Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 4s (2024)

A Novel Hybrid SDR Model for Dengue Prediction using Opt_Recurr Feature Selection Algorithm

Sharlie Vasanthi . N[1], Dr. S. Nagasundaram[2]

[1]Research Scholar , PhD Computer Science, Vels Institute of Science Technology and Advanced Studies , Chennai 600117

Associate Professor, Department of Computer Science and Technology

Women ‘s Christian College

Chennai-6, nsharlie74@gmail.com

[2]Research supervisor, Assistant Professor, Department of Computer Applications, Vels Institute of Science, Technology and Advanced Studies, Chennai 600117,

snagasundaram.scs@velsuniv.ac.in

Article History:

Received: *26-04-2024*

Revised: *12-06-2024*

Accepted: *25-06-2024*

Abstract:

Dengue is a vector borne disease , which can be fatal at times . Early detection dengue of dengue is vital as no vaccines have been developed for dengue yet..The process of eliminating irrelevant and redundant features from the data set facilitates the optimal features selection . This study is proposed to select the Optimal features by using Opt_Recur algorithm from the Dengue dataset , a hybrid SDR model which makes prediction with better accuracy when compared to the conventional classifiers like the Support Vector Machine (SVM), Decision Tree(DT) and the Random Forest (RF) classifiers.

Keywords: Dengue, Hybrid SDR Model, Opt_Recurr algorithm, Feature selection, SVM. DT, RF.

1. Introduction

Dengue fever is a viral disease that spreads quickly by the mosquito in warm weather. It is transferred by a female mosquito known as 'Aedes aegypti.'. Dengue cases have increased dramatically in recent years over the world. However, the actual number of dengue infections is either never recorded or is classified incorrectly. Dengue fever is a fatal disease that is widespread by viral infections. It is a rapidly spreading tropical virus infection with an increased death rate[6]

According to the World Health Organization , Dengue is a fatal disease so early diagnosis is a must. There are many machine learning methods help physicians to early diagnose the disease, which makes accurate prediction using the classification techniques which can save up time.[13].The machine learning techniques use statistical methods to select optimal features from the dataset . Optimal features helps not only to make accurate prediction it also speeds up the execution time.

There are many machine learning algorithms that help to make predictions of the disease In this paper it is put forth that traditional classifiers like the SVM, Decision tree and Random Forest has applied on the optimal features were selected by the Opt-Recur Algorithm . This study makes comparison among the prominent classifier like SVM, Decision tree and Random Forest ,and also with proposed hybrid SDF model in terms of their accuracy rate, Precision, Recall and F1-score. The ROC curve is finally used to evaluate the performance measurement.[14].

2. Related Work

The weighted score evaluation values were first combined using the Weighted Score Arithmetic Averaging (WSAA) operator. Secondly, the functions for scoring were computed. Third, based on the score values, the best option is chosen. Ultimately, it is possible to compute the score values and get the ranking outcomes. Then, the J48 appears to be the best classifier out of the four traditional classifiers: Naïve Bayes (NB), Decision Tree (J48), Multi Layer Perceptron (MLP), and Support Vector Machine (SVM) [2] has concluded from the study.

Naïve Bayes, KNN, and J48 feature selection techniques are used in PCA and Wrapper. Different ANN models have been created for every method of feature selection. Of the four feature selection techniques, PCA produces an ANN with a greater accuracy. In summary, PCA works better with the provided dataset. The two most expressive characteristics selected by all wrapper feature selection methods are myalgia and retro-ocular pain [4] has been proposed in the study..

The important features were chosen using the features selection technique. Z-Score was used to standardize the data. To address the imbalance issue in the dataset, the SMOTE+ENN hybrid technique was utilized to divide the dataset into training and testing sets using cross-validation techniques such 10-fold and Holdout cross-validation. Following that, machine learning models were created, and the accuracy, F1-score, precision, recall, and AUC scores were used to assess how well they performed [6].

In [10] concluded that when compared to LDA and KNN, the chosen models DT, CART, and NB demonstrated great accuracy. But when it comes to accuracy, recall, sensitivity, and specificity—all of which are determined by looking at the classification matrix—DT was discovered to be the best. Patients with dengue fever: this is the group from which we can also highlight the importance of prompt medical attention for those exhibiting warning indicators. Out of the pool of classifiers that were chosen, the suggested model produced good classification characteristics including 99.8% accuracy, 0.86 precision, and 1.0 recall, which indicated promise.

3. Methodology

Feature Selection

A crucial stage in data processing before putting the information into a learning system is feature selection. Removing redundant and irrelevant input improves the machine learning algorithm's performance.

