses KNN As The Classifier. The Advantage Of This App Also Provides The Recommended Medicine For That Particular Disease. A User-Friendly Environment For The Users. The Issues Are Using KNN Only Will Not Give The Best Prediction Rate, There Are Algorithms Which Are Far Superior Than KNN Such As RF. But All Of Them Will Have Their Own Pros And Cons. Therefore, Comparing And Selecting The Suitable Classifier Will Be The Best Method To Do That.

## 2 ABOUT THE DATASET

The Dataset Used Here Is The PIMA Indian Dataset For The Prediction Of Diabetes Available In Kaggle For The ML Analysts To Work With The Data. This Dataset Is Originally From The National Institute Of Diabetes And Digestive And Kidney Diseases.

The Objective Of The Dataset Is To Diagnostically Predict Whether Or Not A Patient Has Diabetes, Based On Certain Diagnostic Measurements Included In The Dataset. Several Constraints Were Placed On The Selection Of These Instances From A Larger Database. The Dataset Consists Of 9 Columns They Include Number Of Pregnancies In Case Of Female Patients, Blood Sugar Level, Insulin, BMI, Age Etc. The PIMA Indians Dataset In Shown In Fig 1 Below.

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 5 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 6 | 3 | 78 | 50 | 32 | 88 | 31.0 | 0.248 | 26 | 1 |
| 7 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 8 | 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 9 | 8 | 125 | 96 | 0 | 0 | 0.0 | 0.232 | 54 | 1 |
| 10 | 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | 0 |
| 11 | 10 | 168 | 74 | 0 | 0 | 38.0 | 0.537 | 34 | 1 |
| 12 | 10 | 139 | 80 | 0 | 0 | 27.1 | 1.441 | 57 | 0 |
| 13 | 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 | 59 | 1 |
| 14 | 5 | 166 | 72 | 19 | 175 | 25.8 | 0.587 | 51 | 1 |
| 15 | 7 | 100 | 0 | 0 | 0 | 30.0 | 0.484 | 32 | 1 |
| 16 | 0 | 118 | 84 | 47 | 230 | 45.8 | 0.551 | 31 | 1 |
| 17 | 7 | 107 | 74 | 0 | 0 | 29.6 | 0.254 | 31 | 1 |
| 18 | 1 | 103 | 30 | 38 | 83 | 43.3 | 0.183 | 33 | 0 |
| 19 | 1 | 115 | 70 | 30 | 96 | 34.6 | 0.529 | 32 | 1 |

Fig 1 – Pima Indians Diabetes Dataset

### *2.1 Data Normalization*

As Said Before The Missing Values In Each Columns Are Filled With The Mean Of That Particular Column Or Attribute[8]. Figure 2 Below Shows The Number Of Missing Values In The Entire Dataset.

```
Pregnancies                 111
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
dtype: int64
```

Fig 2 - Number Of Missing Values

Figure 3 Below Shows The Percentage Of The Values Missing In Each Attribute Column.

```
Pregnancies                 14.453125
Glucose                      0.651042
BloodPressure                4.557292
SkinThickness               29.557292
Insulin                     48.697917
BMI                          1.432292
DiabetesPedigreeFunction     0.000000
Age                          0.000000
dtype: float64
```

Fig-3 Percentage Of Missing Values

From Figure 3 We Can See That Mostly 50% Of The Patients Are Not Tested With Insulin Levels From In The Dataset[9]. This Maybe Because Of Two Reasons Maybe The Doctors Only Took Tests For Those Patients Who Are Not Looking Well And Ignored The Other Patients Or They Would Have Taken A Prior Diagnosis And Then Decided To Take Tests On Patients Based On The Results[10]. However, This May Cause A Major Problem During Prediction Phase Because The Model Would Have Only Be Trained For 50% Of The Patient's Insulin Experiences[11][14]. This Problem Can Be Solved By Just Ignoring That Attribute, But Ignoring That Attribute As Well Can Also Cause Problems So We Are In A State To Select The Attributes Which Will Give Us The Most Accurate Prediction Result. This Method Of Selecting Attributes Is Called As Attribute Selection.

### *2.2 Attribute Selection*

This Is The Phase Where We Select The Attributes That Will Give Us The Most Accurate Prediction Results Or In Other Words The Attribute Which Is More Correlated Towards The Output Attribute[12]. To Know That We Will Be In Need Of Graphical Representation Of Insulin Vs. Output Columns In The Dataset. Figure 4 Below Shows The Graph Plotted Between Insulin And Outcome Column.

