Journal of Computer Science

Original Research Paper

A Collaborative Prediction of Presence of Arrhythmia in Human Heart with Electrocardiogram Data using Machine Learning Algorithms with Analytics

Selvaraj Rajalakshmi and Kuthadi Venu Madhav

Department of CS and IS, Faculty of Science, BIUST, Botswana

Article history
Received: 16-11-2018
Revised: 23-01-2019
Accepted: 21-02-2019

Corresponding Author:
Kuthadi Venu Madhav
Department of CS and IS,
Faculty of Science, BIUST,
Botswana
Email: kuthadiv@biust.ac.bw

Abstract: Human heart is the major organ of human being which could fail the other systems in the body at the same time. Hence predicting heart disease is one of the challenging researches that requires meticulous analysis of heart rhythms properly. The irregular heart rhythms or beat is referred to as the Arrhythmia where heart rhythms with low or high rates comparing to the normal heart beat rate which ranges from 60 to 100 beats per minute. The heartbeat can be monitored and identified with the electrical disorder disease called Arrhythmia. This is very deadly when untreated for a long time as mortality rate is extremely high. Hence a prediction system is required to identify the irregular nature of heart and predict the heart problem in the future. The major objective of this research paper is to predict the presence of arrhythmia which is caused as a result of electrical imbalance and irregular heart beat in human being. The prediction is formulated with the help of essential parameters from electrocardiogram like age, gender, height, weight, BMI, QRS duration, P-R interval, Q-T interval, T interval, P interval, QRS, T, P, QRST, J values which will help the prediction of Arrhythmia in human to the best. The dataset sample is collected from UCI Repository based on electrocardiogram report values and pre-processed using Mat lab. The data is converted into test data and prediction is expected to be completed using Machine Deep learning Algorithms as they could be the best models for disease or syndrome predictions. Finally, the Analytics is carried out using Rapid Miner Studio where machine learning algorithms is applied and results obtained. The research will be a starter for futuristic research on automatic prediction of heart disease in human beings with various other parameters.

Keywords: Arrhythmia Prediction, Electrocardiogram, Machine Learning, Deep Learning, Heartbeat, Electrical Imbalance

Introduction

Arrhythmias is a stage where the heart of human beings beats at an improper state which may be either too high or too low. Arrhythmia is a collection of situations in human heart where the heart beat rhythm is irregular in nature which may result in too fast or too slow condition. The normal hear beat rhythm of up and down is 60 to 100 beats per minute in adult whereas it varied with children or old age people (Mohanty *et al*., 2017). If the heart rate is very high that is more than 100 beats per minute, it is known as tachycardia and at the same time if it is very less than 60 beats per minute, it is known as bradycardia. In general, this is not a disease whereas it is considered as a disorder by medical world as it has no symptoms. In some cases, there may be a pause in the rhythm of the heart which is known as heart attack and that indicates the matured stage of Arrhythmia disorder. Hence it is very important to identify and analyse the presence of Arrhythmia in the early stage itself and prevent the human beings from getting affected with the higher levels of cardiac problems. The moderate levels of heart rhythm may cause the presence of stroke or failure of heart in many cases. According to medical terminology, four major arrhythmia types (Shirai *et al*., 2017) are identified. They are 1.Extra beats-It indicates premature contractions of arteries or ventricles with contractions in the junctions of the heart.2.Supraventricular tachycardia's, It includes fibrillation of the arteries, fluttering of arteries paroxysmal stage of supraventricular tachycardia. 4. Ventricular arrhythmias-this stage includes

the fibrillation of the ventricles and tachycardia in the ventricles and finally Brady arrhythmia, this indicates the slowness in pumping of the heart or flow of blood from the heart or slow heart rate less than 60 beats per minutes.

The major objective of this research paper is to identify the type of arrhythmia stage and indicate the methods to Thus the prediction of such a disorder is required to preserve the life of human beings and to ponder the needs of the people who need analysing their heartbeat nature for future life. This heart disorder is a silent killer and somehow it has to be treated before it hurts the life of the individuals (Brian *et al*., 2017) in the future with high levels of problems like heart attack, stroke or heart failure. Thus, the proposal for arrhythmias prediction is formulated with the help of data mining algorithms that could always bring knowledge from available data. The dataset is collected from secondary sources from UCI Repository and then architecture is proposed. Before presenting the outcomes of the research, few of the earlier research areas are explored for making the research more perfect and scope for the future perspectives.

Literature Review

Gandhi and Singh (2015) conducted a survey that gives details on various data mining techniques and methods that gives knowledge about heart disease prediction in the future perspectives. The author analysed the various data mining methods like naïve Bayesian, neural networks and few decision tree algorithms that can be applied with medical data and knowledge could be extracted through the same.

Purusothaman and Krishnakumari (2015) analysed and provided a better and simple analysis on the various prediction models that could predict heart disease in an effective manner. The efficiency of each of the technique is analysed and results provided with risk prediction model. The survey is conducted with all the recent techniques that could bring better outcomes in the research. Finally a hybrid approach model is achieved as the best model compared with the remaining models.

Verma *et al*. (2016) produced a novel hybrid model that could present a CAD diagnosis which includes risk factors to be predicted using correlation-based subset feature selection method. The author used k-means clustering algorithm with swarm optimisation. Various other supervised learning algorithms such as multinomial logistic regression, multi-layer perceptron model, c4.5 etc. are used to find the results. The dataset collected from Indira Gandhi medical college, Shimla with 26 features and 335 instances from the cardiology department and the accuracy achieved is 88.4%. The result is tested with cleaver land heart dataset as well. The author mentioned that accuracy can be improved in the future.