The efficiency and efficacy of many current feature selection techniques are severely hampered by the growth in the dimensionality of data. This study examines several popular feature selection strategies and examines how well they may be applied to attain high machine learning algorithm performance, which in turn enhances the classifier's predicted accuracy.

As a result, the classifiers processes the data more quickly and accurately, which also improves accuracy. The classification accuracy can be significantly impacted by irrelevant information, such as noisy data. By using feature selection approaches, data handling can be enhanced while successfully lowering costs. Feature selection techniques are frequently employed to boost a classifier's capacity for generalization. To obtain the best accuracy, we compare the dataset's result with significant attributes.

Filter, wrapper, and embedding methods are the three categories into which feature selection techniques fall. The data is preprocessed using filter algorithms. In order to compute and forecast the target feature, these take into account the correlation between features. To find the high rank feature, a variety of statistical tests are run on the characteristics [3].

The proposed Opt_Recur Algorithm runs through a series of steps which makes accurate prediction of desired features which are needed to make accurate prediction of Dengue .

Opt_Recur Algorithm:

1. Input all the attributes from the dataset.
2. Find the correlation coefficient of all the attributes.
3. Select the attributes which are highly correlated by ranking them and filtering the attributes above the threshold .
4. The selected attributes are passed into RFE algorithm which further eliminates the undesired attributes and iterates until the desired attributes are reached.

The Opt_Recur algorithm selects optimal features which are need , it ranks all the attributes according to the scores of the correlation coefficient and the heat map which visualizes the correlation scores of these variables is shown in Fig 1.

The Pearson correlation coefficient is used for correlation analysis in appropriate feature selection. Values range from -1 to 1, and it can be used to calculate the correlation between two variables. Features are generally regarded as associated if their Pearson correlation coefficient is higher than 0.3. [19]. In this study those features with the correlation coefficient less than 0.3, have been filtered out .Those attributes which are above the threshold is further passed on to the Recursive Feature Elimination Algorithm which selects the optimal features which are needed for making accurate prediction .

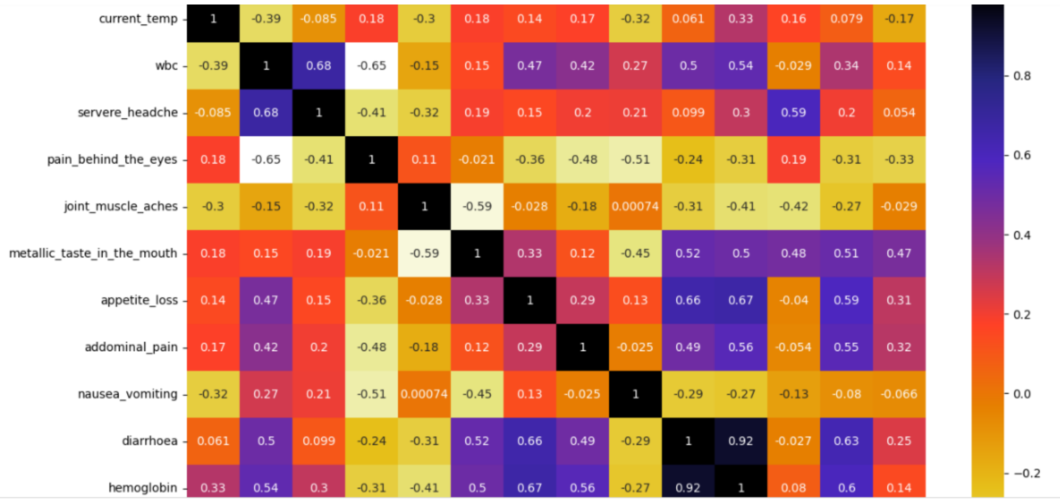

Fig 1. Heat Map of the ranked attributes for feature selection

Hybrid SDR Model

After the optimal features are selected the dataset is split into testing data and the training data . The base model is chosen which makes accurate prediction on the test and train data. In the Proposed SDR hybrid model we use have used Support Vector Machine , Decision Tree and Random Forest Classifiers as the base model , each of these base model make accurate prediction on data. The predictions which have been made are fed into the meta model . The meta model is the Logistic regression which combines all the predictions made from traditional classifiers . The Architecture of the proposed model is shown .

The SDR hybrid Model can be represented mathematically as Y_Hybrid. let ($Y1,Y2,Y3\dots Y_n$) be represented as the Predictions of the base model . Y1 represents prediction of Support Vector Machine , Y2 represents prediction of Decision Tree , Y3 represents prediction of Random Forest are the base models. The meta model which is the Logistic regression which combines all the base models are represented as Y_meta .

$$Y\_Hybrid= Y\_meta(Y1,Y2,Y3\dots Y_n) \quad (1)$$

The Proposed SDR Hybrid Model are finally compared with other Conventional classifiers like as Support Vector Machine , Decision Tree and Random Forest in terms of accuracy rate, Precision rate , Recall rate and F1-score . The ROC curve which makes additional measurement of proposed model.

