International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Earlier Forecasting of Diseases and Assessment of Risk Using a Novel Deep-Learning Approach

[1]Ragavendra U., [2]Rahul Bhatt, [3]Neetha S. S., [4]Harjinder Singh

Submitted:18/04/2023 **Revised:**11/06/2023 **Accepted:**22/06/2023

Abstract: Globally, chronic kidney disease (CKD) is a problem with mortality rate and high morbidity. Rapid responses and better patient outcomes depend on early detection and precise risk assessment. To enable earlier CKD forecasting and risk assessment, this research suggests an innovative strategy incorporating randomized Gaussian-search Aquila optimization with Deep Neural Network (RGAO-DNN) network. The model intends to boost the network's efficiency and increase its capacity to capture complicated temporal connections in CKD data by integrating RGAO. An extensive dataset containing measures from CKD patients is used to assess the suggested approach. To deal with errors, normalize the data, and tackle group disparity problems, the min-max normalization methodology is used. The suggested technique is trained on the cleaned information, allowing it to recognize significant risk variables for the development of CKD and learn from temporal patterns. The effectiveness of the strategy is assessed using several measures. The experimental findings show that the suggested technique works better than other approaches regarding CKD risk evaluation and early prediction.

Keywords: *Chronic kidney disease (CKD), patient outcomes, risks, early prediction, randomized Gaussian-search Aquila optimization with Deep Neural Network (RGAO-DNN)*

1. Introduction

Back in the 1950s, oral communication dominated amongst people. But as technology advanced, people became increasingly fixated on it. The key question is still, "Why are people more obsessed with technology?" The answer is simple to understand. Increased manufacturing demand fuels faster growth in product, trade, and business perspective[11]. The IoT appliances (connected together) produced vast amounts of data that must be efficiently examined using cutting-edge technologies and methodologies.

The grouping of various items into clusters of like-minded things is what is meant by clustering. Unlike the objects in the other groups, the objects in the cluster/group are comparable to one another. The program is widely relevant in fields including marketing, the internet, earthquake research, aerospace, biology, and insurance, among others[12]. On the other side, A classification strategy was utilized to divide the given data into various classes based on their commonalities when the information was rendered with class or category labels. Recognition of speech and handwriting, biometric identification, document classification, and more uses for classification are available.

The kidney plays a crucial role in the human body in absorbing and eliminating all harmful and unnecessary substances, commonly wastes, through the egestion. In India, a million new cases of chronic kidney disease (CKD) are reported each year. Although it was unpredictable, Finding CKD was essential at an initial time because its symptoms develop gradually and are not specific to the condition[8]. The kidneys purify wastes and extra fluid from the blood, which are ultimately eliminated in feces. A few indications or symptoms of CKD will be present in its early stages.

One of the most widely used types of ML(machine learning) in the healthcare industry is classification. For each data point, the classification model displays the result's class.

The methods used for classification include Support Vector Machines(SVM), Decision Trees(DT), estimated glomerular filtration rate (eGFR). To anticipate the development of the disease at an early stage, The relationship between several CKD risk factors is demonstrated using KNN [13]. A developing region of Al was ML, which deals with the analysis of massive amounts of variable data. It evolved from research into pattern

1Professor, School of Engineering & Technology, Jaipur National University, Jaipur, india, Email Id: dean.associate@jnujaipur.ac.in
2Assistant Professor, School of Engineering and Computer, Dev Bhoomi Uttarakhand University, Uttarakhand, India, Email Id: socse.rahul@dbuu.ac.in
3Assistant Professor, Department of Computer Science and IT, Jain(Deemed-to-be University), Bangalore-27, India, Email Id: neetha.s.s@jainuniversity.ac.in
4Assistant Professor, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India, Email id: harjinder.mca07@gmail.com

recognition, including handwriting, speech, and computational learning theory. ML uses a variety of algorithms, methods, and techniques to analyze and predict the data. In order to determine if the patient has CKD or not in this scenario, ML might be helpful. ML achieves this by training a predictive model utilizing historical patient data on CKD.

This study employs the min-max normalization algorithm to deal with mistakes, normalize the data, and address group disparity issues. The suggested method may identify important risk factors for the onset of CKD and learn from temporal patterns because it has been trained on the cleaned data. Several metrics are used to gauge the strategy's effectiveness. The results of the experiments demonstrate that the suggested method performs better in terms of CKD risk assessment and early prediction than other methods.