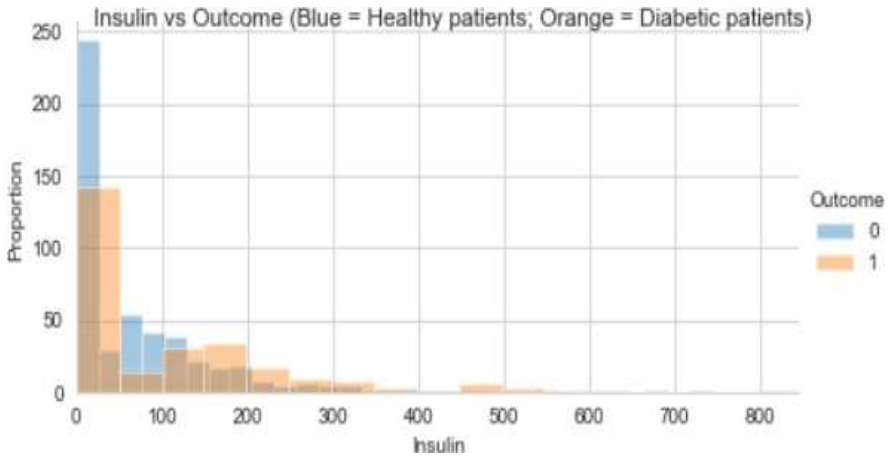


Fig 4- Insulin Vs Outcome

From The Above Graph We Come To Know That Presence Of Insulin Attribute In The Dataset Will Bias The Result Towards The Outcome 0(Healthy Patients)[13]. Due To This There Will Be A Great Reduction In The Prediction Results And The Accuracy Of The Model Will Be Reduced. So We Must Be Sure That No Other Column Has The Same Properties. To Ensure This We Use A Special Type Of Graphical Representation Called Heat-Maps Which Plots The Graph Based On The Correlation Co Efficient Of Each Attribute In The Dataset. Figure 5 Below Represents The Heat-Map Diagram For The Dataset Used.

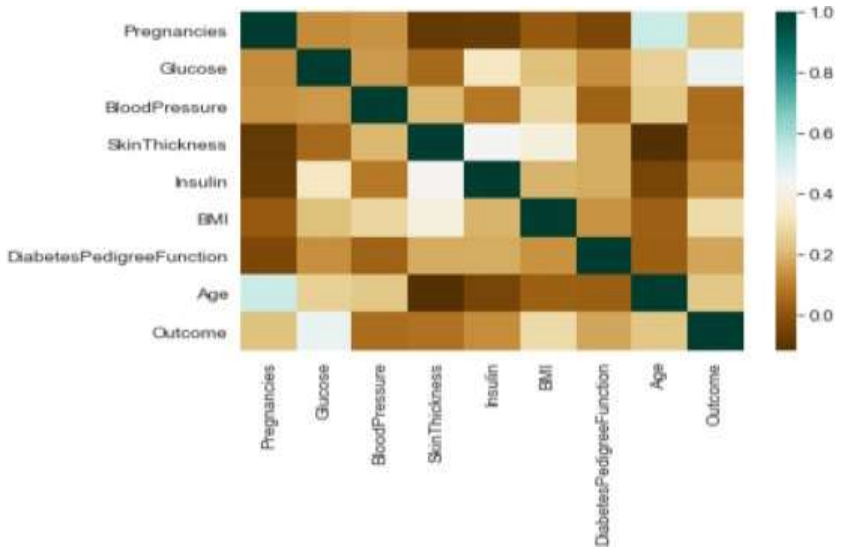


Fig-5 Heat-Map Diagram

From The Above Heat-Map We Realize That Not Only Insulin But Also Skin Thickness Is Less Correlated Towards The Outcome So We Don't Need Those Attributes In Our Dataset, We Simply Remove Them From Out Dataset. Figure 6 Below Shows The Dataset After Removing Insulin And Skin Thickness Columns.

| | Pregnancies | Glucose | BloodPressure | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|
| **0** | 6 | 148 | 72 | 33.6 | 0.627 | 50 | 1 |
| **1** | 1 | 85 | 66 | 26.6 | 0.351 | 31 | 0 |
| **2** | 8 | 183 | 64 | 23.3 | 0.672 | 32 | 1 |
| **3** | 1 | 89 | 66 | 28.1 | 0.167 | 21 | 0 |
| **4** | 0 | 137 | 40 | 43.1 | 2.288 | 33 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **763** | 10 | 101 | 76 | 32.9 | 0.171 | 63 | 0 |
| **764** | 2 | 122 | 70 | 36.8 | 0.340 | 27 | 0 |
| **765** | 5 | 121 | 72 | 26.2 | 0.245 | 30 | 0 |
| **766** | 1 | 126 | 60 | 30.1 | 0.349 | 47 | 1 |
| **767** | 1 | 93 | 70 | 30.4 | 0.315 | 23 | 0 |

768 rows × 7 columns

Fig-6 Dataset After Removing Insulin And Skin Thickness

### *2.3 Selecting Classifiers*

Classifiers Are Methods Or Models Used To Train And Test Our Data In The Dataset. Selecting The Best Classifier Suitable For A Problem Is Very Important Because The Accuracy And Prediction Results Depend Upon The Training And The Testing Given On The Dataset. Here We Used Cross Validation Method In Order To Select The Best Classifier For Our Problem.

