

# Multi Disease Prediction Using HDO Machine Learning Approach

Rutuja A Gulhane<sup>1</sup>, Sunil R Gupta<sup>2</sup>

<sup>1</sup>Research Scholar, Department of CSE  
PRMIT&R, Badnera, MH, India  
gulhanerutuja@gmail.com

<sup>2</sup>Assistant Professor, Department of CSE  
PRMIT&R, Badnera, MH, India  
sunilguptase@gmail.com

**Abstract**— Several machine learning approaches can do predictive analytics on vast volumes of information in various sectors. Predictive analytics in health care is a challenging task. Still, it may ultimately aid physicians in making timely judgments about the health and handling of patients based on vast amounts of information. Breast cancer, diabetes, and heart-related disorders cause numerous fatalities worldwide, yet most of these decreases are attributable to an absence of appropriate screenings. The lack of remedial substructure and a short doctor-to-population proportion contribute to the issue above. Following WHO recommendations, physicians' ratio to affected persons should be in some range; India's doctor-to-public proportion indicates a doctor scarcity. Heart, cancer, and diabetes-related disorders pose a significant danger to humanity if not detected initially. Thus, early detection and identification of these disorders may save many lives. Using classification methods based on machine learning, the focus of this effort is to anticipate dangerous illnesses. Diabetes, heart disease, and breast cancer are discussed in this study. To make this effort easy and accessible to the general community, a web application for therapeutic tests has been developed that use machine learning to create illness predictions. In this study, a web application is created for illness prediction that employs the notion of machine learning-based forecasts for illnesses such as breast cancer, diabetes, and cardiovascular sickness.

**Keywords**- Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Hybrid meta-heuristic Technique.

## I. INTRODUCTION

The use of AI in healthcare has come a long way in the last 50 years. When combined with the vast volumes of data created by healthcare systems, developments in processing power have allowed us to go beyond the limitations of early AI, which focused on creating algorithms that could make decisions that were previously only possible for humans. Appointment scheduling, drug discovery, and disease diagnosis are just some of the many contemporary healthcare applications of AI. Primary care physicians' manual labor has decreased as AI aid has increased, according to a 2016 study. Disease prediction using machine learning AI models is not novel; there are multiple examples of this already. The reliance on data is the area of uncertainty and has not been examined sufficiently. How much information is required for a project is not always made obvious, and instead just recommendations are made. There is no foolproof method; instead, rule of thumbs and educated guesses are the best bets [1]. The "one in ten rule" is a rule of thumb that suggests having ten times as many data points as features to reduce the likelihood of over fitting in regression situations. When an AI is overfit, it is modelled too closely on the training data, leading to subpar performance on the test data. Outliers exist

in all data, and training models to accurately identify them might be counterproductive when introducing new information. A learning curve, a graph showing how a model improves with added events, can be used to get a sense of how much data is required to solve a problem. However, that is by no means exhaustive. Several techniques exist for optimizing models at varying numbers of events and features, including model selection, feature selection, and parameter tweaking. Because of their inherent ability to learn complicated non-linear correlations between input and output properties, non-linear algorithms also require a larger data set. Let's pretend a linear method works well with hundreds of occurrences per class and produces satisfactory results. In this instance, a non-linear method, such as 'random forest' or an artificial neural network, may require thousands of events per class. A important factor in deciding how much data is required is the degree of correlation between the features [2]. For example one is trying to determine whether something is human or not, and one of the qualities being considered is the number of fingers it has. The AI would need to observe fewer events if it were to make the prediction that having 10 fingers is highly correlated with being human. In truth, machine learning is an inductive process. The model can only reproduce what it has

previously observed, thus if a given scenario is absent from the training data, the model is unlikely to provide support for it.

The primary cause is the approximately one million annual deaths attributable to several diseases. In the United States, several diseases account for one out of every three deaths. (US) Nearly half of the 9.2 million persons who suffered from multiple assaults in 2016 did so suddenly and without warning. Multi illness has only one symptom: unexpected death. A multi attack occurs when the multi's internal organs or muscles begin to die, yet the multi still manages to stay alive for a while. The sections of the multi will malfunction severely, and that may cause health problems such a greater risk of sudden cardiac arrest. According to current predictions, India will have a higher prevalence of multi-disease burdens than any other country. In India, several causes account for one out of every five deaths. Over the next three years, one in three people will pass away, with a disproportionate number of young people dying (especially in India). When compared to the West, the onset of several diseases in India occurs between one and one and a half decades earlier [3]. About 45,000,000 people, according to estimates, will be impacted by many issues. More and more children in India are falling victim to many diseases. The future seems bleak if young people are as glued to technology as millions of people today. The prevalence of hypertension has been steadily increasing over the past five decades, especially in urban areas as opposed to rural ones [4]. The rate is 30% in major cities but 10%-50% in rural areas. The risk of developing several diseases increases with the prevalence of sedentary lifestyles.

