REFERENCES

Abdar, M., Acharya, U. R., Sarrafzadegan, N., & Makarenkov, V. (2019). NE-nu-SVC: A new nested ensemble clinical decision support system for effective diagnosis of coronary artery disease. *IEEE Access*, *7*, 167605–167620.

Abdellatif, A., Abdellatef, H., Kanesan, J., Chow, C. O., Chuah, J. H., & Gheni, H. M. (2022). An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods. *IEEE Access*, *10*, 79974–79985.

Abirami, R. N., & Raj, P. D. (2020). Cardiac arrhythmia detection using ensemble of machine learning algorithms. In *Soft Computing for Problem Solving* (Vol. 1057, pp. 475–487). Springer.

Alam, M. Z., Rahman, M. S., & Rahman, S. M. (2019). A random forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, *15*, 100180.

Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. (2017). Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm. *Computers in Biology and Medicine*, *141*, 19–26.

Alsahaf, A., Petkov, N., Shenoy, V., & Azzopardi, G. (2022). A framework for feature selection through boosting. *Expert Systems with Applications*, *187*, 1–11.

Al-Tashi, Q., Rais, H., & Jadid, S. (2019). Feature selection method based on grey wolf optimization for coronary artery disease classification. In *Recent trends in data science and soft computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology* (pp. 257–266). Springer International Publishing.

An, N., Ding, H., Yang, J., Au, R., & Ang, T. F. (2020). Deep ensemble learning for Alzheimer's disease classification. *Journal of Biomedical Informatics*, *105*, 103411.

Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, *2*(10), 5370–5376.

Archika, Jain, Somwanshi, D., Singh, R., & Singh, G. (2022). Hybrid layered classification model for heart disease diagnosis. *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, 643–647.

Babič, F., Olejár, J., Vantová, Z., & Paralič, J. (2017). Predictive and descriptive analysis for heart disease diagnosis. In *2017 federated conference on computer science and information systems (FEDCSIS)* (pp. 155-163). IEEE.

Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: A survey. *Journal of Electrical Systems and Information Technology*, *10*, 40.

Badriyah, T., Sakinah, N., Syarif, I., & Syarif, D. R. (2020). Machine learning algorithm for stroke disease classification. In *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)* (pp. 1–5). IEEE.

Bai, Y., Li, Y., Shen, Y., Yang, M., Zhang, W., & Cui, B. (2022). AutoDC: An automatic machine learning framework for disease classification. *Bioinformatics*, *38*(13), 3415–3421.

Bösner, S., Becker, A., Hani, M. A., Keller, H., Sönnichsen, A. C., Haasenritter, J., & Donner-Banzhoff, N. (2010). Accuracy of symptoms and signs for coronary heart disease assessed in primary care. *British Journal of General Practice*, *60*(575), e246–e257.

Cavalcanti, G. D., Oliveira, L. S., Moura, T. J., & Carvalho, G. V. (2016). Combining diversity measures for ensemble pruning. *Pattern Recognition Letters*, *74*, 38–45.

Cenitta, D., Arjunan, R. V., & Prema, K. V. (2022). Ischemic heart disease prediction using optimized squirrel search feature selection algorithm. *IEEE Access*, *10*, 122995–123006.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Cios, K. J., Swiniarski, R. W., Pedrycz, W., & Kurgan, L. A. (2007). The knowledge discovery process. In *Data Mining* (pp. 9–24). Springer.

Crockett, D., & Eliason, B. (2016). *What is data mining in healthcare?* Insights: Health Catalyst.

Cruz, R. M., Zakane, H. H., Sabourin, R., & Cavalcanti, G. D. (2017). Dynamic ensemble selection vs k-nn: Why and when dynamic selection obtains higher classification performance? In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1–6). IEEE.

Cunningham, P., & Carney, J. (2000). Diversity versus quality in classification ensembles based on feature selection. In *Machine learning: ECML 2000: 11th European Conference on Machine Learning* (pp. 109–116). Springer.

Dahal, K., & Gautam, Y. (2020). Argumentative comparative analysis of machine learning on coronary artery disease. *Open Journal of Statistics, 10*, 694–705.

De Jesus Silva, L. F., Cortes, O. A. C., & Diniz, J. O. B. (2023). A novel ensemble CNN model for COVID-19 classification in computerized tomography scans. *Results in Control and Optimization, 11*, 100215.

DeHan, C. (2018). *BoostARoota*. GitHub. https://github.com/chasedehan/BoostARoota

Dhar, J. (2021). Multistage ensemble learning model with weighted voting and genetic algorithm optimization strategy for detecting chronic obstructive pulmonary disease. *IEEE Access, 9*, 48640–48657.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*, 139–141.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000* (pp. 1–15). Springer.

Dinh, A., Miertschin, S., & Young, A. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, *19*, 211.

Divya, T., & Agarwal, S. (2014). Feature selection based least square twin support vector machine for diagnosis of heart disease. *International Journal of Bio-Science and Bio-Technology*, *6*(2), 69–82.

Divya, R., & Shantha Selva Kumari, R. (2021). Alzheimer's disease neuroimaging initiative: Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification. *Neural Computing and Applications*, *33*(14), 8435–8444.

Dolatabadi, A. D., Khadem, S. E. Z., & Asl, B. M. (2017). Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. *Computer Methods and Programs in Biomedicine*, *138*, 117–126.

Doppala, B. P., Bhattacharyya, D., Chakkravarthy, M., & Baik, N. (2022). A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques. *Journal of Healthcare Engineering*, *2022*, 1–13.

Dong, Y., Zhou, S., Xing, L., Chen, Y., Ren, Z., Dong, Y., & Zhang, X. (2022). Deep learning methods may not outperform other machine learning methods on analyzing genomic studies. *Frontiers in Genetics*, *13*, 992070.

Duda, R. O., & Hart, P. E. (2006). *Pattern classification*. John Wiley & Sons.

Durairaj, M., & Ranjani, V. (2013). Data mining applications in healthcare sector: A study. *International Journal of Scientific and and Technology Research*, *2*(10), 29–35.

Essa, E., & Xie, X. (2021). An ensemble of deep learning-based multi-model for ECG heartbeats arrhythmia classification. *IEEE Access*, *9*, 103452–103464.

Essadik, I., Nouri, A., Touahni, R., Bourcier, R., & Autrusseau, F. (2022). Automatic classification of the cerebral vascular bifurcations using dimensionality reduction and machine learning. *Neuroscience Informatics*, *2*(4), 100108.

Feng, P., Ma, J., Sun, C., Xu, X., & Ma, Y. (2018). A novel dynamic Android malware detection system with ensemble learning. *IEEE Access*, *6*, 30996–31011.

Figueira, A. (2016). Predicting grades by principal component analysis: A data mining approach to learning analytics. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)* (pp. 465–467). IEEE.

Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. *IEEE Access*, *7*, 144777–144789.

Gadekallu, T. R., & Khare, N. (2017). Cuckoo search optimized reduction and fuzzy logic classifier for heart disease and diabetes prediction. *International Journal of Fuzzy System Applications (IJFSA)*, *6*(2), 25–42.

Gamal, A., Elattar, M., & Selim, S. (2022). Automatic early diagnosis of Alzheimer's disease using 3D deep ensemble approach. *IEEE Access*, *10*, 115974–115987.

Garate-Escamila, A. K., El Hassani, A. H., & Andres, E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, *19*, 100330.

García-Ordás, M. T., Bayón-Gutiérrez, M., Benavides, C., Aveleira-Mata, J., & Benítez-Andrades, J. A. (2023). Heart disease risk prediction using deep learning techniques with feature augmentation. *Multimedia Tools and Applications*, *82*, 1–15.

Giacinto, G., & Roli, F. (2000). A theoretical framework for dynamic classifier selection. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000* (Vol. 2, pp. 8–11). IEEE.

Giacinto, G., & Roli, F. (2001). Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, *19*(9-10), 699–707.

Gnanambal, S., Thangaraj, M., Meenatchi, V. T., & Gayathri, V. (2018). Classification algorithms with attribute selection: An evaluation study using WEKA. *International Journal of Advanced Networking and Applications*, *9*(6), 3640–3644.

Gokulnath, C. B., & Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Computing*, *22*, 14777–14787.

Gopal, V. N., Al-Turjman, F., Kumar, R., Anand, L., & Rajesh, M. (2021). Feature selection and classification in breast cancer prediction using IoT and machine learning. *Measurement*, *178*, 109442.

Gunduz, H. (2021). An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification. *Biomedical Signal Processing and Control*, *66*, 102452.

Gupta, R., Kumar, H. S., Arora, & Raman, B. (2019). MIFH: A machine intelligence framework for heart disease diagnosis. *IEEE Access*, *8*, 14659–14674.

Hassannataj Joloudari, J., Azizi, F., Nematollahi, M. A., Alizadehsani, R., Hassannatajjeloudari, E., Nodehi, I., & Mosavi, A. (2022). GSVMA: A genetic support vector machine ANOVA method for CAD diagnosis. *Frontiers in Cardiovascular Medicine*, *8*, 760178.

Hossain, M. M., Swarna, R. A., Mostafiz, R., Shaha, P., Pinky, L. Y., Rahman, M. M., & Iqbal, M. S. (2022). Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease. *Machine Learning with Applications, 9*, 100330.

Huang, C., Huang, X., Fang, Y., Xu, J., Qu, Y., Zhai, P., & Li, J. (2020). Sample imbalance disease classification model based on association rule feature selection. *Pattern Recognition Letters*, *133*, 280–286.

Ilyas, Q. M., & Ahmad, M. (2021). An enhanced ensemble diagnosis of cervical cancer: A pursuit of machine intelligence towards sustainable health. *IEEE Access*, *9*, 12374–12388.

Iyer, T. J., Kishan, B., & Nersisson, R. (2021). Prediction and classification of cardiac arrhythmia using a machine learning approach. In V. L. N. Komanapalli, N. Sivakumaran, & S. Hampannavar (Eds.), *Advances in automation, signal processing, instrumentation, and control. i-CASIC 2020. Lecture Notes in Electrical Engineering* (Vol. 700). Springer.

Jain, S., Raghuvanshi, R., & Ilyas, M. (2017). A survey paper on overview of basic data mining tasks.

Jayaraman, V., & Sultana, H. P. (2019). Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification. *Journal of Ambient Intelligence and Humanized Computing*, *10*(1), 1–10.

Jiang-hong, M. A. (2002). Data mining and knowledge discovery in database. *Chinese Journal of Engineering Mathematics*, *19*(1), 1–13.

Kabir, M., Shahjahan, M., Murase, K., & Barbosa, H. J. C. (2013). Ant colony optimization toward feature selection. In *Ant colony optimization—Techniques and applications* (pp. 3–44). InTech.

Kaur, T., Gandhi, T. K., & Panigrahi, B. K. (2021). Automated diagnosis of COVID-19 using deep features and parameter free BAT optimization. *IEEE Journal of Translational Engineering in Health and Medicine*, *9*, 1–9.

Kavitha, R., & Kannan, E. (2016). An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. In *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)* (pp. 1–5). IEEE.

Keles, M. K., & Kilic, U. (2022). Classification of brain volumetric data to determine Alzheimer's disease using artificial bee colony algorithm as feature selector. *IEEE Access*, *10*, 82989–83001.

Khade, S., Gite, S., Thepade, S. D., Pradhan, B., & Alamri, A. (2021). Detection of iris presentation attacks using hybridization of discrete cosine transform and haar transform with machine learning classifiers and ensembles. *IEEE Access*, *9*, 169231–169249.

Khare, A., Bhandari, S., Singh, S., & Arora, A. (2012). ECG arrhythmia classification using Spearman rank correlation and support vector machine. In K. Deep, A. Nagar, M. Pant, & J. Bansal (Eds.), *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011), Advances in Intelligent and Soft Computing* (Vol. 131). Springer.

Khemphila, A., & Boonjing, V. (2011). Heart disease classification using neural network and feature selection. *21st International Conference on Systems Engineering*, 406–409.

Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems*, *12*(1), 242–252.

Kiliç, Ü., & Keleş, M. K. (2018). Feature selection with artificial bee colony algorithm on Z-Alizadeh Sani dataset. *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 1–3.

Kilicarslan, S., Celik, M., & Sahin, Ş. (2021). Hybrid models based on genetic algorithm and deep learning algorithms for nutritional anemia disease classification. *Biomedical Signal Processing and Control*, *63*, 102231.

Kim, J., Seo, K., & Chung, K. (1997). A systematic approach to classifier selection on combining multiple classifiers for handwritten digit recognition. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition* (Vol. 2, pp. 459–462). IEEE.

Kim, S. B., & Rattakorn, P. (2011). Unsupervised feature selection using weighted principal components. *Expert Systems with Applications*, *38*(5), 5704–5710.

Ko, A. H., Sabourin, R., de Souza Britto Jr, A., & Oliveira, L. (2007). Pairwise fusion matrix for combining classifiers. *Pattern Recognition*, *40*(8), 2198–2210.

