Machine Learning based Cardiovascular Disease Pattern Prediction Technique for Remote Healthcare Monitoring Systems

Subhasini SV, S. Raja Mohamed

M.E. Computer Science and Engineering, KIT- KalaignarKarunanidhi Institute of Technology, Coimbatore, TN, India

Head, Computer Science and Engineering, KIT- KalaignarKarunanidhi Institute of Technology, Coimbatore, TN, India

Abstract: *Wearable devices and the various applications of Wearable devices are increasing greatly which encourages Researchers to focus much on Internet of Medical Things (IoMT). The IoMT is playing a major and considerable role for reducing mortality rates as IoMT is helping diseases pattern well in advance. As we know, the Cardiovascular disease threatens us as mortality is relatively high which is observed from the available literature survey. It was noticed from the statistical report that there were relatively high rate of Heart Diseases registered. The Healthcare System is needed for the better Cardiovascular Diseases prediction techniques, which will help Medical Practitioners to predict this disease well in advance that will facilitate early prevention, detection, and fruitful treatment to Patients. This will save human life and can reduce mortality rate. As we know, Machine Learning Models and the Internet of Medical Things jointly enabled methodologies for supporting healthcare services particularly for Cardiovascular Disease Pattern Prediction, classification and accurate Diagnosis. It was noticed that there were two IoMT based Models proposed recently. They are Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT) and Hybrid Random Forest-Linear Model(HRFLM). These two Models were implemented and Analyzed their performances during Training and Testing Processes in terms of Prediction Accuracy, Precision, Sensitivity, Specificity, FScore and Average Processing Time(ms) and it was noticed that the Hybrid Random Forest-Linear Model(HRFLM) is performing well in terms of accuracy and time complexity where as Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT) is performing well in terms of Precision, Sensitivity, Specificity and FScore. To maximize Classification and Prediction accuracy better, this work is proposed an efficient Ensemble Support Vector Classifier – Weighted Random Forest called Ensemble SVC-WRF (E-SVC-WRF) and implemented for analysis. From the experimental results, it was noticed that the proposed Ensemble Support Vector Classifier – Weighted Random Forest called Ensemble SVC-WRF (E-SVC-WRF) is performing well as compared with Hybrid Random Forest-Linear Model(HRFLM) and Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT) in terms of Prediction Accuracy, Precision, Sensitivity, Specificity, FScore and Average Processing Time(ms).*

KEYWORDS: **Cardiovascular Diseases, Patterns Prediction, Internet of Medical Things, SVM, Weighted Random Forest, Machine Learning**

1. INTRODUCTION

As we know, Heart Diseases happening due to Coronary Artery Disease, Pulse Rate Issue, Blood Vessel Diseases, Arrhythmias issues etc. There were a few common conditions for Heart Diseases that are high blood pressure, heart failure, heart attack, Congenital heart conditions, unstable angina etc. As we know, People calling heart diseases as cardiovascular diseases. The heart diseases is also called as cardiovascular diseases (CVD)[3,4,5,10,11].

The Cardiovascular Diseases CVD refers various reasons and conditions for heart diseases particularly blocks in blood vessel. As blood is blocked by blocks in blood vessels, it causes heart's muscle damage, valve damage and pain in chest lead to heart attack or stroke.

Health care industries consider that Cardiovascular Diseases are the primary reasons to heart attack and human death in the world. That is the reason why most of the Researchers are focusing Technologies and proposing techniques through ML / DL to predict Cardiovascular Diseases well in advance. This better prediction is mandatory as for as health care data analysis are concerned [2,3,8,10].

The intelligent techniques through Soft Computing, Artificial Intelligence, Machine and Deep Learning helps researchers to extract real patterns and information from the clinical datasets that is helping Medical Practitioners for taking better predictions and wise decisions.

From the literature survey, it was noticed that the Cardiovascular Diseases is considered as the prime reasons which leads cause of human death. It is also noticed that around 6,10,000 human beings are losing their life due to Cardiovascular Diseases in US. That is 25% of deaths due to Heart Diseases. As far as Indian statistical reports concerned, there were predicted that in our population, there were 27% of deaths due to Heart Diseases as compared with Global Average 23.5%. [1,2,9,14]. This is the common disease for both Male and Female.