The goal is to develop a reliable prediction algorithm by examining key features in order to provide the most accurate results possible when compared to scientific findings based on EMR data. Every day, the healthcare industry generates a mountain of data about health and illness, including patient treatment records. Decisions that affect the patient's health require the identification of patterns in the data. By developing efficient machine-learning-based algorithms and methodologies for processing and analyzing biological data, we can improve performance and accuracy. The approach is based on EMRs, which have been shown to be effective illness detectors and predictors. The primary objective is to enhance the efficiency and value of illness prediction and risk assessment systems.

The study is comprised of five sections which are as follows, Section 1 contains an introduction. Section 2 contains the related work that is performed in developing the model for predicting multi diseases. Section 3 includes the methodology along with the dataset and evaluation metrics. The experimental results are stated in section 4. Lastly, section 5 consists of the conclusion.

## II. RELATED WORK

Related efforts in creating the proposed model for forecasting numerous diseases are outlined here. The following are some of the findings from a review of the relevant literature that informed the design of the proposed system.

The highest reported accuracy (74.4% for the cataloguing method Arbitrary Forest) and the lowest reported accuracy (71.3% for the KNN algorithm) can both be found in the literature. In [5], the authors discussed how data mining methodologies and diabetes sub classification can be used to reduce the risk of diabetes disease by better understanding how people's lifestyle choices contribute to the development of the condition. The authors recorded data from Google Forms surveys using Naive Bayes and support vector machine techniques, with the latter showing 64.92% accuracy and the former showing 60.44% accuracy.

The authors [6] examined the use of machine learning algorithms for categorizing brain tumors. KNN, RF, SVM, and LDA machine learning algorithms were used to study and evaluate multifocal, multicentric, gliomatosis features of brain images. The results show that the support vector machines algorithm outperforms the competition by a factor of 90%.

Using the Framingham data set Kaggle, the authors [7] discussed cardiac diseases in depth and applied clustering methods as Naive Bayes, SVM, K-Nearest Neighbors and Decision Tree. Many different forms of AI for predicting the risk of cardiovascular infections were evaluated by the researchers. The KNN characterization method has the highest accuracy (83.60) in this research. [8] The proposed research uses an Information Extraction framework based on Evolutionary Learning to probe cardiovascular illness. (java programming technique for creating the development prototypical for data mining matters). This work is guaranteed to be accurate to within 86.7% of the time. [9] Using a Naive Bayes classifying strategy, the authors illustrate how they mined classified information from a cardiac disease data set in order to answer complex questions.

[10, 11] Presentation on cancer classification using machine learning techniques focused on early diagnosis as a means to save lives. Using the Pima dataset, it is discovered that cataloguing methods could accurately predict cases of diabetes with an accuracy of 76%. By using the KNN technique, predicted the occurrence of breast growth with an accuracy of 83.3% and accurately predicted cancer with an accuracy of 92.20% by using Random Forest models.

## III. PROPOSED APPROACH

This section consists of the methodology adopted by the proposed work. The aim is to develop a multi disease

prediction system to detect heart disease, diabetes and breast cancer in a single platform using the machine learning models. Combining Red Deer Algorithm (RDA) and Dragonfly Algorithm (DA) methods, a new meta-heuristic approach is presented called Hybrid Metaheuristic Algorithm (HDO). The Efficient Recurrent Neural Network (E-RNN) model is implemented by this algorithm HDO by enhancing the RNN's "number of hidden neurons" and epoch count. The improved accuracy and precision help them diagnose heart disease and breast cancer with a higher rate of success.

