# Performance Analysis of Supervised Machine Learning Algorithms on Medical Dataset

**Amit Juyal, Chetan Pandey, Janmejay Pant, Ankur Dumka, Vikas Tomar**

***Abstract*: Machine learning (ML) algorithms are designed to perform prediction based on features. With the help of machine learning, system can automatically learn and improve by experience. Machine learning comes under Artificial intelligence. Machine learning is broadly categorized in two types: supervised and unsupervised. Supervised ML performs classification and unsupervised is for clustering. In present scenario, machine learning is used in various areas. It can be used for biometric recognition, hand writing recognition, medical diagnosis etc. In medical field, machine learning plays an important role in identifying diseases based on patient's features. Presently, doctors use software application based on machine learning algorithm in various disease diagnosis like cancer, cardiac arrest and many more. In this paper we used an ensemble learning method to predict heart problem. Our study described the performance of ML algorithms by comparing various evaluating parameters such as F-measure, Recall, ROC, precision and accuracy. The study done with various combination ML classifiers such as, Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF) algorithm to predict heart problem. The result showed that by combining two ML algorithm, DT with NB, 81.1% accuracy was achieved. Simultaneously, the models like Support Vector machine (SVM), Decision tree, Naïve Bayes, Random Forest models were also trained and tested individually.***

***Keywords: Cardiovascular Disease, Ensemble Learning, Machine Learning. Naïve Bayes, Decision Tree.***

## I. INTRODUCTION

In present scenario cardiovascular disease [1] become a common and dangerous disease. If it is not diagnosed on time it may be dangerous for the heart patient. This disease effects so instantly to the heart patient that the person hardly get time to recover. So it is important to diagnose this disease [15] as early as possible to start treatment on time. To discover hidden patterns from medical data to predict disease is challenging task and machine learning [4] can perform well in identifying hidden patterns from medical dataset and predict heart disease [2] accurately.

Analysis of medical related data and processing [13] [17] the outcomes are highly beneficial for mankind.

Since last two decades researchers from different areas specially medical and computer science has extracted so many useful information which not only helping doctors in understanding the post problems related to patients but also providing scope for inventors in the process of discovery of tools and techniques which was and will be beneficial for human beings. Although in prior research there is lack of quality data and also technology was not so advanced. However as time moves forward, the technology related to mining of medical data improves [11] [14] which allows researcher to collect data in wider area. Other technologies like Hadoop for big data and machine learning algorithms significantly raises the probability of development of new ways to summarize the medical data and extract useful information out from it. However since last decade there is dramatic change observe in this area as now researchers not only using existing techniques but also sharing and suggesting improved process of analysis of medical data. After analyzing past paper authors of this paper discovered that Naive Bayes [1] will be highly accurate and here authors make use of other algorithms and found that the later gives better accuracy against Decision tree and K- Nearest Neighbor. However some authors investigate about quality data and its attribute which plays a vital role in knowledge discovery. One of the paper [2] identify that bagging algorithm got highest accuracy of 85.03%. To achieve this authors investigated different classifiers on medical data which has less attribute. Since last few years many researchers initiated to applying ensemble learning instead of a machine learning over a medical dataset. Since this time groups of different machine learning algorithms are implemented, the outcomes of it was in the form of more informative and accurate information. Adding to that, some authors [3] used Random Forest along with Support Vector Machine (SVM) for their study. According to them [3] a cardiovascular-patient classifier is proposed to classify heartbeat in which ECG signals was used to obtain from Hypertrophic cardiomyopathy (HCM) patient and from non-HCM patients.

Apart from ensemble learning, Neural Networks are also widely used by researchers for classification of data and obtaining knowledge out of it. Taking this into account, researchers proceeded with such models, for example convolutional neural network (CNN), for the categorization of a data on the basis of attributes provided by it. Authors in a paper [4] analyzed Heart rate variability (HRV) and Pulse transit time variability (PTTV) on both non effected or effected patients by using a model based on CNN.