Kolukisa, B., Hacilar, H., Goy, G., Kus, M., Bakir-Gungor, B., Aral, A., & Gungor, V. C. (2018). Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease. In *Proceedings of 2018 IEEE International Conference on Big Data (Big Data)* (pp. 2232–2238).

Kotseva, K., De Bacquer, D., Jennings, C., Gyberg, V., De Backer, G., Rydén, L., Amouyel, P., Bruthans, J., Cifkova, R., Deckers, J. W., & De Sutter, J. (2017). Time trends in lifestyle, risk factor control, and use of evidence-based medications in patients with coronary heart disease in Europe: Results from 3 EUROASPIRE Surveys, 1999–2013. *Global Heart*, *12*(4), 315–322.

Kuncheva, L. I., & Whitaker, C. J. (2001). Ten measures of diversity in classifier ensembles: Limits for two classifiers. In *A DERA/IEE Workshop on Intelligent Sensor Processing* (Ref. No. 2001/050).

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, *51*(2), 181.

Lakshmi, B. N., & Raghunandhan, G. H. (2011). A conceptual overview of data mining. In *Innovations in Emerging Technology (NCOIET), 2011 National Conference on* (pp. 27–32). IEEE.

LaRosa, J. C. (2001). Prevention and treatment of coronary heart disease: Who benefits? *Circulation*, *104*(14), 1688–1692.

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

Lee, S. (2015). Feature selection based on the center of gravity of BSWFMs using NEWFM. *Engineering Applications of Artificial Intelligence*, *45*, 482–487.

Lee, Y. W., Choi, J. W., & Shin, E. H. (2021). Machine learning model for diagnostic method prediction in parasitic disease using clinical information. *Expert Systems with Applications*, *185*, 115658.

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access*, *8*, 107562–107582.

Li, N., Yu, Y., & Zhou, Z. H. (2012). Diversity regularized ensemble pruning. In *ECML/PKDD* (Vol. 1, pp. 330–345).

Liu, R., & Yuan, B. (2001). Multiple classifiers combination by clustering and selection. *Information Fusion*, *2*(3), 163–168.

Maghawry, E., Gharib, T. F., Ismail, R., & Zaki, M. J. (2021). An efficient heartbeats classifier based on optimizing convolutional neural network model. *IEEE Access*, *9*, 153266–153275.

Maldonado, S., Weber, R., & Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, *181*(1), 115–128.

Mandal, M., Singh, P. K., Ijaz, M. F., Shafi, J., & Sarkar, R. (2021). A tri-stage wrapper-filter feature selection framework for disease classification. *Sensors*, *21*(16), 5571.

Manonmani, M., & Balakrishnan, S. (2020). Review of optimization-based feature selection algorithms on healthcare dataset. In *Emerging Research in Data Engineering Systems and Computer Communications* (pp. 239–245). Springer.

Mitra, M., & Samanta, R. K. (2013). Cardiac arrhythmia classification using neural networks with selected features. *Procedia Technology*, *10*, 76–84.

Ottom, M.A., & Alshorman, W. (2019). Heart diseases prediction using accumulated rank features selection technique. *Journal of Engineering and Applied Sciences*, *14*, 2249–2257.

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

Mustaqeem, S., Anwar, S. M., Majid, M., & Khan, A. R. (2017). Wrapper method for feature selection to classify cardiac arrhythmia. In *Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3656–3659).

Muthukaruppan, S., & Er, M. J. (2012). A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. *Expert Systems with Applications*, *39*(14), 11657–11665.

Naganjaneyulu, S., & Rao, B. S. (2018). A novel feature selection based classification algorithm for real-time medical disease prediction. In *2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing* (pp. 275–282). IEEE.

Nagarajan, S. M., Muthukumaran, V., Murugesan, R., Joseph, R. B., Meram, M., & Prathik, A. (2022). Innovative feature selection and classification model for heart disease prediction. *Journal of Reliable Intelligent Environments*, *8*(4), 333–343.

Nainwal, A., Kumar, Y., & Jha, B. (2022). Arrhythmia classification based on improved monarch butterfly optimization algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 5100–5109.

Nassif, A. B., Mahdi, O., Nasir, Q., Talib, M. A., & Azzeh, M. (2018). Machine learning classifications of coronary artery disease. In *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1–6).

Nawaz, M. S., Shoaib, B., & Ashraf, M. A. (2021). Intelligent cardiovascular disease prediction empowered with gradient descent optimization. *Heliyon*, *7*(5), e06948.

Oliveira, L. S., Morita, M., & Sabourin, R. (2006). Feature selection for ensembles applied to handwriting recognition. *International Journal of Document Analysis and Recognition (IJDAR)*, *8*(4), 262–279.

Padhy, N., Mishra, D., & Panigrahi, R. (2012). *The survey of data mining applications and feature scope*. arXiv preprint arXiv:1211.5723.

Pandey, S. K., Sodum, V. R., Janghel, R. R., & Raj, A. (2020). ECG arrhythmia detection with machine learning algorithms. In K. Raju, R. Senkerik, S. Lanka, & V. Rajagopal (Eds.), *Data engineering and communication technology. Advances in intelligent systems and computing* (Vol. 1079). Springer.

Panthong, R., & Srivihok, A. (2015). Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. *Procedia Computer Science*, *72*, 162–169.

Parvin, H., MirnabiBaboli, M., & Alinejad-Rokny, H. (2015). Proposing a classifier ensemble framework based on classifier selection and decision tree. *Engineering Applications of Artificial Intelligence*, *37*, 34–42.

Pérez-Gállego, P., Castano, A., Quevedo, J. R., & del Coz, J. J. (2019). Dynamic ensemble selection for quantification tasks. *Information Fusion*, *45*, 1–15.

Prabhakar, S. K., Rajaguru, H., & Lee, S. W. (2020). A framework for schizophrenia EEG signal classification with nature inspired optimization algorithms. *IEEE Access*, *8*, 39875–39897.

Prince, J., Andreotti, F., & De Vos, M. (2018). Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data. *IEEE Transactions on Biomedical Engineering*, *66*(5), 1402–1411.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*, 1–11.

Qummar, S., Khan, F. G., Shah, S., Khan, A., Shamshirband, S., Rehman, Z. U., & Jadoon, W. (2019). A deep learning ensemble approach for diabetic retinopathy detection. *IEEE Access*, *7*, 150530–150539.

Rajadevi, R., Devi, E. R., Shanthakumari, R., Latha, R. S., Anitha, N., & Devipriya, R. (2021). Feature selection for predicting heart disease using black hole optimization algorithm and XGBoost classifier. In *2021 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1–7). IEEE.

Ramprakash, P., Sarumathi, R., Mowriya, R., & Nithyavishnupriya, S. (2020). Heart disease prediction using deep neural network. In *2020 International Conference on Inventive Computation Technologies (ICICT)* (pp. 666–670). IEEE.

Rani, U. D. (2017). A survey on data mining tools and techniques in medical field. *International Journal of Advanced Networking & Applications*, *8*(5), 51–54.

Reddy, N. S. C., Nee, S. S., Min, L. Z., & Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. *International Journal of Innovative Computing*, *9*(1).

Riajuliislam, M., Rahim, K. Z., & Mahmud, A. (2021). Prediction of thyroid disease (hypothyroid) in early stage using feature selection and classification techniques. In *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)* (pp. 60–64). IEEE.

Rogova, G. (1994). Combining the results of several neural network classifiers. *Neural Networks*, *7*(5), 777–781.

Rohini, M., & Surendran, D. (2019). Classification of neurodegenerative disease stages using ensemble machine learning classifiers. *Procedia Computer Science*, *165*, 66–73.

Roli, F., & Giacinto, G. (2002). Design of multiple classifier systems. In *Hybrid methods in pattern recognition* (pp. 199–226).

Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., & Fuster, V. (2020). Global burden of cardiovascular

diseases and risk factors, 1990–2019. *Journal of the American College of Cardiology*, *76*(25), 2982–3021.

Rückstieß, T., Osendorfer, C., & Van der Smagt, P. (2011). Sequential feature selection for classification. In *Australasian Joint Conference on Artificial Intelligence* (pp. 132–141). Springer.

Ruta, D., & Gabrys, B. (2001). Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems.

Ruta, D., & Gabrys, B. (2005). Classifier selection for majority voting. *Information Fusion*, *6*(1), 63–81.

Sabab, S. A., Munshi, M. A. R., & Pritom, A. I. (2016). Cardiovascular disease prognosis using effective classification and feature selection technique. In *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)* (pp. 1–6). IEEE.

Sawhney, R., Mathur, P., & Shankar, R. (2018). A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis. In *International Conference on Computational Science and Its Applications* (pp. 438–449). Springer.

Seifert, J. W. (2004). *Data mining: An overview*. National security issues, 201–217.

Mirjalili, S., & Lewis, A. (2016). The Whale Optimization Algorithm. *Advances in Engineering Software*, *95*, 51–67.

Shafi, A. S. M., Rahman, M. B., Anwar, T., Halder, R. S., & Kays, H. E. (2021). Classification of brain tumors and auto-immune disease using ensemble learning. *Informatics in Medicine Unlocked*, *24*, 100608.

Hera, S. Y., Amjad, M., & Saba, M. K. (2022). Improving heart disease prediction using multi-tier ensemble model. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *11*(1), 41.

Shahi, M., & Gurm, E. R. K. (2017). Heart disease prediction system using data mining techniques-A review. *Heart Disease*, *3*(4).

Shankar, K., Lakshmanaprabu, S. K., Gupta, D., Maseleno, A., & De Albuquerque, V. H. C. (2020). Optimal feature-based multi-kernel SVM approach for thyroid disease classification. *The Journal of Supercomputing*, *76*, 1128–1143.

Shipp, C. A., & Kuncheva, L. I. (2002). Relationships between combination methods and measures of diversity in combining classifiers. *Information Fusion*, *3*(2), 135–148.

Singh, J., Bagga, S., & Kaur, R. (2020). Software-based prediction of liver disease with feature selection and classification techniques. *Procedia Computer Science*, *167*, 1970–1980.

Singh, V. K., Maurya, N. S., Mani, A., & Yadav, R. S. (2020). Machine learning method using position-specific mutation based classification outperforms one hot coding for disease severity prediction in haemophilia 'A'. *Genomics*, *112*(6), 5122–5128.

Skalak, D. B. (1996). The sources of increased accuracy for two proposed boosting algorithms. In *Proceedings of the American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop* (Vol. 1129, pp. 1129–1133).

Sugianela, Y., & Ahmad, T. (2020). Pearson correlation attribute evaluation-based feature selection for intrusion detection system. In *IEEE International Conference on Smart Technology and Applications* (pp. 1–5). IEEE.

Sun, Z., Wang, C., Zhao, Y., & Yan, C. (2020). Multi-label ECG signal classification based on ensemble classifier. *IEEE Access*, *8*, 117986–117996.

Supriya, M., & Deepa, A. J. (2020). A novel approach for breast cancer prediction using optimized ANN classifier based on big data environment. *Health Care Management Science*, *23*, 414–426.

Tang, E. K., Suganthan, P. N., & Yao, X. (2006). An analysis of diversity measures. *Machine Learning*, *65*, 247–271.

Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. *Health and Technology*, *10*, 1137–1144.

Tsymbal, A., Pechenizkiy, M., & Cunningham, P. (2005). Diversity in search strategies for ensemble feature selection. *Information Fusion*, *6*(1), 83–98.

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 281–316.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1).

Verma, L., Srivastava, S., & Negi, P. C. (2016). A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *Journal of Medical Systems*, *40*(1), 1–7.

Vijayashree, J., & Sultana, H. P. (2018). A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. *Programming and Computer Software*, *44*(6), 388–397.

Wang, J., Liu, C., Li, L., Li, W., Yao, L., Li, H., & Zhang, H. (2020). A stacking-based model for non-invasive detection of coronary heart disease. *IEEE Access*, *8*, 37124–37133.

Wang, M., Jie, B., Bian, W., Ding, X., Zhou, W., Wang, Z., & Liu, M. (2019). Graph-kernel based structured feature selection for brain disease classification using functional connectivity networks. *IEEE Access*, *7*, 35001–35011.

Wang, M., Wei, Z., Jia, M., Chen, L., & Ji, H. (2022). Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records. *BMC Medical Informatics and Decision Making*, *22*(1), 1–13.

Wazery, Y. M., Saber, E., Houssein, E. H., Ali, A. A., & Amer, E. (2021). An efficient slime mould algorithm combined with k-nearest neighbor for medical classification tasks. *IEEE Access*, *9*, 113666–113682.

Wijaya, S. H., Pamungkas, G. T., & Sulthan, M. B. (2018). Improving classifier performance using particle swarm optimization on heart disease detection. In *2018 International Seminar on Application for Technology of Information and Communication* (pp. 603–608). IEEE.