Taking advantages of RDA and DA methods, a new HDO algorithm is proposed. In this work, RDA is selected due to its fast arrival at global optimum solutions, improved convergence rate, and superior balancing between the exploration and exploitation phases. However, RDA has issues with discovering global solutions and fine-tuning a variety of "controlling parameters, fewer execution speeds" [14]. Therefore, a novel HDO algorithm is presented that uses the DA technique's features to increase harmony between the algorithm's phases, broaden the range of possible solutions, and boost its performance. There are fewer restrictions on how you can do it, therefore it's easier. In an effort to provide optimal solutions at a manageable convergence rate [12], DA with RDA is employed.

The hybrid metaheuristic (HDO) algorithm is suggested based on the sensing area, where the random vector uniformly distributed in the range [0, 1] used in DA is used for updating the solutions  $rnd > 0.5$ . Here, the random number  $rnd$  is formulated by determining the ratio between the best fitness and the mean of best fitness solutions in Eq. (1). When this condition holds, either the DA method or the RDA algorithm is used to update the answers and solutions, respectively [13].

$$rnd = \frac{M^{best}(j)}{\text{mean}(M^{best}(j))} \quad (1)$$

Here, the mean of best fitness solutions  $M^{best}(j)$  is known, and the best fitness solutions are given  $M^{best}(j)$ . If  $rnd > 0.5$  is satisfied, then the solutions are updated using DA technique.

The dragonfly's instinct to find food drives DA. In addition, it takes into account "dynamic and static behaviors of dragonflies." To adjust the location of swarm members, five rules are used: "control cohesion," "alignment," "separation," "attraction (towards food sources), and "distraction (towards external opponents). The number of dragonflies can be calculated with Eq. (2).

$$M_i^{j+1} = M_i^j + \Delta M_i^{j+1} \quad (2)$$

$$\Delta M_i^{j+1} = (soSo_i + avAV_i + csCS_i + hsHs_i + enEn_i) + \zeta \cdot \Delta M_i^j \quad (3)$$

In the equations as mentioned above, the cohesion weight is derived  $cs$ , iteration counter is referred to as  $j$ , the position of dragonflies is given as  $M_i^j$  and inertia weight is known as  $\zeta$ , the  $i^{th}$  dragonfly food source is indicated as  $Hs_i$ , velocity or step vector is termed as  $\Delta M_i^{j+1}$ , the cohesion and alignment of the  $i^{th}$  dragonfly are represented as  $CS_i$  and  $AV_i$ , respectively, and the food factor is suggested as  $hs$  and  $i^{th}$  dragonfly enemy position is specified as  $En_i$ , the alignment weight is noted as  $av$  and the  $i^{th}$  dragonfly's separation and the separation weight are identified as  $So_i$  and  $so$ , respectively, and the enemy factor is given as  $en$ . These constraints are formulated as follows.

$$So_i = -\sum_{t=1}^{Nj} (M - M_t) \quad (4)$$

$$AV_i = \frac{\sum_{t=1}^{Nj} Vo_t}{Nj} \quad (5)$$

$$CS_i = \frac{\sum_{t=1}^{Nj} M_j}{Nj} - M \quad (6)$$

$$Hs_i = Fo_m - M \quad (7)$$

$$En_i = Ey_m + M \quad (8)$$

Here, the position of the food source is noted as  $Fo_m$ ,  $Ey_i$  denotes the enemy's position, the  $t^{th}$  neighboring individual's position is given as  $M_j$ , the current individual's position is indicated  $M$ , the velocity is represented as  $Vo_i$   $t^{th}$  at the adjacent individual, and the count of neighboring individuals is considered as  $Nj$ .

The RDA method is used to update the solutions if the condition  $rnd \leq 0.5$  holds. As a result of "abnormal mating behavior in a breeding season," RDA is being applied for Scottish red deer. The population of Red Deers (RDs) is first established, and then split into subpopulations such as "hinds and male RDs." Additionally, a group of female RDs can be referred to as a harem. The hunting behavior of RDs is inferred from observations of their aggressive and roaring behaviors.

The best results can be achieved through roaring by male RDs. Every male RD's position is swapped and then equalized in Eq. (9).

$$M_{new} = \begin{cases} M_{old} + bs_1 \times (((ub - lw) * bs_2) + lw), & \text{if } bs_3 \geq 0.5 \\ M_{old} - bs_1 \times (((ub - lw) * bs_2) + lw), & \text{if } bs_3 < 0.5 \end{cases} \quad (9)$$

Here, the current position and updated position of male RD are correspondingly specified as  $M_{old}$ ,  $M_{new}$  constraints,  $bs_1$ ,  $bs_2$  and  $bs_3$  are generated among 0 and 1 through a uniform distribution in a random way [15].