Thus, it is the major issue to be addressed. But however, it seems it is really challenging prediction to predict Cardiovascular Diseases in advance as it was observed that there were numerous factors are involving for this disease. Pulse Rate, LDL, HDL, BP and Diabetes are considered as the risk factors. As many risk factors mentioned above causes heart diseases, the healthcare industries and medical practitioners need modern intelligent techniques such as Soft Computing, Artificial Intelligence, Machine and Deep Learning with Data Analytics. This will help researchers to design and develop intelligent model for better prediction.

It is noted that heart disease is developing due to stressful life style. Medical Practitioners have been diagnosing based on patient’s life style and earlier diagnosis report. As a result, more researchers are focusing the health sector for better CVD risk in advance. Artificial Intelligence, Machine Learning and Deep Learning were introduced better prediction by removing noise and extracting useful information from datasets. These techniques are sued to remove unwanted information from the datasets by applying dimensionality reduction techniques. This will help Medical Practioners to take wise decision with better disease prevention treatment. From the literature survey, it was noticed that a few techniques proposed for diagnosing cardiac diseases. [1,2,8,13]. It is also noticed that though numerous methods and techniques proposed for this purpose, we need still better classifiers for the best prediction with highest prediction accuracy.

From the Literature Survey, it was noticed that there were still various intelligent techniques proposed namely pattern recognition, classification and predictive methods for better cardiovascular problems prediction. It is also noted that ML/DL based classifiers like kNN, DT, SVM, RF, ANN and linear and Nonlinear Regression techniques to predict Cardiovascular diseases.

2. RELATED WORKS

Padmaja et al. [11] was proposed an efficient Machine Learning Technique, which is proposed for better cardiovascular diseases patterns. It considered as useful method for Medical Practitioners to understand and predict the actual level of heart diseases.

Geetha [10] was providing an Machine Learning Technique in association with Artificial Neural Networks (ANN) for predicting Cardiovascular diseases patterns prediction. The author used 13 features and predict the patterns.

The author implemented and shown this methods performance and noted the efficiency of the classification accuracy 70%.

Gao et al. [7] has introduced an ML based Classifier which is the hybrid model implemented and tried to achieve better classification accuracy. The Kaggle dataset was considered for implementation and analysis. This Model ensemble both bagging and boosting approaches.

As both PCA and LR Models, it selected required information from datasets for better classification. From the results, it was noticed that the classification accuracy was 98.6%.

Agrahara [4] has introduced CVD Pattern Prediction and analyzed effectively. It compares Regression, Random Forest, Neural Networks, kNN and Decision Tree. It predicted that the accuracy was 98.02%, better precision and less MSE.

Xiaoming Yuan and et. al.[14] designed a Fuzzy GBDT technique which combines Fuzzy and Boosting DT for better classification prediction. It is much suitable to avoid overfitting. It was noticed from the experimental results that it got better classification accuracy.

Mohan et al. [9] has introduced an effective Random Forest Model called Random Forest Hybrid with a Linear Model (HRFLM). This is the integrated model with Linear Model and Random Forest Model. The accuracy was predicted and measured as 88.7%. This work considered the dataset named Hungarian from the database UCI.

From the literature survey, it was noticed that the above mentioned IoMT based Models were relatively better. ie we would like to implement and analysis both the Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT) [14] and Hybrid Random Forest-Linear Model(HRFLM) [9] as it was predicted as much suitable of IoMT. This paper discusses the characteristics and features of the proposed Ensemble Support Vector Classifier – Weighted Random Forest called Ensemble SVC-WRF (E-SVC-WRF) and the existing Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT) and Hybrid Random Forest-Linear Model(HRFLM) in the following sections.

2.1 Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT)

Bootstrap Aggregation proposed by [14] Xiaoming Yuan and et. al. is used for reducing the DT Variance. For training this technique, the features were taken in random pattern. All the predictions were found average and it is considered as the best robust model.

The learning approach [1,4,14] is obtaining better area under the curve ie AUC and this greatly reduced the variance by bagging technique.

Xiaoming Yuan and et.al. [14] utilized the great ML technique GBDT. The GBDT diagram is as shown in the Fig. 1. From the experimental report, it is clearly noticed that it reduced loss function and achieved better accuracy which is demonstrated in the equation 1.

