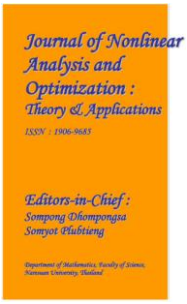
Journal of **N**onlinear **A**nalysis and **O**ptimization

Vol. 14, Issue. 2 : 2023

ISSN : **1906-9685**



EARLY PREDICTION OF MAJOR DISEASES USING MACHINE LEARNING

[1]Shaik Mohammed Imran, T Ashmitha Reddy, K Sanjana Reddy, A Shirisha, R Yagnitha
[1]Assitsant Professor, [2,3,4,5]UG Students, Dept. of CSE (AI & ML), Malla Reddy Engineering College for Women (Autonomous), Hyderabad, India. E-Mail: imransk@gmail.com

ABSTRACT
There are several hospitals around the world that have cutting-edge diagnostic tools. But even with modern technology, some patients may not receive the right care and end up dying. The main cause of this is lack of time, which our medical systems struggle to manage. We developed a project that uses machine learning techniques to identify people with significant diseases like heart disease, kidney disease, liver disease, malaria, pneumonia, and diabetes disease before time on so as to appropriate therapies may be delivered to them. We took the datasets from GitHub, cleaned them up, and selected the best algorithm for each one. We were able to predict heart disease with 98.52% accuracy, kidney disease with 98.73% accuracy, diabetes with 80.55% accuracy, liver disease with 90.0% accuracy, malaria with 80.33% accuracy, and pneumonia with 80.00% accuracy. We applied the Random Forest method to all three models. In order to facilitate user interaction, we ultimately constructed a web application using Flask.

INTRODUCTION
Patients frequently lose their lives as a result of delayed medical care. Due to a lack of time, healthcare organizations are unable to prioritize which patients to treat first. However, the healthcare sector also produces a vast amount of information about the health of its clients. This data allows for the extraction of very deep insights. So, combining this data and cutting-edge machine learning methods, we came up by the "Multiple Disease Prediction System" project. In our study, three machine learning models will be combined to detect people with diabetes, renal disease, and heart disease at an early stage, allowing such patients to receive treatment first. For our project, we first determined the type of data that would be needed and then understood the issue statement. For our three machine learning models, we obtained three distinct datasets from Kaggle. After gathering data, we carefully analyzed it and visualized it for easier comprehension. The data was then cleaned by encoding categorical features and imputing null values. The next phase involved splitting the dataset keen on training and testing sets, with 80% of the data being worn to train the machine learning model and the enduring 20% being worn to test it. Following that, we tested various classification methods on all three datasets, including “Logistic regression, Random Forest classifier, Support Vector Machine classifier, and XGBOOST classifier”.
For all three datasets, we discovered that the Random Forest classification technique outperformed the others. We accomplish 98.52% testing precision on the dataset for heart disease, 98.73% testing accuracy on the dataset for renal disease, and 80.55% testing precision on the dataset for diabetes using the Random Forest algorithm. So, for simple user interaction, we constructed a web application using Python's Flask framework and dumped the model using the Pickle library of Python. Additionally, we

http://doi.org/10.36893/JNAO.2023.V14I2.0182-0189

used Tableau to construct visualization dashboards for each of the three datasets in order to enhance the project.

LITERATURE SURVEY

The study underlines how dangerous diabetes is and how it can cause a variety of illnesses, including blindness among others. Diabetes is one of the most lethal diseases in the world. In this work, machine learning techniques were used to identify the diabetic condition since it is straightforward and adaptable to forecast if the patient has an illness or not. The researchers wanted to develop a system that would accurately diagnose a patient's diabetes in order to aid them. Here, they compared the accuracy of four primary algorithms: Decision Tree, Nave Bayes, SVM, which were utilized in this study. The accuracy of these algorithms was 85%, 77%, and 77.3%, respectively.

In order to determine whether the disease was correctly identified or not, they also applied the ANN algorithm after the training phase. Here, they compared all of the models' accuracy, F1 score support, and precision [1].