Finally, the solutions are updated until reaching the final answers. The “pseudo-code” of the developed hybrid metaheuristic (HDO) is given in Algorithm 1.

breast cancer illnesses. This data collection consists of the thirty-one medical forecaster characteristics and one goal feature [18].

Here, the prototypical is skilled in forecasting whether or not an individual has heart disease using some standard inputs like age, chest pain, maximum heart rate, and so on.

TABLE I. CONSIDERED DATASET FOR PREDICTION

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

### Data Analysis

It is a tidy, straightforward collection of facts. Nevertheless, the significance of a few column headings is obscure. Here is their meaning:

- age: In years, the age of person
- sex: The sex of the person, (Value 1: male, Value 0: Female)
- cp: Experiencing chest discomfort (Value 1 = typical angina, Value 2 = atypical angina, Value 3= non-anginal pain, Value 4 = asymptomatic)
- trestbps: The blood pressure of a person at rest (mm Hg on entry to the Hospital)
- chol: The individual’s cholesterol, measured in mg/dl
- fbs: The individual’s fasting blood glucose (> 120 mg/dl, 1 = true; 0 = false)
- restecg: Resting electrocardiographic record (0 = normal, 1 = having ST-T wave irregularity, 2 = viewing possible or certain left ventricular hypertrophy by Estes’ standards)
- thalach: The individual’s extreme heart rate attained
- exang: Workout persuaded angina (1=yes;0=no)
- oldpeak: ST sadness persuaded by workout relation to relaxation (‘ST’ narrates to situations on the ECG design)
- slope: the gradient of the topmost workout ST section (Value 1 = upsloping, Value 2 = flat, Value 3 = downsloping)
- ca: The amount of key vessels (0-3)
- thal: a blood illness named thalassemia (3: normal, 6: fixed defect, 7: reversible fault)
- target: Heart illness (0: no, 1: yes)

### B. Evaluation Metrics

The model will use variety of assessment measures such as:

**Algorithm 1:** Proposed HDO Algorithm

Initialization of the population  
 Computation of the fitness among individuals  
 While  $j < j_{\max}$   
   if  $rnd > 0.5$   
     **Solutions are updated based on DA.**  
       Update the solution vector by Eq. (2).  
       Update the velocity vector by Eq. (3).  
   Else  
     **Solutions are updated based on RDA.**  
       for every male RD  
         Utilize Eq. (8) for roaring  
         Update the better positions  
       end for  
       Update stage and commanders using Eq. (9)  
       for every male commander  
         Use Eq. (3) for fighting among stages and commanders.  
         Update position of commanders and stages  
       end for  
       for every stage  
         Determine the distance through Eq. (8)  
         Update Eq. (9) for the mating stage with selected hind  
       end for  
       Use a roulette wheel strategy for choosing the next generation.  
     end if  
   Update the best solutions  
 end while  
 Return optimal solutions

### A. Dataset

Heart Dataset: The UCI "Cardiac Disease Dataset" is used to forecast the prevalence of heart illnesses. This data collection contains thirteen medical predictor characteristics and a single goal feature [16]. The attributes are shown in Table I as follows age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal.

Diabetes Dataset: A dataset from Kaggle is used to forecast the prevalence of Diabetic illnesses. This information set contains the eight medical forecaster characteristics and a single goal feature. The characteristics are Blood Pressure, Prenatal Period, Glucose, Skin Thickness, BMI, Insulin, Phase, and Diabetes Pedigree [17].

Breast Cancer Dataset: Kaggle "Breast Cancer Wisconsin (Diagnostic) Data Set" is used to forecast the incidence of

- Accuracy: This can be calculated using the following formula and indicates how close an extent is to the true value.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$

- Precision: Based on the consistency of data from several measurements, you may calculate this using the following formula.

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

- Recall: This may be calculated using the method and indicates how well the model classifies the relevant outcomes.

