The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2918 - 2926

Improving Classification Accuracy Using Hybrid Machine Learning Algorithm on Clinical Datasets

Dr. R. Thriumalai Selvi
Associate Professor

Sujdha C.
Research scholar

PG & Research Department of Computer Science
Government Arts college for Men, Nandanam, Chennai – 35.

Abstract

In the present era, maintaining a healthy and disease-free life is complex due to multiple personal and environmental impacts. Early identification and diagnosis will help human beings lead a sustainable life. However, to achieve this, health care data has to be processed in an efficient manner with more accuracy. Thus, the impacts of diseases or future impacts can be predicted or detected and proper medication can be provided by the physicians. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. In this paper we have studied and implemented many traditional machine learning algorithms which are K-Nearest Neighbor algorithm (KNN), Support Vector Machine (SVM), Decision Tree (DT) AND Artificial Neural Network (ANN). Based on output we have implemented hybrid model by combining the above mentioned algorithms to archive more accuracy. Accuracy measure has been used to compare the effectiveness and performance of the individual algorithms and the proposed hybrid approach. We have observed that the classification accuracy has been improved for different clinical datasets with the proposed hybrid model using a stacking classifier technique.

Keywords: Hybrid Machine learning, Clinical datasets, Heart disease, Liver disease, Diabetics, Breast cancer.

1. Introduction

Machine learning algorithms have been successful in many domineering activities around the globe. Machine learning algorithms are commonly associated with mathematics and reasoning that can be easily predicted from a given dataset. By employing machine learning algorithms, complicated diseases can be diagnosed in a very comprehensible manner. There has been a lot of study done on machine learning algorithms in healthcare for disease prediction. Thousands of researchers around the globe have learned a lot about predicting diagnoses by using machine learning algorithms. There are numerous machine learning algorithms that are usually classified according to their learning style (supervised learning, unsupervised learning, or semi-supervised learning) or similarity in form or function. (i.e., classification, regression, decision trees, clustering, deep learning, etc.). Machine learning in medical diagnosis is currently a new trend for big medical data applications. The majority of medical diagnosis techniques are systematised using an intelligent data classification strategy.

Artificial intelligence (AI) based on machine learning (ML) and deep learning (DL) has played critical roles in analysing medical data to aid in illness diagnosis and treatment selection. It is used to identify patterns in clinical data automatically and then reason about the clinical data to predict early risk for patients with conditions such as heart disease (Saleh et al., 2022;

Cardoso et al., 2018) and COVID-19 (Spagnuolo et al., 2020; Alouffi et al., 2021). Authors have recently used ensemble learning to improve the performance of these models in the healthcare area (Melin et al., 2020). Ensemble learning combines the decisions of different base classifiers to improve the end decision using a variety of techniques such as voting or averaging (Sagi and Rokach, 2018). There are three types of ensemble algorithms: boosting (Freund and Schapire, 1996), stacking (Rajagopal et al., 2020), and bagging (Bühlmann, 2012). Because it is based on a meta learner, which learns from data how to weight the base classifiers and combine them in the best way to optimise the performance of the resulting model, stacking ensembles is considered the best method for creating ensemble models. Ensemble stacking uses a meta-learner to combine the decisions of a collection of heterogeneous base models (Rajagopal et al., 2020).

Machine learning technology is one of the most exciting areas of AI, and many companies are attempting to leverage it for their purposes. ML is becoming increasingly popular. It uses algorithms to facilitate data driven learning and can be used in scenarios ranging from business to healthcare. Healthcare is constantly changing due to the constant development of new technology and ideas. ML could assist medical professionals in some of these new scenarios.

In this study, we proposed an optimized ensemble stacking model that merged the k-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT) and Artificial Neural Network (ANN) classification algorithms to enhance the performance of the various disease prediction. The proposed model has been tested using four different clinical datasets like Breast Cancer, Heart Disease, Diabetes and Liver disorder. Further, comparative study is made between individual algorithms and proposed hybrid algorithm to prove the improvement in prediction accuracy on clinical datasets. The proposed algorithm shows enhanced performance compared to the individual classifiers and assist the physician in diagnosis. To proceed and implement the proposed hybrid algorithm, related work on medical datasets is discussed in section 2. The methodology used for building proposed hybrid machine learning algorithm using python Programming is discussed in Section 3. Section 4 discusses comparative study and the Section 5, includes conclusion and recommendation for future work.