**Revised Manuscript Received on February 15, 2020.**

***** Correspondence Author

**Amit Juyal***, School of Computing, Graphic Era Hill University, Dehradun, India. Email: amitjuyal26@gmail.com

**Chetan Pandey**, School of Computing, Graphic Era Hill University, Dehradun, India. Email: schetanpandey@gmail.com

**Janmejay Pant**, Dept. of Computer Science, Graphic Era Hill University, Bhimtal, India., Email: jpant@gehu.ac.in

**Dr. Ankur Dumka**, Dept. of CSE, Graphic Era Deemed to be University, Dehradun, India. Email: ankurdumka2@gmail.com

**Vikas Tomer**, Dept. of CSE, Graphic Era Deemed to be University, Dehradun, India. Email: vikas.tomer123@gmail.com

However here [4] authors figured out that by using support vector machine the proposed model can achieve around 90% of accuracy by incorporating HRV and PTTV factors for $P(M/N) = \frac{(P(N/M) * P(M)}{P(N)}$ classification. Another paper [5] proposed a model based on SVM along with differential algorithm. Authors proposed a model using Support vector machine and differential algorithms to predict blood glucose level by observing only continuous glucose monitoring data regard less other factors like meal, insulin level, and emotional condition. There are some research which are based on predicting something on the basis of some information. For instance authors in one paper [6] proposed an intelligent ML based predictive model to diagnosis heart disease. Here they tested their model on Cleveland heart disease dataset and proposed a system that can able to classify people who suffered from heart disease from normal people. Nowadays machine learning algorithms are widely used in medical research either to analyzing some data to conclude something or to estimate some results out of it. Since there is a dramatic variations in the technology since last decades, medical images are also gathered along with medical data records. Authors of [7] and [8] focuses more on medical images and use machine learning algorithms for their proposals. As mentioned in [7], ML in artificial intelligence proved to be one of the promising techniques to provide self-interpretation of images, quality extraction of data and predictions for a patient related to previous medical records. Here medical records may be in the form of images [8] since by using deep learning system, it is accurate to make assessments regarding either approaches towards a system in which predictions are based on the processing of images (for example injury images),] or recommending some medical facilities to them.

## II. METHODS

Decision tree supervised ML [18] is a popular and well known. Decision tree has a hierarchical structure. In which each node represents features and each link or branch represents a decision rule. Leaf represents actual classified class. Generally decision tree [11] is used for classification problem. Decision tree can be implemented based on available and popular algorithms like ID3, C4.5, CART and J48. Two steps involved in this techniques first build a tree. Secondly trained and tested the tree to the dataset.

Support Vector machine (SVM) is a supervised ML model used for both classification and regression. SVM draws hyperplane between data-points to perform classification. Hyperplane should be drawn in such a manner that it best fit between two classes. The nearest data point to the hyperplane are known as support vectors. The distance between support vectors and hyperplane are known as margin. So in SVM the main objective is to draw a best fit hyperplane in N-dimensional space. Hyperplane that has maximum margin can best classify two classes.

Naïve Bayes, well-known supervised ML algorithm, Naïve Bayes (NB) commonly used for classification. In this method classification [12] performed on Bayes theorem. According to NB theorem, probability of an event happening can be find of the probability of some other event has already occurred.

In mathematically we can express theorem as: Finding probability of an event M when event N has already occurred.

Ensemble machine learning [9] is a technique of combing more than one machine learning algorithm to produce more accurate prediction. In this method multiple classifiers are involved in classification. There are many types of ensemble methods, BAGGing or Bootstrapp, Boosting, Voting and Stacking etc. In our study we have used Voting technique. Voting is an ensemble machine learning algorithm. In this method multiple models participate in prediction [10]. The output of one model is consider as vote. So the final output or prediction is based on majority of the votes of models. To understand voting concept consider a scenario to predict tomorrow will be rain or not. Suppose output of one model says tomorrow will 65% chance of raining. If we combine more than one model and majority of model says tomorrow will 80% chance of raining. Then we can more confident that tomorrow will be rain.

Random Forest is an example of ensemble machine learning. It is used for both classification and regression. It combines more than one Decision tree for class prediction. It takes output or prediction of all the decision tree and produces final model prediction based on majority of prediction. Suppose RF consist of 5 decision tree after testing 3 out of 5 DT predicting tomorrow will be rain and 2 DT predicting no rain. So in this case majority is predicting tomorrow will be rain will become model's prediction.