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

#### IV. EXPERIMENTAL RESULTS

This section shows the experimental results carried out during the work which is done by using the laptop equipped with i5 CPU & the python code. The methods used are different from the exiting approaches and precisions are computed with a cv factor of 10. Using the bar graphs, accuracy of the respective illness is shown. The sickness datasets are separated into exercise and divided into the training and the testing datasets for the classification purpose [19].

The following Fig. 1 shows the frequency of occurrences of the heart disease depends on gender and the Fig. 2 shows the performance analysis of the novel approach hybrid metaheuristic technique (HDO) with the existing machine learning approaches.

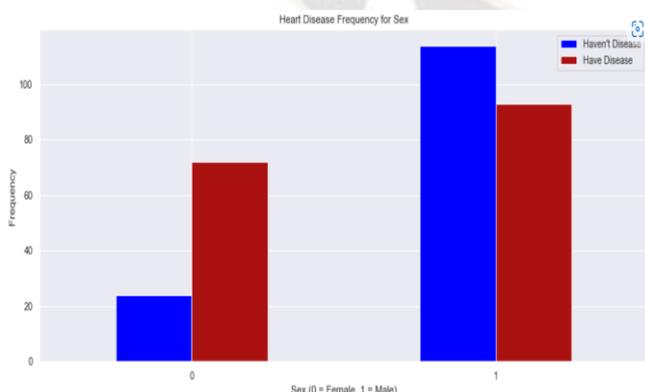


Figure 1. Frequency of Occurrence of Heart Disease depends on Sex.

TABLE II. PERFORMANCE ANALYSIS OF THE PROPOSED APPROACH

Disease	LR	SVM	KNN	HDO (Novel Approach)
Heart Disease	82.50	78.87	83.84	85.30
Diabetes	77.60	75.64	75.52	79.12
Breast Cancer	94.55	91.38	92.55	95.67

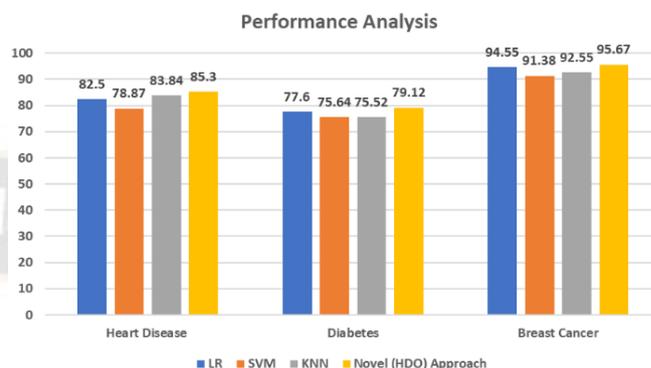


Figure 2. Performance analysis of Novel Approach (HDO) with exiting ML Approaches.

The proposed work surpasses the precisions of previously proposed models for predicting the diseases. The most excellent accuracy levels reached by using the hybrid meta heuristic technique for heart disease, diabetes and breast cancer are 85.30 %, 79.12 %, and 95.67 % as shown in Table II, utilizing the Logistic Regression for Diabetes & Breast Cancer and the K-Nearest Neighbor for Heart Disease.

#### V. CONCLUSION

The comprehensive analysis of the existing healthcare systems has considered only one disease at a time. For example, one system is used to test for the diabetes, another is used to test for the diabetes retinopathy, and another is used to predict the heart disease and so on. The most systems concentrate on a certain illness. An organisation must use a variety of models when analysing the health reports of its patients. The strategy under the existing system is helpful for the analysis of certain illnesses. The proposed multi disease prediction system is useful for analysing more than one disease on a single website with the highest accuracy. Because of this the user doesn't need to traverse different websites which saves time as well.

The approach consolidates diabetes, cardiovascular illness, and breast cancer into a sole stage by organizing competent representations utilizing a lightweight, flask API outline. Three cataloging methods are employed to train the models, with the hybrid meta heuristic technique (HDO) demonstrating the high accuracy levels for predicting diabetes, breast cancer, and heart disease. The hybrid metaheuristic algorithm's most significant degree of precision is determined by selecting the highest value among 1 to 21 neighbors. In the future, it will be able to extend the work by adding new illnesses learned by

machine learning models and diseases that deep learning models train.