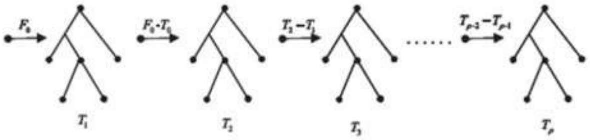


Fig. 1 Schematic diagram of the GBDT Technique14]

As shown in the equation 1, the weighted coeffients are a,b,c….ρ. the F(x) is the sum of all trees outcomes make better conclusion.

$$F(x) = a.F_0(x) + b.F_1(x) + c.F_2(x) + \cdots..+k.F(x) \ (1)$$

This work was found to be best technique for the better distribution in ML Techniques. It is able to execute and process different data types to predict better disease patterns.

The author [14] has implemented this classifier in parallel which is able to predict patterns in parallel. Thus the complexity is reduced in the datasets. It addresses the overfitting issues as well.

2.1.1 Fuzzy-GBDT: Fuzzy Logic Integrates GBDT Algorithm

Data Fuzzification: As we know, if we needed better classification accuracy, that will lead to more process and data complexity. When executing diagnosis processes, we can notice that the prediction report might be the same for different patients whereas all the attributes are common.

This might be lead to complexity. This issue was solved in this model as Fuzzy Logic was employed. The datasets used under hierarchical model and narrates degree for each and every attributes.

As this model employs membership function, its complexity is much reduced [2,7,14]. The Membership function was shown in the Fig. 2.

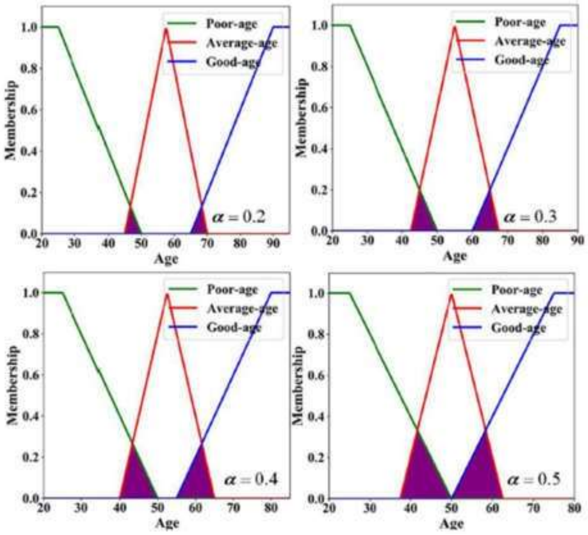


Fig. 2 Effect of different values α in membership [14]

$$\mu_1 = \begin{cases} 0, & > a + \sigma \\ \frac{a+\sigma-S}{\sigma}, & a \le x \le a+\sigma \\ 1, x < a & \end{cases} \tag{2}$$

$$\mu_2 = \begin{cases} 0, & x < \theta \text{ or } x > a + (2 - \ ) \\ \frac{S-\theta}{\frac{\sigma}{2}}, & \theta \le x \le \theta + \frac{\sigma}{2} \\ \frac{a+(2-\alpha)\alpha-S}{\frac{\sigma}{2}}, & \theta + \frac{\sigma}{2} < x < \theta + \sigma \end{cases} \tag{3}$$

$$\mu_3 = \begin{cases} 0, \ x >< +(2-2\alpha)\sigma & \\ \frac{S-}{\sigma}, & a + (2-2\alpha)\sigma \le x \le b \\ 1, & x > b \end{cases} \tag{4}$$

The fuzzification with different slope is shown in the Fig. 2. In the data samples, the age distribution is from 0 to 90. The slope can be correlated with age by adjusting α and the data set has 14 attributes.

2.2 Hybrid Random Forest-Linear Model(HRFLM)

Mohan et al. [9] has introduced an effective Random Forest Model called Random Forest Hybrid with a Linear Model (HRFLM). This is the integrated model with Linear Model and Random forest Model. The accuracy was predicted and measured as 88.7%. This work considered the dataset named Hungarian from the database UCI. As shown in the Fig. 3, the HRFLM considers 14 attributes such as AGE, SEX, CP, BPS, FBS, REST ECG, SLOPE, CA, THAL. These Features took from UCI Dataset is feeding to Linear Method in association with Random Forest is feed again to HRFLM. This will predict the pattern based on the input and will be classified. As mentioned, the classification accuracy was 88.7%.

