Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)
IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2

COMPARATIVE STUDY OF MACHINE LEARNING CLASSIFICATION TECHNIQUES TO PREDICT THE CARDIOVASCULAR DISEASES USING HRFLC

[1]PAVITHRA V AND [2]JAYALAKSHMI V

School of Computing Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), India
[1]vpavithra.1989@gmail.com
[2]jayasekar1996@yahoo.co.in

2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432105

Abstract— Due to life style changes more people are facing health issues even at the younger age and Cardiovascular diseases (CAD) was one of the major issues in it. The death rate due to heart disease is higher compared to other health issue. Early prediction of the heart disease (CAD) can help to control the death rate. By adopting to AI, the detection can be faster and accurate which will save the patient life. In this paper the important Features for predicting Heart disease are decided and the Features are applied to different classification model to find the best model out of it. Feature selection is decided based on HRFLC algorithm which is the hybrid algorithm using Random Forest classifier, ADABOOST algorithm and statistical technique Pearson Correlation coefficient. The feature selected based on HRFLC is applied to different Machine Learning models and best model is decided based on different Evaluation Metrices.

Keywords-CAD,ClassificationTechniques,ADABOOST, Gradient Boosting,Random Forest,Svn,,XG Boost

I INTRODUCTION

. In today's world the decision is made based on Data as data is available in large amount due to raise in technology and it is readily accessible. Though the data the available the integrity or privacy of data is maintained by using proper security measures. For AI the data is the source engine and models are built based the data available, the better the data the better in creating Accuracy and realistic model without Bias and Variance. Data mining is the technique which is used to perform the extraction of the data and pre-process with proper pre-processing technique. In medical field huge amount of medical data are not processed properly and each medical record will have some hidden information in it. For prediction we need to explore the data properly. To analyze and explore Data mining tools are used which is useful in extracting and pre-processing of the data from the large dataset.

Cardiovascular disease account for major death rate in the world according to report published by World Health organization in the year 2017.According to the report gender plays a role in it as Men are infected comparatively higher than Women. Detecting the disease earlier will avoid death rate but such detection is difficult as the reason for the disease are different in each scenario. Diabetes, excessive sweating, blood pressure and heart contraction etc(Obasi & Omair Shafiq, 2019)[1].Based on the symptoms there are variety of type for heart disease like arteries blockage, improper circulation of blood and development of heart is not of expected size in the mother womb. (Pavithra & Jayalakshmi, 2020)[2].

The problem to detect the patient is affected by Cardio Vascular disease will be under Supervised learning technique and binary classification problem. Based on the life cycle of the Machine learning algorithm Exploratory data analysis is performed on the data set where unnecessary data like anomaly in the dataset, duplicate data are removed from it. After EDD selection of import features is done from the data set. During feature selection the features which are not correlated and features which are of higher importance in helping for the prediction of results are selected. There are various types of Filer methods like Filter, Wrapper and Embedded methods. The above steps will help the algorithm to be robust and computationally faster as unwanted data are removed [3]. In this paper we try to identify the important feature using new HRFLC technique for a Cardio Vascular Disease (Heart Disease) problem where the accuracy of the algorithm improved using it.

II RELATED WORK

Karthick et al., focuses on an enhance study for predicting a cardiovascular disease occurring in human where age factor is considered for classification. the author focuses on predicting and detecting a cardiovascular disease for adults within an age of 50 years. In the recent survey of united states 10% of people are affected by [4] cardiovascular diseases and it is most common cause of death. cardiovascular diseases not only affect adult people it also affects new born, teenager, toddlers because of some risk factors. for predicting heart disease killana sowmjanya et al., using machine learning and classification algorithm. Machine learning alogorithms help in diagnosing the disease in medical organization in earlier stage [5]. IN this paper they use different classification algorithm KNN, RF, TPOT DT to predict the heart diseases. For best classification rule for diagnosing a heart disease was focused by azhar hussein et al.,.[6] To generate classification rule they uses particle swarm optimization technique 1) rules based on pso algorithm 2) based on their accuracy the rules are encoded and they are improved and finally the results are compared with c4.5 algorithm. predicting and diagnosing the heart diseases using classification technique c. sowmiya et al., focuses on Apriori algorithm with support vector machine to diagnosis a heart disease