Paper organization: Section 2- related works, section 3- methodology includes dataset, Min-max normalization, Randomized Gaussian-search Aquila optimization, DNN based forecasting, section 4- result includes analysis, section 5- conclusions were demonstrated on this paper.

2. Related Works

The effectiveness and caliber of medical care could be enhanced by utilizing EHRs to anticipate the onset of diabetes. In order to develop a prediction, In the study [1] used a dl approach, which combines a lot of features, a deep feed-forward neural network, and the power of a generalized linear method. Experiments demonstrate that the suggested method achieves acceptable predictive performance and fairly comprehensible feature analysis; it might be used to supplement established epidemiological models for pandemic surveillance at the national level, such as COVID-19. The outcomes and conclusions from the DL method may help academics and policymakers come up with efficient mitigation and response plans [2].The best indications of MDD included having public health insurance and being pleased with one's living position. Marijuana use and having current immunization records were the two main signs of GAD. According to the research[3], ML approaches can detect GAD and MDD from EHR data with a relatively successful prediction performance. For two research locations located in Iran and India, The autoencoder models and multilayer perceptron (MLP) are combined with a novel hybrid model in the paper to produce susceptibility maps. From the 2 cases, the predictor variables for the mapping of flood susceptibility were nine and twelve factors, respectively [4].To generate predictions for Alzheimer's disease, the OASIS data was used, and the performance of ML models was gauged using metrics including Recall, Accuracy, Precision, and F1-score. Clinicians can diagnose these disorders using the proposed classification approach. At initial diagnosis, the ML algorithms have the potential to drastically lower the annual mortality rates of Alzheimer's disease. [5].Article[6] used the COVID-19 Kaggle data to do multilayer perceptron, vector autoregression, and linear regression to predict the rate of COVID-2019 cases in India epidemiological example of the disease. The potential COVID-19 trends in India were anticipated using data from Kaggle. [6].

The Sino-French New City Branch of Tongji Hospital, Wuhan, admitted 183 patients with severe COVID-19 infection; clinical, initial laboratory findings and demographic were used to create the predictive models. Methods of ML were employed[7].

Article [8] applied a novel hybrid functional ML system to predict the spatial distribution of landslides in the Sarkhoon watershed, Iran. To predict landslides, they created a new ensemble model called the ABSGD model that combines a functional approach with an AdaBoost meta classifier and stochastic gradient descent (SGD). 20 landslide conditioning factors were included in the model, and they used the LSSVM method to rank them.The research [9] suggested an attribute integration and ensemble deep learning system for forecasting of coronary artery disease. To create useful healthcare data, the feature fusion approach first merges the derived features from sensor data with electronic medical records. Second, by removing unnecessary redundant features and choosing the most crucial ones, the information gain technique reduces the computational load and improves system performance. The conditional probability technique computes a distinct feature weight per each class, further boosting system efficiency. The article's [10] goal was to create an ML model to predict FLD that would help medical professionals categorize high-risk patients, and come up with a managed FLD and new diagnosis. To determine the classification models, FLD, including artificial neural networks (ANN), logistic regression (LR), random forest (RF), and Naive Bayes (NB), were created. Utilizing the area under ROC, four models' analysis was compared. Using a DL system, article [11] developed the VHP (Virus Host Prediction) to anticipate probable hosts for viruses. According to their prediction, 2019-nCoV and other human coronaviruses, particularly (SARS-CoV) are highly infectious.On dataset creation, a historical study was done to investigate the significant blood biomarkers for estimating illness mortality. In order to find important biomarkers that predict a patient's demographic, clinical and patient outcomes were examined using ML technologies. For assessing the mortality risk among COVID-19 patients, a nomogram was created in the article[12].

In the study [13], for the diagnostic uses, the most precise ML classifiers were sought after. Several supervised ML algorithms were employed, and their efficacy was assessed in the prediction of heart illness. All employed algorithms were predicted to have feature relevance ratings for every feature, with the exception of KNN and MLP.The three top-performing algorithms (three different versions of SVM) were utilized in the remaining portions of the study after the initial test, ten conventional ML techniques. The performance of these algorithms was enhanced by data standardization and preprocessing. Additionally, a stratified along with a genetic algorithm and particle swarm optimization, 10-fold cross-validation was used. [14].The main result was the risk of developing a severe illness, which was indicated by ICU, mechanical ventilation, and multi-organ failure. Comparing machine-learning models to the most effective techniques available, they showed excellent efficacy in predicting important COVID-19. A comparison of the APACHE II risk-prediction score and three different machine-learning algorithms that were employed to predict patient deterioration was made in the article [15].