2. as the heart plays a significant part in living beings that is the paper's major goal. Therefore, it is imperative that heart-related disease be accurately diagnosed and predicted because doing so can result in fatalities. Artificial intelligence and machine learning thus aid in the prediction of all kinds of natural disasters. In order to determine the accuracy of machine learning for estimating heart disease by means of k-nearest neighbor, decision tree, linear regression, and SVM, they employ the UCI repository dataset for training and testing in this study. The accuracy of the algorithms SVM (83%), Decision Tree (79%), Linear Regression (78%), and K-Nearest Neighbour (87%) was also compared [2].

3. According to the system, liver illnesses are a leading cause of death in India and are also regarded as a serious health problem worldwide. Thus early liver disease detection is challenging. So, we can accurately diagnose liver illness using an automated programme that uses machine learning methods. They employed and compared the SVM, Decision Tree, and Random Forest algorithms for quantitative measurement, calculating metrics for precision, accuracy, and recall 95%, 87%, and 92% accuracy rates, respectively [3].

Problem definition

Even though modern healthcare systems are equipped with cutting-edge and powerful diagnostic tools, many people still die because they don't receive timely care. Time is one resource that healthcare systems lack, thus they are unable to prioritise which patients should receive treatment first. As a result, a patient in need might not receive care in time, endangering his life.

Many of us need a routine checkup once in a while so that we can detect any of the problem in our immune system or our whole body. A routine checkup for many diseases in the hospitals are expensive so we need a simple and effective disease prediction system. Machine learning algorithms are efficient enough to predict the diseases. Each and every information related to a particular disease is fed to the data models. This project concentrates on detecting multiple diseases at once by making it easy for the users to get diagnosed.

FLASK Functionalities

The micro-framework's categories include flask. Micro-frameworks often have minimal or no dependence on outside libraries. Pros and downsides exist for this. The framework's advantages include its light weight, low dependency requirements, and low need for security bug monitoring. Its disadvantages include the need for occasional independent work and the need to add plugins to the list of dependencies.

PICKLE Functionalities

When serializing and desterilizing a Python object structure, pickle is primarily employed. It involves transforming a Python object into a byte stream in order to accumulate it in a file or database, keep

programme state consistent across sessions, or send data over the network. By unpickling the stream, the pickled byte stream can be utilized to recreate the initial object chain of command. The entire procedure is comparable to Java or.Net object serialization.

The pickle module first generates an occurrence of the original object while a byte stream is unpickled, and it then fills the instance with the appropriate data. To do this, just the information unique to the first object instance is present in the byte stream. However, simply having the data might not be enough. The pickled byte stream provides instructions for the unpickler to rebuild the original object structure as well as instruction operands that aid in populating the object structure in order to correctly unpickle the object.

RANDOM FOREST CLASSIFIER Functionalities

Random forest is a supervised machine learning technique. This method creates a number of decision trees and selects predictions from each tree using a few randomly selected subsets of the training set. The top candidate is then determined by the random forest algorithm through voting. An assortment of decision trees are produced by the Random forest classifier using a portion of the data used for training that is selected at random. It simply consists of an assortment of decision trees (DT) chosen at random across a subset of the training set, having those decision trees being utilized to make the ultimate forecast. This classification technique will be trained and tested using IRIS flower datasets. A model will be constructed to classify the many sorts of flowers.

Naive Bayes classifier Functionalities:

The theory underlying the Naive Bayes classifiers and how they work are covered in this article.

Naive Bayes detectors are a subset of algorithms for classification based on the Bayes theorem. Instead of getting a single approach, it is a family of techniques that are all predicated on the notion that each combination of characteristics that is categorized is independent of one another.

Let's think about a dataset first.

Take a look at a hypothetical dataset that details the weather requirements for a round of golf. Each tuple determines whether the weather is suitable for playing golf ("Yes") or not ("No") based on the current conditions.