Wong, M. T., He, X., Nguyen, H., & Yeh, W. C. (2012). Particle swarm optimization based feature selection in mammogram mass classification. In *2012 International Conference on Computerized Healthcare (ICCH)* (pp. 152–157). IEEE.

Wu, Z. H., Z. J. X., & Chen, Y. J. S. F. (2001). Genetic algorithm based selective neural network ensemble. In *IJCAI-01: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*.

Xu, S., Zhang, Z., Wang, D., Hu, J., Duan, X., & Zhu, T. (2017). Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework. In *IEEE 2nd International Conference on Big Data Analysis* (pp. 228–232).

Yang, F., Du, J., Lang, W., Lu, L., Liu, C., Jin, C., & Kang, Q. (2020). Missing value estimation methods research for arrhythmia classification using the modified kernel difference-weighted KNN algorithms. *BioMed Research International*, *2020*, 1–9.

Yılmaz, E. (2013). An expert system based on fisher score and LS-SVM for cardiac arrhythmia diagnosis. *Computational and Mathematical Methods in Medicine*, *2013*, 1–6.

Yin, Z., Sulieman, L. M., & Malin, B. A. (2019). A systematic literature review of machine learning in online personal health data. *Journal of the American Medical Informatics Association*, *26*(6), 561–576.

Zhenya, Q., & Zhang, Z. (2021). A hybrid cost-sensitive ensemble for heart disease prediction. *BMC Medical Informatics and Decision Making*, *21*(1), 73.

Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, *137*(1-2), 239–263.

Zunaidi, W. H. A. W., Saedudin, R. R., Shah, Z. A., Kasim, S., Seah, C. S., & Abdurohman, M. (2018). Performances analysis of heart disease dataset using different data mining classifications. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(6), 2677–2682.

ANNEXURE I

Attributes in Z-Alizadeh Sani Heart Dataset

S.No	Features	Values	S.No	Features	Values
1	Age	30-86	29	Nonanginal CP	1,2,3,4
2	Weight	48-120	30	Exertional CP (Exertional Chest Pain)	Yes, no
3	Length	140-188	31	Low Th Ang (Low Threshold Angina)	Yes, no
4	Sex	M,F	32	Q Wave	Yes, no
5	BMI (Body Mass Index Kg/m2)	18-41	33	ST Elevation	Yes, no
6	DM (Diabetes Mellitus)	Yes, no	34	ST Depression	Yes, no
7	HTN (Hyper Tension)	Yes, no	35	T inversion	Yes, no
8	Current Smoker	Yes, no	36	LVH (Left Ventricular Hypertrophy)	Yes, no
9	Ex-Smoker	Yes, no	37	Poor R progression (poor R wave progression)	Yes, no
10	FH (Family History)	Yes, no	38	BBB	Yes, no
11	Obesity (MBI > 25)	Yes, no	39	FBS (Fasting Blood Sugar in mg/dl)	62-400
12	CRF (Chronic Renal Failure)	Yes, no	40	Cr (Creatine in mg/dl)	0.5-2.2
13	CVA (Cerebrovascular Accident)	Yes, no	41	TG (Triglyceride in mg/dl)	37-1050
14	Airway Disease	Yes, no	42	LDL (Low Density Lipoprotein in mg/dl)	18-232
15	Thyroid Disease	Yes, no	43	HDL (High Density Lipoprotein in mg/dl)	15-111
16	CHF (Congestive Heart Failure)	Yes, no	44	BUN (Blood Urea Nitrogen in mg/dl)	6-52

S.No	Features	Values	S.No	Features	Values
17	DLP (Dyslipidemia)	Yes, no	45	ESR (Erythrocyte Sedimentation Rate in mm/h)	1-90
18	BP (Blood Pressure mmHg)	Yes, no	46	HB (Hemoglobin in g/dl)	8.9-17.6
19	PR (Pulse Rate ppm)	90 190	47	K (Potassium in mEq/lit)	3.0-6.6
20	Edema	50-110	48	Na (Sodium in mEq/lit)	128-156
21	Weak Peripheral Pulse	Yes, No	49	WBC (White Blood Cell in cells/ml)	3700-18,000
22	Lung Rales	Yes, No	50	Lymph (Lymphocyte in %)	7-60
23	Systolic Murmur	Yes, No	51	Neut (Neutrophil in %)	32-89
24	Diastolic Murmur	Yes, no	52	PLT (Platelet in 1000/ml)	25-742
25	Typical Chest Pain	Yes, no	53	ET-TTE(Ejection fraction in %)	15-60
26	Dyspnea	Yes, no	54	Region RWMA(Region Wall Motion Abnormality)	0,1,2,3,4
27	Function Class	Yes, no	55	VHD (Valvular Heart Disease)	1-4
28	Atypical	Yes, no	56	Cath	0-1

ANNEXURE II

Institution Human Ethics Certificate

INSTITUTIONAL HUMAN ETHICS COMMITTEE

Avinashilingam

Institute for Home Science and Higher Education for Women
(Deemed to be university under Category 'A' by MHRD, Estd. u/s 3 of UGC Act 1956) Re-accredited with 'A^{++}' Grade by NAAC.
Recognised by UGC Under Section 12 B
Coimbatore- 641043, Tamil Nadu, India

Chairman
Dr.Sudha Ramalingam
Director – Research and Innovation
Professor- Community Medicine,
PSG Institute of Medical Sciences & Research, Coimbatore

Member Secretary
Dr. A Thirumani Devi
Professor
Department of Food Science and Nutrition

Members
Mr.K. Arulmoli (Legal Expert)
Dr. Subashini K.Sripathi
Dr. A Saraswathy(Medical Officer)
Ms. D. Kavitha
Dr. A R Sudamani Ramasamy
Dr. G. Victoria Naomi
Dr. Judith Justin
Dr. Anitha Subash
Dr. K. Sampath Rani

05.01.2023

To
Ms. Anuradha, P.
Department of Computer Science
Avinashilingam Institute for Home Science and
Higher Education for Women
Coimbatore- 641043

Dear Anuradha,

Ref: Your proposal No. IHEC/22-23/CS-01 entitled "Prediction of Risk of Heart Disease Using Machine-Learning Techniques" submitted for approval of IHEC on 21.11.2022.

The Institutional Human Ethics Committee of our University hereby grants approval to your research proposal No. IHEC/22-23/CS-01 entitled Prediction of Risk of Heart Disease Using Machine-Learning Techniques". The Approval number for the same is AUW/IHEC/CS-22-23/XMT-01.

We wish you all the best in your research endeavours.

Regards

5.1.23

Dr. A Thirumani Devi
Member Secretary

Publications

- Anuradha. P and Vasantha Kalyani David,“Improved Heart Diseases Risk Prediction using Optimized Super Learner Ensemble Model”, International Journal of Intelligent Engineering and Systems, Vol.17, No.5, 2024, DOI: 10.22266/ijies2024.1031.25 (scopus- indexed)

- Anuradha. P and Vasantha Kalyani David, "Feature Selection by ModifiedBoostARoota and Classification by CatBoost Model on High Dimensional Heart Disease Datasets," *International Journal of Computer Theory and Engineering*, vol. 14, no. 4, pp. 141-148, 2022. doi: 10.7763/IJCTE.2022.V14.1321. (scopus-indexed)

- Anuradha. P and Vasantha Kalyani David, “Super Learner Model In Prediction Of Heart Attack Based On Cardiac Biomarkers”, *Indian Journal of Computer Science and Engineering*, Vol. 12. No.6, 2021. doi: 10.21817 /indjcse/ 2021/ v12i6/ 211206076. (Scopus-indexed journal till 2022)

- Anuradha. P and Vasantha Kalyani David, "Feature selection using ModifiedBoostARoota and prediction of heart diseases using Gradient Boosting algorithms," *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021, pp. 19-23, doi: 10.1109/ICCCIS51004.2021.9397154. (paper is scopus indexed)(IEEE conference)

- Anuradha. P and V. K. David, "Feature Selection and Prediction of Heart diseases using Gradient Boosting Algorithms," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 711-717, doi: 10.1109/ICAIS50930.2021.9395819. (paper is scopus indexed)((IEEE conference)

- Anuradha. P, Vasantha Kalyani David, "Feature Selection Using Whale Swarm Algorithm And A Comparison Of Classifiers For Prediction Of Cardiovascular Diseases", *International Journal Of Research And Analytical Reviews (IJRAR)*, Volume.6, Issue 2, pp.123-130, June 2019, Available at http://www.ijrar.org/IJRAR1ANP018.pdf

- Anuradha. P and Vasantha Kalyani David, “Leveraging The Power of Hybrid Machine Learning Algorithms to Predict Cardiovascular Diseases - A Review”, *i-manager’s Journal on Computer Science*,(2017), 5(3), 60-67. https://doi.org/10.26634/jcom.5.3.14018

Avinashilingam Institute for HomeScience and Higher Education for Women
(Deemed to be University Estd. u/s 3 of UGC Act 1956, Category 'A' by MHRD
Re-accredited with A++ Grade by NAAC. CGPA 3.65/4, Category I by UGC
Coimbatore - 641 043, Tamil Nadu, India

Appendix L2

(Item No 5 of Check List) Details of Research Publications

S.No	Article	Journal	Other Details Vol/No/Page No/ Year	Published in UGC-CARE / Scopus Indexed/ Web of Science
1	FEATURE SELECTION BY MODIFIEDBOOSTARDOTA AND CLASSIFICATION BY CATBOOST MODEL ON HIGH DIMENSIONAL HEART DISEASE DATASETS	INTERNATIONAL JOURNAL OF COMPUTER THEORY AND ENGINEERING	VOL.14, NO.4, PP.141-148 2022.	SCOPUS INDEXED
2	IMPROVED HEART DISEASES RISK PREDICTION USING OPTIMISED SUPERLEARNER ENSEMBLE MODEL.	INTERNATIONAL JOURNAL OF INTELLIGENT ENGINEERING AND SYSTEMS (INASS PUBLISHER)	VOL.17, ISSUE 5. (PAPER ACCEPTED TO BE PUBLISHED IN VOL.17, ISSUE 5) (ON AUGUST 31ST 2024)	SCOPUS INDEXED

*Proof of list of Journals from Internet to be attached along with copies of reprints.

Scholar : Anuradha

Supervisor : Dr. Kalyani David

Checked By: 18/7/2024
HoD/Dean of Respective School

The scholar Miss. Anuradha, P (Reg.No. 17PHCSP007) has published/ got acceptance for her research paper in the following journals:
1. International Journal of Computer Theory and Engineering - indexed in Scopus and
2. International Journal of Intelligent Engineering and Systems - indexed in Scopus - she got acceptance from this journal.
This may be considered.

16.07.24

Received: April 28, 2024. Revised: July 2, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Improved Heart Diseases Risk Prediction Using Optimized Super Learner Ensemble Model

Anuradha P[1]* **Vasantha Kalyani David[1]**

[1]*Department of Computer Science,*
Avinashilingam Institute for Home Science and HigherEducation for Women, Coimbatore, T.N, India
* Corresponding author's Email: anujith72@gmail.com

Abstract: Cardio Vascular Diseases (CVD) has become a serious concern for humans as fatalities rate due to CVD are increasing at an alarming pace. With the aid of machine learning techniques, heart illnesses can be predicted much earlier, and therapy or dietary changes can prevent deaths. By combining predictions from various individual models, the machine learning technique known as ensemble learning improves forecasting accuracy and resiliency. In this work, a Super Learner Ensemble Model is used where the base learners are a diverse combination of linear, probabilistic, bagging, boosting and stacking models. To improve the performance of the Super Learner Ensemble Model, an Optimized Super Learner Ensemble Model (OSLEM) is proposed, where optimal selection of base learners in the ensemble is done based on the pairwise disagreement accuracy diversity measure of classifiers in each best fitness whale obtained by different iterations of Whale Optimization Algorithm (WOA). ModifiedBoostARoota (MBAR), a wrapper feature selection technique is used to choose the most significant features of six different heart datasets and the proposed OSLEM modelled on the selected features exhibits high performance compared to other existing ensemble models.

Keywords: Heart diseases, Super learner, Ensemble, Prediction, Optimization, Classifier.

1. Introduction

1.1 Ensemble learning

Classifier ensembles have become increasingly prevalent in recent years due to the better efficacy of ensemble models compared to the single classifiers. By utilizing the collective outcomes of the classifiers, the ensemble model seeks to reduce any inaccuracies or biases that may be present in individual models. This method not only improves accuracy but also offers resistance to data uncertainty. Ensemble learning has proven to be a strong tool in many domains, providing more robust and reliable forecasts by successfully combining predictions from numerous models.

Latha C.B.C. et al. [1], tested individual classifiers such as Multilayer Perceptron, SVM, PART, Bayes Net, C4.5, and Naive Bayes as well as combinations of these classifiers using ensemble approaches such as stacking, boosting, majority voting, and bagging. When comparing the levels of accuracy produced by the various approaches, it was found that the majority voting classifier produced higher levels of accuracy.