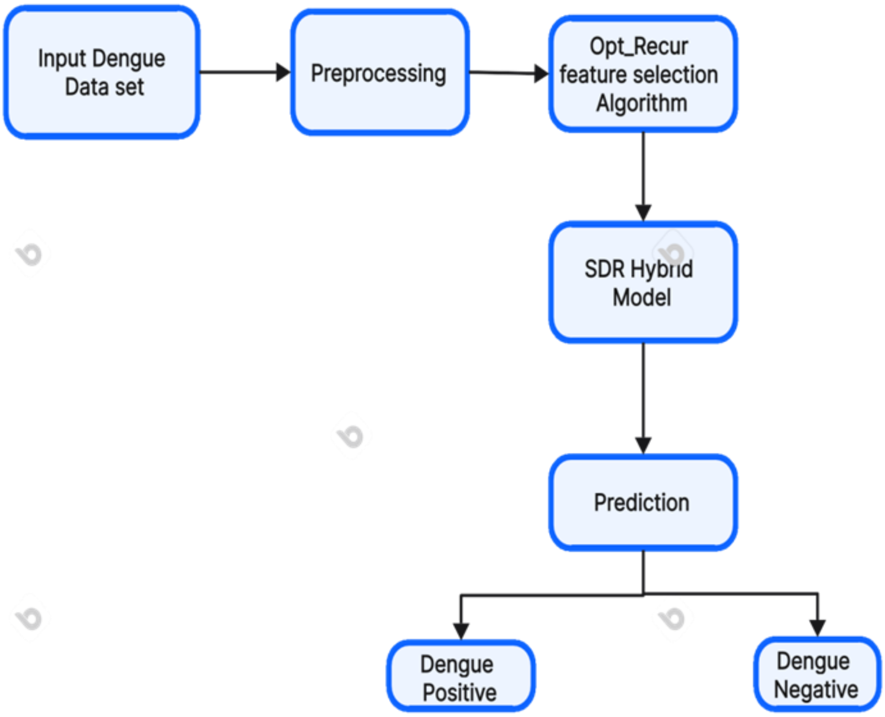

Fig 2. The proposed Architecture of the Hybrid model

4. Results and Discussions

In this proposed study , the data set is first preprocessed then the dengue dataset is validated by splitting the input data set into test and train data. Before applying the any feature selection techniques the dengue data set is used to make prediction using conventional classifiers such as SVM , DT and RF . The measurement of the prediction is made in terms of the accuracy rate , Precision rate , Recall rate and F1-Score .

With the exception of time, the confusion matrix aids in the calculation of all metrics. The components of the confusion matrix are false positive (FP), false negative (TN), false positive (TP), and false-negative (FN). The most significant prediction in health care statistics is a false negative [6]. The performance measures that are used to evaluate can be represented mathematically.

$$\text{Accuracy} = \left\{ {(TP+TN)} \big/ {(\text{TP+FP+TN+FN})} \right\} X\,100 \tag{2}$$

$$\text{Precision} = \left\{ {TP} \big/ {(\text{TP+FP})} \right\} X\,100 \tag{3}$$

$$\text{Recall} = \left\{ {TP} \big/ {(\text{TP+FP})} \right\} X\,100 \tag{4}$$

$$\text{F1 Score} = \left\{ {Precision\ x\ Recall} \big/ {(Precision + \ Recall} \right\} X\,100 \tag{5}$$

The performance evaluation of the conventional classifiers like the SVM , Decision Tree and the Random Forest , then in the hybrid SDR model without applying feature selection Opt_Recur algorithm is shown in Table 1. The proposed Opt_Recur feature selection approach is applied to standard classifiers such as Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF), and the results are compared in terms of Accuracy, Precision, Recall, and F1 score. With the proposed hybrid SDR model after that is shown in Table 2

Table 1. Performance Evaluation of SVM, DT,RF Without applying Opt_Recur Feature Selection Algorithm

Classifier	Accuracy	Precision	Recall	F1-Score
SVM	77%	92%	72%	81%
DT	80%	86%	92%	89%
RF	94%	88%	85%	87%
Hybrid SDR	96%	91%	92%	96%

Table 2. Performance Evaluation of SVM, DT,RF after applying Opt_Recur Feature Selection Algorithm

Classifier	Accuracy	Precision	Recall	F1-Score
SVM	85%	100%	77%	87%
DT	93%	90	100%	95%
RF	96%	96%	92%	96%
Hybrid SDR	96%	100%	93%	97%

The proposed study depicts that hybrid SDR model performs well when compared to other conventional classifiers like the SVM , DT and the RF before applying the Opt_Recur Feature Selection algorithm with an Accuracy rate of 96% , Precision rate of 91% , Recall rate of 92% and F1 score 96%.