System Architecture

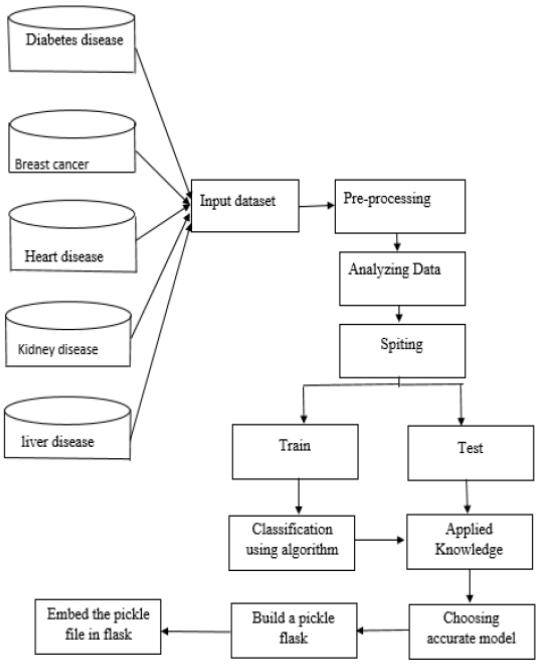


Fig.1. System model

http://doi.org/10.36893/JNAO.2023.V14I2.0182-0189

IMPLEMENTATION

Both academic and commercial use of OpenCV (Open Source Computer Vision Library) is free because it is released under a BSD licence. It has C++, Python, and Java interfaces and supports Windows, Linux, Mac OS, iOS, and Android. The development of OpenCV placed a strong emphasis on real-time applications and optimum processing effectiveness. The library be able to take improvement of multi-core processing because it was created in C/C++ that was optimized. When Open CL is enabled, It can benefit from the supporting heterogeneous compute platform's hardware amplification. A group of computer vision and machine learning technologies that are open source and free. The creation of an established framework for computer vision applications using OpenCV helped to speed the incorporation of artificial intelligence into products. Companies can easily use and modify the code because to OpenCV's BSD licencing. The package provides a variety of shared or static libraries as a result of OpenCV's modular nature.

Open CV-Python

Because of its simplicity accessibility and legible code, Guido van Rossum's general-purpose programming language, Python, rapidly became well-known. A smaller amount of code is needed for the coder to convey his ideas while losing readability.

Python is lesser than other languages similar to C/C++. But another crucial aspect of Python is its simplicity in C/C++ extension. This feature makes it easier for us to develop computationally demanding C/C++ code and convert it into a Python wrapper so that we may utilize the wrapper as a Python module.

Tensor Flow:

The general-purpose programming language Python was developed by Guido van Rossum and immediately became well-known for its usability and legible code. The developer may convey his ideas in less code while maintaining readability.

Python is lesser than other languages similar to C/C++. But another crucial aspect of Python is its simplicity in C/C++ extension. This feature makes it easier for us to develop computationally demanding C/C++ code and convert it into a Python wrapper so that we may utilize the wrapper as a Python module.

KNN Algorithm

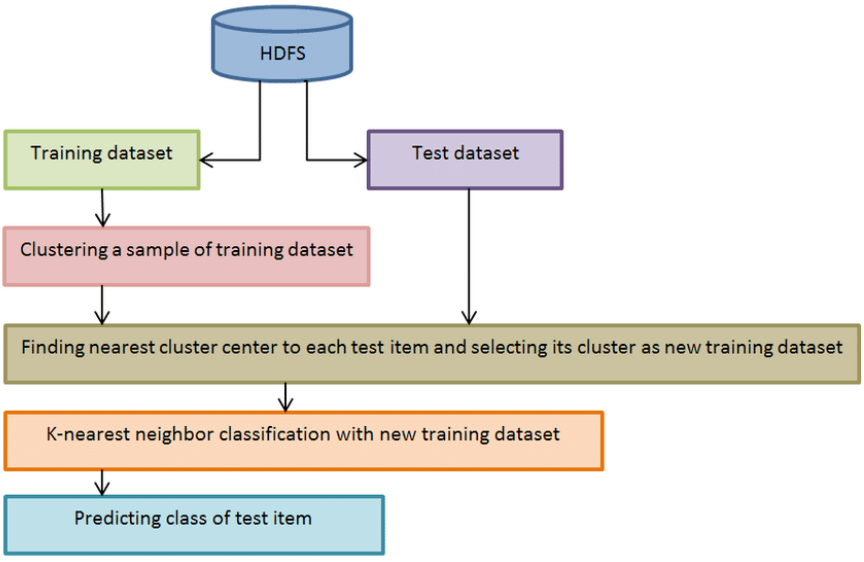


Fig.2. KNN Algorithm

The K-NN algorithm functions as follows:

a. Begin by choosing the K value, for instance, k=5.

http://doi.org/10.36893/JNAO.2023.V14I2.0182-0189

b. The Euclidean distance between the locations will then be determined. The formula is as follows.
c. The closest neighbor's Euclidean distance is then determined.
d. Compute how many data points there are in every category.