2. Related Work

In this section related work done on medical datasets using hybrid machine learning algorithms is discussed. The researchers strived to improve the accuracy by using different combinations of machine learning algorithms to build the hybrid algorithm. The performance of individual classifiers can be enhanced by using the hybridization method (Dahiya, 2015). It is an important and latest area of research as compared to individual learning approaches (Malhotra, 2003). Hybrid and ensemble methods in machine learning have attracted a great attention of the scientific community over the last years (Lughofer, 2013). Both ensemble models and hybrid methods make use of the information fusion concept but in slightly different way. In case of ensemble classifiers, multiple but homogeneous, weak models are combined (Kajdanowicz and Kazienko, 2010), typically at the level of their individual output, using various merging methods, which can be grouped into fixed (e.g., majority voting), and trained combiners (e.g., decision templates) (Kuncheva, 2007). Hybrid methods, in turn, combine completely different, heterogeneous machine learning approaches (Castillo et al., 2007).

In literature, there are different ways of classifying the training/test instances into one of the predefined categories, like (1) Individual models, (2) Hybrid models and (3) Ensemble based models. Individual approach involves using a single statistical or machine learning technique for classification. The hybrid and ensemble models are efficient and robust because they combine the complementary features of more than one learning technique and overcome the weakness of individual techniques. The hybrid models can be stand alone, transformational, tightly coupled or fully coupled (Bahrammirzaee, 2011). As per (Tsai and Chen, 2010) hybrid models are of 4 types: Classification combined with Classification, Classification combined with Clustering, Clustering combined with Clustering and Clustering combined with Classification. Ensemble learning uses various base classifiers combined using a particular strategy of combination such as bagging, boosting, voting, etc. Abhishek and Kumar in (Col and Lal, 2017) developed a hybrid classifier algorithm by merging Decision Tree and Naïve Bayes algorithms which will classify the Fitness data set. The classification accuracy of the Hybrid Classifier has enhanced by 15.79 % and 3.6 % as compared to Decision Tree and Naïve Bayes classifier. In (Yu et al., 2014) authors designed a general hybrid adaptive ensemble learning framework (HAEL), and apply it to address the limitations of random subspace-based classifier ensemble approaches (RSCE).

The experiments on the real-world datasets from the KEEL dataset repository for the classification task and the cancer gene expression profiles showed that: 1) HAEL works well on both the real-world KEEL datasets and the cancer gene expression profiles and 2) it outperforms most of the state-of-the-art classifier ensemble approaches on 28 out of 36 KEEL datasets and 6 out of 6 cancer datasets. Artificial neural network has the highest performance when compared with Decision tree algorithm. In addition, they found that the large datasets can easily be trained and tested in using these algorithms to predict the diseases that are expected according to the datasets. In (Patil et al., 2010), a new system was proposed for breast cancer classification. The new system uses a hybrid of K-means and Support Vector Machine (SVM). The proposed algorithm was compared with different classifier algorithms. The experimental results showed the effectiveness of the proposed algorithm and how it can obtain better results. In (Güzel and Engineering, 2013) researcher proposed using k Nearest neighbor algorithm (kNN) and Naïve Bayes with imputation techniques which was used instead of removing the values that are missing from the Mammographic Mass data. The system was evaluated using different performance criteria such as accuracy, sensitivity, and specificity and ROC analysis.

Ramana et al. (2011) develop a classification model to predict liver disease diagnosis using five popular classification algorithms and evaluate the performance of each model in terms of accuracy, precision, sensitivity and specificity. The study showed that the performances of all classifiers are better in one dataset (AP Liver dataset) as opposed to the other (BUPA Liver dataset) due to highly significant attributes such as total count of bilirubin, direct bilirubin and indirect bilirubin in the AP dataset A hybrid algorithm was presented (Michelakos et al., 2010) to combine the cAnt-Miner2 and the mRMR feature selection algorithms. The proposed algorithm was experimentally compared to cAnt Miner2, using some public medical data sets to demonstrate its functioning. The experiments were very promising and the proposed approach is better in terms of accuracy, simplicity and computational cost than the original cAnt-Miner2 algorithm. Another study in (Shouman et al., 2012) demonstrated that the effectiveness of an unsupervised learning technique which is k-means clustering in improving supervised learning technique which is naïve bayes. The results showed that integrating k-means clustering with naïve bayes with different initial centroid selection could enhance the naïve bayes accuracy in diagnosing heart disease patients. In (Abed et al., 2016) authors hybrid the genetic algorithm and the knearest neighbor algorithm in order to design efficient classifier model for breast cancer classification and they achieved high classification performance.