### *2.4 Cross-Validation*

Cross Validation Is A Process Of Dividing The Dataset In To K-Folds, At Each Time Consider One Fold As The Testing Set And Others As Training Set. Implement A Classifier Method On The Training Set And Test The Resulting Model Using The Kth Testing Set. The Advantage In Using Cross Validation Is That Each And Every K-Fold In The Dataset Will Get A Chance To Be In Both Training And Testing Datasets. Generally K Value Is Set At 10 Which Is Experimented By Data Scientists To Be The Moderate Value In Which The Bias Of The Model Is Less And Variance Is Moderate. Figure 7 Below Shows The Cross Validation Score Of Linear Regression, Random Forest And KNN. These Scores Signify That Random Forest Will Be The Suitable Method To Proceed.

```
K-Fold Cross-Validation Accuracy:

LR: 0.751403
RF: 0.769376
KNN: 0.711687
```

Fig-7 Cross Validation Scores

## 3. MODULES

*a. Web Application*

After Training And Testing The Models We Have To Create A Web App Which Will Integrate All The Ipynb Files And The HTML Files And Creates A Web App Which First Will Be Tested In Local Server And Later Be Deployed In HEROKU. First A Folder Must Be Created And The Model That We Have Trained Will Be Loaded Into A Sub Folder Called Model. Names Are Given To The Folders For The HEROKU Server To Easily Recognize Them And Extract Or Compile The Programs Included Within Them. An App.Py Python File Must Be Created In Order To Call The Model Whenever An User Interacts With The HTML Web Page, So That The Model Can Test And Predict The Result Form It's Already Trained Model. The HTML And CSS Files Must Be Loaded Into A Sub Folder Named Templates. The Pickle Library From Python Is Used To Load The ML Model Into The Web Page. A File Named Requirements Is Created And The Required Packages That Must Be Installed By The Python IDE Are Listed There. A File Called Procfile Is Created And The Web Processes That Are To Be Performed During Startup Are Listed There. The Frame Work For The Web App Looks Like The Figure 8 Shown Below.

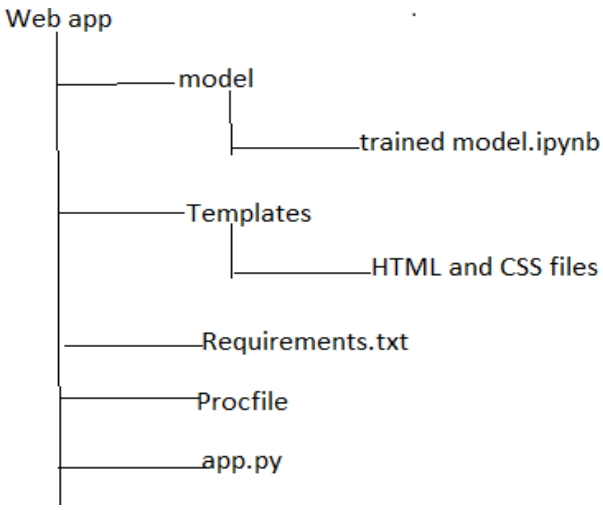


Fig-8 Framework Web App

*3.2 Local Server*

The Above Framework Is Converted Into A Running App By Using Anaconda Server. By Creating A Virtual Environment In Anaconda Prompt, Changing Our Directory Into The One Containing The Above Mentioned Files Using Flask Run Command Will Create An Instantly Running Web App Which Runs In Our Local Host Server.

*b. Heroku Global Server*

The Same Procedure Can Be Carried Out To Create A Web App On A Global Server Using Some Of The Built In Functions Of HEROKU And Github Terminal. Using Git Add And Git Commit Command To Add All The Files Mentioned In The Git Repository

Git Push Heroku Master Command Is Used To Push All The Files In The GIT Repository To The HEROKU'S Cloud Server, It Then Compiles The Files And Folders One By One. A Simple Command Heroku Open Will Launch The Web App On The Global Server. The Input Page Of The Web App Is Given In Figure 9.