3. Methodology

This paper suggests a fresh strategy focused on the RGAO-DNN network, which combines randomized Gaussian-search Aquila optimization. By using RGAO, the model aims to improve the network's effectiveness and expand its capability to grasp intricate temporal relationships in CKD data. To evaluate the proposed method, a sizable dataset containing measurements from CKD patients was used.

3.1 Dataset

Disease records were chosen as the work's data source; the sample set was gathered at the general hospital in Yobe State's Gashua Local Government Area. It contains 400 patient records with the following parameters: sodium, potassium, bicarbonate, urea, creatinine, urea acid, albumin, age, gender, categorization, chloride that is binary-classified into non-CKD and CKD disease.

3.2 Min-max normalization

Numeric features are typically transformed to a certain range using the min-max normalization approach, also referred to as feature scaling or rescaling. When you wish to scale the values of your data between a minimum and maximum value, which is typically between 0 and 1, it is really helpful. Equation (1) represent the Min-max normalization formula.

The following steps are part of the min-max normalization methodology:

1. Determine the maximum and minimum values of the feature in your dataset that you want to normalize.

2. Decide on the normalization range that you want to use; it is usually between 0 and 1.

3. All of the feature's values should be normalized using the min-max algorithm.

$$Y' = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \times (max_{new} - min_{new}) + min_{new} \quad (1)$$

Where, Y- original value of the feature, Y_{min}= minimum value of the feature in the dataset, Y_{max}= maximum value of the feature in the dataset, $max_{new}$and min_{new} represents the desired range for the normalized values.

3.3 Randomized Gaussian-search Aquila optimization

Gaussian-search with randomization, A novel technique called Aquila optimization, has been used to predict diseases in the past. To improve the accuracy of disease prediction, this method integrates components of randomized search, Gaussian distribution, and the Aquila optimization algorithm. Using a randomized search approach, a diverse group of candidate solutions are produced in the first step of this procedure. This approach tries to cover a substantial amount of the search space related to illness forecasting by investigating a wide range of potential solutions. This first phase encourages the examination of various patterns and causes leading to the occurrence of disease while preventing convergence to less-than-ideal remedies. The Gaussian distribution is used to model the fitness landscape of the problem once the candidate solutions have been developed. The odds that various solutions will be optimal or substandard can be described using the Gaussian distribution. The algorithm can focus its search efforts by identifying regions with a higher likelihood of containing optimum solutions by taking into account the fitness landscape.

The second of the genuine AO algorithm's four techniques, the Levy flight function effect, causes Aquila search to be inadequate in the solution space and to frequently settle within local optima. The third, on the other hand, results in a limited regional exploitation capability because the parameter remains constant. In order to improve on the second and third methods of the original AO, we develop a SCF with an absolute value (abs) that lowers with the number of iterations. This allows us to slow down Aquila's search rate as iterations go. The ROBL and GM techniques are additionally included to enhance the exploration and exploitation phases, respectively. Below, more information on the IAO is provided.

3.3.1 Factors Controlling Search (FCS)

The basic search step size and direction of Aquila are controlled by the search control factor. The Aquila will move less frequently as iteration goes on, which will improve the search's precision. As a result, the FCS's abs get smaller with each iteration (t), which is defined in equation (2) and (3)

$$FCS = L_v.\exp\,(1-\tfrac{t}{T}) \times D \qquad (2)$$

$$D = \begin{cases} 1, if\ n < 0.5 \\ \quad -1, else \end{cases} \qquad (3)$$

Where $L_{v=}$ Constant, n= random number between 0 and 1,

$\exp(1-\frac{t}{T})$ uses a number of iterations to regulate the Aquila's flight speed.

D= Direction control factor, It is employed to manage the Aquila's flying path.

3.3.2 Better narrated exploration

The Aquila thoroughly searches the target solution space using various directions and speeds in the second enhanced approach before attacking the prey. Full search with a brief glide attack is the name of the enhanced technique.

An illustration of the enhanced strategy's mathematical expression formula is shown in equation (4).

$$Iy_2(t+1) = Y_n(t) + FCS(t) \times (Y_{best}(t) - Y(t)) \times n_6 \times (z-y) \quad (4)$$

Where, $IY_2(t+1) =$ next iteration of t's solution

$Y_n\,(t) =$ Aquila population was chosen at random.