Portela *et al.* (2015) predicted the probability of a patient to experience heart attack in the next few hours using data mining techniques with classification algorithms. The result of this research indicated high sensitivity of 95% which provided good results that is acceptable for the future research outcomes.

Rodger (2015) developed a software for managing patient information through hybrid Hadoop hive and to manipulate the databases. The communication of data with transfers are managed using repetitions of data and based on the less errors in the system. The warehouse infrastructure is also built with query, summarise and also conduct the analysis of the heart problems. The data is collected from ships and based on body injuries in the shipyard. The analysis is made based on mortality and survival of the patients who experienced a terrible problem in the clinical trials.

Vilar *et al*. (2015) matched the drug data and the symptoms that are the major causative agents to arrhythmia. The author also found similarities between the chemical structure, side effects of the drugs used, targets related to chemotherapy and other interactions. The results are verified using performance measures like sensitivity and specificity with precision factors giving positive results. The research was conducted with 100 candidates using precision enhancement in similarity to the proportional reporting ratio.

Sudhakar and Manimekalai (2018) conducted a study on healthcare industry and the impact on heart disease predictions. The author also studied the relationships between attributes based on the symptoms detected in the patients. The comparative study on the various algorithms like decision tree, naïve Bayesian, neural networks, genetic algorithm etc. is analysed and compared in the research.

Gomes *et al*. (2018) presented a combination of data mining techniques like clustering, feature selection and others to generate a more effective and efficient algorithm that could predict the occurrence of heart disease in the future. The author used the traditional dataset posted in UCI and used Random forest tree algorithm to achieve an overall accuracy of 88% more than other research methods.

Liu *et al*. (2015) analysed the electronic health records of patients and identified heart disease using clinical problems, vital signs and symptoms, blood test and reports. The research work suggested a predictive modelling with a temporal graph that indicates the presence or absence of heart disease for the patient. The graph and charts are the major indicators of the disease. If it is related to noisy and irrelevant data, then it is considered as heart problem whereas if the graph or chart is straight without noise it is

safe. This framework is a new dimension for further research on heart problems predictions.

Thus, various reviews are analysed and found that the prediction research on arrhythmias is of less in nature and needs more further research in the future perspectives. Hence for the social need of this research, this research is conducted with the architectural framework.

Arrhythmia Prediction Architecture

Medical Diagnosis is commonly available for predicting the problem of Arrhythmias disorder but automatic prediction is the need of the day as it may provide insights into the results much better than other medical diagnosis (Tashkandi and Wiese, 2018). The dataset is collected as a secondary resource from UCI Repository warehouse and utilised for the research. The dataset comprised of the numerical dataset of the heart rhythms that consists of 452 instances with 279 attributes. Among them feature selection is made and a sample of 10 attributes are selected for prediction as indicated in Table 1. The system is designed using data mining algorithms and its process with analytics. The entire work is formulated into 5 different stages for prediction of arrhythmias disorder in human beings.

Stage 1: Initial Pre-Processing of Dataset

The secondary dataset is a numerical dataset obtained from UCI Repository where the Electrocardiogram report in image format is converted to numerical data. Hence it is bound to errors. Hence it is refined in Fig. 1 using pre-processing techniques like normalisation, finding missing values and removing it, remove irrelevant data and remove inconsistent data from the whole dataset. Basically, three major pre-processing operations are carried out. They are removing Missing Values in the dataset, Inconsistent Values and Irrelevant values.

After pre-processing is completed, the raw dataset is converted to training dataset and is stored in excel format for further process.

Stage 2: Feature Selection Based on Attributes

The attributes that are capable of predicting the arrhythmias disorder has to be identified using this feature selection where right attributes have to be selected and formed as a testing set for prediction process. The selection of right feature for disease prediction under three categories as shown in Fig. 2:

1. Class Attributes (e.g. Age, Gender, BMI) to classifiy data
2. Predictive Attributes (Eg. P-R, Q-T interval) to predict the results

The attribute selection is generally based on the following criteria:

1. The relevance of the attribute
2. The Atomicity of the attribute
3. The Ability to predict the results rather than descriptive in nature
4. The Class attribute to group or test other attributes

Stage 3: Prediction Using Deep Learning Algorithms

This stage deliberates the identification of arrhythmias using the current testing dataset and provides the results based on the accuracy levels of the inputs used for tz.

The application of machine learning algorithms to identify the presence of arrhythmia under three stages as indicated in Fig. 3:

1. **Brady Arrhythmia** – Slow heart rate of the patient
2. **Ventricular Arrhythmia** – Fast heart beat in lower chamber of heart
3. **Supraventricular Tachycardias** – Fast heart beat in upper part of the heart
4. **Extra Beats** – abnormal electrical rhythm in any one of the lower ventricles to upper part.

A Combination of Fuzzy based clusters and classification algorithm (Fatima and Pasha, 2017) is proposed for completing this prediction of Arrhythmia using predictable factors or attributes applied to the testing Excel dataset created. The True Positive values depicting the correctness of the prediction with presence of Arrhythmia is combined with the identification of non-presence of arrhythmia is utilised for the algorithm as shown in the Fig. 4.