978-0-7381-1327-2/21/$31.00 ©2021 IEEE

[7]. M. A Jabbar et al., proposed [8] a new classification model Hidden Navies Bayes for predicting heart diseases. The data are collected from UCI repository they selected 14 feature ,276 instances from the data set. Implementation are done using WEEKA 6.4 tool. The performance evaluation is calculated and showed how the accuracy level was improved in HNB. Different classification techniques are compared for predicting a disease. Two models were built and compared 1) single model to test data 2) combined model to test hybrid data. These models are applied in different classification algorithm, result is compared hybrid approach of the model reach the accuracy level [9]. Nida Khateeb et al., focuses on K nearest algorithm for finding heart disease. The data are collected from UCI REPOSITORY here [10] .1) 14 features was selected without any reduction technique 2) with the help of feature reduction technique 7 features was selected and comparison are done between them. Predictive analysis in the health care system by using different techniques and the usage in the big data analytics [11].

III PROPOSED SOLUTION

. In this paper the features are selected based on the HRFLC MODEL which is a hybrid model which combines the statistical and Machine models to find the important features. The methods used are Pearson Correlation Coefficient, Random Forest Classifier and ADABOOST algorithm. Based on these methods the significant features are selected by comparing the importance of feature in prediction the target value and checking multicollinearity between the features based on threshold value which helps to reduce duplicate features and also eliminate the unnecessary features. The proposed method will be applied to different machine models and evaluation metrices is used to select the better model. The evaluation metrices used are shared in the paper and comparison results are captured based on it.

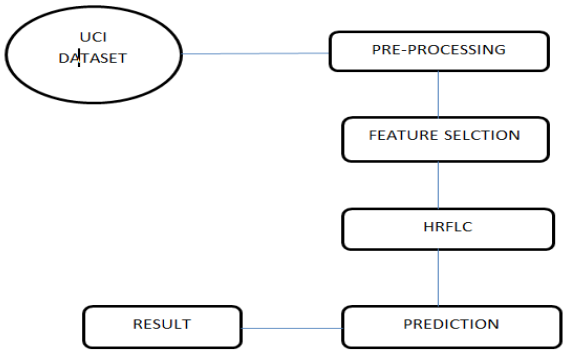

Figure 1. Prediction of heart diseases with HRFLC

IV CLASSIFICATION MODELLING WITH FEATURE SELECTION TECHNIQUES

Classification modelling comes under Supervised Learning as the results of the models can be compared with actual results. Based on the Performance and evaluation matrices like accuracy the models can be tuned further and results will be improved further.

A. Data Preprocessing

UCI dataset is used for this experiment, the dataset contains 303 records and 13 features. From 303 records 270 records are used for the experiment after preprocessing of data is performed where duplicate records and missing value records are removed.

B. FEATURE SELECTION

Feature selection is important step in the life cycle of Machine learning project as it will increase the performance and Accuracy of the models. The dataset used for the experiment is from UCI dataset. Out for 13 features from the dataset 9 features are identified based on proposed “Hybrid Feature selection Technique (HRFLC) for Prediction of Cardio vascular Disease”

'resting_blood_pressure', 'chest','age', 'thal', 'maximum_heart_rate _achieved', 'slope', 'number_of_major_vessels','serum_cholestoral', 'oldpeak',

C. LOGISTIC REGRESSION

Logistic Regression is the basic supervised classification model. The Target will take discrete value based on the input features. The model is similar to linear regression but the line will be s Shaped curve and the probability of patient having heart disease or not is decided based on Maximum like hood [12]. The best line is selected based on Maximum like hood. The equation of logistic regression is below

$$Y=1/1+e^{-z} \quad (1)$$

Which is the sigmoid function. The value of z is calculated based on the linear equation

$$z=y+mx+c \quad (2)$$

when the value of liner regression line is applied to sigmoid function the value transforms from 0 to1 where 1 indicates positive and 0 indicates negative.

