IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11231

Disease Prediction and Classification using Machine Learning Approach

Nikhil Giramkar[1], Shubham Rane[2], Abhishek More[3], Rupesh Bodkhe[4], Shraddha Khonde[5]

[1,2,3,4] UG – Computer Engineering, Modern Education Society's College of Engineering, Pune, Maharashtra

[5] Assistant Professor, Computer Engineering, Modern Education Society's College of Engineering, Pune, Maharashtra

Abstract: Development in Machine Learning algorithms has led to early detection and prediction of fatal diseases. Certain websites help patients to identify such diseases. Many pharmaceutical companies use advanced data mining techniques to extract the data from pathological reports of patients to generate statistical reports and decide their drug supply and marketing strategies. Our software bridges the needs of patients as well as pharmaceutical companies by predicting fatal diseases like brain tumour, diabetes mellitus, lung cancer and cardiovascular (heart) diseases for the patient and generating analytical reports from patient's data which provide a bird's eye view of prevalence of a disease within an area for the pharma sector.

Keywords: Machine Learning, Diabetes Mellitus, Lung Cancer, Cardiovascular Disease, Brain Tumour

1. INTRODUCTION

Early prediction of fatal diseases assists in better medical treatments of patients or even prevention. Advancement in application of Machine Learning in healthcare and pharmaceutical sector has helped patients and pharma companies in several aspects. Prediction of a fatal disease for patients at initial stages helps them decide whether to consult a doctor. Generating analytics report from patients' database in a specific region helps pharma companies decide the quantity of drugs to be manufactured for a certain location according to number of patients. It also helps them decide which medical aid to advertise according to majority type of patients in a region. Hence, directly impacting their production and marketing strategies.

When ML models of high accuracy are used, the predictions turn out to be accurate hence making the system reliable for patients as well as pharma sector as their sales are improved. The proposed work acts as a middle man for both the types of users. The project performs prediction of four fatal diseases namely, diabetes mellitus, brain tumour, lung cancer and cardiovascular diseases for patients and generates analytics report in terms of graphs, plots and other visualization techniques on real patient database for pharmaceutical companies to give an overview of types of patients living in an area along with statistical metrics.

Brain tumour is the accumulation of abnormal cells in the brain. This condition is difficult to cure because the brain has a complex structure and the tissues are interconnected with each other in a complicated network. An efficient and robust segmentation of brain tumour is still a challenging task. Brain abnormalities have been analysed for tumor recognition by using data mining, image processing and machine learning techniques. Abnormalities are analyzed using computed tomography (CT), image cryptography, Magnetic Resonance Imaging (MRI) and electroencephalogram (EEG) related data. MRI is most often used for the detection of tumors, lesions, and other abnormalities in soft tissues, such as the brain. The model used in proposed work classifies the MRI images of brain as Pituitary tumor, Glioma, Meningioma or non-tumor image.

Uncontrollable growth of cancerous cells in lungs is termed as Lung Cancer. It is a fatal disease as it accounts for 25% of cancer deaths globally. Computed Tomography (CT) scan of lungs is used to find severity of cancer. Medical image segmentation and classification of these scans play an important role in medical research field. The proposed work predicts lung cancer via symptom-based lung cancer dataset. As this method, is cost efficient and easy to respond for patients as well.

Insulin is responsible for controlling the glucose levels in body. Diabetes Mellitus (Type-2 Diabetes) is caused due to insufficient secretion of insulin by Islets of Langerhans (special cells in pancreas) or inability of the body to respond to insulin produced. The World Health Organization (WHO) ranked Diabetes Mellitus at 9th position for the "Top 10 Leading causes of Death Globally" in the year 2021 which was ranked 15th in the year 2000 [1]. The current project has predicted diabetes mellitus to an accuracy of around 98% using Early-Stage Risk Prediction Diabetes Dataset by UCI ML repository which contains records of symptoms faced by people and the prediction class.

Cardiovascular (heart) diseases (CVD) are the number one cause of death globally. W.H.O. reported that 17.9 million people died from CVDs in 2016, constituting 31% of all deaths globally. Out of them, 85% were due to heart attack and stroke [2]. Heart attack also known as Myocardial Infarction (MI), occurs when blood flow decreases or stops to a part

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11231

of the heart, causing damage to the heart muscles. The proposed work uses latest Kaggle dataset to detect any cardiovascular disease by taking pathological report readings as input from user. This detection helps patients to cross check the reports given by pathology labs.