3. PROPOSED AN ENSEMBLE SUPPORT VECTOR CLASSIFIER – WEIGHTED RANDOM FOREST CALLED ENSEMBLE SVC-WRF (E-SVC-WRF)

In this proposed model, this research work introduced Ensemble Support Vector Classifier – Weighted Random Forest Called Ensemble SVC-WRF (E-SVC-WRF). In other words, the Support Vector Classifier is employed for selecting and classifying diseases patterns by optimizing hyperplane and Weighted Random Forest is employed for predicting patterns for cardiovascular disease diagnosis based on patterns classification. Further, the score of Weighted Random Forest can be calculated by repeating average weights for different sizes of clusters.

As Support Vector Classifier is designed as the non-linear model by introducing RBF kernel function for patterns, it removes irrelevant patterns, noise, redundancy as well which select the features and patterns effectively. As this model effectively reducing data set size, the prediction accuracy is better than existing model. The working processes of the proposed model is shown in the Fig. 4.

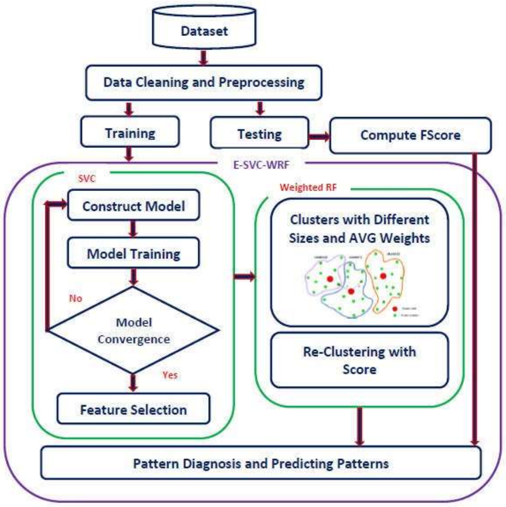


Fig. 4. Proposed Model : E-SVC-WRF

Algorithm

Input dataset

```
For ∀ attribute_featuresfeatures :
For each attributed sampling :
Process and run DT
Classify patterns and features f1,f2…fn of Dataset
Get the number of leaf nodes ln1, ln2,…lnn
Spilit the Dataset D into sd1, sd2,… sdn
```

Output : SVC based Patterns Partitions with Sizes and Weights$_{AVG}$

```
For ∀ features do
Partition R(p1,p2,p3…pn)
Classify the Datasets with Clusters
Classify the Clusters with Sizes and Weight Scores
```

$$\arg\min_{\omega} (\frac{1}{k}\sum_{i=1}^{n} max\{0, 1 - y_{if}(x_i)\} + C_\omega T_\omega)$$

Output :

C(R1(p1,p2,…pn)

R2(p1$_{WC}$,p2WC,p3$_{WC}$, … pn$_{WC}$

```
For ∀ features error minimization do
Min[C1(R1(p1,p2,…pn))
Min[C2(R2(p1_WC,p2WC,p3_WC, … pn_WC))]
Min[C2(R1(p1,p2,…pn))
Min[C2(R2(p1_WC,p2WC,p3_WC, … pn_WC))]
…
Min[Cn(R1(p1,p2,…pn))
Min[Cn(R2(p1_WC,p2WC,p3_WC, … pn_WC))]
```

Output : AttributesFeatures with Classified Attributes F(d1,d2,d3,..dn)

Extracted Features Fo (Training)

$$\sum_{0}^{n} F1(k) = d + m_1x_1 + m_2x_2 + m_3x_3 + \ldots. + m_nx_n$$

$$\sum_{0}^{n} F1(0) = Gaik + \sum_{0}^{n} \mathrm{w}_i$$

Extracted Features F1 (Testing)