***D. DECISION TREE*:**

Decision tree is one of the supervised learning technique which is simple and tree based selection. Decision tree classifies the data by sorting them from root to terminal/leaf nodes and leaf nodes provides the classification results. The
model is recursive and nature and assumes initially all the data as root node, feature values are assumed to be categorical and the ordering of leaf is done through statistical model. Decision tree use multiple algorithms to decide how the root node is selected and traverse through the subsequent nodes.

On each iteration from the given data set it calculated the Entropy and Information gain to split the features. The lower the Entropy gain or higher the information gain helps to find the best split of features in the decision tree and the process is iterated multiple times until we reach the leaf node.

E. ENTROPY

978-0-7381-1327-2/21/$31.00 ©2021 IEEE

Entropy is used to calculate the randomness in the input dataset, so the higher the entropy higher the uncertainty in splitting the data, Entropy is calculated based on the formula below where the degree of uncertainty is reduced on each split.

$$H(X) = -\Sigma (p_i * \log_2 p_i) \quad (3)$$

F. INFORMATION GAIN:

Information gain provided information how much maximum gain is achieved on splitting the tree based on particular node. During splitting feature with higher information gain is selected as root node and subsequently sub nodes are selected.

Information Gain= Entropy(previous)-ΣEntropy(j, current) **(**4)

G. GINI INDEX:

Gini Index can be used to split the node to measure the impurity in the data and it Is calculated based on the formula below

$$I_{G(n)} = 1 - \sum_{i=1}^{J} (Pi)^2 \quad \textbf{(5)}$$

H. K NEAREST NEIGHBORS CLASSIFIER:

K Nearest classifier algorithm classifies problem based on the distance from the current point to be nearest point. Based on the classification results of the nearest points the current point is classified. The distance is calculated based on Euclidean distance and value of "K" have to provide as input based on Cross validation.

Euclidean Distance = sqrt (x2-x1)2 – (y2-Y1)2 (6)

K Nearest Neighbor Algorithm:

1) Select 'K' value for the number of points to which distance to be measured.
2) Calculate the distance for the "K" points selected.
3) Based on target classifier for the nearest points the classifier for the current point is predicted.

I. Random Forest Classifier:

Random forest classifier is an ensemble technique algorithm for regression and binary classification problems. Random forest come under bagging technique [13]. where the data set is split to multiple trees and target is calculated based on the consolidated results from different decision tree.

1) Based on the number of the records randomly split the data in to K different tress
2) The data can be redundant and will be available in different tress
3) For each tree based on the records selected the base node is identified using gini index which provides maximum information gain.

   Information gain(n) = $1 - \sum_{i=1}^{J} (Pi)^2$ (7)
4) Select subsequent node based on the gini index on the remaining nodes.
5) Select a threshold value below which further node branching is not done.
6) Repeat the steps for all the trees
7) Decision classification is made based on the maximum voting result observed from different trees. .

J. ADABOOST

Adaptive boosting algorithm is a boosting technique which provides increased accuracy and the technique is prone from anomalies in the data. It combines many weak classifiers and finally create a strong classifier [14]. In random Forest tress are created parallelly but in Ada boost the model is created sequentially from weak classifier to strong classifier. The trees in the Adaboost are called stumps and it will have only leaf node and the stump is calculated on the Gini Index. Each stump is provided higher weigh in the subsequent classification.

Adaboost feature selection steps:

1) Provide equal weightage for all the features by finding the Mean weight

   Mean weight = 1/Total no of samples

2) Select the 1st stump after calculating the Gini Index

3) Find the total error of the stump for the wrongly classified data.

   Amount of say = ½ log (1- total error/Total error)

4) Based on the incorrect classified samples increase the weightage of it and new stump is created.

   New Sample weight = Sample Weight x e amount of say

5) Decrease the weightage of the correctly classifies samples done in step2

   New Sample weight = Sample Weight x e- amount of say.

6) Based on the new weights create again the Mean weight

7) Now again the stump is selects with new weights and steps are repeated

.

8) Similar to Decision tress voting is done to classify the data.