The evaluation metrics after applying the Opt_Recur Feature Selection algorithm on the conventional classifiers like the SVM ,DT and the RF .The hybrid SDR model showed better performance than others classifiers with Accuracy of 96% , Precision of 100% , Recall of 93% and F1 score 97%.

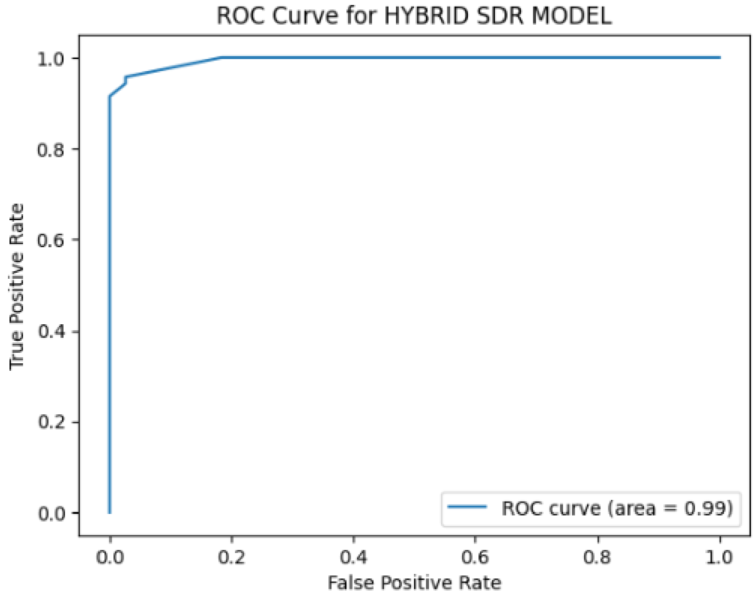

Fig 3. ROC Curve for the hybrid SDR model

The performance of the classification models is typically estimated diagrammatically using the Receiver Operating Characteristic (ROC) curve, which spans all practical thresholds. By tracing the FPR on the x-axis against the TPR on the y-axis, a ROC curve is produced. Since ROC is unbiased toward both classes, it is significant when training results show a change in the number of instances of either class. The optimal classifier has a range under the ROC that is near to 1.

[16]. ROC curve of the proposed hybrid SDR model shows the value of .99 is proved to be the best classifier.

5. Conclusion

Dengue is one deadliest disease in the world , the proposed study helps in the accurate prediction at the early stage will save human life . The proposed Opt_Recur algorithm is feature selection algorithm which make optimal selection feature which are needed for making accurate prediction. The hybrid SDR algorithm which is found to produce better efficiency even when compared to traditional classifiers like the Support Vector Machine (SVM) , Decision Tree(DT) and the Random Forest(RF). The ROC curve also provided the hybrid model had out performed the traditional classifiers .

References

[1] Rian Budi Lukmantoa, Suharjitoa, Ariadi Nugrohoa, Habibullah Akbara, ‘Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine ’, *4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12–13 September 2019* , pp. 48–54, 2019.

[2] S. Appavu alias Balam, , G. Chinthana urugan , M.S. Mohamed Mallick , ‘Improved prediction of dengue outbreak using combinatorial feature selector and classifier based on entropy weighted score based optimal ranking ’, *Informatics in Medicine Unlocked* , JULY2020, doi: https://doi.org/10.1016/j.imu.2020.100400.

[3]. Gopika ,A.Meena kowshalaya , ‘ Correlation Based Feature Selection Algorithm for Machine Learning ’, *Proceedings of the International Conference on Communication and Electronics Systems (ICCES 2018)*, pp. 692–695.

[4] Sruthi Nair ,Abhishek Gupta ,Dr. Vidya Chitre ,Raunak Joshi, ‘Combining Varied Learners For Binary Classification Using Stacked Generalization’, :*2202.08910v1 [cs.LG] 17 Feb 2022* , FEB2022.

[5] R Lathesparan, RMKT Rathnayaka, WU Wickramaarachchi, ‘Finding the Best Feature Selection Method for Dengue Diagnosis Predictions’, h*ttps://www.researchgate.net/publication/360587139*, pp. 123–129, MAY2022.