$y_{\,best}(t)$ =best-attainable solution up until iteration t

3.3.3 Improved Expanded Exploitation

The third enhanced technique involves the Aquila extensively circling the victim before attacking it. This technique, which is formally defined by Equation (5), tries to increase the exploitation capability.

$$Iy_2(a,b) = IA_6 + rand.(XA_b - IA_b) + n_7.\left(Y_n(b) - Y_{best}(b)\right).FCS(t).(1-\tfrac{t}{T}) \quad (5)$$

$IY_{3\,(a,b)\,=}$ Next iteration of t's solution

3.4 DNN-based forecasting

The DNN structure consists of an input, a hidden layer, and an output, simulating the input, weights, and activation functions of a biological neural network. The DNN is a subset of artificial neural networks. In a DNN, the output is referred to as a, while Wi and Xi are the input and weight, respectively. The input data can be used to determine the number of neurons. Three layers were employed in the model's training. "Relu" is utilized as an activation function for input and the hidden layer. Relu produces 0 or 1. One of two output results of KD with Sigmoid as the activation function is available from the output layer. The model's optimizer was a stochastic gradient. Python is the programming language used to create the experiments. Figure 1 shows the structure of DNN model.

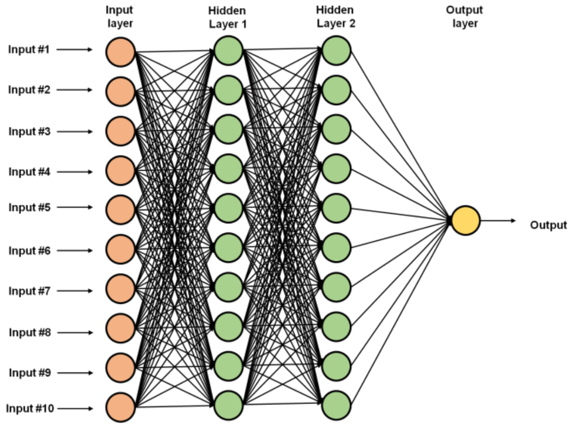


Fig 1: Structure of DNN model

4. Result

The innovative DL approach for disease prediction and risk assessment has produced encouraging results in a number of research. This method has shown enhanced accuracy and efficiency in illness prediction and risk assessment by utilizing the power of DNN. In multiple tests, the DL model beat conventional statistical techniques and other ML algorithms for predicting diseases. More precise forecasts of illness onset, development, and recurrence have been made possible by the model's capacity to understand complex patterns and relationships from vast amounts of medical data.

For instance, the DL model successfully identified persons at risk in research on the prediction of CKD. The program was able to recognize small risk indicators that were previously missed by traditional risk assessment methods by assessing a wide variety of patient data, including demographics, medical history, and physiological measurements. Similar to this, the DL approach has produced encouraging results in the field of CKD. The algorithm could accurately categorize people as being at low, intermediate, or high risk of acquiring particular types of disease by training on a variety of datasets that included patient characteristics, genetic profiles, and medical imaging data. The ability to tailor interventions and create individualized treatment programs to risk assessment may have improved patient outcomes.

4.1 Analysis

Accuracy is employed to determine the proportion of all data points that were successfully anticipated. It is calculated by dividing the total number of correctly predicted outcomes by the total number of forecasts.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (6)$$

The degree to which the reported value matches the property's "true" value is referred to as accuracy. Equation (6) is used for analyzing the result of the Accuracy. Figure 2 depicts the accuracy results. It demonstrates that our suggested technique, RGAO-DNN has a greater accuracy (95.75%) in forecasting.

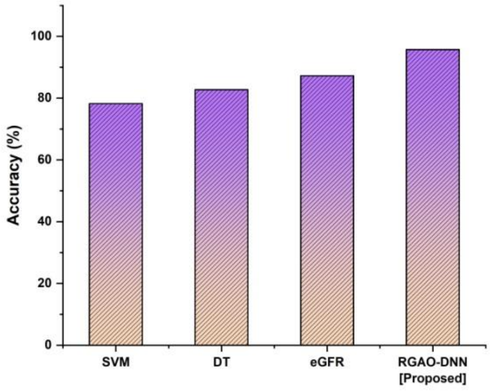


Fig 2 : Accuracy

$$Precision = T\,P/\,(T\,P + F\,P) \quad (7)$$

By using equation (7), we have analyzed the result of the precision. This metric is primarily used to indicate the reliability of the positive samples. Figure 3 depicts the precision result. This demonstrates that our suggested technique, RGAO-DNN has a better performance (92.15%) than existing work in forecasting.