Fig-9 Input Page For The Web App

## 4. CONCLUSION

In This Project We Proposed A Web Application To Predict Diabetes In Patients Which Uses The PIMA Indians Diabetes Dataset. Before Creating The Web App All The Three Classifiers Like KNN, Random Forest, And Linear Regression Are Tested Using Cross Validation. Before That All The Missing And Nan Values Are Taken Care Of By Filling Them Up By The Means Of The Respective Attribute Columns And Splitting The Dataset Into Training And Testing Datasets, The Attribute Selection Was Carried Out Different Graphical Expressions Like Linear Graphs And Heat-Maps Showed That The Attributes Insulin And Skin Thickness Were Less Correlated To The Outcome So They Were Dropped From The Dataset And The Model Was Trained And Tested. Thus The Project's Idea Saves A Lot Of Time For The Doctors To Concentrate On COVID Tested Patients And Other Severely Affected Patients. It Also Saves The Cost, Time And Risk Of Patients Going To Hospitals Especially In Periods Of COVID-19.

## 5. FUTURE WORK

The Web App Can Be Upgraded To Also Predict Many Other Dreadful Diseases By Collecting The Datasets Required For The Prediction Of The Particular Diseases, It Can Save A Lot Of Doctor's Work. The App Can Also Include An Automatic Prescription Providing Mechanism Which Gives The Medicines Required For The Disease Along With The Patient's Results.

## 6. REFERENCES

[1] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes Prediction Using Ensembling Of Different Machine Learning Classifiers. IEEE Access, 1–1. Doi:10.1109/Access.2020.2989857

[2] Mathew, R. B., Varghese, S., Joy, S. E., & Alex, S. S. (2019). Chatbot For Disease Prediction And Treatment Recommendation Using Machine Learning. 2019 3rd International Conference On Trends In Electronics And Informatics (ICOEI). Doi:10.1109/Icoei.2019.8862707

[3] Patil, M., Lobo, V. B., Puranik, P., Pawaskar, A., Pai, A., & Mishra, R. (2018). A Proposed Model For Lifestyle Disease Prediction Using Support Vector Machine. 2018 9th International Conference On Computing, Communication And Networking Technologies (ICCCNT). Doi:10.1109/Icccnt.2018.8493897

[4] Dahiwade, D., Patle, G., & Meshram, E. (2019). Designing Disease Prediction Model Using Machine Learning Approach. 2019 3rd International Conference On Computing Methodologies And Communication (ICCMC). Doi:10.1109/Iccmc.2019.8819782

[5] Wang, Q., Cao, W., Guo, J., Ren, J., Cheng, Y., & Davis, D. N. (2019). DMP_MI: An Effective Diabetes Mellitus Classification Algorithm On Imbalanced Data With Missing Values. IEEE Access, 1–1. Doi:10.1109/Access.2019.2929866

[6] Https://Blog.Cambridgespark.Com/Deploying-A-Machine-Learning-Model-To-The-Web-725688b851c7

[7] Https://Www.Coursera.Org/Lecture/Python-Machine-Learning/Model-Evaluation-Selection- BE2l9

[8] Https://Www.Kaggle.Com/Uciml/Pima-Indians-Diabetes-Database

[9] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease Prediction By Machine Learning Over Big Data From Healthcare Communities. IEEE Access, 5, 8869–8879. Doi:10.1109/Access.2017.2694446

[10] Dahiwade, D., Patle, G., & Meshram, E. (2019). Designing Disease Prediction Model Using Machine Learning Approach. 2019 3rd International Conference On Computing Methodologies And Communication (ICCMC). Doi:10.1109/Iccmc.2019.8819782

[11] Kohli, P. S., & Arora, S. (2018). Application Of Machine Learning In Disease Prediction. 2018 4th International Conference On Computing Communication And Automation (ICCCA). Doi:10.1109/Ccaa.2018.8777449

[12] Sreeja Mole, Sujatha Krishnamoorthy* (2019) An Efficient Gait Dynamics Classification Method For Neurodegenerative Diseases Using Brain Signals, Published In Journal Of Medical System, Springer

[13] K.Venkatachalam, A.Devipriya, J.Maniraj, M.Sivaram, A.Ambikapathy, Iraj S Amiri, "A Novel Method Of Motor Imagery Classification Using Eeg Signal", Journal Artificial

Intelligence In Medicine Elsevier, Volume 103, March 2020, 101787

[14] N. Bacanin, T. Bezdan, K. Venkatachalam, And F. Al-Turjman, "Optimized Convolutional Neural Network By Firefly Algorithm For Magnetic Resonance Image Classification Of Glioma Brain Tumor Grade," *Journal Of Real-Time Image Processing*, Pp. 1-14, 2021.