[6] Bilal Abdualgalil , Sajimon Abraham , Waleed M. Ismael , ‘Early Diagnosis for Dengue Disease Prediction Using Efficient Machine Learning Techniques Based on Clinical Data’, *Journal of Robotics and Control (JRC)*, vol. 3, no. 3, pp. 257–268, May 2022, doi: 10.18196/jrc.v3i3.14387.

[7] Xi Li,Michèle ,Curiger Thomas ,HanneRolf Dornberger, ‘Optimized Computational Diabetes Prediction with Feature Selection Algorithms’, *ISMSI 2023, April 23, 24, 2023, Virtual Event, Malaysia*, pp. 36–43, Jul. 2023, doi: https://doi.org/10.1145/3596947.3596948.

[8] Robson Aleixo,Fabio, Kon,Rudi Rocha, Marcela Santos Camargo, Raphael Y. de Camargo, ‘Predicting Dengue Outbreaks with Explainable Machine Learning’, *http://www.riocomsaude.rj.gov.br/Publico/MostrarArquivo.aspx?C=NqviPkhBljU%3D*.

[9] R Arafiyah, F Hermin, I R Kartika, A Alimuddin and I Saraswati, ‘Classification of Dengue Haemorrhagic Fever (DHF) using SVM, naive bayes and random forest ’, *IOP Conf. Series: Materials Science and Engineering 434* , 2018, doi: 10.1088/1757-899X/434/1/012070.

[10] Supreet Kaur , Dr Sandeep Sharma, ‘Comparative Analysis of Machine Learning Classifiers on Forecasting Dengue Fever Infection ’, *Recent Developments in Electronics and Communication Systems*, pp. 492–497, 2023, doi: doi:10.3233/ATDE221302.

[11] M G Dinesh, D.Prabha , ‘Diabetes Mellitus Prediction System Using Hybrid KPCA GA-SVM Feature Selection Techniques’, *ICDIIS 2020 Journal of Physics: Conference Series*, 2021, doi: 0.1088/1742-6596/1767/1/012001.

[12] Salim G. Shaikh , Dr. B. Suresh Kumar, Dr. Geetika Narang Prof. N.N.Pachpor , ‘Diagnosis of Vector Borne Disease using Various Machine Learning Techniques’, *Intelligent Systems And Applications In Engineering*, pp. 517–526, Feb. 2023.

[13] P.K. Swaraj, G. Kiruthiga, ‘Design And Analysis On Medical Image Classification For Dengue Detection Using Artificial Neural Network Classifier’, *ICTACT JOURNAL ON IMAGE AND VIDEO PROCESSING*, vol. 11, no. 3, pp. 2407–2411, Feb. 2021, doi: 10.21917/ijivp.2021.0343 .

[14] Yosef Masoudi-Sobhanzadeh, Habib Motieghader , Ali Masoudi-Nejad, ‘FeatureSelect: a software for feature selection based on machine learning approaches’, *MC Bioinformatics*, 2019, doi: https://doi.org/10.1186/s12859-019-2754-0.

[15] T.Sajana, M.Navya, YVSSV.Gayathri, N.Reshma, ‘Classification of Dengue using Machine Learning Techniques ’, *International Journal of Engineering & Technology* , pp. 212–218, 2018.

[16] Dhiman Sarma ,sohrab Hossain, Tanni Mittra ,Md. Abdul Motaleb Bhuiya ,Ishita Saha ,Ravina Chakma , ‘Dengue Prediction using Machine Learning Algorithms’, *2020 IEEE Xplore(R10-HTC) | 978*, 2021, doi: 10.1109/R10-HTC49770.2020.9357035.

[17] Rajeev Kapoora,Varinder Kadyanb , Sachin Ahuja, ‘Identification of influential parameter for early detection of Dengue using Machine Learning Approach.’, *proceedings of 5th International Conference on Cyber Security & Privacy (ICCS) 2019* , pp. 129–135.

[18] Rajeev Kapoor, Variender Kadyan, Sachin Ahuja, ‘Weight Based- Artificial Neural Network (W-Ann) For Predicting Dengue Using Machine Learning Approach With Indian Perspective’, *International Journal Of Scientific & Technology Research Volume 9, Issue 02, February 2020* , vol. 9, no. 2, pp. 3290–3298, Feb. 2020.

[19] Fei Zhou ,HonghaiFan, Yuhan Liu , Hongbao Zhang , Rongyi Ji “Hybrid Model of Machine Learning Method and Empirical Method for Rate of Penetration Prediction Based on Data Similarity”, *Applied Science* 2023, https://doi.org/10.3390/app13105870.