K Naive Bayes Classifier:

Naïve bayes classifies the problem based on the Bayes theorem, here each feature is compared with target and they are independent of each other. Data set in the algorithm is classified as Feature matrix and response vector. Feature Matrix are independent variable and Response vector is the target variable [15]. Naïve Bayes theorem used below formula to calculate the response classifier.

978-0-7381-1327-2/21/$31.00 ©2021 IEEE

Likelihood Class prior probability

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)}$$

Posterior probability predictor prior probability

$$P(c|x)=p(x_1|c) * p(x_2|c)*\ldots* p(x_n|c)*p(c)$$

- The posterior probability of class (target) P(c|x) is given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

L. Gradient Boosting Classifier:

Gradient boosting classifier is an example of ensemble boosting technique. It is similar to Adaboost where the weak classifiers are combined to produce the strong classifier but it differs from Adaboost from execution logic, in Adaboost weak classifier weights are updated in sequential step which are classified wrongly, in Gradient boosting the process of updating is based on Loss function/Residuals. Boosting algorithm provides greater accuracy and performance.

Gradient boosting Algorithm.

1) Log of odds is calculated based on the number of target output

   Log(winning/loss)

2) Find the probability of the log of odds by

   $e^{\log(odds)} / 1+ e^{\log(odds)}$

3) Find the residual or loss from the target based on the step 2.

   Residual= Observed-Prediction Observer value is 1 for "yes" target and "0" for "No" target

4) Create the model decision tree based on the independent variables and target variable as Residual we found on the step 3.

5) Apply transformation by using the formula below.

$$\frac{\sum residual}{\sum previous\ probablility\ i*(1-previous\ probabiity\ i)}$$

6) Now update the prediction by combining the initial leaf with new tree by including the learning rate.

   Log(odds) prediction= Log odds (initial leaf) + learning rate + new probability value (step 5)

7) Now repeat step 2 to 6 until we the residual value is near the actual value.

8) The probability calculated in step 2 if greater than threshold value then it will be classified as "Yes" and "No" if below threshold value for the new data.

M . XG Boost Model:

Extreme gradient boosting is most famous model of ensemble boosting technique and it is framework based on gradient boosting which handles missing values, outlier and pruning of trees effectively [16]. It is achieved by using 3 parameters Regularization, Similarity weight and learning rate.
XGboost algorithm:

1) Log of odds is calculated based on the number of target output

   Log (winning/loss)

2) Find the probability of the log of odds by

   $e^{\log(odds)} / 1+ e^{\log(odds)}$

3) Find the residual or loss from the target based on the step 2.

   Residual= Observed-Prediction
   Observer value is 1 for "yes" target and "0" for "No" target

4) Create the decision tree based on the independent variables and target variable as Residual we found on the step 3.
5) Find the similarity weight using formula below

$$\frac{\sum residual\ 2}{\sum previous\ probablility\ i*(1-previous\ probabiity\ i)+\lambda}$$

   Where λ is the pruning value.

6) Find the information gain of each branch by using formula below

   Gain = Sum of Residual of each branch after split – residual value before split.

7) While calculation the residual of each leaf the branch will happen on if residual value is greater than the λ

8) value. This is how pruning is achieved in XG boost

9) Find the information gain the values independent features branching is done.

10) Now to calculate new probability value by applying the below formula value to log of odds formula

978-0-7381-1327-2/21/$31.00 ©2021 IEEE

New probability = Previous probability + learning rate +Similarity weight of the leaf

11) The steps from step 3 to 9 are repeated again and again until the residual value is minimal.

IV EVALUATION PARAMETERS

For selecting best machine learning model 4 important metrices are used Recall, F1 score, Precision and accuracy. Based on the business need any of the metrices can be used. Based on the evaluation parameter the accuracy rate can be calculated:

Precision:
It is the defined as the positive data point selected by the model and the actual positive data from the dataset.

$$P=Tp/Tp+Fp$$

Recall:
Recall (R) is defined as the number of positive data point over total positives in the actual data set

$$R=Tp/Tp+Fn$$

F1Score:
This is the contribution of both Precision and recall and hence it is defined as the harmonic mean of it.

$$F1=2P\times RP+R$$

V RESULTS AND OBSERVATION:

Applying HRFLC model on different machine learning algorithm produced below results. As observed Naive Bayes algorithm is producing better results while comparing with other machine learning Models followed by SVN algorithm.