3. Dataset

We used the Breast Cancer, Diabetes and Heart Disease dataset collected from UCI repository. Breast Cancer dataset includes 699 observations, 9 independent features and one dependent variable as the class label for predicting Breast Cancer. The diabetes dataset includes 769 observations, 8 independent features and one dependent variable as the class label for predicting diabetes. The Heart Disease dataset includes 1026 observations, 13 independent features and one dependent variable as the class label for predicting Heart Disease. The Liver Disease dataset collected from Kaggle dataset which contains 584 observations, 10 independent features and one dependent variable as the class label for predicting Liver Disease.

4. RESEARCH METHOD

This section contains the methodology used for proposing the new hybrid algorithm.

Step-1: Data Collection: The source of data for proposed algorithm is University of California (UCI) repository for machine learning and the Kaggle dataset

Step-2: Developing Proposed Hybrid Machine Learning Algorithm

The major activities undertaken in this phase are:

Identifying the requirements of hybrid algorithm to be developed.

Proposing the methodology used for developing the algorithm and implementing using Python programming language.

Evaluating the results obtained by hybrid algorithm.

Proposed Hybrid Machine Learning approach

```
1. Input dataset D
2. Pre-process the dataset
     replace missing values
3. Apply KNN, SVM, DT and ANN algorithm on the pre-processed dataset
     Classification Accuracy of these classifiers are observed individually.
4. Combine these classifiers for enhancing the classification accuracy and the results are
   observed for different clinical dataset.
5. Output: Hybrid algorithm
6. Evaluating the results obtained by hybrid algorithm.
```

The above proposed hybrid methodology was implemented using python programming language on different clinical dataset. The advantage of the proposed algorithm over the existing algorithms is that our algorithm will give more accuracy to predict the result of test dataset.

5. Results

This study set out to enhance the prediction accuracy of hybrid machine learning algorithm. Hence, this section shows the analysis results and discussion about individual and proposed hybrid machine learning classifiers prediction accuracy on selected medical dataset. Here we have selected four supervised machine learning algorithms, they are KNN, SVM, DT and ANN algorithms. By combining these selected algorithms, the proposed hybrid algorithm was produced. Whose methodology of implementation is already discussed in section 4. The following results shown in the table and graphs are the outcomes of our proposed hybrid algorithm and it gives better prediction performance on selected dataset.

The following table Table1, shows the comparison of prediction accuracy of KNN, SVM, DT and ANN algorithms and the proposed Hybrid algorithm on few clinical datasets. It is observed that the proposed Hybrid approach outperforms the individual classifiers in all the datasets used for this research.

Table 1: Comparison of Prediction accuracy on clinical dataset

Dataset Algorithms	Breast cancer (%)	Diabetics (%)	Heart disease (%)	Liver disorder (%)
KNN	98.64	75.52	74.319	72.09
SVM	81.63	67.71	98.44	62.79
DT	97.28	76.56	90.27	68.60
ANN	97.96	67.71	87.159	72.09
Hybrid	**100.00**	**78.65**	**100**	**73.26**