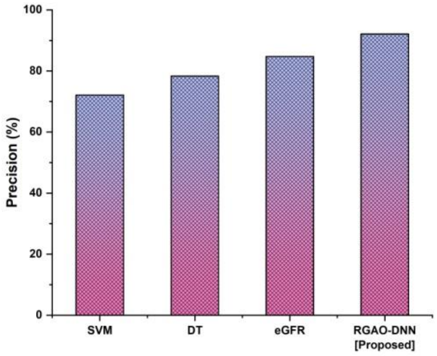


Fig 3: Precision

$$Recall = T\,P/\,(T\,P + F\,N) \quad (8)$$

By using equation (8), we have analyzed the result of the recall. The amount of time required by our suggested algorithm to complete its assigned work is referred to as the computation time. It is represented as seconds. Figure 4 depicts the outcome of computational time. This shows that when compared to the currently in use approaches, SVM, DT, and eGFR, our recommended methodology, RGAO-DNN takes less time (93.25%). This demonstrates that the model we offered will be effective.

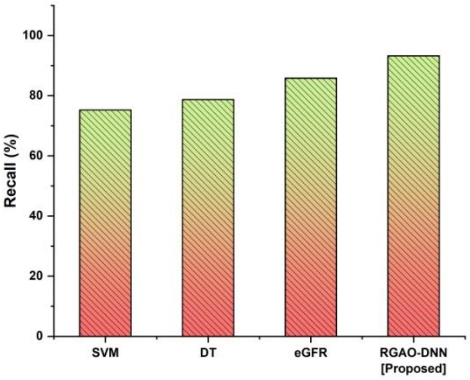
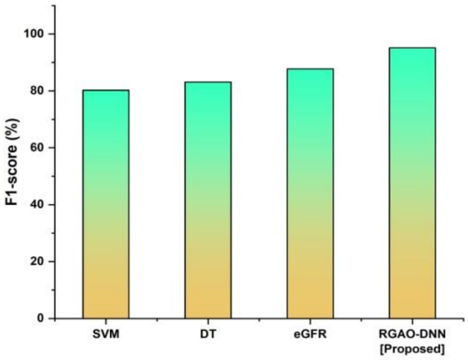


Fig 4: Recall

$$F1\,Score = 2 * (Precision * Recall)\,/\,(Precision + Recall) \quad (9)$$

By using the equation (9), we have analyzed the result of the F1-Score. Figure 5 shows that when compared to the currently in use approaches, SVM, DT, and eGFR, our recommended methodology (RGAO-DNN) performance measure (95.15%) is better and more effective.



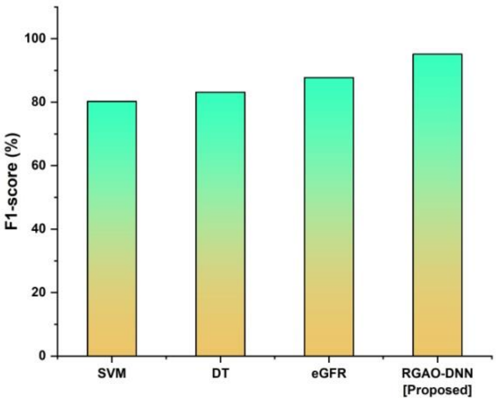


Fig 5: F1-score

5. Conclusion

In summary, utilizing the DL approach for risk assessment and disease prediction has enormous potential in the healthcare industry. The findings of numerous research show the superiority and usefulness of deep neural networks in predicting the onset, progression, and recurrence of disease as well as identifying specific risk factors. This research proposes a novel approach integrating randomized Gaussian-search Aquila optimization with Deep Neural Network (RGAO-DNN) network to enable early CKD predictions and risk assessment. Min-max normalization, Randomized Gaussian-search Aquila optimization and DNN based forecasting were analysed and used in this research to predicting CKD. The suggested method may identify important risk factors for the onset of CKD and learn from temporal patterns because it has been trained on the cleaned data. Several metrics are used to gauge the strategy's effectiveness. The results of the experiments demonstrate that the suggested method performs better in terms of CKD risk assessment and early prediction than other methods.