All these diseases are predicted under a single application hence, the proposed work stands unique as compared to disease prediction websites available on the internet. Analytics reports are generated from patient's database via auto generated SQL queries when pharma clients select the required parameters to view statistics. Hence, a complete healthcare system benefitting both the users has been presented.

2. LITERATURE REVIEW

2.1. Brain Tumour Detection

A comparative study of various researches on brain tumor detection have been mentioned in Table 1.

Table 1. Comparative study of researches in Brain Tumour Detection

Reference	Year	Dataset	Algorithm	Results
K. Usman and K. Rajpoot [3]	2017	MICCAI BRATS data	Random Forest KNN AdaBoostM2 RusBust	Accuracy: 0.90 ± 0.03 Accuracy: 0.88 ± 0.03 Accuracy: 0.89 ± 0.03 Accuracy: 0.90 ± 0.02
L. Lefkovits et. al. [4]	2017	BRATS dataset	Random Forest SVM Adaboost	WT- 0.905; TC-0.887 WT-0.736; TC-0.817 WT-0.720; TC 0.791
S. Deepak, P.M. Ameer [5]	2019	Figshare	InceptionV3(Transfer Learning)-Softmax	Accuracy: 99.4453%
A. Kabir et. al. [6]	2018	IXI dataset, Cancer imaging archive dataset, REMBRANDT dataset, TCGA-GBM data collection	CNN evolved with Generic algorithm	For glioma grades:90.9% For Glioma, Meningioma, and Pituitary tumor: 94.2%

2.2. Diabetes Mellitus Prediction

A comparative study of various researches on prediction of diabetes mellitus have been mentioned in Table 2.

Table 2. Comparative study of researches in Diabetes Mellitus Prediction

Reference	Year	Dataset	Algorithm	Results
M Chakradar et. al. [7]	2020	CALERE dataset	Logistic Regression XGBoost SVM LDA	Accuracy: 97±1(for all)
K. Kannadasn, et.al. [8]	2019	Pima Indian Dataset	DNN Stacked Auto Encoders	Accuracy: 86.26%
P. Juliet et. al. [9]	2019	Pima Indian Dataset	Naive Bayes Decision Tree K Star Logistic Regression SVM	Precision: 0.770, Recall: 0.775; Precision: 0.742, Recall: 0.749 Precision: 0.691, Recall: 0.699; Precision: 0.772, Recall: 0.777 Precision: 0.767, Recall: 0.771;

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11231

S. Dey et. al. [10]	2018	Pima Indian Dataset	SVM + MMS KNN+ MMS GNG+MMS ANN+MMS	Accuracy:78.05% Accuracy: 75.5% Accuracy:79.3% Accuracy:82.35%
D. Pei et.al. [11]	2019	Chinese ethnic population	Decision Tree	Precision:0.94, Recall:0.942

2.3. Cardiovascular Disease Detection

A comparative study of a few researches in detection of cardiovascular (heart) disease has been shown in Table 3.

Table 3. Comparative study of researches in Heart Disease Detection

Reference	Year	Dataset	Algorithm	Results
A. Singh et. al. [12]	2020	UCI Heart Disease dataset	SVM Decision Tree Linear Regression KNN	Accuracy: 83% Accuracy: 79% Accuracy: 78% Accuracy: 87%
D. Krishnani et. al. [13]	2019	FHS dataset	Random Forest Decision Tree KNN	Accuracy: 96.71% Accuracy: 92.1% Accuracy: 91.49%
L Sharma et. al. [14]	2019	PTB database	SWT + KNN SWT + SVM	Accuracy: 98.69% Accuracy: 98.84%
V. Sharma et. al. [15]	2020	UCI Heart Disease dataset	SVM Random Forest Decision Tree Naïve Bayes	Accuracy: 98% Accuracy: 99% Accuracy: 75% Accuracy: 90%
K. Feng et. al. [16]	2019	PTB diagnostic ECG database	Multi-channel CNN and LSTM	Accuracy: 95.4%
L. Ibrahim, et. al. [17]	2020	ECG-VIEW II database	CNN RNN XGBoost	Accuracy: 89.9% Accuracy: 84.6% Accuracy: 97.5%

2.4. Lung Cancer Prediction

A comparative study of various researches on detection of myocardial infarction have been mentioned in Table 4.