A. Experiment on breast cancer dataset

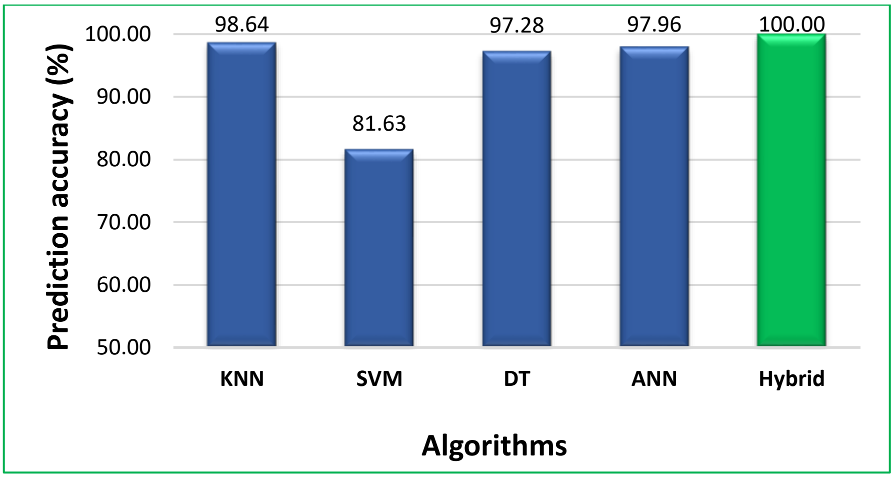

Fig. 1: Prediction accuracy on breast cancer dataset

The above graph in Fig1, shows the prediction accuracy of KNN, SVM, DT and ANN algorithms and the proposed hybrid algorithm on breast-cancer dataset using python programming language. It is observed that the prediction accuracy (100%) of the proposed hybrid algorithm is better than the individual classifiers.

B. Experiment on diabetes dataset

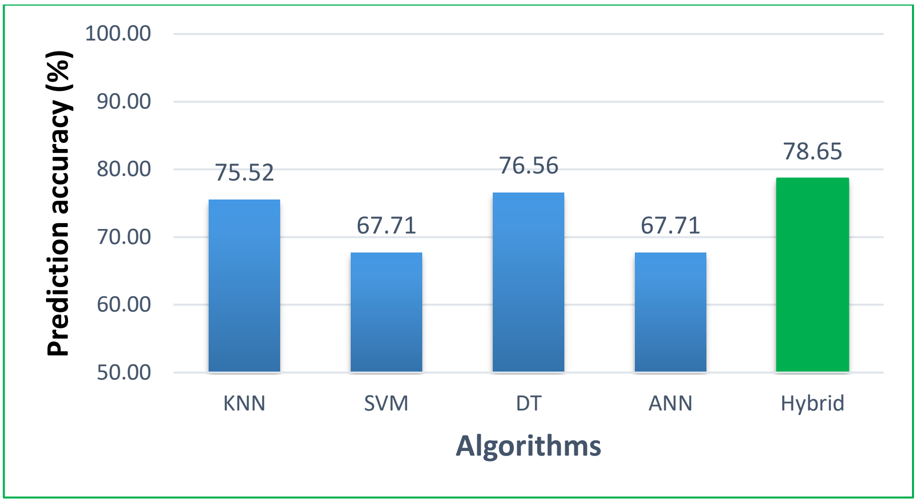

Fig. 2: Prediction accuracy on diabetes dataset

The above graph in Fig 2, shows the prediction accuracy of individual classifiers and the hybrid classifier on diabetes dataset using python ensemble technique. In this case, it can be observed the improvement in prediction accuracy (78.65%) when proposed hybrid algorithm is used.

C. Experiment on Heart Disease dataset

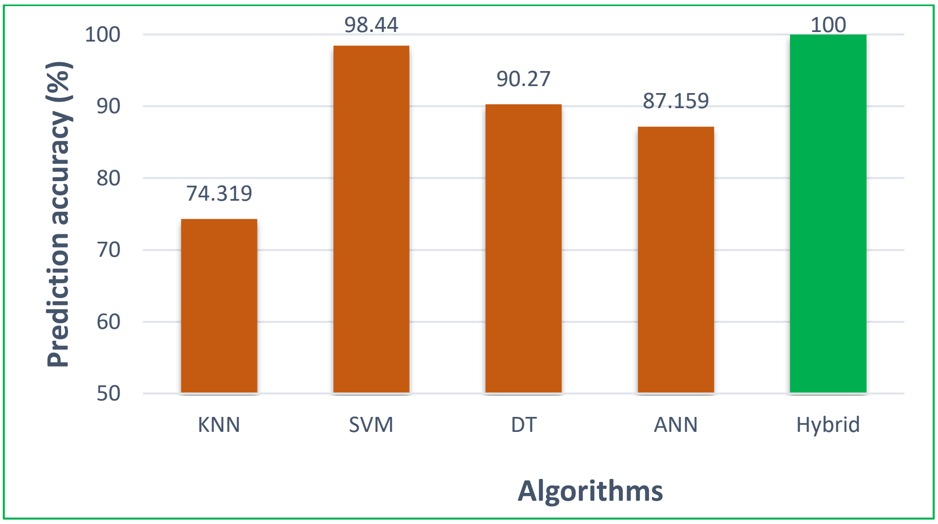

Fig. 3: Prediction accuracy on Heart Disease dataset

The above graph in Fig 3, shows the prediction accuracy of selected individual algorithms and proposed hybrid algorithm on Heart Disease dataset using python programming language. The proposed hybrid algorithm has better prediction accuracy (100%) than the individual algorithms.