Stage 4: Analytics with Performance Measures

The research analysis has to be tested with the accuracy as well as performance measures as indicated in Fig.5. like sensitivity, specificity, F-Score value and Accuracy value (Abdullah *et al*., 2018). The result is checked for both the algorithms to be created and tested for maximum prediction results.

Stage 5: Verification of Results using Rapid Miner

Rapid Miner is a designer tool to directly test the results and also to identify the potential outcomes with varied results. Hence the tool is used to test the results by forming machine learning clusters of data.

The Research architecture is proposed and modelled for implementation in the research tool specifically for prediction of results and checking for accuracy and performance measures.

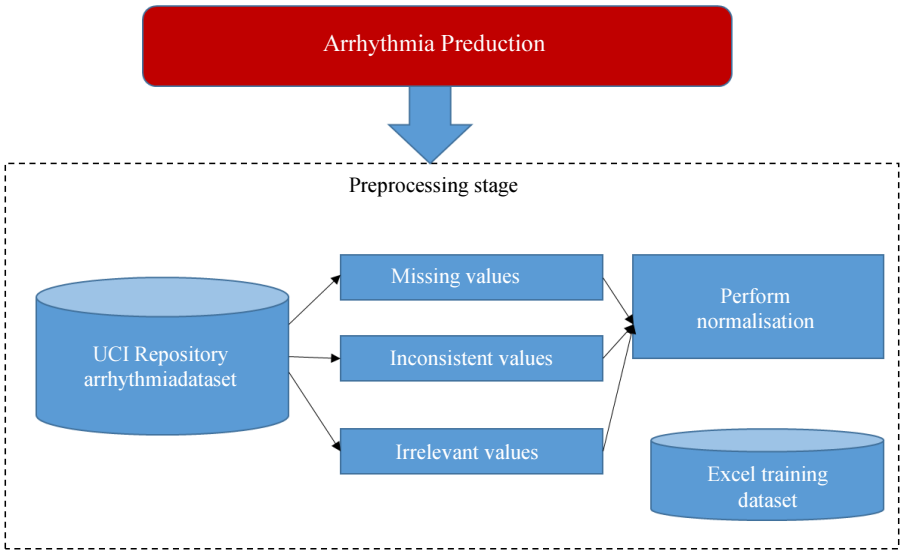


Fig. 1: Stage-1: Pre-processing the raw dataset

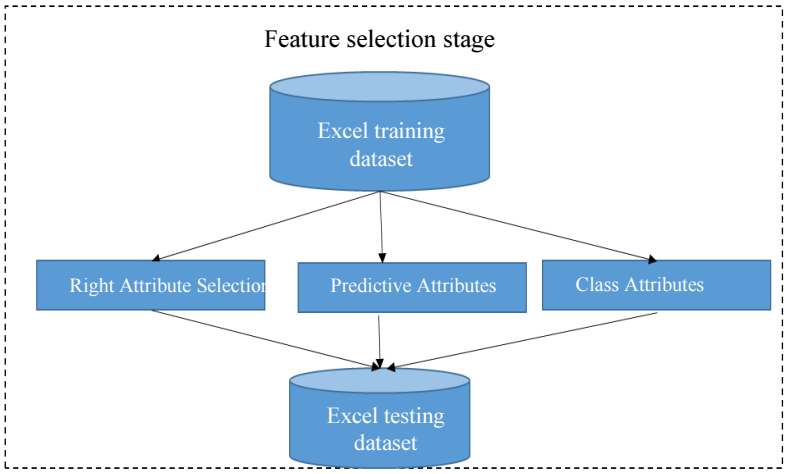


Fig. 2: Stage-2: Feature selection for grouping right attributes

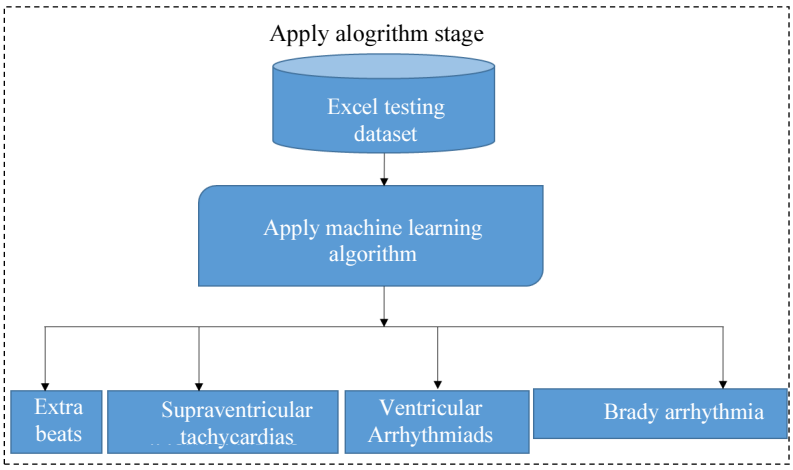


Fig. 3: Stage-3: Apply algorithm to predict the stage of arrhythmia

Algorithm ARFHC (Arrhythmia Fuzzy Hybrid Classifier)
Declare
True_Positive = 0;
True_Negative = 0;
Range [ ]
Begin
For I in 1 to End_of_record
Inspect Age
If Age within Range(x)
If default values (attributes)
Increment True_positive by 1;
Else
Increment False_Negative by 1;
End
End
End For
End