Table 4. Comparative study of researches in Lung Cancer Prediction

Reference	Year	Dataset	Algorithm	Results
I. M. Nasser et.al. [18]	2019	UCL (Symptom dataset)	Artificial Neural Network (ANN)	Accuracy: 96.67%
W. Zhu et. al. [19]	2018	LUNA 16	3D Faster Region Based CNN	Accuracy: 81.42%
M. Nishio et.al. [20]	2018	The Cancer Imaging Archive	XGBoost	Accuracy: 79.7%
D. Moitra el. al. [21]	2020	The Cancer Imaging Archive	1D CNN	Accuracy: 96±3%
S. Makaju et. al. [22]	2017	LIDC-IDRI	SVM	Accuracy: 92%

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11231

3. METHODOLOGY

The proposed work has two interfaces to interact with respective users. The individuals who want to predict or detect any of the given diseases (brain tumour, cardiovascular disease, diabetes mellitus, lung cancer) use the Patient User Interface (UI). This interface allows the person to select the disease for which he/she wishes to know the prediction. Upon selecting the disease, user needs to fill in the details asked in the form displayed. This form contains question regarding the particular disease and user needs to either give integer values (e.g., age) or check certain checkboxes (yes/no) or upload an image file (MRI scan in case of brain tumour). On submitting this form, the backend of the software works on user data to generate predictions and displays the results (positive/negative) to the patient UI user.

When the form is submitted by the user, the data input by them is mapped with attributes of the dataset on which the selected ML model can make predictions. This ML model is selected among various algorithms which built models over training dataset. The algorithm giving the highest accuracy among others have been selected to make predictions. The comparative study of all algorithms for respective diseases has been shown in Results and Discussion section of this paper. After generating the classification results, all the attributes and class of disease are stored in a separate Patient Database. The Patient UI fetches the results from Patient Database to satisfy the user requirements.

The second interface has been developed to interact with individuals (pharma companies) who wish to view the analytics report for a disease in an area. The Pharma Client UI facilitates user to select certain parameters for which automatic SQL queries are generated in the backend which query the patient database and display results in form of graphs and charts to give an overview of demographical prevalence of a disease in a certain geographical region. The system architecture diagram of the proposed work has been given in the figure below (Fig. 1.)