### A. EVALUATION METHOD

Accuracy is a model evaluation technique. It is used to evaluate that overall how much percentage our classifier predicting correctly. It is calculated in percentage which can be found out as True Positive upon total. Accuracy is not perform well when there is skewed data. But if data is not skewed we can take accuracy for model evaluation.

$$\text{Accuracy} = \frac{True\ Positive\ (TP)}{TP+TN+FP+FN(Total)}$$

Precision is another evaluation technique that checks the positive rate of a classifier or a model is how much correct.

$$\text{Precision} = \frac{TP}{TP+FP(predicted\ yes)}$$

Recall also known as True Positive Rate, can be calculated as TP upon actual yes

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-measure, combines precision and recall to evaluate ML models. It is also known as harmonic mean of precision and recall. F- Measure can be calculated as

$$\text{F-measure} = \frac{2*Precision*Recall}{Recall+Precision}\ \text{F-measure} = \frac{2*TP}{2*TP+FP+FN}$$

Receiver Operating Characteristics (ROC), it is a plot of sensitivity against 1-specificity. In this evaluation method a curve is drawn between sensitivity and 1-specificity. The area under the curve is consider as a measure of accuracy of model.

### B. METHODOLOGY

In this paper, the medical dataset has been obtain from UCI repository. Original dataset contains missing values and noise. Data set have 14 attributes like Age, gender, bp etc. and class is identified in disease column. So in this dataset we have 14 columns and 270 rows.

Mostly dataset have noisy data, missing values, and inconsistent data. If we used inconsistent or noisy data to train our model then our model will incorrectly classify. So the data preprocessing is an important step. Steps involves in data preprocessing is figure 1.

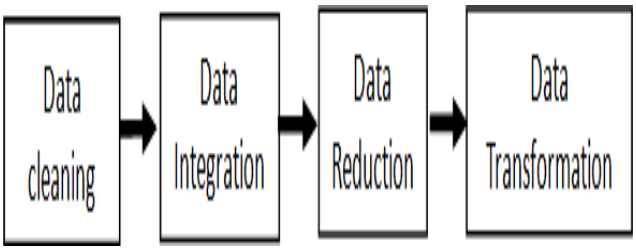


**Fig: 1- Sequence involved in data preprocessing**

To process dataset first we perform data cleaning. In data cleaning first missing and irrelevant parts are identified. In dataset if the multiple values are missing in some rows than we can remove that entry from dataset. We can take average of present values to fill missing values or we can take mean squared values to fill missing values. Another task in data cleaning is to remove noisy data. Suppose age is an attribute its expected value is integer type but in place of valid integer value some string value is present then this is called noise. Noisy data can be replaced by using various method such as mean of all values in column or by using boundary values.

Second step in data preprocessing is data integration. In this step scattered data is integrated into single sequential unit to represent data into same range. In our work we skip this step because our dataset is not scattered.

To handle huge amount of data high computation resource are needed. Analysis becomes tedious task when data is big. So data reduction is an important step when we have huge amount of data. Third step in data preprocessing is data reduction. Data reduction is a technique to handle huge dataset. In our case we skip this step in data preprocessing because we study is based on small dataset.

Data transformation is to transform raw data into dataset according to the need of work. Various techniques are used in Data transformation like smoothing, aggregation and generalization. I our work we have used data transformation to trained our model.

After preprocessing we have split dataset into 70% training and 30% testing datasets.

In this study we have divided our work into two phases. In first phase we trained different ML algorithms such as Naïve Bayes, SVM, Decision tree and Random forest [16] on medical dataset. In which we have studied the performance (accuracy, F-measure, ROC, and precision) of each model. When we implement Naïve Bayes on medical dataset. We observe that Naïve Bayes individually achieved around 79% accuracy. Similarly we have implemented SVM, Decision tree and Random forest. We achieved 70% accuracy for Decision tree, 77.4074% for SVM and 79.2593% for Random forest. Than we form ensemble models using different combinations like NB with DT, NB with SVM, NB with RF, DT with RF, DT with SVM and RF with SVM.

We found that NB with DT achieved better performance in all evaluation criteria. Overall process is shown in following figure:

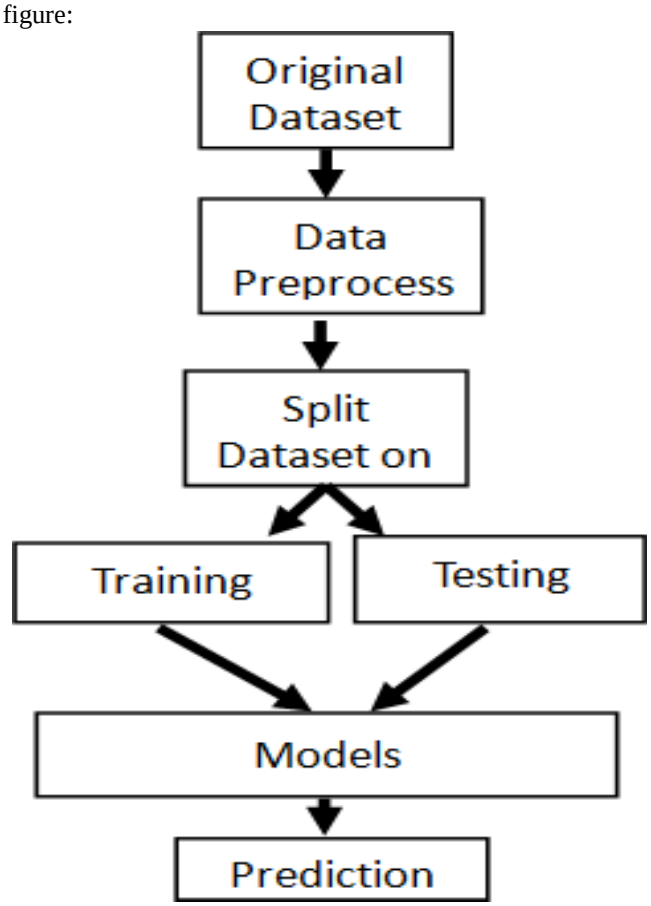


**Fig 2: Flowchart of the present study**

Confusion matrix of Ensemble model (DT+NB)

| Predicted Values | | a=1 | b=2 |
|---|---|---|---|
| Actual values | a=1 | 125 | 25 |
| | b=2 | 26 | 94 |

Observing confusion matrix we can see that model correctly predicted that 125 person not suffering from heart problem and in actual they are not suffering from heart problem while 25 person were incorrectly predicted that they are suffering from heart problem an In actual they are not suffering from heart problem. Model correctly predicted that 94 person suffering from heart problem and in actual they are suffering from heart problem while for 26 person model wrongly predicted that they are not suffering from have heart problem while they actually have heart disease.

## III. RESULT

**Table- I: Model evaluation values of different classifiers**

| Classifiers | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC | Accuracy |
|---|---|---|---|---|---|---|---|
| SVM | 0.77 | 0.232 | 0.774 | 0.774 | 0.774 | 0.542 | 77.4074 |
| Random Forest | 0.7 | 0.328 | 0.701 | 0.7 | 0.694 | 0.735 | 70 |
| Decision Tree (DT) | 0.793 | 0.221 | 0.792 | 0.793 | 0.791 | 0.809 | 79.2593 |
| Naïve Bayes (NB) | 0.793 | 0.218 | 0.792 | 0.793 | 0.792 | 0.891 | 79.2593 |
| Ensemble:NB+DT | 0.811 | 0.194 | 0.811 | 0.811 | 0.811 | 0.891 | 81.1111 |
| Ensemble:NB+RF | 0.796 | 0.215 | 0.796 | 0.796 | 0.796 | 0.861 | 79.6296 |
| Ensemble:NB+SVM | 0.744 | 0.232 | 0.774 | 0.774 | 0.744 | 0.886 | 77.4074 |
| Ensemble:DT+RF | 0.796 | 0.208 | 0.796 | 0.796 | 0.796 | 0.854 | 79.62 |
| Ensemble:DT+SVM | 0.752 | 0.259 | 0.751 | 0.752 | 0.751 | 0.821 | 75.1852 |
| Ensemble:RF+SVM | 0.748 | 0.258 | 0.748 | 0.748 | 0.748 | 0.806 | 74.8148 |

For training and testing heart disease dataset was used. This dataset generally used for implement model based on machine learning. This dataset is available in UCI repository. Data set contain various information like blood pressure, serum cholesterol, age, sex etc. 270 observation present in data set and each observation have 13 attributes. Dataset contains positive and negative observations. These 270 observations split into 70 % training and 30% for testing.

In ML different model evaluation techniques are used. After applying different classifiers on dataset we got different values of model evaluation parameters. By comparing results of different classifiers we can see that ensemble learning model (NB+DT) perform better than all other models.