D. Experiment on Liver disorder dataset

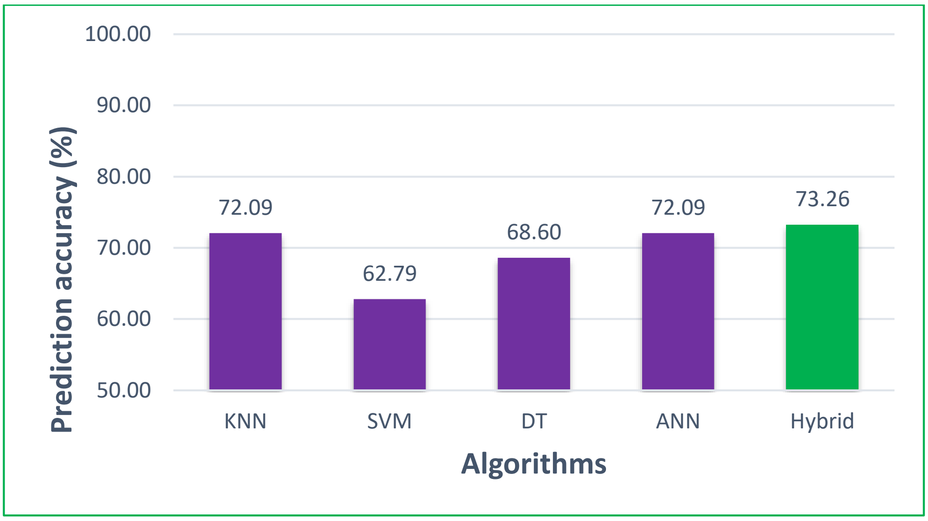

Fig. 4: Prediction accuracy on Liver disorder dataset

The above graph in Fig 4, shows that the performance of selected individual and proposed hybrid classifier using Liver disorder dataset to predict accuracy. It can be observed that prediction accuracy (73.26) outperforms over individual classifiers.

6. CONCLUSION

Machine learning systems make medical professionals faster and smarter in their diagnosis. As a result, it reduces uncertainty in their decisions, thereby reducing costs, risks and saving valuable time. In this study, the proposed ensemble and hybrid algorithm demonstrate that hybrid machine learning techniques perform better than the individual algorithms on selected clinical datasets. The proposed hybrid algorithm composed of KNN, SVM, DT and ANN algorithms. To implement the proposed hybrid algorithm python machine learning tools were used. Selected individual algorithms separately and proposed hybrid algorithm is applied on different clinical datasets. Cross validation test option is used to get better prediction accuracy. The proposed hybrid algorithm outperforms over the selected individual algorithms. The result shows that the hybrid machine-learning algorithm is the key to improve the prediction accuracy of individual machine learning algorithms. In this study, some medical datasets are used to check prediction accuracy of algorithms, so it is better to use various category datasets that have different size for further work. In addition, the hybrid deep learning algorithms can be developed and implemented for improving the classification accuracy further.

References

[1] Abed B. M et al., 2016. "A hybrid classification algorithm approach for breast cancer diagnosis", 2016 IEEE Ind. Electron. Appl. Conf., pp. 269–274.

[2] Alouffi, B.; Alharbi, A.; Sahal, R.; Saleh, H. 2021. An Optimized Hybrid Deep Learning Model to Detect COVID-19 Misleading Information. Comput. Intell. Neurosci. 2021, 9615034.

[3] Bahrammirzaee, A. R. G. A. 2011. "Hybrid Credit ranking intelligent system using expert system and artificial neural networks", Appl. Intell., vol. 34, pp. 28–46.

[4] Bühlmann, P. 2012. Bagging, boosting and ensemble methods. In Handbook of Computational Statistics; Springer: Berlin/Heidelberg, Germany, pp. 985–1022.