Fig. 4: Algorithm applied to identify stage of arrhythmia

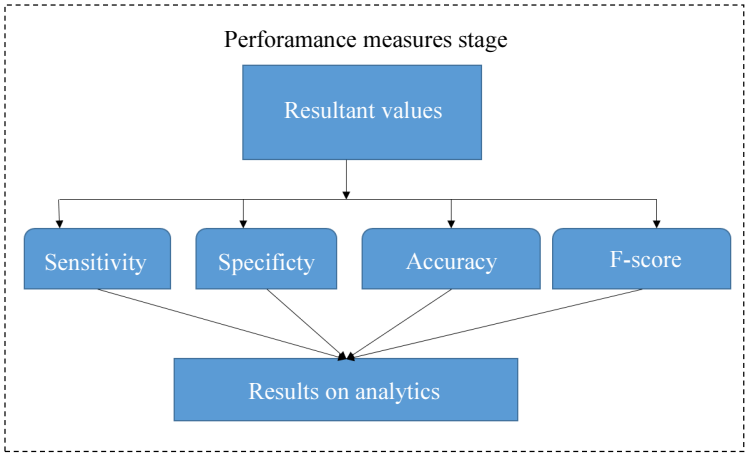


Fig. 5: Stage-4: Verification using performance measure analysis

Implementation

The arrhythmia is a very complicated disease and needs careful analysis to predict the results in an effective manner. Mat lab is an effective tool where GUI based design can promote outcomes with effective methodology. Mat lab 2015 is utilised for predicting the results. In the Initial stage, Pre-processing of Training Dataset is performed using relevant filtering techniques like normalisation, remove missing values and also to select numeric values. The pre-processing of dataset is completed with refined testing set stored in MS Excel format as shown in Fig. 6. for further evaluation. The loaded raw dataset is tested for normalisation of missing values, irrelevant values and inconsistent values to provide better results in the outcome of the research. The created training dataset in excel format is stored in directory for next stage of pre-processing in the research process.

In the second stage, feature selection is carried out to select the best predictive attribute by importing the Trained Excel sheet after pre-processing as represented in Fig. 7. feature selection is carried out to select the best predictive attribute for effective outcome of the results. At the end of the selection process, 11 major predictive attributes are selected for evaluation. The selected attributes are grouped and stored as testing set for predictive analysis is indicated in Fig. 8.