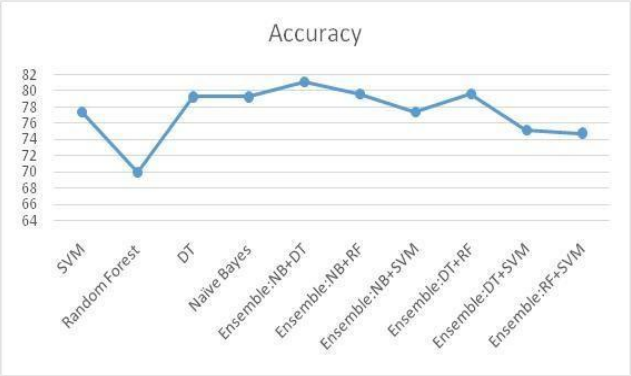


**Fig3: Accuracy graph of different ML models.**

This graph shows different accuracy achieved by ML models. It is clearly shown in the graph that ensemble model (Naïve Bayes with Decision Tree) got highest accuracy 81.11%. Another model which is closed to this accuracy is again an ensemble model (Naïve Bayes with Random Forest), which is getting 79.62%.

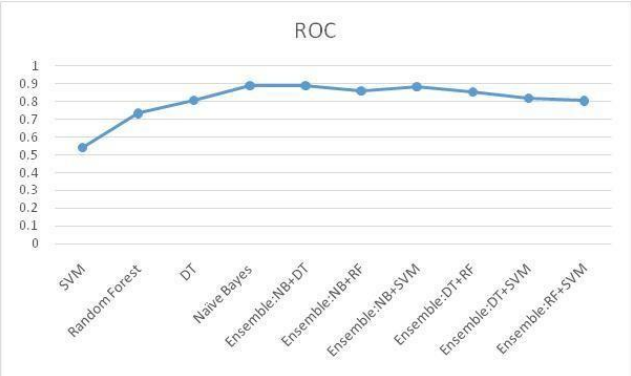


**Fig 4: Showing different percentage achieved by ML models.**

ROC is a technique to evaluate ML model performance. It is plotted with True Positive Rate (TPR) against False positive Rate(FPR).This figure showing around 89% ROC valueachieved by Ensemble model (Naïve Bayes with Decision Tree). By analysis this graph SVM is getting least performance around 54%. Naïve Bayes and Ensemble model (Naïve Bayes with Decision Tree) both of these ML models achieving equal percentage 89%.

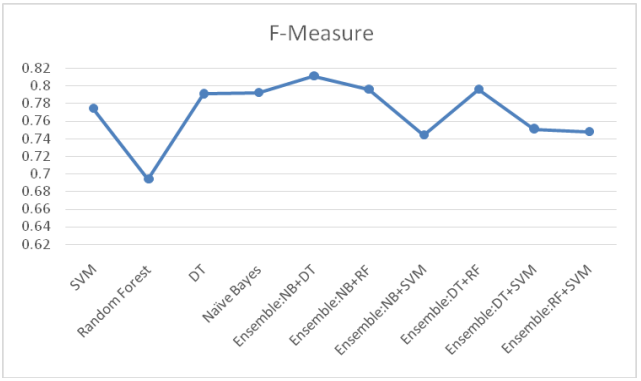


**Fig 5: F-measure or F-Sore of different ML models.**

F-score or harmonic mean of precision and recall is another ML models evaluation technique. It is used to evaluate model's accuracy by calculating mean of precision and recall. By using both precision and recall, F-score give more accurate model performance. It is clearly shown in fig. 5 that Ensemble model (NB with DT) getting better F-score in comparison of other models.

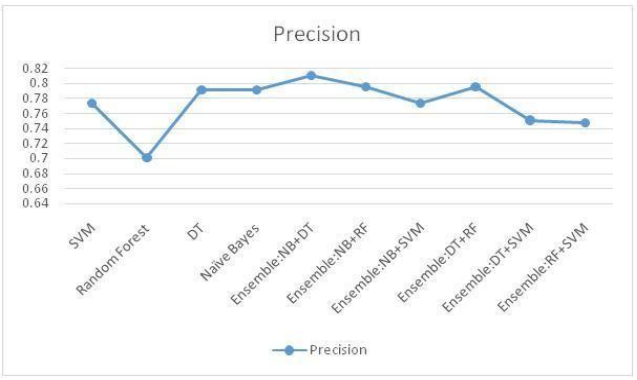


**Fig 6: This figure showing precision score achieved by ML models.**

Precision is used to test the positive result of a model. In our study we observed that Random Forest has least precision rate. In the figure Ensemble model (NB with DT) is getting 81% precision.