Random Forest Algorithm

A random forest can be constructed by merging N decision trees, and then it can be used to make prophecy for every tree that was produced in the primary step. The random forest operates as pursue:

a. Initially it will randomly select K data points from the training set.
b. The decision trees associated with the selected data points (Subsets) are generated after selecting k data points.
c. then deciding on N for the decision trees you wish to construct.
d. doing steps 1 and 2 again.
e. Determining the forecasts from each decision tree, then placing the fresh data in the group of data that has the strongest support.

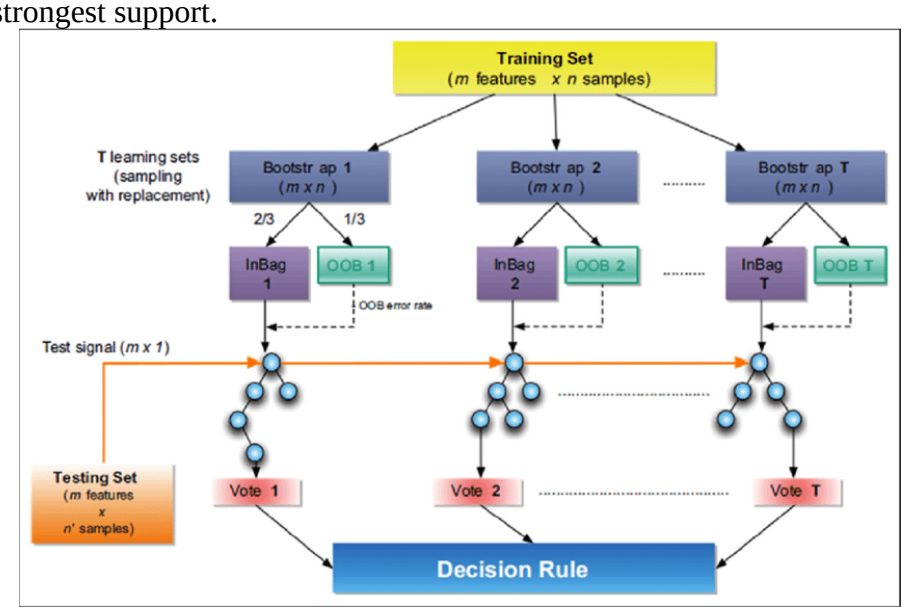


Fig.3. Random forest algorithm

XG Boost Algorithm

The XGBoost algorithm operates as follows:

Step 1: Making a tree with just one leaf first.

Step 2: Following that, we must calculate the regular of the target variable as prediction for the first tree before by means of the appropriate loss function to calculate the residuals. subsequently, for following trees, the residuals are derived from the prediction that was present in the first tree.

Step 3: using the following formula to determine the similarity score: where Hessian = number of residuals; The regularization hyper parameter gradient2 is equal to the squared sum of the residuals.

Step 4: We choose the relevant node by applying the similarity score. More homogeneity is seen when the similarity score is higher.

Step 5: Information gain is calculated using the similarity score. Information gain reveals how much homogeneity is obtained by splitting the node at a specific place and helps to distinguish between old and new similarities.

Step 6: By adjusting the regularization hyper parameter, the aforementioned method's pruning and regularization can be used to create the tree of the appropriate length.

The Decision Tree you created can then be used to forecast the residual values.

Step 8: The learning rate is used to calculate the new set of residuals.