DOI: 10.3844/jcssp.2019.278.287

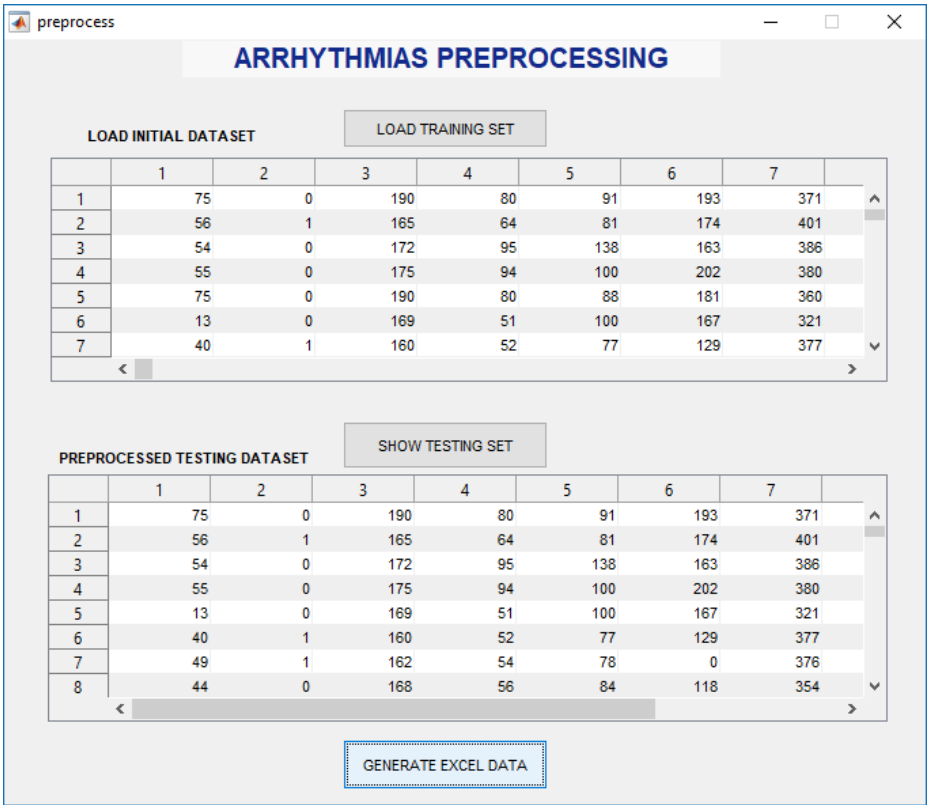


Fig. 6: Implementing pre-processing of dataset

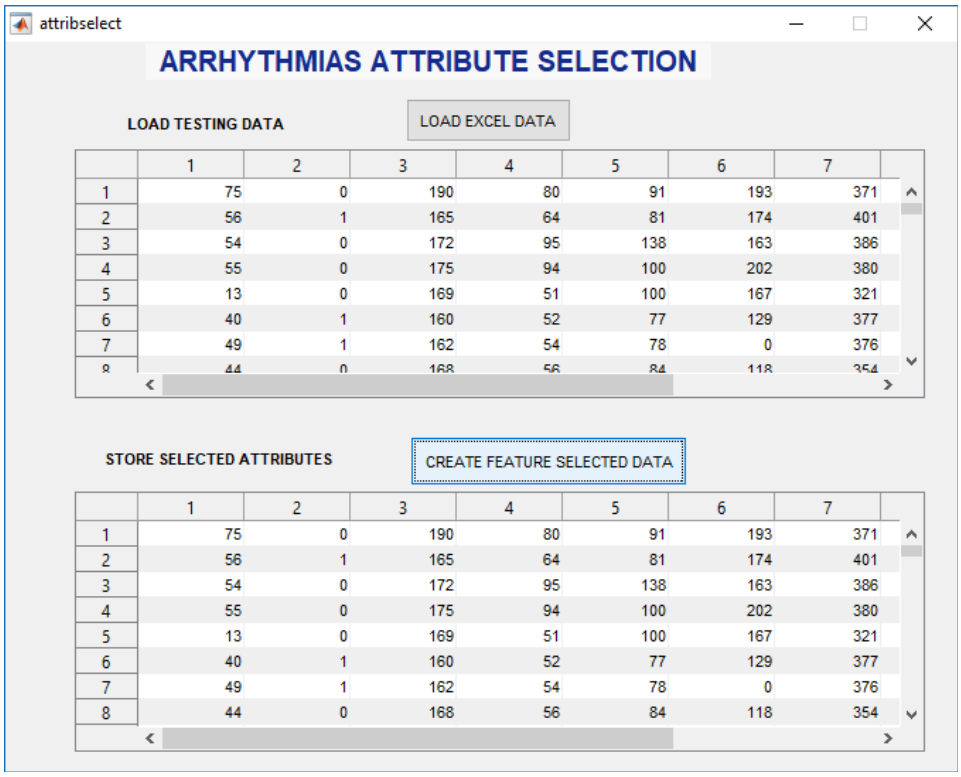
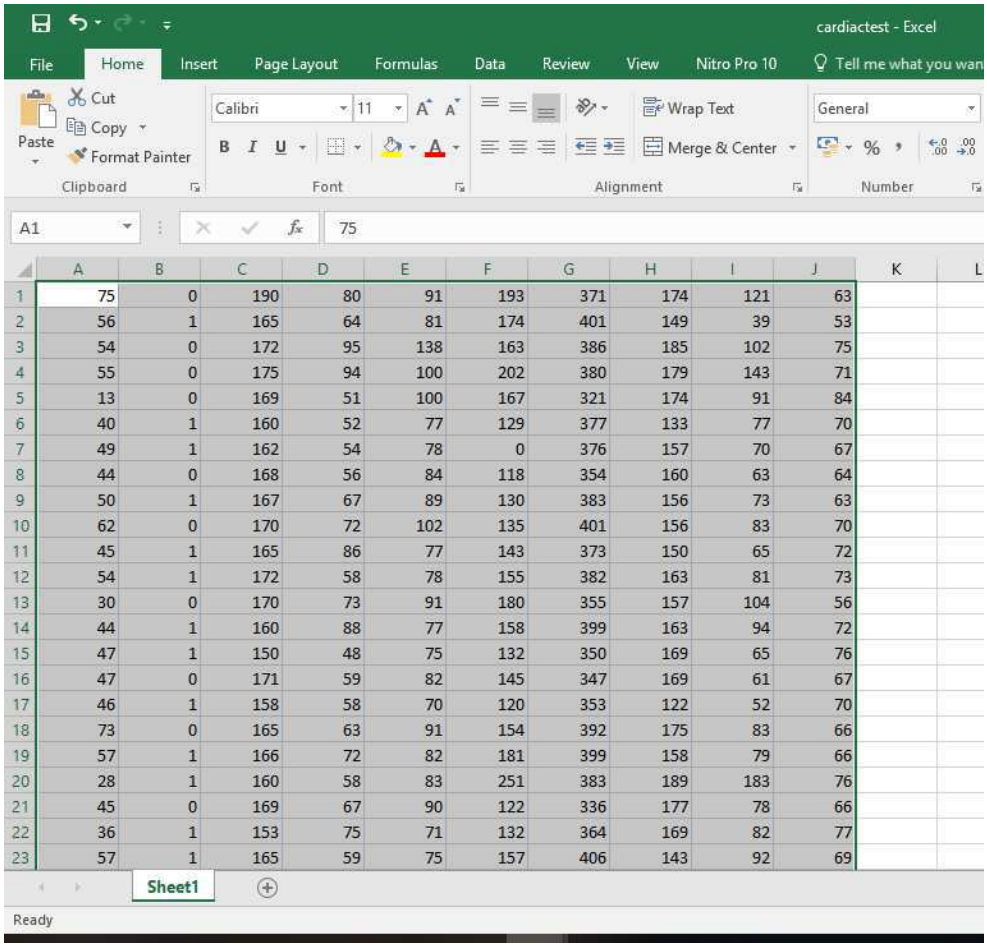