[5] Cardoso, M.R.; Santos, J.C.; Ribeiro, M.L.; Talarico, M.C.R.; Viana, L.R.; Derchain, S.F.M. 2018. A metabolomic approach to predict breast cancer behavior and chemotherapy response. Int. J. Mol. Sci. 19, 617.

[6] Castillo, W., O. Melin, P. Pedrycz, 2007. "Hybrid Intelligent Systems: Analysis and Design (Studies in Fuzziness and Soft Computing)," Springer, pp. 55–64, 2007.

[7] Col, L. and A. Lal, 2017. "Hybrid Classifier for Increasing Accuracy of Fitness Data Set," pp. 1246–1249,

[8] Dahiya, S. 2015. "Credit Modelling using Hybrid Machine Learning Technique," pp. 103–106.

[9] Freund, Y.; Schapire, R.E. 1996. Experiments with a new boosting algorithm. ICML Citeseer, 6, 148–156.

[10] Güzel, C and F. Engineering, 2013. "Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation," AWER Procedia Inf. Technol. Comput. Sci., vol. 4, pp. 401–407.

[11] Kajdanowicz, K., T., Kazienko, P., 2010. "Boosting algorithm with sequence-loss cost function for structured prediction," Springer, pp. 573–580.

[12] Kuncheva, L. 2007. "Combining pattern classifiers: Methods and algorithms", 2004.

[13] Lughofer, E. 2013. "Hybrid and Ensemble Methods in Machine Learning J. UCS Special Issue," vol. 19, no. 4, pp. 457–461.

[14] Malhotra, D. K. M. R. 2003. "Evaluating consumer loans using neural networks," Omega –The Int. J. Manag. Sci., pp. 83–96.

[15] Melin, P.; Monica, J.C.; Sanchez, D.; Castillo, O. 2020. Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: The case of Mexico. Healthcare, 8, 181.

[16] Michelakos, I., E. Papageorgiou, and M. Vasilakopoulos, 2010. "A hybrid classification algorithm evaluated on medical data," Proc. Work. Enabling Technol. Infrastruct. Collab. Enterp. WETICE, pp. 98–103.

[17] Patil, B. M., R. C. Joshi, and D. Toshniwal, 2010. "Hybrid prediction model for Type 2 diabetic patients," Expert Syst. Appl., vol. 37, no. 12, pp. 8102–8108.

[18] Rajagopal, S.; Kundapur, P.P.; Hareesha, K.S. 2020. A stacking ensemble for network intrusion detection using heterogeneous datasets. Secur. Commun. Netw. 2020, 4586875.

[19] Ramana, N. B. V. B. V. 2011. "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," Int. J. Database Manag. Syst., vol. 3, no. 2, pp. 101–114, 2011.

[20] Sagi, O.; Rokach, L. 2018. Ensemble learning: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 8, e1249.

[21] Saleh, H.; Alyami, H.; Alosaimi, W. 2022. Predicting Breast Cancer Based on Optimized Deep Learning Approach. Comput. Intell. Neurosci. 2022, 1820777.

[22] Shouman, M., A. Defence, F. Academy, A. Defence, F. Academy, and R. Stocker, 2012. "Integrating naive bayes and k-means clustering with different initial centroid selection methods in the diagnosis," no. August 2014, p. 125–137.

[23] Spagnuolo, G.; De Vito, D.; Rengo, S.; Tatullo, M. 2020. COVID-19 outbreak: An overview on dentistry. Int. J. Environ. Res. Public Health, 17, 2094.

[24] Tsai, C.F., M.L. Chen, 2010. "Credit Rating by Hybrid Machine Learning Techniques", Appl. Soft Comput., vol. 10, pp. 374–380.

[25] Weissler, E.H.; Naumann, T.; Andersson, T.; Ranganath, R.; Elemento, O.; Luo, Y.; Freitag, D.F.; Benoit, J.; Hughes, M.C.; Khan, F.; et al. 2021. The role of machine learning in clinical research: Transforming the future of evidence generation. Trials, 22, 1–15.

[26] Yu, Z., S. Member, L. Li, J. Liu, and G. Han, 2014. "Hybrid Adaptive Classifier Ensemble", IEEE Trans. Cybern., pp. 1–14.