The nu-SVC algorithm, which uses linear, polynomial, RBF, and sigmoid kernels, was employed by Abdar et al. [2], as the foundation for their suggested NE-nu-SVC model. The Nested Ensemble (NE) approach enables the combination of various ensemble learning methodologies at various model levels. At three separate levels, they used four ensemble learning strategies here. The stacking and bagging approaches were used to integrate the nu-SVC, Stochastic Gradient Descent (SGD), and random forest algorithms at the first level. The voting method was utilised at the second level, while the SMO (Sequential Minimal Optimisation) and Naive Bayes algorithms were applied at the third level.

Additionally, approaches for balancing the dataset and feature selection by GA were used to improve the performance of the new model on heart datasets.

To choose classifiers for an ensemble, Zhi-Hua et al. [3] employed Genetic algorithm based Selective ENsemble (GASEN) method which uses evolved weights that could relate to the fitness of incorporating, neural networks as base classifiers in the ensemble. Analyzing the relationship between the neural network ensemble's capacity for generalization and the correlation of individual networks indicates that, under some circumstances, assembling a specific subset of individual networks is preferable to assembling all of the individual networks.

Empirical studies ascertain that ensembles perform better than single classifiers [4,5]. The main issue with ensemble approaches is that the final ensemble often has large number of classifiers. According to experimental evidence [6], a smaller number of classifiers can still retain the generalization performance of an ensemble. The base learners in an ensemble can be optimized which would improve the efficiency of the ensemble. Several meta-heuristic algorithms based on the hunting behavior of animals are used in optimization.

1.2 Whale optimization algorithm

WOA is a meta-heuristic optimization algorithm developed by Mirjalili, S., & Lewis, A [7], draws inspiration from the humpback whales' hunting techniques. Hunting schools of small fishes near the surface is preferred by humpback whales. Bubble-net feeding, which involves the formation of unique bubbles along a spiral path, is the term used to describe the humpback whales' foraging behaviour.

The spiral bubble-net feeding strategy is mathematically modelled in the WOA in order to achieve optimization. WOA modelled hunting activity by using random search agent to head after the prey. The agent believed to be closest to the target prey is the current best candidate solution, other solutions update their location in relation to the best agent.

$$\vec{D} = \left|\vec{C} \bullet \vec{X}^*(t) - \vec{X}(t)\right| \quad (1)$$

$$\vec{X}(t+1) = \vec{X}^*(t) - \vec{A} \bullet \vec{D} \quad (2)$$

In Eqs. (1) and (2), $\vec{X}$ represents the position vector, $\vec{X}^*$ represents the current best solution, t represents the current iteration. Parameters $\vec{A}$ and $\vec{C}$ are determined using Eqs. (3) and (4).

$$\vec{A} = 2\,\vec{a} \bullet \vec{r} - \vec{a} \quad (3)$$

$$\vec{C} = 2 \bullet \vec{r} \quad (4)$$

The symbol $\vec{r}$ represents a random number in the interval [0,1]. The decreasing enclosing behavior is achieved by $\vec{a}$ dropping linearly from 2 to 0 during the course of exploration and exploitation repetitions.

Based on the hunting strategy of humpback whales, two approaches are intended: 1) Shrinking encircling mechanism: This behaviour is achieved by decreasing the value of 'A' from 2 to 0 over the course of iterations. 2) Spiral updating position: When whales hunt, they simultaneously employ a spiraling course and a shrinking encirclement to secure a prey.

$$\vec{D}' = \left|\vec{X}^*(t)\right| - \vec{X}(t) \quad (5)$$

$$\vec{X}(t+1) = \vec{D}' \bullet e^{bl} \bullet cos(2\pi l) + \vec{X}^*(t) \quad (6)$$

$\vec{D}'$ in Eq. (5) is the distance between the best solution so far and the i^{th} whale, where l takes random value between -1 and 1 , and the profile of the logarithmic spiral is defined by the constant b. Each mechanism is deployed with an equal chance to simulate this behavior.

$$\vec{X}(t+1) = \begin{cases} \vec{X}^*(t) - \vec{A} \bullet \vec{D} & , if\ p < 0.5 \\ \vec{D}' \bullet e^{bl} \bullet cos(2\pi l) + \vec{X^*}(t) & , if\ p \geq 0.5 \end{cases} \quad (7)$$

Where p is a random number of [0,1]. Humpback whales search randomly according to the position of each other.

$$\vec{D} = \left|\vec{C} \bullet \overrightarrow{X_{rand}} - \vec{X}\right| \quad (8)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A} \bullet \vec{D} \quad (9)$$

Where $\overrightarrow{X_{rand}}$ is a location vector selected from the current set at random.

Algorithm: Whale Optimization algorithm (WOA) **(**Mirjalili, S., & Lewis, A., 2016) [7]

Initialize the whale population X_i (i = 1, 2...n)
Calculate the fitness of each search agent
X^*-the best search agent
while (t < maximum number of iterations)
for each search agent
Update A, C, 1, and p
ifl (p<0.5)
if2 (A<1)
Update the position of the current search agent by the Eq. (1)
else if (A≥1)
Select a random search agent (X_{rand})
Update the position of the current search agent by the Eq. (9)
end if2
else ifl (p>0.5)
Update the position of the current search by the Eq. (6)
end ifl
end for
Check if any search agent goes beyond the search space and amend it
Calculate the fitness of each search agent
Update X^* if there is a better solution
t = t+1
end while
return X^*

1.3 ModifiedBoostARoota (MBAR) feature selection algorithm

Feature Selection is used for finding the most pertinent features that contribute to the prediction of target variable of a given dataset. The feature selection algorithm ModifiedBoostARoota[8] is a wrapper method which uses catboost as base model. The algorithm is as follows:

Algorithm **ModifiedBoostARoota (MBAR) [8]**

1. Compute shadow feature (by shuffling original features at random) for each feature in the dataset and merge the shadow features with the dataset to form an extended dataset of 'n' features.
2. Using any Tree-Based models, compute the Feature Importance (FI) of all features in the extended dataset.
3. For each feature compute the rank, r_i is i=1 to n.
4. If the original feature's FI is less than the shadow feature's FI, then both the original and shadow features is removed.
5. If FI of any feature is insignificant then remove that feature.
6. For each feature in the extended dataset, determine its fscore (fs) using,

$$fs_i = \frac{r_i}{FI_i}, i = 1, \dots, n$$

7. Compute weighted harmonic mean as:

$$whm = \frac{\sum r_i}{\sum fs_i}, i = 1, \dots, n$$

8. If $fs_i < whm$, then remove feature $'i'$ from the dataset.
9. If fs of any original feature < fs of its corresponding shadow feature, then eliminate that original feature. Also, if fs of any feature is insignificant then remove that feature.
10. Repeat steps 1 to 9 until in each iteration at least 10% of the features are eliminated or if maximum iterations have not been completed. Else, return the remaining features and stop.

The authors, Anuradha. P and VK David, in their previous work [9], had developed the Super Learner Ensemble Model (SLEM) with a diverse combination of base learners for heart disease prediction. The work in [9] demonstrated that selecting the best base learners will lead to optimal ensemble model performance. Therefore, the objective of this study is to optimize the selection of base learners in the ensemble model. This study proposes Optimized Super Learner Ensemble model (OSLEM), where the Whale Optimization algorithm (WOA) is used for selecting the base learners of the ensemble classifier. Then, disagreement pairwise diversity measure is used to assess the dissimilarity or variation among the predictions made by chosen individual classifiers. It quantifies the level of disagreement or diversity between pairs of classifiers. So, on measuring the disagreement pairwise diversity, the subsets of classifiers with greater diversity measure are chosen as the base models for the OSLEM classifier and further, the classifier is evaluated.

The remaining part of this paper is organized as follows: Section 2 describes the methodology, Section 3 describes the datasets, Section 4 discusses

the results and section 5 gives the conclusion of this work.

2. Methodology

Initially, the significant features in each of the heart disease datasets are selected using ModifiedBoostARoota (MBAR). The base learners considered for the ensemble are a diverse combination of linear, probabilistic, bagging, boosting and stacking models. Support Vector Machine (SVM), Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Random forest (RF) [10], Decision Tree (DT), k-Nearest Neighbors (KNN), Majority Vote Ensemble (MVE), XGboost (XGB) [11, 12], and CatBoost (CatB) [13, 14] are considered for base learners in the Super Learner Ensemble Classifier. Stratified 5-fold cross validation is performed three times, and the average

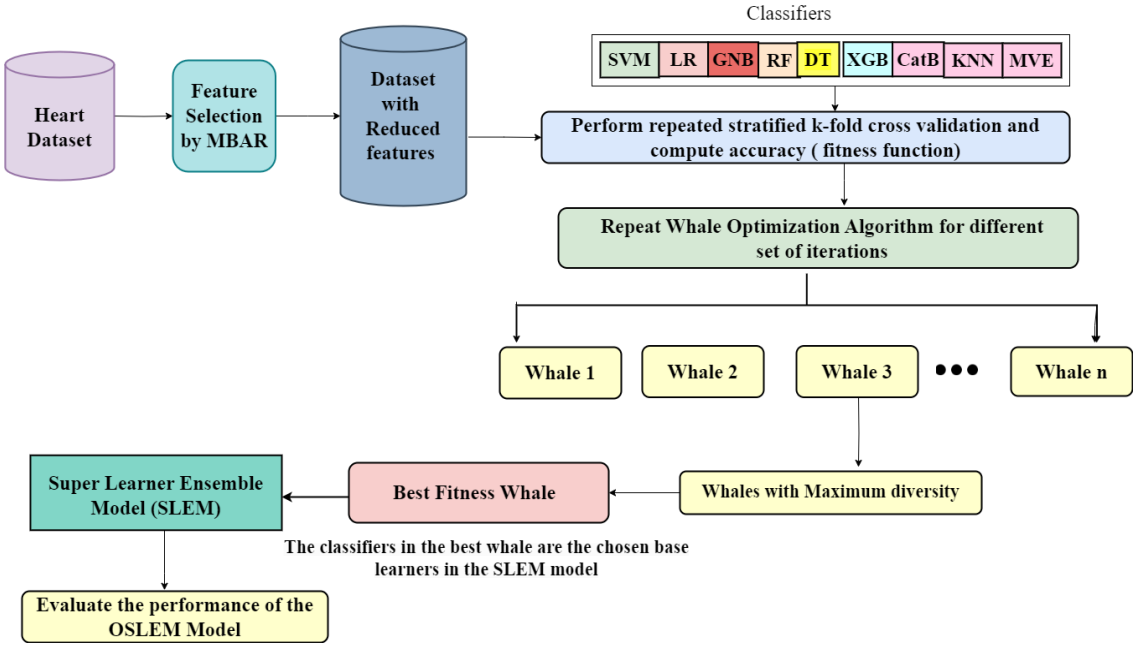

Figure. 1 Optimization of base learners using WOA and diversity measure

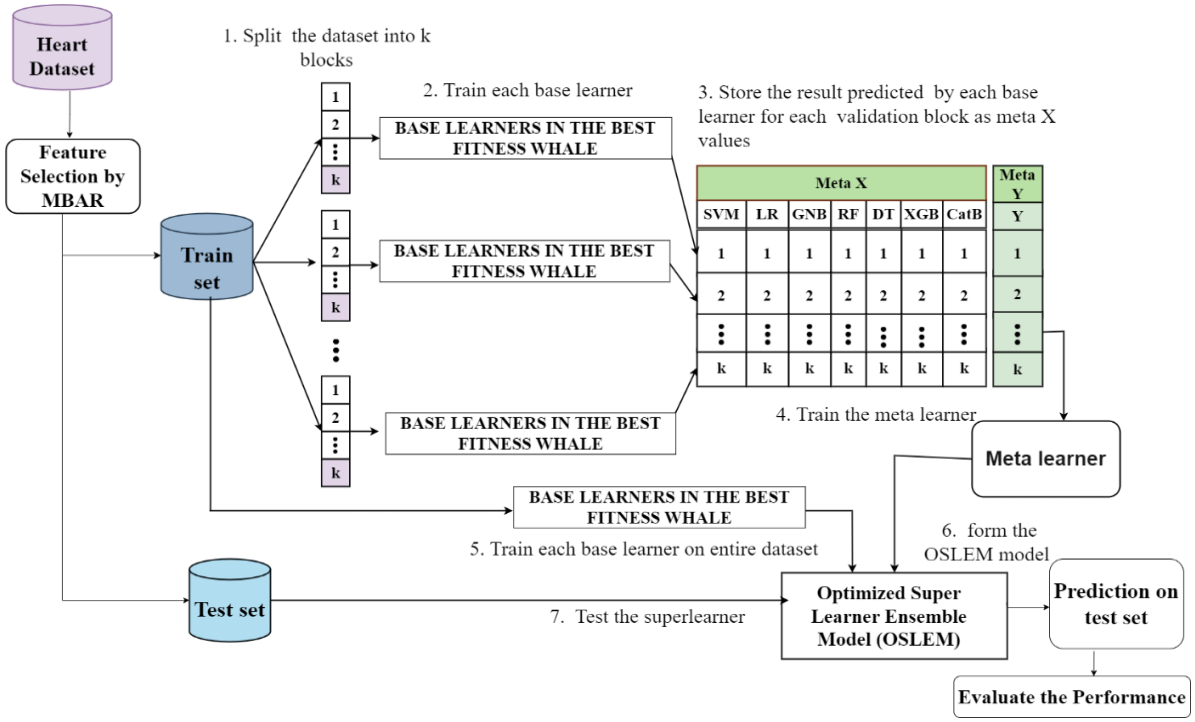

Figure. 2 Proposed Optimized Super Learner Ensemble Model

accuracy of each classifier is determined. The absence or presence of these base classifiers is coded as whales in Whale Optimization Algorithm (WOA). The WOA outputs a whale with good fitness value. This is repeated for different sets of iterations. Then, the pairwise disagreement accuracy diversity value is calculated for each whale, and the most diverse whale is chosen. This whole process is demonstrated in Fig. 1.