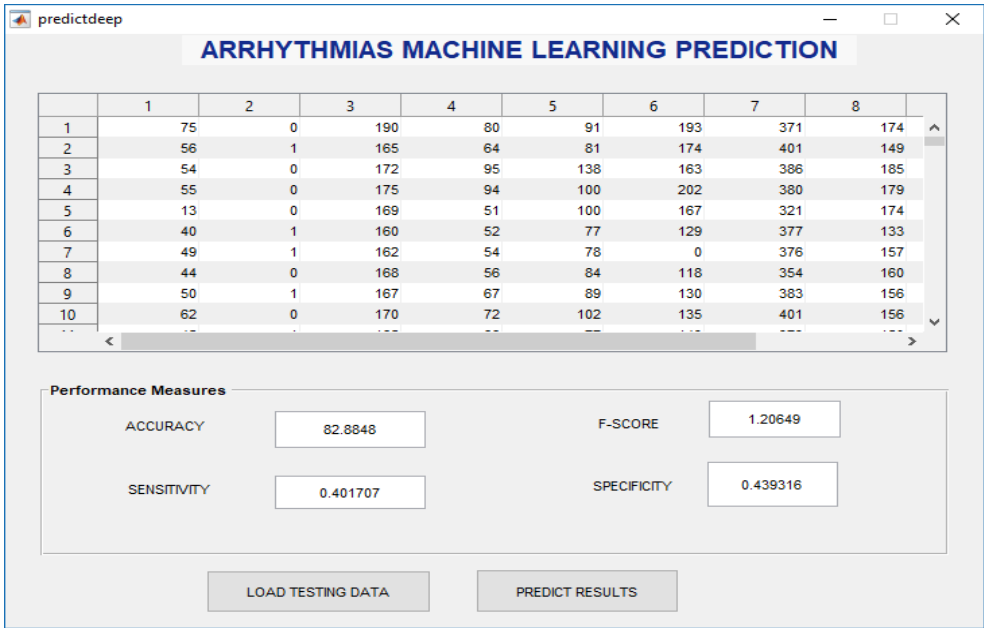
Fig. 7: Created pre-processed dataset in excel format (Training set)

DOI: 10.3844/jcssp.2019.278.287



	A	B	C	D	E	F	G	H	I	J
1	75	0	190	80	91	193	371	174	121	63
2	56	1	165	64	81	174	401	149	39	53
3	54	0	172	95	138	163	386	185	102	75
4	55	0	175	94	100	202	380	179	143	71
5	13	0	169	51	100	167	321	174	91	84
6	40	1	160	52	77	129	377	133	77	70
7	49	1	162	54	78	0	376	157	70	67
8	44	0	168	56	84	118	354	160	63	64
9	50	1	167	67	89	130	383	156	73	63
10	62	0	170	72	102	135	401	156	83	70
11	45	1	165	86	77	143	373	150	65	72
12	54	1	172	58	78	155	382	163	81	73
13	30	0	170	73	91	180	355	157	104	56
14	44	1	160	88	77	158	399	163	94	72
15	47	1	150	48	75	132	350	169	65	76
16	47	0	171	59	82	145	347	169	61	67
17	46	1	158	58	70	120	353	122	52	70
18	73	0	165	63	91	154	392	175	83	66
19	57	1	166	72	82	181	399	158	79	66
20	28	1	160	58	83	251	383	189	183	76
21	45	0	169	67	90	122	336	177	78	66
22	36	1	153	75	71	132	364	169	82	77
23	57	1	165	59	75	157	406	143	92	69

Fig. 8: Resulting Testing set cardiac after feature selection (Testing set)



	1	2	3	4	5	6	7	8
1	75	0	190	80	91	193	371	174
2	56	1	165	64	81	174	401	149
3	54	0	172	95	138	163	386	185
4	55	0	175	94	100	202	380	179
5	13	0	169	51	100	167	321	174
6	40	1	160	52	77	129	377	133
7	49	1	162	54	78	0	376	157
8	44	0	168	56	84	118	354	160
9	50	1	167	67	89	130	383	156
10	62	0	170	72	102	135	401	156