**Fig 7: Recall percentage of different ML models.**

Recall is also known as sensitivity. It is use to test the ML model's positive prediction results, whether class predicted yes is actually positive. In the *figure 7* highest Recall score achieved by Ensemble model (NB+DT). This means that the model predicting 81% correctly.

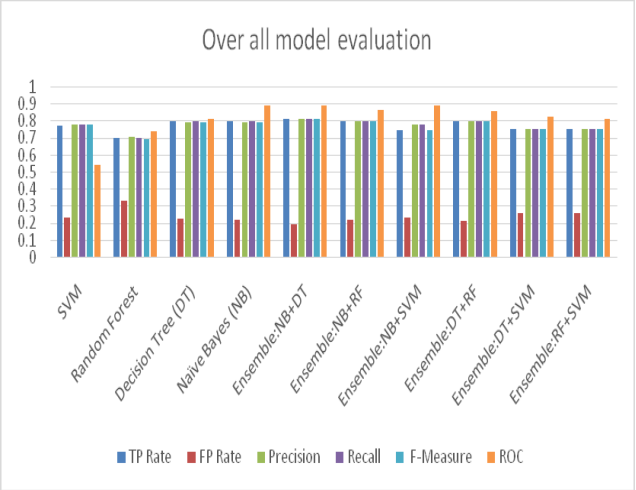


**Fig 8:Overall performance of various machine learning models**

It is clearly shown in *figure 8* that in all ML model evaluation parameter, Ensemble learning (NB with Decision Tree) achieved better results in comparison of all other ML models.

## IV. CONCLUSION

In modern scenario machine learning algorithms are used for classification and prediction in medical field. Medical data contains hidden information that can be useful for knowledge discovery. Datasets contains different type of information in which some information is not useful. Data can have different nature it may be images, numeric data and categorical data. To find which model will perform well in particular data set, a comparative study is useful. In this work, first we split dataset into 70% training and 30% into testing. Than we trained different ML models with training dataset and predict the heart disease. We concluded that NB algorithm with Decision tree, which is an ensemble models achieved better result than all other models. To justify our observation we tested ML models on different evaluation parameter such as precision, recall, F-measure and ROC.

## REFERENCES

1. Peter, T. John, and K. Somasundaram. "An empirical study on prediction of heart disease using classification data mining techniques." In IEEE-International conference on advances in engineering, science and management (ICAESM-2012). IEEE, 2012, pp. 514-518.
2. Chaurasia, V. and Pal, S., 2014. Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, pp.56-66.
3. Rahman, Q.A., Tereshchenko, L.G., Kongkatong, M., Abraham, T., Abraham, M.R. and Shatkay, H., 2015. Utilizing ECG-based heartbeat classification for hypertrophic cardiomyopathy identification. IEEE transactions on nanobioscience, 14(5), pp.505-512.
4. Zhao, L., Liu, C., Wei, S., Liu, C. and Li, J., 2019. Enhancing Detection Accuracy for Clinical Heart Failure Utilizing Pulse Transit Time Variability and Machine Learning. IEEE Access, 7, pp.17716-17724.
5. Hamdi, T., Ali, J.B., Di Costanzo, V., Fnaiech, F., Moreau, E. and Ginoux, J.M., 2018. Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. Biocybernetics and Biomedical Engineering, 38(2), pp.362-372.
6. Haq, A.U., Li, J.P., Memon, M.H., Nazir, S. and Sun, R., 2018. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018.
7. Arman Kilic, " Artificial Intelligence and Machine Learning in Cardiovascular Healthcare", Kilic A, Artificial Intelligence and Machine Learning in Cardiovascular Healthcare, The Annals of Thoracic Surgery (2019), doi: https://doi.org/10.1016/j.athoracsur.2019.09.042.
8. Zahia, S., Zapirain, M.B.G., Sevillano, X., González, A., Kim, P.J. and Elmaghraby, A., 2019. Pressure injury image analysis with machine learning techniques: A systematic review on previous and possible future methods. Artificial Intelligence in Medicine, p.101742.
9. Tripoliti, E.E., Papadopoulos, T.G., Karanasiou, G.S., Naka, K.K. and Fotiadis, D.I., 2017. Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Computational and structural biotechnology journal, 15, pp.26-47.
10. Tithi, Sushmita Roy, Afifa Aktar, Fahimul Aleem, and Amitabha Chakrabarty. "ECG data analysis and heart disease prediction using machine learning algorithms." In 2019 IEEE Region 10 Symposium (TENSYMP). IEEE, 2019, pp. 819-824.
11. Hijazi, S., Page, A., Kantarci, B. and Soyata, T., 2016. Machine learning in cardiac health monitoring and decision support. Computer, 49(11), pp.38-48.
12. Peterkova, Andrea, Martin Nemeth, German Michalconok, and Allan Bohm. "Computing Importance Value of Medical Data Parameters in Classification Tasks and Its Evaluation Using Machine Learning Methods." In Computer Science On-line Conference, Springer, Cham, 2018, pp. 397-405.
13. Rajawat, Pushpendra Singh, Deepak Kumar Gupta, Santosh Singh Rathore, and Avtar Singh.
14. "Predictive Analysis of Medical Data using a Hybrid Machine Learning Technique." In 2018 First International Conference on
15. Secure Cyber Computing and Communication (ICSCCC), IEEE, 2018, pp. 228-233 .
16. Elhoseny, M., Ramírez-González, G., Abu-Elnasr, O.M., Shawkat, S.A., Arunkumar, N. and Farouk, A., 2018. Secure medical data transmission model for IoT-based healthcare systems. Ieee Access, 6,
17. pp.20596-20608.
18. Awan, S.E., Sohel, F., Sanfilippo, F.M., Bennamoun, M. and Dwivedi, G., 2018. Machine learning in heart failure: ready for prime time. Current opinion in cardiology, 33(2), pp.190-195.