	Model	Accuracy	F1Score	Precision	Recall
0	Logistic regression	0.765432	0.732394	0.742857	0.722222
1	Decision tree	0.740741	0.695652	0.727273	0.666667
2	Random Forest	0.740741	0.695654	0.727273	0.666667
3	SVN	0.827160	0.794118	0.843750	0.750000
4	XGboost	0.765432	0.716418	0.774194	0.666667
5	Adaboost	0.777778	0.742857	0.764706	0.722222
6	Gradient Boosting	0.753086	0.705882	0.750000	0.666667
7	KNN	0.790123	0.760563	0.771429	0.750000
8	Naïve Bayes	0.864198	0.845070	0.857143	0.833333

TABLE1. Accuracy score of Different Machine learning Model

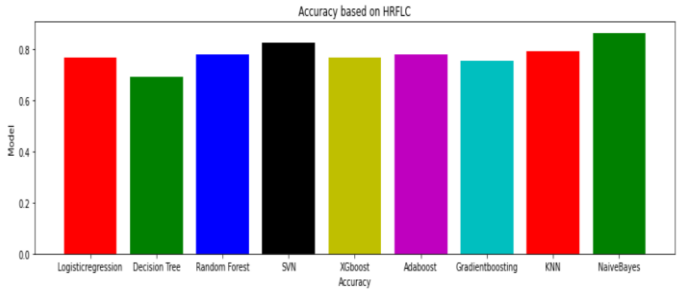

FIGURE 2. Accuracy score of Different Machine learning Model

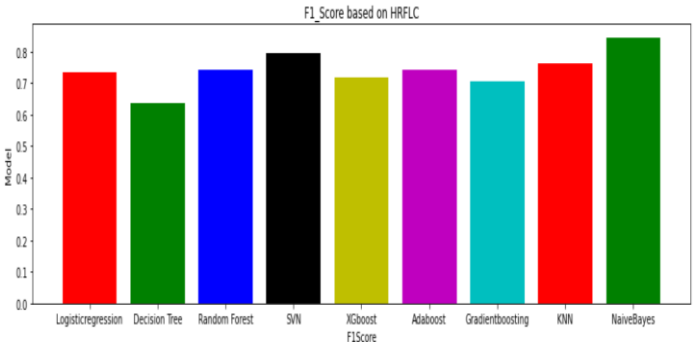

FIGURE 3.F1 score of Different Machine learning Model

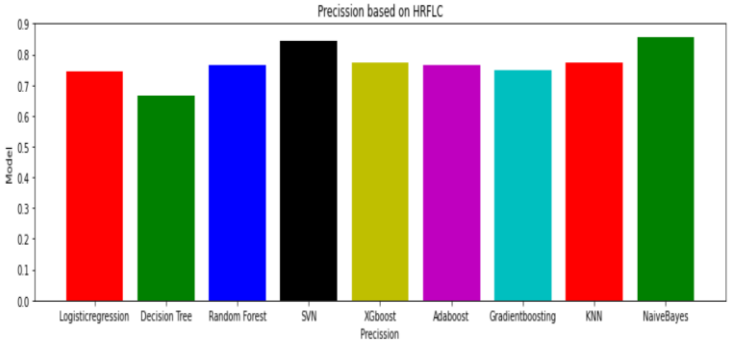

FIGURE 4. Precision score of Different Machine Learning Model

978-0-7381-1327-2/21/$31.00 ©2021 IEEE

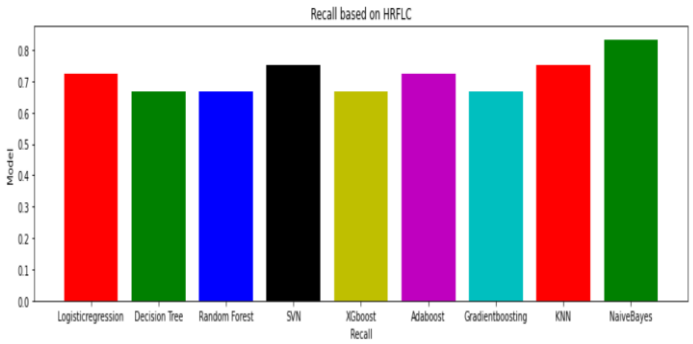

FIGURE 5. Recall score of different Machine Learning Model:

CONCLUSION

This paper is based on the proposed model HRFLC which select best features for the given dataset for Heart disease. Based on the 9 features selected by the model the selected features are implemented on different machine learning algorithm. The dataset is evaluated with different metrices such as F1 score, Precision, Recall and Accuracy and it is observed Naïve bayes algorithm is producing better results than other models. The new feature selection technique to be applied on different medical data set as part of future work to find the accuracy of it.

REFERENCES

1 Obasi, T., & Omair Shafiq, M. (2019). Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 2393–2402.

2 Pavithra, V., & Jayalakshmi, V. (2020). A Review on Predicting Cardiovascular Diseases Using Data Mining Techniques. *Lecture Notes on Data Engineering and Communications Technologies*, *49*, 374–380.

3 Pavithra, V., & Jayalakshmi, V. (2020, June). Review of Feature Selection Techniques for Predicting Diseases. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1213-1217). IEEE.

4 Karthick, D., & Priyadharshini, B. (2018). Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using classification techniques within fifty years of age. *Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018*, *Icisc*, 1182–1186.

5 Sowjanya, K., & Krishna Mohan, G. (2020). Predicting heart disease using machine learning classification algorithms and along with tpot (Automl). *International Journal of Scientific and Technology Research*, *9*(4), 3202–3210.

6 Hassoon, M., Kouhi, M. S., Zomorodi-Moghadam, M., & Abdar, M. (2017). Using PSO Algorithm for Producing Best Rules in Diagnosis of Heart Disease. *2017 International Conference on Computer and Applications, ICCA 2017*, 306–311.

7 M. A. Jabbar and S. Samreen, (2016) Heart disease prediction system based on hidden naïve bayes classifier,2016 *International Conference on Circuits, Controls, Communications and Computing (I4C), Bangalore,* 2016, pp. 1-5.

8 C. Sowmya and P. Sumitra, (2017). Analytical study of heart disease diagnosis using classification techniques,'' in Proc. *IEEE Int. Conf. In tell. Techno. Control, Optima. Signal Process. (INCOS), Mar.* 2017, pp. 1–5

9 Purusothaman, G. and Krishnakumar, P., "A survey of datamining techniques on risk prediction: Heart disease", *Indian Journal of Science and Technology, Vol. 8, No. 12, (2015).*

10 Khateeb, Nida, Usman Muhammad "Efficient Heart Disease Prediction System using K-Nearest Neighbour Classification Technique" *Proceedings of the International Conference on Big Data and Internet of Things, London, United Kingdom, December*20-22, 2017.

11 Smys, S. "Survey on accuracy of predictive big data analytics in healthcare." *Journal of Information Technology 1, no. 02 (2019): 77-86.*

12 Tarawneh, M., & Embarak, O. (2019, February). Hybrid approach for heart disease prediction using data mining techniques. *In International Conference on Emerging Internetworking, Data & Web Technologies (pp. 447-454). Springer, Cham.*

13 Ray, P., Kharke, R. B., & Chauhan, S. S. Cardiovascular Disease Classification Using Different Algorithms. In *Inventive Communication and Computational Technologies* (pp. 189-201). Springer, Singapore.

14 B.Venkatalakshmi, M.V Shivsankar, "Heart Disease DiagnosisUsing Predictive Data mining", IJIRSET Volume 3, Special Issue3, March 2014 ,pp. 1873-1877

15 Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.

16 Amin, Mohammad Shafenoor, Yin Kia Chiam, and Kasturi Dewi Varathan. "Identification of significant features and data mining techniques in predicting heart disease." *Telematics and Informatics* 36 (2019): 82-93.

978-0-7381-1327-2/21/$31.00 ©2021 IEEE