Fig. 2 demonstrates the proposed Optimized Super Learner Ensemble Model (OSLEM). Super Learner Ensemble Model (SLEM) designed with the best fitness whale (base models) as its base Learners and the meta-learner, LR, integrating predictions from various base models, is the proposed Optimized Super Learner Ensemble Model (OSLEM). The proposed OSLEM model is devised to improve the ensemble model's performance by selecting the optimal combination of base classifiers using WOA and pairwise disagreement accuracy diversity measure. The proposed OSLEM's performance is evaluated on test data.

3. Datasets

In this study, five low dimensional and two high dimensional heart disease datasets were used. The features of the low dimensional heart disease datasets are shown in Table 1.

i) Cardiovascular disease dataset taken from Mendeley data [15] has 1000 instances.

ii) Cleveland heart disease dataset taken from UCI Machine Learning Repository, created by Robert Detrano, M.D., Ph.D., V.A. Medical Center, Long Beach and Cleveland Clinic, has 303 instances.

iii) Statlog heart disease dataset taken from UCI Machine Learning Repository has 270 instances.

iv) South African heart disease dataset, hosted at Harvard Dataverse, has 462 instances.

v) The high- dimensional dataset, Arrhythmia heart dataset [16] in the UCI Machine Learning Repository, consists of ECG signals data with 279 attributes and 452 instances. Among the attributes, 206 contained linear values, and the rest are nominal. The instances of the dataset belonged to sixteen groups or classes [16] [17]. Class 1 referred to normal beats.Class 2 to Class 15 referred to different types of Arrhythmias. Unclassified beats were grouped as in Class 16 [16] [17]. There are 245 instances of normal types, and 207 instances of the abnormal types. In this work, these instances are grouped into two classes: normal and arrhythmia.

vi) The next high- dimensional dataset used is the Z-Alizadeh Sani dataset in the UCI Machine Learning Repository that consists of 55 features related to coronary artery disease and 303 instances. The dataset contains ECG, demographic, laboratory,

Table 1. Heart Disease Datasets

Statlog heart dataset	Cleveland heart dataset	South African heart dataset	Cardiac biomarkers dataset	Cardiovascular disease dataset
1. age 2. sex 3. fbs-fasting blood sugar 4. cp-chest pain type 5. chol-serum cholesterol 6. restbp-resting blood pressure 7. restecg-ecg at rest 8. maxheartrate - maximum heart rate 9. angina-exercise induced angina 10. colored vessels - number of major vessels colored 11. slope- slope of the peak exercise ST segment 12. thal -defect type 13. oldpeak- ST depression induced by exercise 14. Target: disease	1. age 2. sex 3. fbs-fasting blood sugar 4. cp-chest pain type 5. chol-serum cholesterol 6. trestbps- resting blood pressure 7. restecg -ecg at rest 8. thalach-maximum heart rate 9. exang-exercise induced angina 10. ca-number of major vessels colored 11. slope- slope of the peak exercise ST segment 12. thal -defect type 13. oldpeak-ST depression induced by exercise 14. Target: Num	1. adiposity 2. obesity 3. sbp: systolic blood pressure 4. tobacco: cumulative tobacco 5. ldl: low density cholesterol 6. famhist: family history 7. typea: type-A behavior 8. alcohol: alcohol consumption 9. age: age at onset 10. Target: class	1. Age 2. Gender: 0- female, 1-Male 3. CKMB- Creatine Kinase Myocardial Band 4. Myoglobin 5. Troponin-i 6. BNP- Brain Natriuretic Peptid 7. D-Dimer 8. ACS_types- heart disease types 9. Target: Disease: 0- no AMI, 1- AMI, 2- heart problems	1. Age 2. Gender 3. Chest Pain 4. Resting BP 5. Serum cholesterol 6. Fasting Blood sugar 7. Resting Electro 8. Max heart rate 9. Exercise angina 10. Old peak 11. Slope 12. No.of Major Vessels 13. Target

echo, symptom and examination data of the patients [18]. A patient is categorized as normal, if his/her diameter narrowing is less than 50%; otherwise she or he has CAD

vii) Cardiac Biomarkers dataset is real world data of patients who presented chest pain and had undergone lab tests, at Specialist hospital, Bangalore, India, in order to confirm whether they had a Myocardial Infarction (MI) or not. The personal details of the patients were not disclosed. The data collected consists of Age, Gender, CKMB, Myoglobin, Troponin-I, BNP, D-Dimer, ACS_types, Disease, shown in Table 1, consists of 192 instances. Abnormal values of the biomarkers indicate MI or heart attack.

4. Results and discussion

Using Ubuntu OS and Python in Jupyter Notebook, the experiment was done on a system with an i5 processor and 4 GB of RAM. For this research work, the experiments were performed on all the datasets mentioned in section 4. First, the datasets are pre-processed and balanced using Synthetic Minority Over Sampling Technique (SMOTE). Then the significant features contributing to the heart disease are selected using ModifiedBoostARoota (MBAR). The super learner framework [19] is adopted for the classification of datasets and SVM, LR, GNB, RF, DT, KNN, MVE, XGB, and CB are considered for base learners in the Super Learner Ensemble Classifier. Stratified 5-fold cross validation is performed three times, and the average accuracy of each classifier is determined.

In WOA, the size of a whale is the number of base classifiers. Classifier occurrences make up each whale, with 'zeros' representing the absence of that classifier from the combination and 'ones' representing its presence. The following Eq. (10), is used for calculating the fitness function of WOA for selecting the subset of classifiers.

$$F(i) = \max_{\vartheta} \left(accuracy\,(X) - \gamma * \frac{selected\ base\ classifier\ (X)}{Total\ number\ of\ base\ classifers\ (Y)} \right) \quad (10)$$

$F(i)$ = fitness function; ϑ is the whale population, $accuracy\,(X)$ depicts classification accuracy of selected base classifier. The sum of the base classifiers is denoted by Y. There are two steps to the Eq. (10). Both the classification accuracy and the fraction of base classifiers used in the final decision are taken into account during the first step. Variable γ has a value between zero and one. If they put γ close to 1, it shows that model correctness is more essential than the number of base classifiers used. If two whales have identical accuracy values, the one with fewer base classifiers is chosen. When the average fitness of an infinite number of populations does not vary, we have the convergence condition.

Disagreement pairwise diversity metric measures the diversity among the predictions of different base

Table 2. Performance accuracy of the classifiers on low-dimensional datasets with features selected by MBAR

Classifiers	Accuracy on datasets (with Feature selection by MBAR)						
	Cardiac Biomarkers Dataset	Cleveland HD Dataset	SA HD Dataset	Statlog HD Dataset	Cardio-vascular Dataset	Arrhythmia Heart dataset	Z-Alizadeh Sani heart dataset
CB	100.0%	85.5%	81.8%	83.3%	99.0%	89.8%	96.1%
GNB	87.50%	86.8%	70.2%	77.8%	94.5%	75.5%	87.40%
XGB	100.0%	84.2%	76.9%	84.4%	97.2%	86.7%	93.1%
RF	100.0%	85.5%	76.8%	83.3%	98.6%	89.8%	92.0%
DT	95.83%	80.3%	76.0%	83.3%	95.1%	71.4%	90.8%
MVE	93.75%	88.2%	71.9%	83.3%	98.3%	82.7%	94.3%
SVM	75.0%	85.5%	72.7%	77.8%	79.7%	73.5%	74.%
KNN	81.25%	82.9%	72.7%	83.3%	83.1%	68.4%	61.0%
LR	93.75%	86.8%	69.4%	77.8%	96.9%	74.5%	94.2%
SLEM	95.00%	85.5%	81.0%	83.3%	98.6%	88.8%	93.1%
OSLEM	**100%**	**93.4%**	**89.2%**	**92.6%**	**99.0%**	**90.0%**	**97.7%**

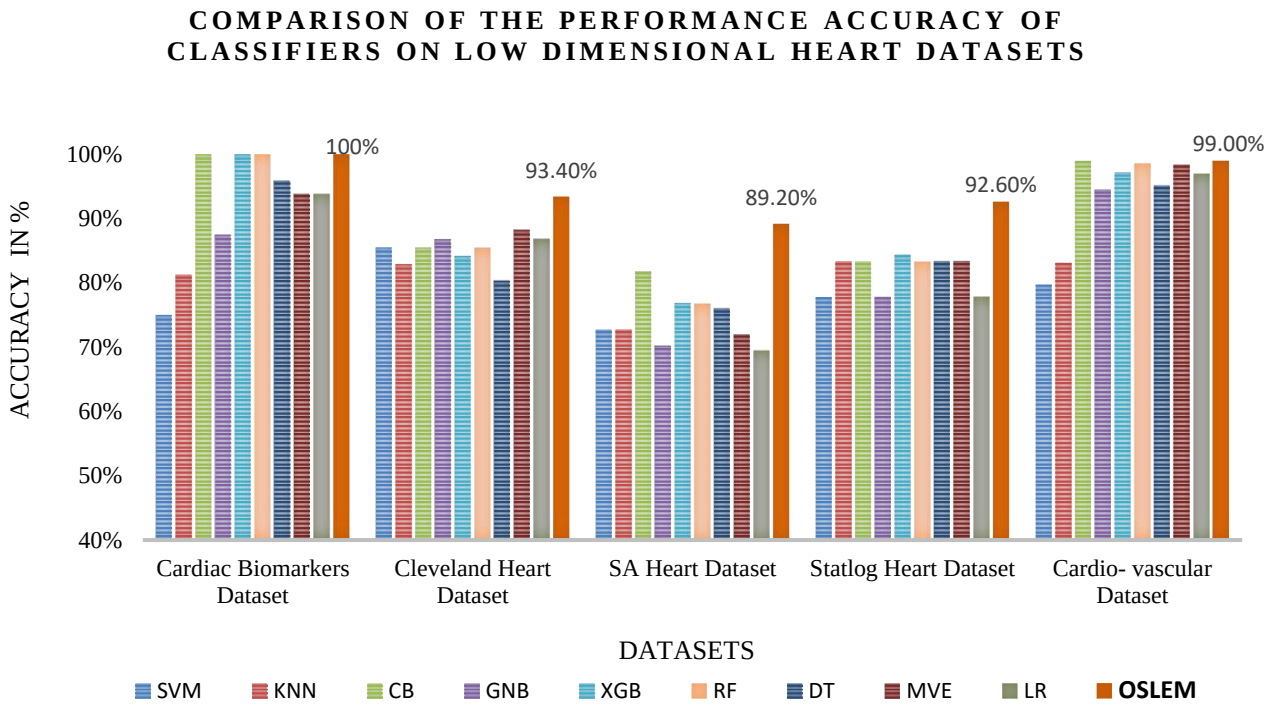

Figure. 3 Performance accuracy of the classifiers on low dimensional heart datasets

Table 3. Performance metrics of proposed MBAR+ OSLEM on both low and high dimensional heart datasets

Dataset	Precision	Recall	f1-score	Accuracy
Cardiac Biomarkers dataset	100%	100%	100%	**100%**
Cleveland heart dataset	95.80%	88.50%	92.00%	**93.40%**
SA heart dataset	92.00%	74.10%	82.10%	**89.30%**
Statlog heart dataset	88.90%	88.90%	88.90%	**92.60%**
Cardio vascular disease dataset	98.50%	99.20%	98.80%	**99. 0%**
Arrhythmia heart dataset	88.00%	91.70%	89.80%	**90.00%**
Z-Alizadeh Sani heart dataset	97.87%	97.87%	97.87%	**97.70%**

Comparison of the Performance Accuracy of OSLEM with Other Classifiers on High Dimensional Heart Datasets