Step 9: After that, go back to step 1 and carry out the procedure for each tree.

http://doi.org/10.36893/JNAO.2023.V14I2.0182-0189

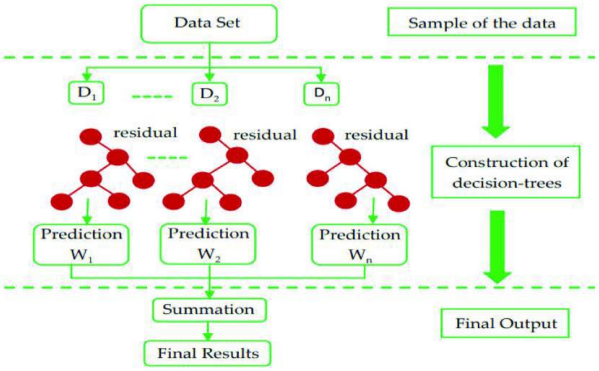
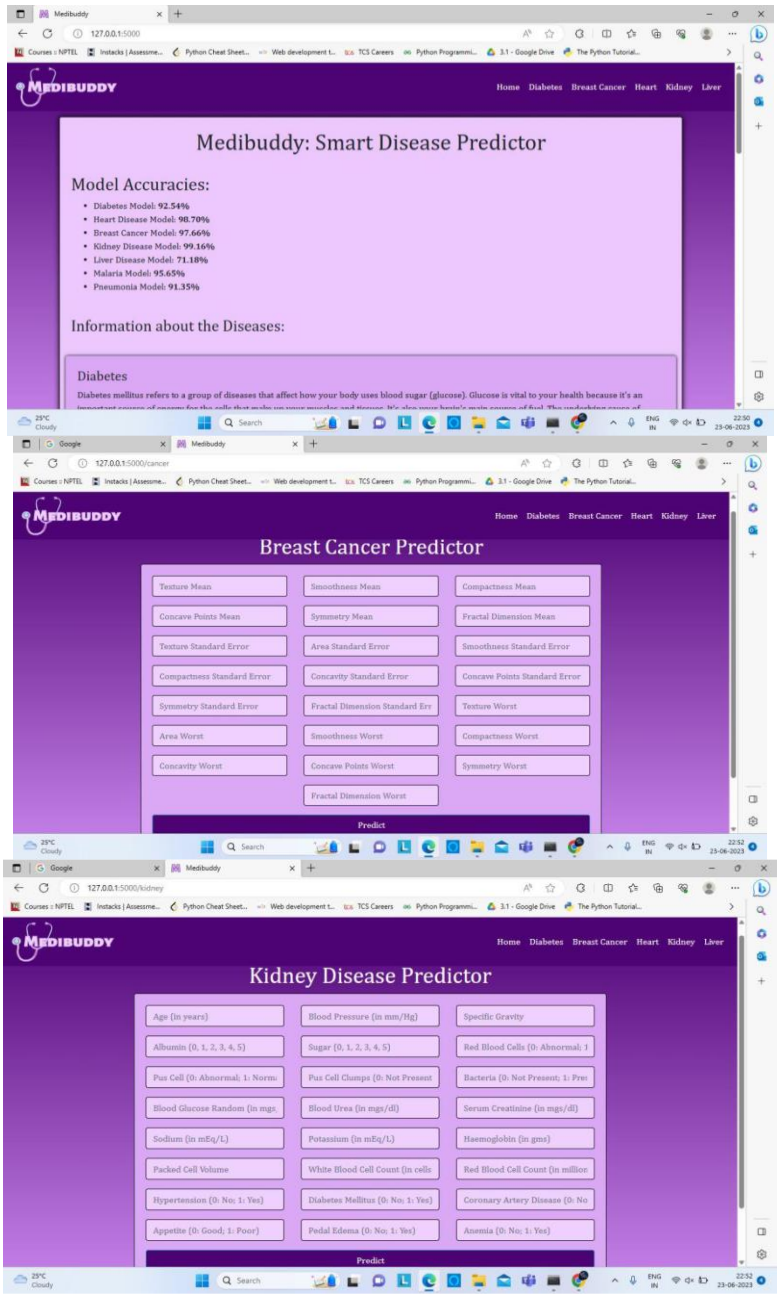
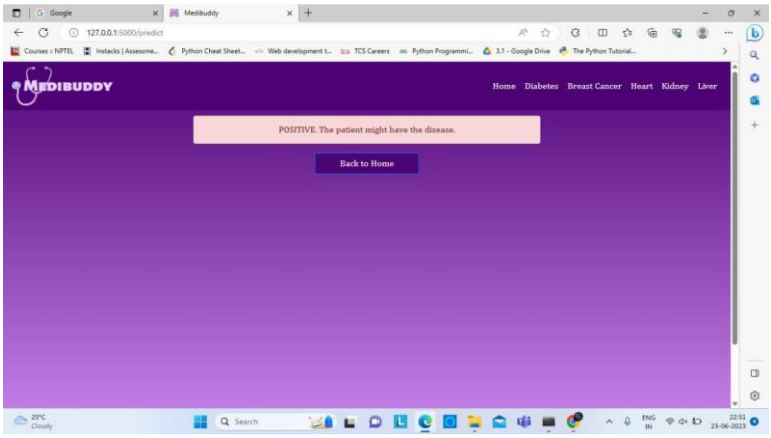