$$\sum_{0}^{n} F1(k) = d + m_1x_1 + m_2x_2 + m_3x_3 + \ldots. + m_nx_n$$

$$\sum_{0}^{n} F1(0) = Gaik + \sum_{0}^{n} \mathrm{w}_i$$

Output : Patterns Diagnosis

4. EXPERIMENTAL SETUP

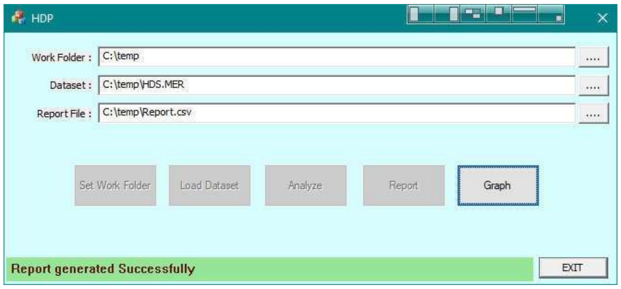
The dataset from repository of Machine Learning UCI was downloaded[1]. The Heart Disease Dataset namely Switzerland, Hungary, Cleveland were used for performance analysis. The downloaded dataset has 303 records and there were 76 attributes. We considered 14 attributes for Cardiovascular Patterns Prediction.

The Datasets were taken from [16] which is used for analysing the above mentioned Classifiers. The experimental set up is created in with BioWeka and simulations are conducted by this project work by using the cleverland Data Sets, Master.MER. This was taken from the above mentioned link for study.

As mentioned earlier, this project Work has developed with the VC++ Professional 2022 and MSVC Tool that is developed for extracting and validating various patterns of Heart Diseases. The standardized and Cleaned Dataset was feeding to BioWeka to analyze the identified Models. Simulations are carried out to evaluate the performances and classification and prediction abilities of the above listed our proposed classifiers.

This work considered 10 different Data Sets grouped together for predicting cardiovascular disease patterns, and each group has 50,000 samples out of 5,00,000 samples used for prediction analysis of the existing model. The experiments were executed again and again for measuring the efficiencies of the classifier.

The performances of the proposed classifier Ensemble Support Vector Classifier – Weighted Random Forest called Ensemble SVC-WRF (E-SVC-WRF and the identified two classifiers Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT) and Hybrid Random Forest-Linear Model(HRFLM) were implemented and carefully analyzed its performances during Training and Testing Processes in terms of Sensitivity, Specificity, Accuracy, FScore and Average Processing Time(ms).



HDP

Work Folder : C:\temp

Dataset : C:\temp\HDS.MER

Report File : C:\temp\Report.csv

Report generated Successfully

Fig. 5. Execution Processes of the proposed E-SVC-WRF

5. RESULTS AND DISCUSSION

As shown at the Fig. 5, Execution Processes of the proposed E-SVC-WRF and the experimental set up is created in with BioWeka and simulations are carried out with cleverland Data Sets, Master.MER.

The proposed Ensemble Support Vector Classifier – Weighted Random Forest called Ensemble SVC-WRF (E-SVC-WRF was trained effectively with the datasets and tested to measure its prediction accuracy. The performance metrics Sensitivity, Specificity, and FScore were calculated by repeating experiments which are shown in the Figures Fig. 6, Fig. 7, and Fig 8.

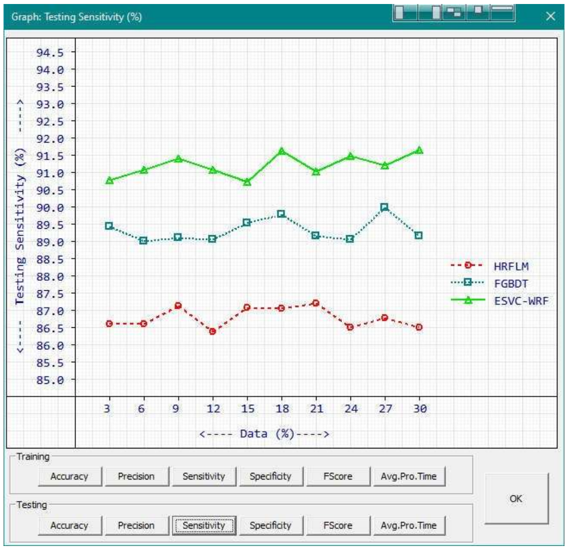


Fig. 6. Performance Analysis – Sensitivity of E-SVC-WRF during Testing

. From the figures, it was noticed that the proposed model was predicting the patterns effectively as compared with the exiting models Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT) and Hybrid Random Forest-Linear Model(HRFLM).