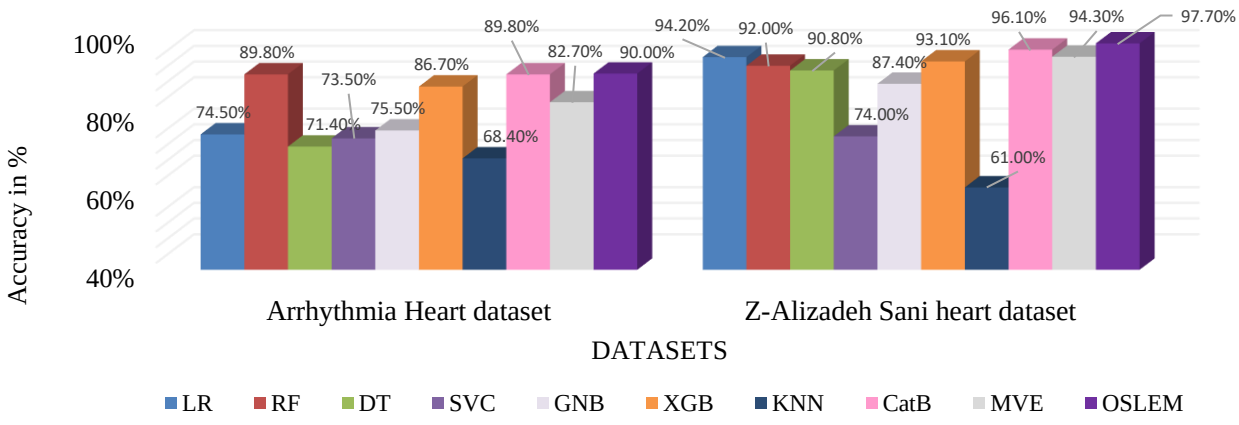

Figure. 4 Performance accuracy of the classifiers on high dimensional heart datasets

Table 4. Performance comparison of the proposed OSLEM with other existing models on cleveland heart dataset

Related Work	Methodology	Precision	Recall	F1-score	Accuracy
Doppala, B. P., 2022 [20]	Ensemble(NB, RF, SVM, XGBoost) with voting	85.00%	90.00%	88.00%	88.24%
Rajadevi, R., 2021 [21]	Black Hole Optimization Algorithm (BHO), XGBoost	90.00%	92.00%	91.00%	92.30%
Zhenya and Zhang, 2021 [22]	ReliefF, ensemble model(RF, LR, SVM, ELM, KNN)	88.67%	89.68%	88.84%	91.60%
Wijaya.H et al., 2018 [23]	Particle Swarm Optimization (PSO), NB	87.77%	88.67%	88.22%	86.67%
Wenxin, X. 2020 [24]	Ensemble model(DT, SVM, ANN)	82.80%	90.80%	87.00%	87.00%
Proposed model	**MBAR with OSLEM**	95.80%	88.50%	92.00%	**93.40%**

Table 5. Performance metrics of proposed model and existing models on statlog heart dataset

Related Work	Methodology	Precision	Recall	F1-score	Accuracy
Zhenya and Zhang, 2021 [22]	ReliefF, ensemble model (RF, LR, SVM, ELM, KNN)	92.6%	92.2%	92.4%	92.4%
Hera, S.Y., et al., 2022 [25]	RF, Multi-Tier Ensemble (MTE) (Stacking (RF, LR, SGD), bagging GBC, ADA)	84.9%	79.2%	81.5%	84.1%
Jikuo Wang et al., 2020 [26]	Stacking(GNB, GB, RF, ET, ADB, MLP, XGB)	94.7%	85.8%	89.1%	90.7%
Proposed Model	**MBAR with OSLEM**	88.9%	88.9%	88.9%	**92.6%**

learners of a whale. The calculation is as follows: Suppose we have N classifiers in a whale, where each classifier makes a prediction on the same set of data points. For each pair of classifiers (i, j), compare their predictions. Let d_{ij} = 1 if the predictions of item 'i' and item 'j' disagree on a data point, and 0 if they agree. Calculate the average disagreement over all pairs of classifiers and all data points. The disagreement measure D is given in Eq. 11 as:

$$D = \frac{2}{N(N-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{1}{M} \sum_{k=1}^{M} dij(k) \quad (11)$$

Applying WOA and pairwise disagreement accuracy diversity measure, the optimal set of classifiers is chosen as base learners for the proposed Optimized Super Learner Ensemble Model (OSLEM).

Then OSLEM is modelled on the selected features of each dataset and the performance is evaluated. Table 2 demonstrates the accuracy attained by individual classifiers on experimenting 3 rounds of stratified cross-validation with 5 folds on the selected low dimensional heart datasets. From Table 2, it is evident that the tree-based learners show significantly higher performance compared to other base learners.

The proposed OSLEM model provides more efficient classification results than other individual classifiers considered. On optimizing the choice of base models in SLEM using WOA, the resultant OSLEM model's performance has improved considerably. Accuracy values achieved by the proposed OSLEM and other classifiers, on low dimensional heart disease datasets with features selected using MBAR, are graphically depicted in Fig. 3. In case of Cardiac Biomarkers Dataset, the OSLEM model attains high accuracy of 100%. Similarly, for Cleveland dataset, SA heart dataset, Statlog Dataset and Cardiovascular disease dataset, the OSLEM model achieves 93.4%, 89.2%, 92.6% and 99.0% prediction accuracy. The proposed OSLEM outperforms other models. Fig. 4 demonstrates that the proposed OSLEM model outperforms the other classifiers when it comes to classification accuracy. For the high-dimensional datasets like Alizadeh Sani dataset and Arrhythmia dataset, the OSLEM achieves 97.7% and 90% prediction accuracy. Table 3 displays the model evaluation metrics of the OSLEM on the heart datasets with features selected by MBAR.

Overall, the model evaluation metric values show that the proposed OSLEM model performs very well on all the Heart disease datasets used in this study.

Tables 4-7 showcases the comparison of the proposed model with other models applied on cleveland, statlog, Arrhythmia and Z-Alizadeh Sani heart disease datasets. Table 4 shows a comparison of the performance of the proposed MBAR+OSLEM with other existing related works on Cleveland heart disease dataset. From Table 4, we can observe that

Table 6. Performance metrics of proposed model and related models on arrhythmia heart dataset

Related Work	Methodology	Specificity	Sensitivity	Accuracy
Iyer, T.J., et al., 2021 [27]	Principal Component Analysis (PCA); XGBoost	92.2%	75.8%	83.20%
Pandey, S.K., et al., 2020 [28]	PCA ; SVM, NB	80.0%	70.0%	89.74%
Mitra, M., & Samanta, R. K., 2013 [29]	Correlation-based Feature Selection; Levenberg-Marquardt	88.4%	86.7%	87.71%
Yilmaz et al., 2013 [30]	Fisher score, Least-squares SVM	82.0%	84.86%	82.09%
R. N. Abirami and P. D. Raj., 2020 [31]	SVM, RF	83.3%	81.3%	85.15%
Proposed Model	**MBAR with OSLEM**	88.0%	91.7%	**90.00%**

Table 7. Performance Metrics of Proposed Model and Related Models on Z-Alizadeh Sani Heart Dataset

Related Work	Classifier	Specificity	Sensitivity	Accuracy
Z. Arabasadi et al., 2017 [32]	Genetic Algorithm, Neural Networks	92.0%	97.0%	93.85%
U.Kilic et al., 2018 [33]	ABC algorithm and SMO	89.43%	89.35%	89.44%
Jikuo Wang et al., 2020 [26]	Stacking(GNB, GB, RF, ET, ADB, MLP, XGB)	94.44%	95.84%	95.43%
Hassannataj .J.J. et al., 2022 [34]	Genetic Support Vector Machine Along With ANOVA (GSVMA)	100%	81.22%	89.45%
Proposed Model	**MBAR+ OSLEM**	**97.5%**	**97.87%**	**97.70%**

the proposed model's precision is 95.80% which is higher than other ensemble models taken into comparison. The Black Hole Optimization (BHO)+XGBoost model [21] has higher recall compared to others, but on observing F1-score and accuracy, the proposed MBAR+OSLEM outperforms other existing models considered.

Table 5 displays a comparison of the performance of the proposed model with other existing ensemble related works on Statlog heart disease dataset. Here, Zhenya and Zhang's model [22] displays higher precision, recall and F1-score, but, the accuracy of the proposed model is narrowly higher than the existing models considered. We can observe that as the number of false positives and false negatives are same, the precision, recall and f1-score values are balanced for the proposed model.

Table 6 displays the performance metrics of proposed and existing models on Arrhythmia heart dataset. The specificity of PCA+XGBoost model [27], is higher than others whereas the accuracy and sensitivity of the proposed model is higher than the existing models considered.

Table 7 exhibits a comparison of the proposed and existing models on Z-Alizadeh Sani heart dataset. The specificity of GSVMA model [34], is 100%, but the proposed model's specificity is higher than the other three models considered. Also, the proposed model's performance is higher than the existing models considered in terms of sensitivity and accuracy.

5. Conclusion

Mortality rate due to cardiovascular diseases can be averted if the disease is predicted at an early stage. Several Machine Learning (ML) algorithms are used for early prediction of heart diseases. Recently, ensemble classifiers are gaining importance as they demonstrate better performance compared to individual ML models. In this work, Support Vector Machine (SVM), Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Random forest (RF), Decision Tree (DT), k-Nearest Neighbors (KNN), Majority Vote Ensemble (MVE), XGboost (XGB), and CatBoost (CatB) are considered as base learners in the Super Learner Ensemble Model (SLEM). These base learners considered are a diverse combination of linear, probabilistic, bagging, boosting and stacking models. In SLEM, repeated stratified k-fold cross validation is done and the predictions of these base learners are integrated to train the meta-learner, LR. In order to improve the performance of the Super Learner Ensemble model (SLEM), an Optimized Super Learner Ensemble Model (OSLEM) is proposed in this study, where optimal selection of base learners in the ensemble is done based on the pairwise disagreement accuracy diversity measure of classifiers in each best fitness whale, obtained by different iterations of Whale Optimization Algorithm

(WOA). Both low and high dimensional heart disease datasets are used in this study. ModifiedBoostARoota (MBAR), a wrapper feature selection technique is used to choose the most significant features of the heart datasets and the proposed OSLEM is modelled on the selected features. On evaluating the performance metrics, the Optimized Super Learner Ensemble Model yields an accuracy of 93.4% ,89.3%, 92.6%, 99%, 90% and 97.7% on Cleveland heart disease, South African heart, Statlog heart disease, Cardiovascular diseases, Arrhythmia heart and Z-Alizadeh Sani heart datasets. The proposed model (MBAR with OSLEM) on comparing with existing related works on heart datasets, demonstrates high performance in terms of specificity, sensitivity and accuracy. In future, other optimization algorithms can be experimented to choose base learners and improve the performance of the ensemble model.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Anuradha.P had conducted the research, analyzed the performance of the models and prepared the manuscript. Dr. Vasantha Kalyani David had guided towards the research work. Both the authors approve the final version.

Availability of data and materials

The datasets used in this study are available online at UCI Machine Learning Repository, Mendeley Data and Harvard Dataverse.

References

[1] Latha CBC, Jeeva SC, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", *Informatics in Medicine Unlocked*, Vol.16, 2019, pp.100203.

[2] M. Abdar, U. R. Acharya, N. Sarrafzadegan, and V. Makarenkov, "NE-nu-SVC: a new nested ensemble clinical decision support system for effective diagnosis of coronary artery disease", *IEEE Access*, Vol.7, 2019, pp.167605-167620.

[3] Z. H. Z. J. X. Wu, & Y. J. S. F. Chen, "Genetic algorithm based selective neural network ensemble", In: *Proc of the Seventeenth International Joint Conference on Artificial Intelligence, Seattle, Washington*, 2001.

[4] T. G. Dietterich, "Ensemble methods in machine learning. In Multiple Classifier Systems: First International Workshop", *MCS 2000 Cagliari, Italy*, pp. 1-15, 2000.

[5] R. O. Duda and P. E. Hart, *Pattern classification*, John Wiley & Sons, 2006.

[6] Z. H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all", *Artificial intelligence*, Vol. 137, pp. 239-263, 2022.

[7] S. Mirjalili and A. Lewis, "The whale optimization algorithm", *Advances in engineering software*, Vol. 95, pp. 51-67, 2016.

[8] P. Anuradha and V. K. David, "Feature selection using ModifiedBoostARoota and prediction of heart diseases using gradient boosting algorithms", In: *Proc of 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp.19-23, 2021.

[9] P. Anuradha and Vasantha Kalyani David, "Super Learner Model in Prediction of Heart Attack based on Cardiac Biomarkers", *Indian Journal of Computer Science and Engineering*, Vol. 12, No. 6, pp. 1702–12, 2021, doi:10.21817/indjcse/2021/v12i6/211206076.