Fig. 9: Prediction of arrhythmia with good accuracy

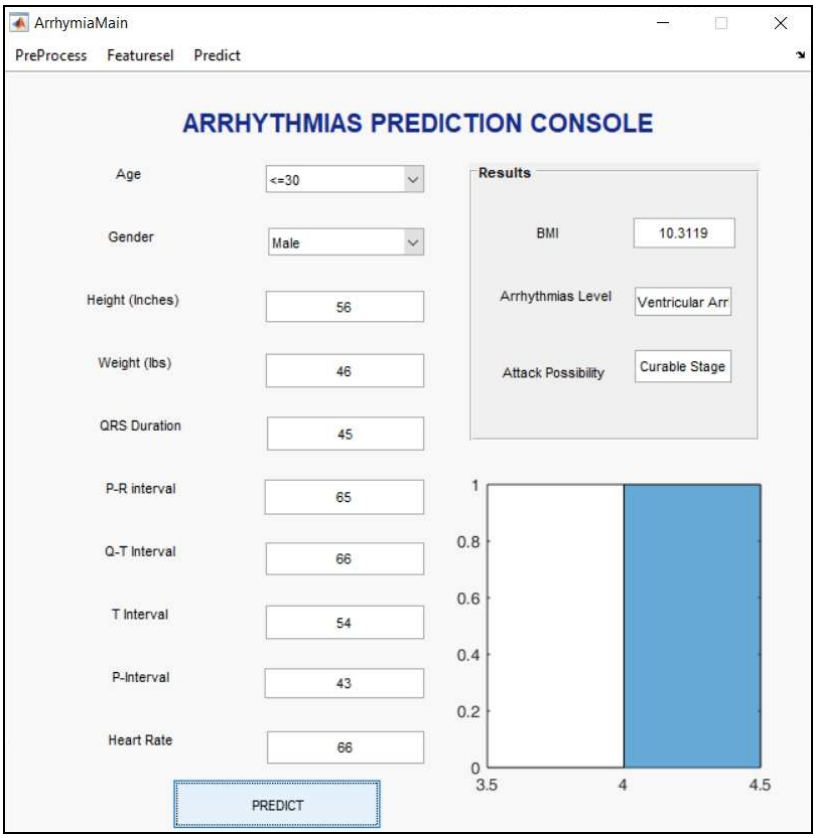


Fig. 10: Prediction of arrhythmia using ARFC algorithm

Table 1: Feature attributes selected for predictive analysis

Attribute	Normal Values	Explanation	
Age	<30, 30-50, 50- 70, >70		
Gender	Male, Female		
Height	In centimetres		
Weight	In Kgs		
BMI			Body Mass Index = (height / weight* weight) * 703
QRS duration	In msec		Average value of QRS
P-R interval	In msec		Average value of P and Q raise
Q-T interval	In msec		Average of Q raise and T down
T interval	In msec		Average of T value
P interval	In msec		Average of P value
Heartrate	Low High	60-100	

Table 2: Results of MATLAB Implementation

Sensitivity	0.40
Specificity	0.43
F-Score	1.21
Accuracy	82.8

Table 3: Cluster verification of results obtained in Matlab

Cluster	Meaning	Items
Cluster_0	Respondents with arrhythmia positive	195
Cluster_1	Respondents with arrhythmia negative	256
Total		**451**

After creating the Final dataset for Prediction, the dataset is loaded into the predictor module where the newly formed hybrid algorithm is applied and the results are identified as shown in Fig. 9 after the conclusive research.

The Result for this Prediction analysis is clear that the Prediction for Arrhythmias is tested using performance measures like accuracy, sensitivity, specificity and F-Score Analysis as given in Table 2.

After confirmation since the results are well above 80% it is accepted to rely on the 11 attributes for effective prediction of arrhythmias in human heart. Hence the final module presenting the prediction is designed and produced with result.

The Fig. 10. indicates the final results obtained after processing the result for arrhythmia prediction under four levels of disease. The result outcomes the BMI, Arrhythmia leavel and attack possibility for the patient given the input for the 11 features selected for

prediction. Thus the results are obtained efficiently using Novel Arrhythmia Fuzzy Hybrid Classifier (ARFC) Algorithm. After forming the results, the analysis is made using performance measures like sensitivity which tests the results as true if the result is identified correctly as true along with specificity which identifies the negative results as wrong respectively. Based on the sensitivity and specificity analysis, the Accuracy and the F-Score results are also predicted for the research.

Results and Discussion

The prediction is further verified using rapid miner studio where one of the hybrid algorithms using k-means clustering segmentation, the result is finalised. The design part in rapid miner is tested to analyse the correctness and accuracy of the research and the verification of results obtained through the algorithm. The rapid miner studio process comprises of designing the verification system using the resultant dataset obtained from the Matlab implementation. The results are formed using K-means supervising algorithm to find the clusters formed in cases affected with arrhythmias and not affected with arrhythmias respectively. The results obtained are presented as Table 3.

The result thus obtained indicated in Fig. 11. that cases ported with arrhythmias is less compared to cases not affected with arrhythmias.Thus, the overall conclusion also confirmed that the results are convincing and it indicated that the attributes selected for predicting arrhythmias is relevant to finding the outcome and is reliable to predict the presence or absence of arrhythmias in heart of human beings.

Conclusion

The outcomes of this research provide an insight into the future generation of machines that serves as expert systems to predict the human disorders like arrhythmias which is very hard to predict using normal methods. The research is completed in a protective environment and the results are verified using analysis. The research conveyed that the prediction accuracy is much higher compared to previous research works. The future implications can be creating the expert system in a common machine that could be used by patients to identify the disorder even without systems and software but by utilising android-based systems along with the IOT based systems.