Fig.4. XG boost algorithm

Results



http://doi.org/10.36893/JNAO.2023.V14I2.0182-0189



CONCLUSION

In conclusion, machine learning techniques have shown great potential in the field of disease prediction. By utilizing large datasets and advanced algorithms, machine learning models can effectively analyze and extract meaningful patterns and relationships from various types of data, including medical records, genomic information, lifestyle factors, and environmental factors. The use of machine learning for multiple disease prediction offers several advantages. First, it enables early detection and intervention, which can significantly improve patient outcomes and reduce healthcare costs. By identifying individuals at high risk for multiple diseases, healthcare providers can implement targeted preventive measures and personalized treatment plans.

Furthermore, machine learning models can handle complex and high-dimensional data, allowing for the integration of diverse factors that contribute to disease development. This holistic approach enhances the accuracy and robustness of disease predictions, taking into account not only genetic predisposition but also lifestyle choices, environmental exposures, and other relevant variables. However, it is important to acknowledge some limitations and challenges associated with machine learning-based disease prediction. The performance of these models heavily relies on the quality and representativeness of the data used for training. Issues such as data bias, missing data, and data quality can impact the accuracy and generalizability of the predictions. In summary, while machine learning holds great promise for multiple disease prediction, its successful implementation requires careful consideration of data quality, interpretability, and ethical considerations.

FUTURE SCOPE

We plan to extend this project to accommodate further information about roads such as follows:

To facilitate smart transport, a real-time map is made available to all users with the most recent information about these potholes. With this information, drivers can be forewarned and their locations can be communicated with local authorities for prompt repair.

Creating a system to map road conditions would aid drivers in making better decisions in addition to identifying potholes.

The severity of a pothole might be categorized as a subsequent feature. Governments would be able to set priorities while correcting potholes if they can distinguish between deep and shallow craters.

REFERENCES

[1] Priyanka Sonar, Prof. K. Jaya Malini," DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES", 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC)

[2] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3)

http://doi.org/10.36893/JNAO.2023.V14I2.0182-0189

[3] A. Sivasangari, Baddigam Jaya Krishna Reddy, Annamareddy Kiran, P. Ajitha," Diagnosis of Liver Disease using Machine Learning Models" 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)

[4] A. S. Abdullah and R. R. Rajalaxmi, ''A data mining model for predicting the coronary heart disease using random forest classifier,'' in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25

[5] M. S. Dr Vijayarani1, ''Liver disease prediction using SVM and Naïve Bayes algorithms,'' Int. J. Sci., Eng. Technol. Res., vol. 4, no. 4, pp. 816–820, 2015

[6] L. Breiman, ''Random forests,'' Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001

[7] I. Jenhani, N. B. Amor, and Z. Elouedi, ''Decision trees as possibilistic classifiers,'' Int. J. Approx. Reasoning, vol. 48, no. 3, pp. 784–807, Aug. 2008.

[8] S. Vijayarani, S. Dhayanand, and M. Phil, ''Kidney disease prediction using SVM and ANN algorithms,'' Int. J. Comput. Bus. Res. (IJCBR), vol. 6, no. 2, pp. 2229–6166, 2015.

[9] N. Al-milli, ''Backpropogation neural network for prediction of heart disease,'' J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

http://doi.org/10.36893/JNAO.2023.V14I2.0182-0189