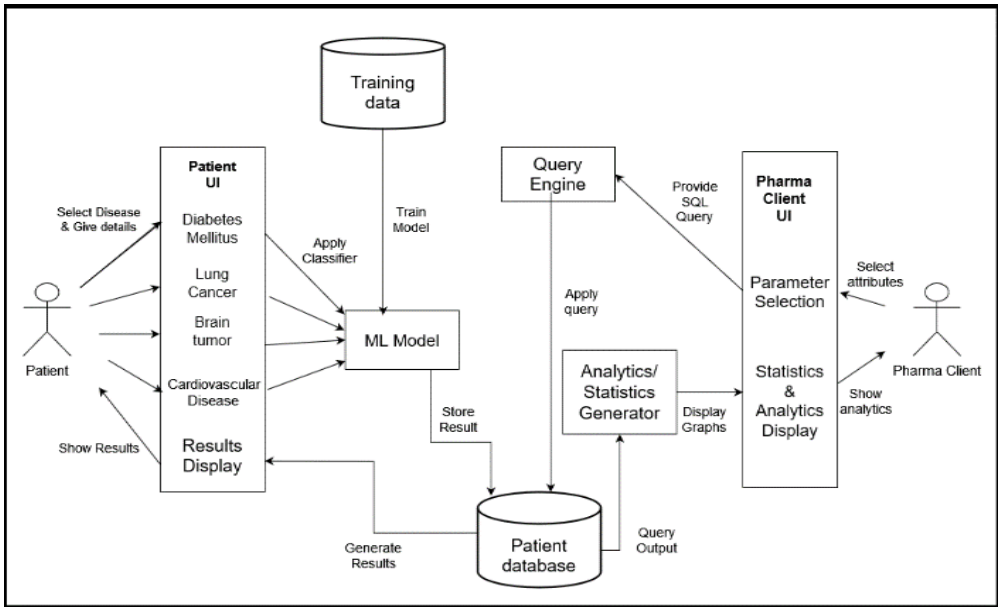

Fig 1. System architecture diagram of proposed work

4. RESULTS AND DISCUSSION

4.1. Diabetes Mellitus Prediction

Early-Stage Risk Prediction Diabetes Dataset from UCI ML Repository [23] was chosen over the PIMA Indian Diabetes Dataset (PIDD) for training various algorithms in the proposed work. Most of the Type-2 Diabetes Prediction researches in the world have used the PIMA Indian Diabetes Dataset but there are certain issues regarding this dataset. It contains an attribute for "No. of Pregnancies" and no attribute considering "Gender", which indicates that this dataset is centred around Diabetes Mellitus Prediction in women. Hence, a generalized dataset which considers all genders was needed. UCI ML Repository contains a Diabetes Dataset which considers Gender and Age as well. Hence this dataset is applicable for people of all age groups and sexuality. The dataset has 17 attributes and 520 instances as shown in Table 5.

Table 5. Information about Diabetes dataset

Attribute	Data
Age	0-100
Sex	Male/Female

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11231

Polyuria	Yes/No
Polydipsia	Yes/No
Sudden Weight Loss	Yes/No
Polyphagia	Yes/No
Genital Thrush	Yes/No
Visual Blurring	Yes/No
Itching	Yes/No
Irritability	Yes/No
Delayed healing	Yes/No
Partial Paresis	Yes/No
Muscle stiffness	Yes/No
Alopecia	Yes/No
Obesity	Yes/No
Weakness	Yes/No
Class	Positive/Negative

This dataset was used to train over multiple ensemble techniques. Ensemble algorithms create multiple sub-models and combine their outcomes to get improvised results. Comparison of accuracies on test data for each algorithm have been shown in Table 6.

Table 6. Comparison of accuracies of Ensemble Algorithms in Diabetes Mellitus Prediction

Ensemble Technique	**Accuracy**
CATBoost	97.12%
AdaBoost	94.23%
GBM	95.19%
Bagging Meta Estimator	**98.07%**

The results show that Bagging Meta-Estimator gave the highest accuracy of all other Ensemble techniques. Hence, it was chosen to predict Diabetes Melllitus on the given dataset.

4.2. Lung Cancer Prediction

Lung Cancer Dataset chosen for the project was taken from data.world's website [24]. This dataset is unique in comparison with other common datasets like UCI ML repository's lung cancer dataset or Kaggle's dataset as it contains features which would help in predicting Lung cancer based on symptoms. It is the most recent dataset available on the internet where features like gender and age are considered making it generalized for people of all type and age groups. The dataset has 15 attributes and 310 instances as shown in Table 7.

Table 5. Information about Data.world dataset for lung cancer

Attribute	**Data**
Age	0-100
Gender	M/F
Smoking	Yes/No
Yellow Fingers	Yes/No
Anxiety	Yes/No
Peer Pressure	Yes/No
Chronic Disease	Yes/No
Fatigue	Yes/No
Allergy	Yes/No
Wheezing	Yes/No
Alcohol Consuming	Yes/No
Coughing	Yes/No
Shortness Of Breath	Yes/No
Swallowing Difficulty	Yes/No
Chest Pain	Yes/No
Class	Positive/Negative

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11231

ML algorithms like Logistic Regression, AdaBoost, Linear Discriminant Analysis (LDA), Random Forest and Light Gradient Boosting (Light GBM) were trained on this dataset. Out of which Light GBM was chosen for its high accuracy. A comparison of accuracies of all algorithms has been shown in Table 8.

Table 8. Comparison of accuracies of classifiers in lung cancer prediction

Classifiers	Accuracy
Logistic Regression	87.03%
Ada Boost Classifier	92.59%
Linear Discriminant Analysis	87.96%
Random Forest	91.6%
Light GBM	**97.22%**

4.3. Cardiovascular Disease Detection

Kaggle's Heart Failure Prediction dataset [25] was used. This dataset is a combination of five different heart disease datasets that are already available independently. These datasets are combined over 11 common features which makes it the largest heart disease dataset available for research purpose. There is total 918 observations with 12 attributes as shown in Table 9.

Table 9. Information about heart failure prediction dataset

Attributes	Data
Age	0-100
Sex	Male/Female
Chest Pain Type	TA/ATA/NAP/ASY
Resting BP	0-200
Cholesterol	0-600
Fasting BS	Yes/No
Resting ECG	Normal/LVH/ST
Max HR	60-202
Exercise Angina	Yes/No
Old Peak	-2.6 – 6.2
ST Slope	Up/Down/Flat
Heart Disease	Yes/No

This dataset was used for training different classification algorithms. Comparison of performance of different algorithms is shown in Table 10.