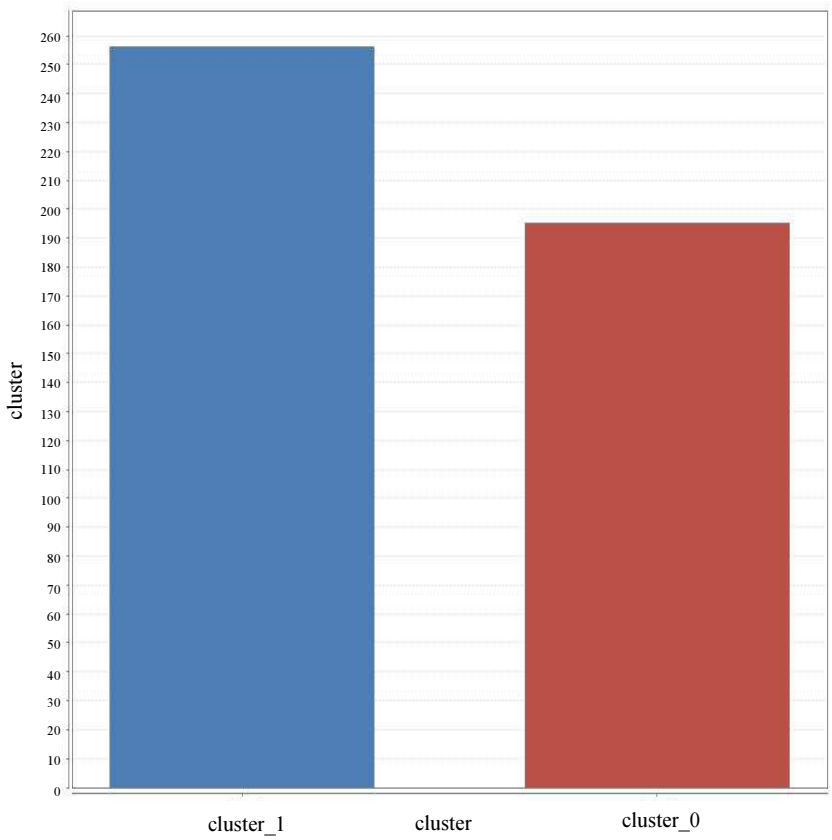


Fig. 11: Verification of results design in rapid miner studio

Acknowledgement

This project was funded by the BIUST research initiation grant, grant no- DVC/RDI/2/1/161(39). We thank our colleagues from BIUST who provided technical and financial support that greatly assisted the research.

References

Abdullah, M., F. Al-Anzi and S. Al-Sharhan, 2018. Hybrid multistage fuzzy clustering system for medical data classification. Proceedings of the International Conference on Computing Sciences and Engineering, Mar. 11-13, IEEE Xplore Press, Kuwait, pp: 1-6. DOI: 10.1109/ICCSE1.2018.8374213

Brian, A., N. Sabna and G.G. Paulson, 2017. ECG based algorithm for detecting ventricular arrhythmia and atrial fibrillation. Proceedings of the International Conference on Intelligent Computing and Control Systems, Jun. 15-16, IEEE Xplore Press, Madurai, India, pp: 506-511. DOI: 10.1109/ICCONS.2017.8250773

Fatima, M. and M. Pasha, 2017. Survey of machine learning algorithms for disease diagnostic. J. Intelli. Learn. Syst. Applicat., 9: 1-16. DOI: 10.4236/jilsa.2017.91001

Gandhi, M. and S.N. Singh, 2015. Predictions in heart disease using techniques of data mining. Proceedings of the International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Feb. 25-27, IEEE Xplore Press, Noida, India, pp: 520-525. DOI: 10.1109/ABLAZE.2015.7154917

Gomes, C., A. Cardoso, T. Silveira, D. Dias and E. Tuler *et al*., 2018. Combining data mining techniques to enhance cardiac arrhythmia detection. Proceedings of the International Conference on Computational Science, Springer, Cham. pp: 321-333.

Liu, C., F. Wang, J. Hu and H. Xiong, 2015. Temporal phenotyping from longitudinal electronic health a graph Based framework. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 10-13, IEEE Xplore Press, Sydney, NSW, Australia, pp: 705-714. DOI: 10.1145/2783258.2783352

Mohanty, S., P. Mohanty, C. Trivedi, C. Gianni and J.D. Burkhardt *et al*., 2017. 61Clinical characteristics and arrhythmia-profile in older women with atrial fibrillation undergoing catheter ablation. Eur. Heart J. DOI: 10.1093/eurheartj/ehx501.61

Portela, F., M.F. Santos, J. Machado, A. Abelha and F. Rua *et al*., 2015. Real-Time Decision Support Using Data Mining to Predict Blood Pressure Critical Events in Intensive Medicine Patients. In: Ambient Intelligence for Health, Bravo, J., R. Hervás and R. V. Villarreal (Eds.), Springer, Cham, pp: 77-90.

Purusothaman, G. and P. Krishnakumari, 2015. A survey of data mining techniques on risk prediction: Heart disease. Indian J. Sci. Technol. DOI: 10.17485/ijst/2015/v8i12/58385

Rodger, J.A., 2015. Discovery of medical big data analytics: Improving the prediction of traumatic brain injury survival rates by data mining patient informatics processing software hybrid hadoop hive. Informat. Med. Unlocked, 1: 17-26. DOI: 10.1016/j.imu.2016.01.002

Shirai, Y., M. Goya, S. Ohno, M. Horie and S. Doi *et al*., 2017. Elimination of ventricular arrhythmia in catecholaminergic polymorphic ventricular tachycardia by targeting catecholamine-sensitive area: A dominant-subordinate relationship between origin sites of bidirectional ventricular premature contractions. Pacing Clinical Electrophys., 40: 600-604. DOI: 10.1111/pace.13006

Sudhakar, K. and D.M. Manimekalai, 2014. Study of heart disease prediction using data mining. Int. J. Advan. Res. Comput. Sci. Software Eng., 41: 1157-1160.

Tashkandi, A.I. and L. Wiese, 2018. Efficient in-database patient similarity analysis for personalized medical decision support systems. Big Data Res., 13: 52-64. DOI: 10.1016/j.bdr.2018.05.001

Verma, L., S. Srivastava and P.C. Negi, 2016. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J. Med. Syst., 40: 178. DOI: 10.1007/s10916-016-0536-z

Vilar, S., T. Lorberbaum, G. Hripcsak and N.P. Tatonetti, 2015. Improving detection of arrhythmia drug-drug interactions in pharmacovigilance data through the implementation of similarity-based modeling. PloS one, 10: e0129974. DOI: 10.1371/journal.pone.0129974