References

[1] Nguyen, B.P., Pham, H.N., Tran, H., Nghiem, N., Nguyen, Q.H., Do, T.T., Tran, C.T. and Simpson, C.R., 2019. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Computer methods and programs in biomedicine, 182, p.105055.

[2] Ramchandani, A., Fan, C. and Mostafavi, A., 2020. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. Ieee Access, 8, pp.159915-159930.

[3] Nemesure, M.D., Heinz, M.V., Huang, R. and Jacobson, N.C., 2021. Predictive modeling of depression and anxiety using electronic health records and a novel ML approach with artificial intelligence. Scientific reports, 11(1), pp.1-9.

[4] Ahmadlou, M., Al-Fugara, A.K., Al-Shabeeb, A.R., Arora, A., Al-Adamat, R., Pham, Q.B., Al-Ansari, N., Linh, N.T.T. and Sajedi, H., 2021. Flood susceptibility mapping and assessment using a novel deep learning model combining multilayer perceptron and autoencoder neural networks. Journal of Flood Risk Management, 14(1), p.e12683.

[5] Kavitha, C., Mani, V., Srividhya, S.R., Khalaf, O.I. and Tavera Romero, C.A., 2022. Early-stage Alzheimer's disease prediction using ML models. Frontiers in public health, 10, p.240.

[6] Bommi, K. ., & Evanjaline, D. J. . (2023). Timestamp Feature Variation based Weather Prediction Using Multi-Perception Neural Classification for Successive Crop Recommendation in Big Data Analysis. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2s), 68–76. https://doi.org/10.17762/ijritcc.v11i2s.6030

[7] Sujath, R.A.A., Chatterjee, J.M. and Hassanien, A.E., 2020. An ML forecasting model for the COVID-19 pandemic in India. Stochastic Environmental Research and Risk Assessment, 34, pp.959-972.

[8] Hu, C., Liu, Z., Jiang, Y., Shi, O., Zhang, X., Xu, K., Suo, C., Wang, Q., Song, Y., Yu, K. and Mao, X., 2020. Early prediction of mortality risk among patients with severe COVID-19, using ML. International journal of epidemiology, 49(6), pp.1918-1929.

[9] Tien Bui, D., Shahabi, H., Omidvar, E., Shirzadi, A., Geertsema, M., Clague, J.J., Khosravi, K., Pradhan, B., Pham, B.T., Chapi, K. and Barati, Z., 2019. Shallow landslide prediction using a novel hybrid functional ML algorithm. Remote Sensing, 11(8), p.931.

[10] Ali, F., El-Sappagh, S., Islam, S.R., Kwak, D., Ali, A., Imran, M. and Kwak, K.S., 2020. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion, 63, pp.208-222.

[11] Wu, C.C., Yeh, W.C., Hsu, W.D., Islam, M.M., Nguyen, P.A.A., Poly, T.N., Wang, Y.C., Yang, H.C. and Li, Y.C.J., 2019. Prediction of fatty liver disease using ML algorithms. Computer methods and programs in biomedicine, 170, pp.23-29.

[12] Guo, Q., Li, M., Wang, C., Wang, P., Fang, Z., Tan, J., Wu, S., Xiao, Y. and Zhu, H., 2020. Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm. BioRxiv,

pp.2020-01.

[13] Chowdhury, M.E., Rahman, T., Khandakar, A., Al-Madeed, S., Zughaier, S.M., Doi, S.A., Hassen, H. and Islam, M.T., 2021. An early warning tool for predicting mortality risk of COVID-19 patients using ML. Cognitive Computation, pp.1-16.

[14] Basaligheh, P. (2021). A Novel Multi-Class Technique for Suicide Detection in Twitter Dataset. Machine Learning Applications in Engineering Education and Management, 1(2), 13–20. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/14

[15] Ali, M.M., Paul, B.K., Ahmed, K., Bui, F.M., Quinn, J.M. and Moni, M.A., 2021. Heart disease prediction using supervised ML algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 136, p.104672.

[16] Chen, J.I.Z. and Hengjinda, P., 2021. Early prediction of coronary artery disease (CAD) by ML method-a comparative study. Journal of Artificial Intelligence, 3(01), pp.17-33.

[17] Assaf, D., Gutman, Y.A., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., Shilo, N., Epstein, A., Mor-Cohen, R., Biber, A. and Rahav, G., 2020. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Internal and emergency medicine, 15, pp.1435-1443.