## REFERENCES

- [1] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, & M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals", *Information Sciences*, Vol. 415–416, pp. 190–198, 2017.
- [2] S. Ahmed, K. Y. Choi, J. J. Lee, B. C. Kim, G. R. Kwon, K. H. Lee, & H. Y. Jung, "Ensembles of Patch-Based Classifiers for Diagnosis of Alzheimer Diseases", *IEEE Access*, Vol. 7, pp. 73373–73383, 2019.
- [3] A. Al-Zebari, & A. Sengur, "Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection", In: *Proc of 1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC*, pp. 2–5, 2019.
- [4] P. Bagga, & R. Hans, "Applications of mobile agents in healthcare domain: A literature survey", *International Journal of Grid and Distributed Computing*, Vol. 8, No. 5, pp. 55–72, 2015.
- [5] R. Chitra, V. Anuja Kumari, "Classification Of Diabetes Disease Using Support Vector Machine", *International Journal of Engineering Research and Applications*, Vol. 3, No. 2, pp. 1797–1801, 2013.
- [6] G. Cinarer & B. G. Emiroglu, "Classification of Brain Tumors by Machine Learning Algorithms", In: *Proc of 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT*, 2019.
- [7] L. Meng, S. Ding, N. Zhang, & J. Zhang, "Research of stacked denoising sparse autoencoder", *Neural Computing and Applications*, Vol. 30, No. 7, pp. 2083–2100, 2018.
- [8] I. M. Najim Adeen, A. M., Abdulazeez, & D. Q. Zeebaree, "Systematic review of unsupervised genomic clustering algorithms techniques for high dimensional datasets", *Technology Reports of Kansai University*, Vol. 62, No. 3, pp. 355–374, 2020.
- [9] S. M. Naqi, M. Sharif, & A. Jaffar, "Lung nodule detection and classification based on geometric fit in parametric form and deep learning", *Neural Computing and Applications*, Vol. 32, No. 9, pp. 4629–4647, 2020.
- [10] L. Rajabion, A. A. Shaltoolki, M. Taghikhah, A. Ghasemi, & A. Badfar, "Healthcare big data processing mechanisms: The role of cloud computing", *International Journal of Information Management*, Vol. 49, pp. 271–289, 2019.
- [11] M. Sameti, R. K. Ward, J. Morgan-Parkes and B. Palcic, "Image Feature Extraction in the Last Screening Mammograms Prior to Detection of Breast Cancer", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 3, No. 1, pp. 46–52, 2009.
- [12] J. Mohammad and M. H. Bayati Chaleshtari, "Using dragonfly algorithm for optimization of orthotropic infinite plates with a quasi-triangular cut-out", *European Journal of Mechanics A/Solids*, Vol. 66, pp.1–14, 2017.
- [13] A. M. Fathollahi-Fard, M. Hajiaghahi-Keshteli, and R. Tavakkoli-Moghaddam, "Red deer algorithm (RDA): a new nature-inspired meta-heuristic", *Soft Computing*, Vol. 24, pp. 14637–14665, 2020.
- [14] H. Saleh, S. F. Abd-el ghany, H. Alyami, and W. Alosaimi, "Predicting Breast Cancer Based on Optimized Deep Learning Approach", *Computational Intelligence and Neuroscience*, Vol. 8, pp. 1–11, 2022.
- [15] G. Huang, Z. Liu, L. van der Maaten, Q. Kilian Weinberger, "Densely Connected Convolutional Networks", *Computer Vision and Pattern Recognition*, 2016.
- [16] S. Vijayarani, & S. Dhayanand, "Data Mining Classification Algorithms for Kidney Disease Prediction", *International Journal on Cybernetics & Informatics*, Vol. 4, No. 4, pp. 13–25, 2015.
- [17] H. Tjahjadi, K. Ramli, "Noninvasive blood pressure classification based on photo plethysmography using K-nearest neighbors algorithm: A feasibility study", *Information*, Vol. 11, No. 2, pp. 1–18, 2020.
- [18] D. Pedrozo, F. Barajas, A. Estupiñán, K.L. Cristiano, D.A. Triana, "Data analysis for a set of university student lists using the k-Nearest Neighbors machine learning method", *Phys. Conf. Ser.*, Vol. 1514, No. 1, pp. 1–8, 2020.
- [19] A. Assegie, P.S. Nair, "Handwritten digits recognition with decision tree classification: a machine learning approach", *Int. J. Electr. Comput. Eng.*, Vol. 9, No. 5, pp. 1–4, 2019.