19. Thirumalai, Chandrasegar, Anudeep Duba, and Rajasekhar Reddy. "Decision making system using machine learning and Pearson for heart attack." In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), vol. 2. IEEE, 2017, pp. 206-210.
20. Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, pp.81542-81554.
21. Procházka, A., Charvátová, H., Vaseghi, S. and Vyšata, O., 2018. Machine learning in rehabilitation assessment for thermal and heart rate data processing. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26(6), pp.1209-1214.

## AUTHORS PROFILE

**Amit Juyal,** is pursuing Doctorate in Computer Science from Graphic Era Deemed to be University, Dehradun. He received his M.Tech. (CSE) in Computer Science from Graphic Era Deemed to be University. His areas of interest are Machine Learning; Soft computing .He has published various papers in the different journals with good impact factor. Currently he is working as Assistant professor in Graphic Era Hill University Dehradun. He has an experience of around 11 years of teaching.

**Chetan Pandey,** is an Assistant Professor in the School of Computing, Graphic Era Hill University, Dehradun (UK), India. He has received Master in Technology in CSE from Graphic Era Deemed to be University, Dehradun (UK), India. His research interests are Image Processing and Machine Learning.

**Janmejay Pant,** is pursuing his Doctorate in Information Technology from Kumaun University Nainital and done M.tech in Information Technology from Graphic Era deemed to be University, Dehradun. Currently he is working as Assistant Professor in Dept. of computer science at Graphic Era Hill University, Bhimtal Campus. He has total Academic teaching experience of more than 10 years with more than 25 publications in reputed National and International SCOPUS, UGC Approve Journals and conferences. His research area includes Machine Learning, Soft Computing, and Data Mining and Deep Learning. He is an active member of CSI. He secured First Rank in Uttarakhand State Eligibility Test (USET) in Computer Science and Applications in 2018. He also received Research Award at University Level in September 2019.

**Dr. Ankur Dumka,** Dr. is working as associate professor in Graphic Era Deemed to be University, Dehradun. He is having a long experience of 8+ years in the field of industry and academics. He is associated with smart city Dehradun as academic expert committee member and co-ordinator in terms of projects related to IT. He is having 40 + international papers in reputed conference and journals. He has contributed 3 books with reputed publisher like Taylor and Francis and IGI global and currently working on 2 more approved book. He had also contributed 15 chapters for reputed publishers. He is also editorial board members of many reputed conference and journals including scopus and IEEE. He is also guest editor of IJNCDS (scopus Indexed inderscience Journal), guest editor for IJNCR (ACM digital library -IGI global journal) and many more. He is associated with many societies and organization for welfare of educationalist societies.

**Vikas Tomer,** is working as assistant professor in Graphic Era Deemed to be University, Dehradun. His area of interest are machine learning and data analytics. He has published various research paper with good impact factor.