[10] L. Breiman, "Random forests", *Machine Learning*, Vol.45, No.1, pp.5–32, 2001.

[11] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", In: *Proc of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016.

[12] J. Brownlee, "XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn", *Machine Learning Mastery*, 2016.

[13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features", In: *Proc. of 32nd Conference on Neural Information Processing Systems*, Montreal, Canada, 2018.

[14] A. Marz, "Catboost Lss an Extension of Catboost to Probabilistic Forecasting", *arXiv*:2001.02121, 2020.

[15] B. P. Doppala, and D. Bhattacharyya, "Cardiovascular_Disease_Dataset", *Mendeley Data, V1*, 2021.

[16] Arrhythmia Dataset, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/arrhythmia, 1998.

[17] M. Ayar and S. Sabamoniri, "An ECG-based feature selection and heartbeat classification model using a hybrid heuristic algorithm," *Informatics in Medicine Unlocked*, Vol.13, 2018, pp.167-175.

[18] Z-Alizadeh Sani Dataset, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani, 2017.

[19] V. Laan, J. Mark, Polley, C. Eric, and E. Hubbard Alan, "Statistical Applications in Genetics and Molecular Biology", *The Berkeley Electronic Press*, Vol. 6, No.1, 2007.

[20] B. P. Doppala, D. Bhattacharyya, M. Janarthanan, and N. Baik, "A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques", *Journal of Healthcare Engineering*, Vol. 2022, Article ID 2585235, 13 pages, 2022.

[21] R. Rajadevi, E. R. Devi, R. Shanthakumari, R. S. Latha, N. Anitha, and R. Devipriya, "Feature selection for predicting heart disease using black hole optimization algorithm and xgboost classifier", In: *Proc of International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-7, 2021.

[22] Q. Zhenya and Z. Zhang, "A hybrid cost-sensitive ensemble for heart disease prediction", *BMC Medical Informatics and Decision Making*, pp. 21-73, 2021.

[23] S. H. Wijaya, G. T. Pamungkas, &M. B. Sulthan, "Improving classifier performance using particle swarm optimization on heart disease detection", In: *Proc. of 2018 International Seminar on Application for Technology of Information and Communication*, pp.603-608, 2018.

[24] X. Wenxin, "Heart Disease Prediction Model Based on Model Ensemble", In: *Proc of 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China, pp.195-199, 2020.

[25] S. Y. Hera, M. Amjad & M. K. Saba, "Improving heart disease prediction using multi-tier ensemble model", *Network Modeling Analysis in Health Informatics and Bioinformatics*, Vol.11, No.41, 2022.

[26] J. Wang, C. Liu, L. Li, W. Li, L. Yao, H. Li, and H. Zhang, "A stacking-based model for non-invasive detection of coronary heart disease", *IEEE Access*, Vol. 8, 2020, pp. 37124-37133.

[27] T. J. Iyer, B. Kishan, and R. Nersisson, "Prediction and Classification of Cardiac Arrhythmia Using a Machine Learning Approach", *In: Proc. of Advances in Automation, Signal Processing, Instrumentation, and Control, i-CASIC 2020*, Vol. 700, pp.603, 2021.

[28] S. K. Pandey, V. R. Sodum, R. R. Janghel, and A. Raj, "ECG Arrhythmia Detection with Machine Learning Algorithms", In: *Proc. of Data Engineering and Communication Technology,Advances in Intelligent Systems and Computing*, , pp.1079 2020.

[29] M. Mitra and R. K. Samanta, "Cardiac arrhythmia classification using neural networks with selected features, *Procedia Technology*, Vol.10, pp.76-84, 2013.

[30] E. Yılmaz, "An expert system based on fisher score and LS-SVM for cardiac arrhythmia diagnosis," *Computational and Mathematical Methods in Medicine*, Vol. 6, 2013.

[31] R. N. Abirami and P. D. Raj, "Cardiac arrhythmia detection using ensemble of machine learning algorithms", *Soft Computing for Problem Solving*, Vol.1057, pp.475–487, 2020.

[32] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network -Genetic algorithm", *Comput Methods Programs Biomed*, Vol.2017, pp.19-26, 2017.

[33] Ü. Kiliç and M. K. Keleş, "Feature Selection with Artificial Bee Colony Algorithm on Z-Alizadeh Sani Dataset", In: *Proc. of 2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp.1-3,2018.

[34] J. H. Joloudari, F. Azizi, M. A. Nematollahi, R. Alizadehsani, E. Hassannatajjeloudari, I. Nodehi, and A. Mosavi, "GSVMA: A Genetic Support Vector Machine ANOVA Method for CAD Diagnosis", *Frontiers in Cardiovascular Medicine*, Vol.8, pp.760178, 2022.

Feature Selection by ModifiedBoostARoota and Classification by CatBoost Model on High Dimensional Heart Disease Datasets

Anuradha. P and Vasantha Kalyani David

***Abstract*—As heart disease is the leading cause of mortality worldwide, early detection and prevention of the disease would reduce the mortality rate. Various Machine Learning Algorithms are employed in the classification and prediction of diseases. For accurate prediction, Feature Selection algorithms are employed to choose features that have a significant association with the disease or target variable. This would reduce computing time and improve the prediction performance. In this paper, ModifiedBoostARoota (MBAR) algorithm was used for Feature Selection, and classifiers CatBoost, XGBoost, Decision Tree, Extra Trees Classifier, Support Vector Classifier, Logistic Regression, K Nearest Neighbors, Naive Bayes, and Random Forest were applied on UCI Arrhythmia dataset and UCI Z-Alizadeh Sani dataset. Synthetic Minority Over Sampling Technique (SMOTE) was used to balance the dataset. A comparison of the performance of the models on the imbalanced and balanced datasets shows that MBAR with CatBoost classifier gives better accuracy of 92.76% on the balanced Z-Alizadeh Sani dataset and 86.33% on the balanced Arrhythmia dataset.**

***Index Terms*—Heart disease, feature selection, CatBoost, classification.**

I. Introduction

Heart disease is the leading cause of human deaths globally. According to the World Health Organization, Cardio Vascular Diseases were responsible for 38 percent of the 17 million premature deaths (below 70 years of age) caused by noncommunicable diseases in 2019. The four main types of heart diseases are i) Heart failure, ii) heart valve disease, iii) Cardiac Arrhythmia and iv) coronary artery disease.

i) If a valve in the heart is damaged or diseased, it leads to heart valve disease. ii) When the heart muscle becomes weak or when heart chambers are not filled with sufficient blood, the heart will not be able to pump the adequate blood required for the body. This condition is called heart failure.

iii) Cardiac Arrhythmia indicates an abnormality in the sequence of electrical impulses, causing the improper beating of the heart [1]. Arrhythmias may be harmless or life-threatening. The heart's electrical activity can be recorded using Electrocardiography (ECG or EKG), which can help diagnose Arrhythmias [1]. To predict Arrhythmia, analysis of each heartbeat of the ECG records might be done for long hours or days [2]. To detect abnormalities quickly and correctly like Arrhythmia in ECG, Machine Learning algorithms can be used, which would be a support for the medical practitioner. iv) Coronary Artery Disease (CAD) arises due to the accumulation of plaque inside the lining of the coronary arteries that would block blood flow to the heart [3].

High blood pressure, diabetes, low HDL cholesterol, family history, high LDL cholesterol, and smoking are the traditional risk factors for CAD [4]. Machine Learning Algorithms when applied much earlier in life on these risk factors can predict whether an individual is likely to get heart disease or not. In case, if the prediction is positive then, preventive measures to avoid CAD would be to adopt a healthy lifestyle, which includes good nutrition and physical activity [4].

A. Machine Learning Algorithms

Machine learning techniques are devised to predict the target/ output/ dependent variable, for the given input/ predictor variables [5]. Various Machine learning algorithms are available and focus of all research works would be to choose the right algorithm that would best suit for the specific dataset. For supervised learning where the output is known, various algorithms namely Linear Regression, Random Forest, Logistic Regression, CatBoost, Support Vector Machines, K Nearest Neighbors, Decision Trees, XGBoost etc., are widely used.

B. Feature Selection

In datasets, especially in high dimensional datasets, not all features contribute to the prediction of the target or outcome variable. So, selecting the features that are highly associated with the target/class variable would highly contribute to effective prediction as well as save computing time.

C. Synthetic Minority Over-Sampling Technique (SMOTE)

In an imbalanced dataset, all classes will not have an equal number of instances. The classifiers perform better on balanced datasets compared to imbalance datasets. N.V. Chawla *et al*., in their paper on SMOTE, showed that better classifier performance can be achieved by over-sampling the abnormal/ minority class and under-sampling the normal/ majority class. [6].

The objective of this work is to focus on the limitations mentioned by the authors in their previous research work in [7]. A feature selection technique called ModifiedBoostARoota algorithm (MBAR) was devised and

Manuscript received April 6, 2022; revised July 10, 2022.

Anuradha. P and Vasantha Kalyani David are with the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University, Coimbatore, India (e-mail: anujith72@gmail.com).

DOI: 10.7763/IJCTE.2022.V14.1321

applied only on low dimensional heart disease datasets in [7]. The authors had earlier mentioned that MBAR was not applied on high-dimensional datasets due to time constraint.

Therefore, in this work, MBAR is applied on high dimensional heart disease datasets; namely, Arrhythmia dataset and Z-Alizadeh Sani dataset available in the UCI Machine Learning repository where the performance of the classifiers is compared when applied on these datasets with and without selected features. Also, the performances of the selected classifier on the imbalanced and balanced datasets are compared.

The following subsections consist of related work discussion and methodology in Section II, brief description of the datasets in Section III, results and discussion in Section IV, limitations in Section V and conclusion in Section VI.

II. Related Literature Survey and Methodology

On reviewing the feature selection and classification techniques used in earlier research works done on the Arrhythmia dataset, it is observed that A. Mustaqeem *et al*. [8] had accomplished Feature Selection by creating shadow features of each feature based on z-score feature importance by Random Forest Classifier. Those features with a z-score less than the maximum value of shadow features were eliminated [8]. On applying repeated ten-fold cross-validation, those authors found that Multi-layer perceptron gave higher accuracy of 78.2% compared to other classifiers [8].

A. Mustaqeem *et al*., in their work in [9], applied Support Vector Machine (SVM) based methods including one-against-all (OAA), one-against-one (OAO), and error-correction code (ECC). The OAO method when used with 80/20 data split, achieved an accuracy rate of 81.11% and on 90/10 data split, the accuracy obtained was 92.07%.

Khare *et al*., in [10], employed Spearman Rank Correlation for selecting features and Principal Component Analysis (PCA) was used for feature extraction. Then SVM was employed for classification, which gave an accuracy of 85.98%.

Fei Yang *et al*. [11] used an advanced approach for missing-value imputation called Robust Principle Component Analysis (RPCA) along with Zero, Mean, and PCA imputation methods. They modified KDF-WKNN by a correction factor. This modified kernel Difference-Weighted KNN (MKDF-WKNN) classification algorithm was used to manage the imbalance datasets problem and an accuracy of 73% was achieved [11].

Ersen Yılmaz had designed an expert system where feature selection was implemented by F-score and classification was done using Least Squares Support Vector Machines (LS-SVM), in which, Gaussian radial basis function was used as the kernel. The accuracy obtained was 82.09% [12].

Jadhav *et al*. used momentum learning rule with back-propagation algorithm which yielded 82.22% classification accuracy [13].

M. A. Khan and Y. Kim applied the hybrid model, principal components analysis (PCA) with LSTM for classification and attained a classification accuracy of 93.5% [14].

Mitra and Samanta employed correlation-based feature selection (CFS) with linear forward selection search [15]. On applying the Incremental back-propagation neural network (IBPLN) and Levenberg-Marquardt (LM) model, a classification accuracy of 87.71% was obtained [15].

Shimpi *et al*., found that Support Vector Machine classifier yielded a better accuracy of 91.2% compared to other models considered in their work [16].

Ayar *et al*., applied the hybrid model, genetic algorithm along with Decision Tree, for Classification. This hybrid model when applied on two-classes achieved an accuracy of 86.96% [17].

The review of the classification works done on Z-Alizadeh Sani dataset are as follows:

Kolukisa *et al*. applied the linear discriminant analysis and the SVM algorithm, which yielded an accuracy of 92.74% [18].

Kolukisa *et al*. in [19] devised an adaptive ensemble classifier consisting of Logistic Regression, k-Nearest Neighbor, Linear Discriminant Analysis, Support Vector Machine, Naïve Ba yes classification algorithms and obtained 88.38% accuracy [19].

Gupta *et al*. designed a computational intelligent system, C-CADZ, using fixed analysis of mixed data (FAMD) and Binary Bat Algorithm (BBA) for feature extraction, after which an accuracy of 97.37% was achieved by applying an ensemble model of Random Forest and Extra Tree classifier [20].