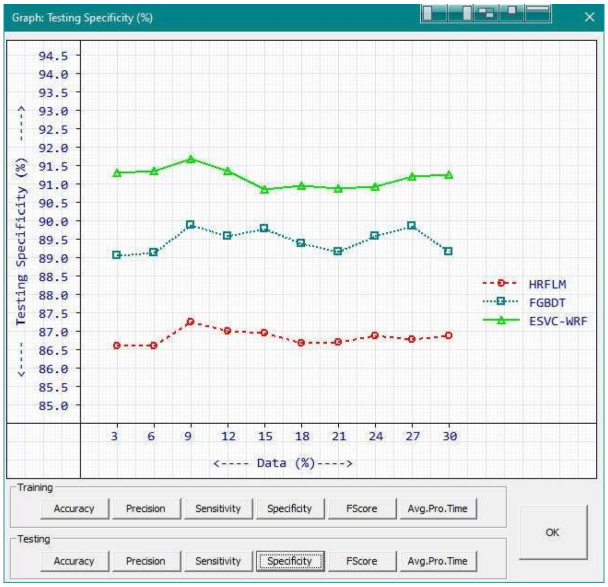


Fig. 7. Performance Analysis – Specificity of E-SVC-WRF during Testing

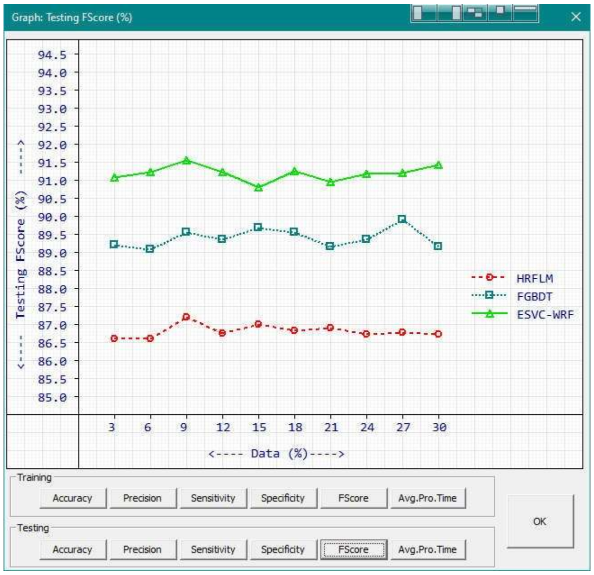


Fig. 8 Performance Analysis – FScore of E-SVC-WRF during Testing

The proposed model is achieving better classification and prediction in terms of True Positive Rate and True Negative Rate as well.

From the Fig. 9, it was observed that during training and Testing processes, the proposed ESVC-WRF is achieving the best Classification Accuracy. It is also noted from the experimental report that the proposed model has less time complexity.