Table 10. accuracy comparison of classifiers on heart failure dataset

Classifiers	Accuracy
SVM	72.46%
Random Forest	88.04%
KNN	68.47%
XGBoost	86.23%
CAT Boost	88.04%
CAT Boost with Optuna	**91.66%**

4.4. Brain Tumor Detection

The dataset used for Brain tumor detection was taken from Kaggle Repository [26]. This dataset consists of over 3000+ pre-processed brain tumor MRI images in jpg format hence providing a good variety of data. It classified the images into 4 categories namely, No Tumor, Pituitary Tumor, Meningioma Tumor and Glioma Tumor. A sample of the dataset has been shown in Fig. 2.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11231

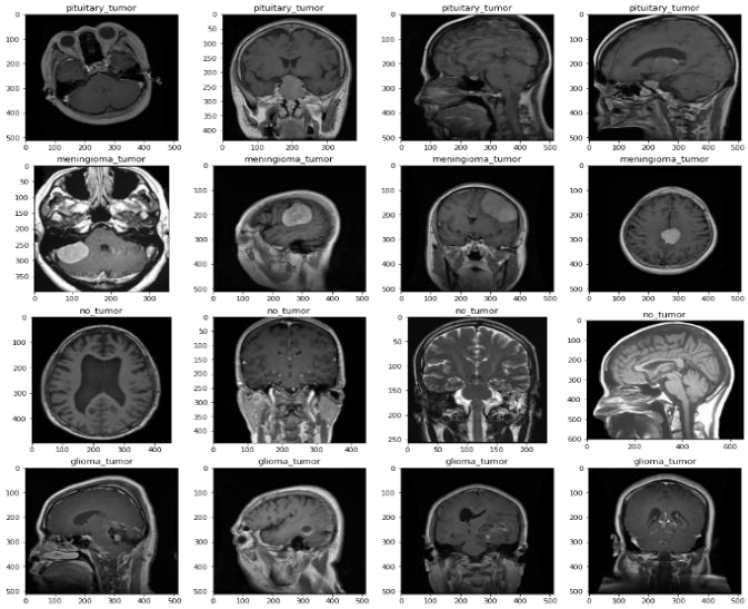

Fig. 2. Sample data from selected Brain Tumor MRI dataset.

Multiple deep learning techniques were applied on the dataset and their accuracies were compared to choose the best model as shown in Table 11.

Table 11. Accuracy comparison of classifiers on Brain Tumour MRI dataset

Algorithm	Accuracy
Keras sequential model with RMSprop	91%
EfficientNetB2 model with Adam optimizer	78.68%
VGG16 model	78%
EfficientNetB0 model with Adam optimiser	82.22%
DenseNet201 model with Adam	86%
EfficientNetB1 model with Adam	**95%**

The results show that EfficientNetB1 gave the highest accuracy of all other Deep Learning techniques. Hence, EfficientNetB1 was for the project.

CONCLUSION

The proposed work functions to meet the needs of patients and pharmaceutical clients. The accuracy achieved for each disease prediction model ensures the effectiveness of the work. The project stands out from the websites available on the internet which predict singular disease as it combines four different fatal diseases under one software as well as generates analytics report from the real patient data which can help in deciding business strategies for pharma companies. The ML models chosen for prediction provide the highest accuracy as well as the selected datasets are unique in their respective way as discussed in Results and Discussion section.

The project can be extended in future to add more disease prediction models. For now, it is limited to predicting only four diseases with good accuracy. With development in new algorithms and hyperparameter tuning, accuracy can be further improved for these models. Analytics report generated can also be extended as per the user's demand. The project can be further integrated with government's healthcare websites for betterment of the country and improving the health index. Further, collaboration with various hospitals and private doctors can be done to display the contact details of doctors to patients who are tested positive for any disease. An AI based diet plan advisor can also be integrated to suggest balanced diet to all users.

DOI: 10.17148/IJARCCE.2022.11231

REFERENCES

1. Peter Urban. "Top 10 Causes of Death Worldwide", AARP, https://www.aarp.org/health/conditions-treatments/info-2020/world-health-organization-data.html (accessed January, 2022).
2. Editor, "Cardiovascular diseases", World Health Organisation, https://www.who.int/health-topics/cardiovascular-diseases/ (accessed January, 2022).
3. K. Usman and K. Rajpoot, "Brain tumor classification from multi-modality mri using wavelets and machine learning", in Pattern Anal. Appl, 2017, pp. 871–881.
4. L. Lefkovits, et. al., "Comparison of classifiers for brain tumor segmentation", in International Conference on Advancements of Medicine and Health Care through Technology; 2017, pp. 195-200.
5. S. Deepak, P.M. Ameer, "Brain tumor classification using deep CNN features via transfer learning", in Computers in Biology and Medicine, 2019, vol. 111, doi: 10.1016/j.compbiomed.2019.103345.
6. Amin Kabir Anaraki, Moosa Ayati, Foad Kazemi, "Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms", in Biocybernetics and Biomedical Engineering, 2019, vol. 39, pp. 63-74, doi: 10.1016/j.bbe.2018.10.004
7. M. Chakradar, A. Aggarwal, "A Machine Learning Based Approach for the Identification of Insulin Resistance with Non-Invasive Parameters using Homa-IR", in International Journal of Emerging Trends in Engineering Research, 2020, pp. 2055-2064.
8. K. Kannadasan,et. al, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks", Clinical Epidemiology and Global Health, 2019, pp. 530-535.
9. P. Juliet, T. Bhavadharani, "An Improved Prediction Model For Type 2 Diabetes Mellitus Disease Using Clustering And Classification Algorithms", in International Research Journal of Engineering and Technology (IRJET), 2019, pp. 1179-1186.
10. S. K. Dey, A. Hossain and M. M. Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm", 21st International Conference of Computer and Information Technology (ICCIT), 2018, doi: 10.1109/ICCITECHN.2018.8631968.
11. D.Pei, C.Zhang, et.al, "Identification of Potential Type II Diabetes in a Chinese Population with a Sensitive Decision Tree Approach", in Journal of Diabetes Research, 2019, doi: 10.1155/2019/4248218.
12. A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," in International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.
13. D. Krishnani, A. Kumari, A. Dewangan, A. Singh and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 367-372, doi: 10.1109/TENCON.2019.8929434.
14. L. Sharma, Dev, Sunkaria, and R. Kumar, "Inferior myocardial infarction detection using stationary wavelet transform and machine learning approach," in Signal, Image and Video Processing, 2018, pp. 199–206.
15. V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," in 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.
16. K. Feng, X. Pi, H. Liu, K. Sun, "Myocardial infarction classification based on convolutional neural network and recurrent neural networ," in Appl. Sci, 2019, doi: 10.3390/app9091879.
17. L. Ibrahim, M. Mesinovic, K. W. Yang and M. A. Eid, "Explainable Prediction of Acute Myocardial Infarction Using Machine Learning and Shapley Values", in IEEE Access, 2020, pp. 210410-210417, doi: 10.1109/ACCESS.2020.3040166.
18. I. M. Nasser and S. S. Abu-Naser, "Lung Cancer Detection Using Artificial Neural Network", in International Journal of Engineering and Information Systems (IJEAIS), 2019, pp. 17-23.
19. W Zhu, C Liu, W Fan and X. Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification", in IEEE Winter Conf. on Applications of Computer Vision (WACV), 2018, pp. 673-681.
20. M. Nishio, et. al., "Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization", PLoS One, 2018, doi: 10.1371/journal.pone.0195875.
21. D Moitra and R.K. Mandal, "Classification of Non-Small Cell Lung Cancer using One Dimensional Convolutional Neural Network", in 2020 Expert Systems with Applications, 2020, doi: 10.1016/j.eswa.2020.113564.
22. S Makaju, PW Prasad, A Alsadoon, A.K. Singh and A. Elchouemi, "Lung cancer detection using CT scan images", in Procedia Computer Science, 2017, pp.107-114, doi: 10.1016/J.PROCS.2017.12.016.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11231

23. Early-Stage Diabetes Risk Prediction Data Set, UCI Machine Learning Repository, Jan. 2022. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.#

24. Lung Cancer Dataset, data.world, Jan. 2022. [Online]. Available: https://data.world/sta427ceyin/survey-lung-cancer

25. Heart Failure Prediction Dataset, Kaggle Repository, Jan. 2022. [Online]. Available: https://www.kaggle.com/fedesoriano/heart-failure-prediction

26. Brain Tumor Classification (MRI), Kaggle Repository, Jan. 2022. [Online]. Available: https://www.kaggle.com/sartajbhuvaji/brain-tumor-classification-mri