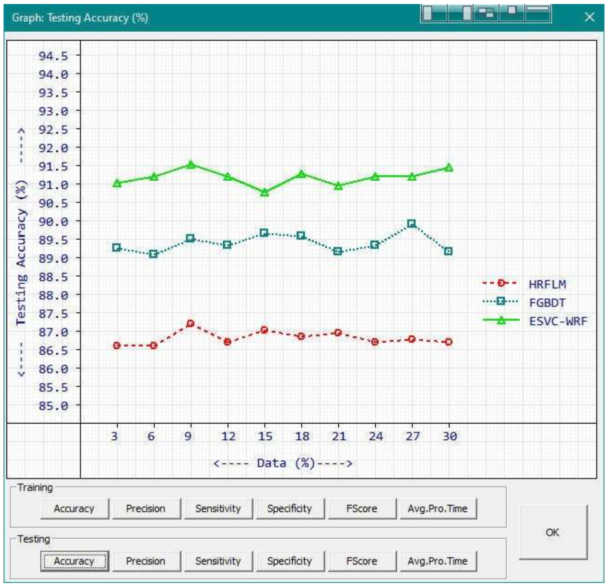


Fig. 9. Performance Analysis – Accuracy of E-SVC-WRF during Testing

6. CONCLUSION

The recently proposed Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT) and Hybrid Random Forest-Linear Model(HRFLM) Classifiers were Studied thoroughly and implemented carefully to analyze its performances during Training and Testing Processes in terms of Prediction Accuracy, Precision, Sensitivity, Specificity, FScore and Average Processing Time(ms). From the experimental results, it was noticed that the Hybrid Random Forest-Linear Model(HRFLM) is performing well during Training and Testing Processes in terms of Prediction Accuracy, Precision, Sensitivity, Specificity, FScore and Average Processing Time(ms) as compared with Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT). However, it is predicted that the Hybrid Random Forest-Linear Model(HRFLM) is not suitable for better prediction when Datasets have more different patterns with very less dissimilarities. To maximize Classification and Prediction accuracy better, this work is proposed an efficient Ensemble Support Vector Classifier – Weighted Random Forest called Ensemble SVC-WRF (E-SVC-WRF) and implemented for analysis. From the experimental results, it was noticed that the proposed Ensemble Support Vector Classifier – Weighted Random Forest called Ensemble SVC-WRF (E-SVC-WRF) is performing well as compared with Hybrid Random Forest-Linear Model(HRFLM) and Bagging-Fuzzy-Gradient Boosting Decision Tree (FGBDT) in terms of Prediction Accuracy, Precision, Sensitivity, Specificity, FScore and Average Processing Time(ms).

REFERENCES

[1] K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, “Classification models for heart disease prediction using feature selection and PCA,” Informatics in Medicine Unlocked, Vol. 19, 2020.

[2] Lakshmanarao, Y. Swathi, and P. S. S. Sundareswar, “Machine learning techniques for heart disease prediction,” International Journal of Scientific & Technology Research, Vol. 8, No. 11, 2019.

[3] Abdallah Abdellatif , Hamdan Abdellatef, Jeevan Kanesan, Chee-Onn Chow, Joon Huang Chuah, and Hassan Muwafaq Gheni, “An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods,” IEEE Access, 2022.

[4] Agrahara A. “Heart Disease Prediction Using Machine Learning Algorithms,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology Vol. 6(4), pp. 137–49, 2020

[5] Alotaibi FS. “Implementation of machine learning model to predict heart failure disease,” International Journal of Advanced Computer Science and Applications, Vol. 10(6), pp.261–268, 2020.

[6] G. Eason, B. Noble, and I. N. and Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, 2021.

[7] Gao X-Y, Amin Ali A, Shaban Hassan H, and Anwar EM. “Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method,” Complexity 2021.

[8] Ghulab Nabi Ahmad, Hira Fatima, Shafiullah, Abdelaziz Salah Saidi and Imdadullah, “Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV,” IEEE Access, pp. 80151 – 80173, 2022

[9] Mohan S, Thirumalai C, and Srivastava G. “Effective heart disease prediction using hybrid machine learning techniques,” IEEE Access, Vol. 7:Pp. 81542–81554, 2019.

[10] P. D. C. Geetha S, Kalaivani V, Haritha CJ, and Preetha G. “Prediction Techniques of Heart Disease and Diabetes Disease using Machine Learning,” Turkish Journal of Computer and Mathematics Education Vol. 12(10):3316–25, 2022.

[11] Padmaja B, Srinidhi C, Sindhu K, Vanaja K, Deepika N, and Patro EKR. “Early and Accurate Prediction of Heart Disease Using Machine Learning Model. Turkish Journal of Computer and Mathematics Education (TURCOMAT) Vol. 12(6), Pp. 4516–28, 2022.

[12] S. Goel, A. Deep, S. Srivastava, and A. Tripathi, ``Comparative analysis of various techniques for heart disease prediction," in Proc. 4th Int. Conf. Inf. Syst. Comput. Netw. (ISCON), Mathura, India, pp. 88-94,2019

[13] V. A. S. Hernndez et al., “A practical tutorial for decision tree induction: Evaluationmeasures for candidate splits and opportunities,” ACM Computational Survey, Vol. 54, No. 1, pp. 1–38, 2021.

[14] Xiaoming Yuan , Jiahui Chen, Kuan Zhang , Yuan Wu and Tingting Yang, “A Stable AI-Based Binary and Multiple Class Heart Disease Prediction Model for IoMT,” IEEE Transactions on Industrial Informatics, Vol. 18, No. 3, 2022.

[15] https://alphaxsalt.medium.com/machine-learning-basics-e1d1be0eff1

[16] https://github.com/VenkateshBH99/Hybrid-Random-Forest-Linear-Model#readme