Arabasadi *et al*. [21] focused on the concept that CAD occurs if one of the left circumflex (LCX) or left anterior descending (LAD) or right coronary (RCA) arteries is stenotic [22]. By using hybrid Neural Network-Genetic algorithms model, those authors achieved 93.85% accuracy. Dahal *et al*. compared five classifiers and observed that the SVM model's prediction was more effective with an accuracy of 89.47% [23].

Cuvitoglu and Isik used Principal Component Analysis (PCA) t-test for feature selection, where five classifiers were compared and Artificial Neural Networks (ANN) yielded an Area-Under-the-Curve value of 93% [24].

Alizadeh Sani *et al*. in [25], used cost-sensitive algorithms along with base classifiers of Support Vector Machine (SVM), K-Nearest Neighbors (KNN), C4.5, Sequential Minimal Optimization (SMO), and Naïve Bayes with ten-fold cross-validation, and better accuracy of 92.09% was achieved by Sequential Minimal Optimization (SMO) [25].

The summarized form of related work on both datasets can be seen in Table II and Table V, where the proposed model is also compared with the related work. Table II shows that almost all the authors have used feature selection on Arrhythmia dataset. In the future, as an extension of these related works, tree-based models can be experimented on both datasets.

Fig. 1 depicts the methodology adopted in this work. The ModifiedBoostARoota (MBAR) algorithm is used for Feature Selection and Classifiers CatBoost, XGBoost, Logistic Regression, Decision Tree, Support Vector Classifier, K Nearest Neighbors, Extra Trees Classifier, Gaussian Naive Bayes and Random Forest were applied on

UCI Arrhythmia dataset and Z-Alizadeh Sani dataset. The stratified 10-fold cross validation accuracy score with three repeats of all the above-mentioned classifiers is compared. Also, on splitting the datasets as 70% train and 30% test sets, the precision, recall, f1 score and AUC score of all these classifiers are analyzed and the best performing classifier is selected. The selected model is applied on the two unbalanced feature selected datasets, SMOTE-balanced feature selected datasets and on the two datasets with no feature selection. The performances are consequently evaluated. The ModifiedBoostARoota feature selection algorithm's performance on high dimensional datasets is assessed [7].

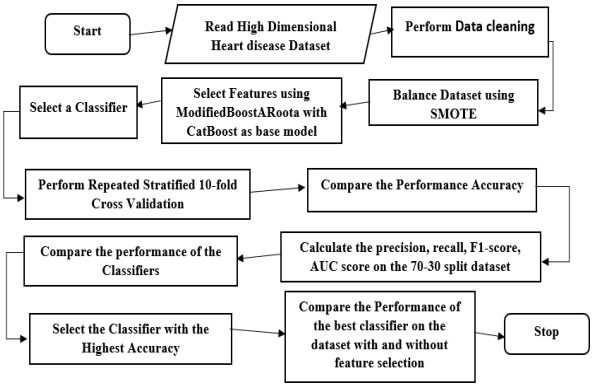

Fig. 1. Methodology.

III. Data Sets

Two datasets were used in this work. The first high-dimensional dataset, the Arrhythmia heart dataset [26] in the UCI Machine Learning Repository, consists of ECG signals data with 279 attributes and 452 instances. Among the attributes, 206 contained linear values, and the rest are nominal. The instances of the dataset belonged to sixteen groups or classes [17], [26]. Class 1 referred to normal beats. Class 2 to Class 15 referred to different types of Arrhythmias. Unclassified beats were grouped as in Class 16 [17], [26]. There are 245 instances of normal types, and 207 instances of the abnormal types. In this work, these instances are grouped into two classes: i) normal and ii) arrhythmia.

The second dataset used is the Z-Alizadeh Sani dataset [27] in the UCI Machine Learning Repository that consists of 54 features related to coronary artery disease and 303 instances. The dataset contains ECG, demographic, laboratory, echo, symptom and examination data of the patients [27]. A patient is categorized as normal, if his/her diameter narrowing is less than 50%; otherwise she or he has CAD.

IV. Results and Discussion

ModifiedBoostARoota algorithm (MBAR) is a wrapper method for a feature selection devised by the authors (Anuradha and David) in their previous work [7], which is mentioned in Fig. 2. MBAR was developed by modifying BoostARoota (BAR) algorithm. BAR was published in Python package Index (PyPI)) and devised by Chasedehan [7]. Catboost is used as the base model in MBAR [7]. In this article, MBAR algorithm is used for feature selection. The experiment was carried out using python on a system with 4 GB RAM and ubuntu operating system.

Algorithm ModifiedBoostARoota [7]:

1. Compute shadow feature (by shuffling original features at random) for each feature in the dataset and merge the shadow features with the dataset to form an extended dataset of 'n' features.
2. Using any Tree-Based models, compute the Feature Importance (FI) of all features in the extended dataset.
3. Assign rank, r_i for all features $i = \text{to} 1\ to\ n$.
4. If FI of original feature < FI of the corresponding shadow feature, then eliminate that original feature and its shadow feature.
5. If FI of any feature is insignificant then remove that feature.
6. Compute fscore for each feature in the extended dataset,
$fs_i = \frac{r_i}{FI_i}$, i=1 to n
7.Compute weighted harmonic mean,
$$m = \frac{\sum r_i}{\sum fs_i}, i = 1\ to\ n$$
8. For any feature $'i'$ in the extended dataset, if $i < whm$, eliminate the feature $'i'$.
9. If fs of any original feature < fs of its corresponding shadow feature, then eliminate that original feature. Also, if fs of any feature is insignificant then remove that feature.
10. Repeat steps 1 to 9 until in each iteration at least 10% of the features are eliminated or if maximum iterations have not been completed. Else, return the remaining features and stop.

Fig. 2. ModifiedBoostARoota algorithm for feature selection.

Initially, in the Arrhythmia dataset, missing values was filled with mean values 36, 49, 37, -14, 75 in columns c10, c11, c12, c13, c14. The normal class was defined as 0 and all other classes in the target variable were grouped as 1. There are 245 instances of class 0 and 207 instances of class 1 [26].

Applying Catboost classifier on Arrhythmia dataset with all features, the stratified ten-fold cross validation accuracy score with three repeats was 83.93%. After balancing the dataset with SMOTE, we get 245 instances of both classes. Then, applying Catboost classifier on the dataset with all features, the stratified ten-fold cross validation accuracy score with three repeats was 85.44%.

Using ModifiedBoostARoota algorithm (MBAR) for feature selection on the unbalanced dataset, one gets 64 features being selected. Upon applying various classifiers namely XGBoost, Logistic Regression, Catboost, Decision Tree Classifier, Gaussian Naive Bayes, Extra Trees, K Nearest Neighbors, Random Forest and Support Vector Classifier, it was observed that Catboost yields highest accuracy of 85.77%.

TABLE I: Performance of the Classifiers by Repeated Stratified K-Fold CV on Arrhythmia Dataset

Classifiers	Accuracy
XGBoost	84.27%
Logistic Regression	73.20%
Decision Tree Classifier	72.86%
Gaussian Naïve Bayes	73.61%
K Nearest Neighbors	71.84%
Random Forest	86.06%
Extra Trees	85.51%
Support Vector Classifier	75.37%
CatBoost	86.33%

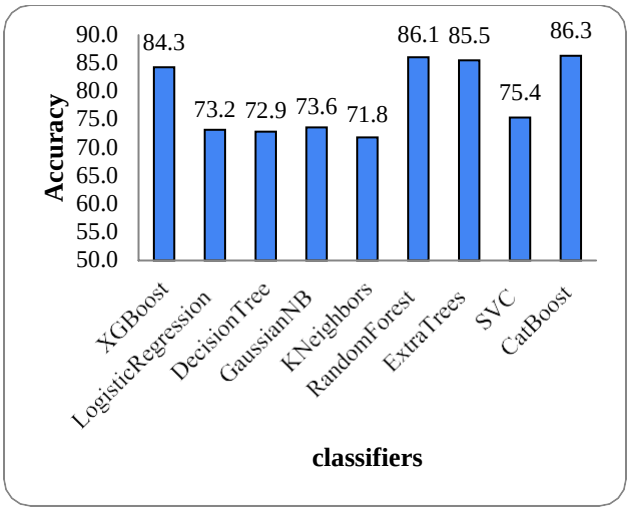

Fig. 3. Comparison of Classifiers by repeated stratified k-fold CV on Arrhythmia dataset after SMOTE and Feature Selection using MBAR.

After balancing the dataset with SMOTE and selecting features using MBAR, Table I displays the accuracy obtained by various classifiers after performing stratified ten-fold cross validation with three repeats. On comparing the performance of all classifiers, CatBoost gives the highest accuracy of 86.33%. Fig. 3 shows that the Tree-Based models performed better on Arrhythmia dataset.

TABLE II: A COMPARISON OF THE MODELS APPLIED ON ARRHYTHMIA DATASET BY VARIOUS AUTHORS

Author	Classifier on Arrhythmia dataset	Accuracy
Mustaqeem *et al*.	FI by RF+MLP	78.2
Mustaqeem *et al*.	Wrapper FS+SVM(OAO)	92.07
Khare *et al*.	Rank corr + PCA + SVM	85.98
Yang *et al*.	MKDF-WKNN	73.01
Yilmaz *et al*.	Fscore+LSSVM	82.09
Jadhav *et al*.	BPNN	82.22
Ayar *et al*.	GA+DT	86.96
Mitra *et al*.	CFS+IBPNN+LM	87.71
Shimpi *et al*.	PCA+ SVM	91.2
Anuradha and David	MBAR+Catboost	85.77
Anuradha and David	MBAR+Catboost (balanced with SMOTE)	86.33

Table II shows a comparison of the models applied on Arrhythmia dataset by various authors. On comparing the performance of other models proposed by various authors detailed in section II, Fig. 4 shows that the proposed model, MBAR+Catboost (balanced with SMOTE) performs generally on par with all models; however, more accurately than those of the Mustaqeem *et al*.'s first method, Khare *et al*., Yang *et al*., Jadhav *et al*. and Yilmaz *et al*. used on Arrhythmia dataset.

Performing 70-30 split of the balanced Arrhythmia dataset with features selected by MBAR, and analyzing the performance of various classifiers, one gets CatBoost classifier displaying higher performance compared to the other classifiers. Table III shows the precision, recall and f1-score of the classifiers considered in the comparison. It shows that Random Forest and CatBoost have very close values of precision, recall and f1-score.

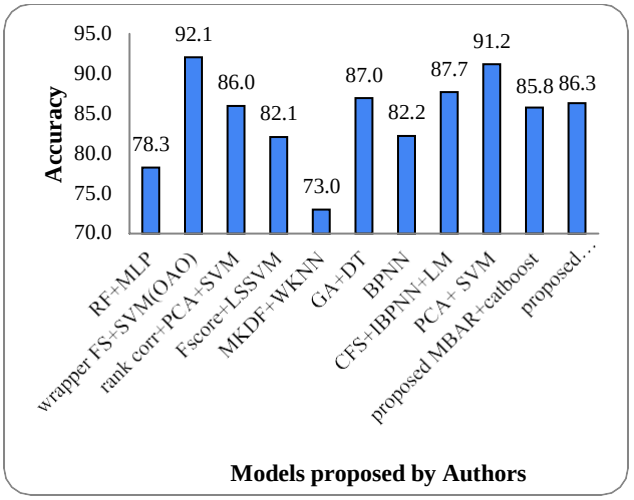

Fig. 4. Comparison of models by various authors on Arrhythmia dataset.

TABLE III: PERFORMANCE OF CLASSIFIERS AFTER SMOTE, FEATURE SELECTION BY MBAR AND APPLYING TRAIN-TEST SPLIT ON ARRHYTHMIA DATASET

Classifiers	Class	Precision	Recall	F1-score	Accuracy
XGBoost	0	0.79	0.83	0.81	80.95%
	1	0.83	0.79	0.81	
Logistic Regression	0	0.67	0.72	0.69	69.39%
	1	0.72	0.67	0.69	
Decision Tree	0	0.68	0.76	0.72	70.75%
	1	0.75	0.66	0.70	
Gaussian NB	0	0.68	0.93	0.79	75.51%
	1	0.90	0.59	0.71	
K Nearest Neighbors	0	0.67	0.89	0.76	73.47%
	1	0.85	0.59	0.70	
Random Forest	0	0.82	0.82	0.82	82.31%
	1	0.83	0.83	0.83	
Extra Trees	0	0.81	0.80	0.81	81.63%
	1	0.82	0.83	0.82	
CatBoost	0	0.82	0.85	0.83	83.67%
	0	0.85	0.83	0.84	

Fig. 5 shows the Receiver operating characteristic curve of classifiers, considered in this study, on Arrhythmia dataset.

The AUC score of CatBoost (CB) is higher than other classifiers. In the Z-Alizadeh Sani dataset [27], the target variable has 216 instances of class 1 and 87 instances of the class 0.

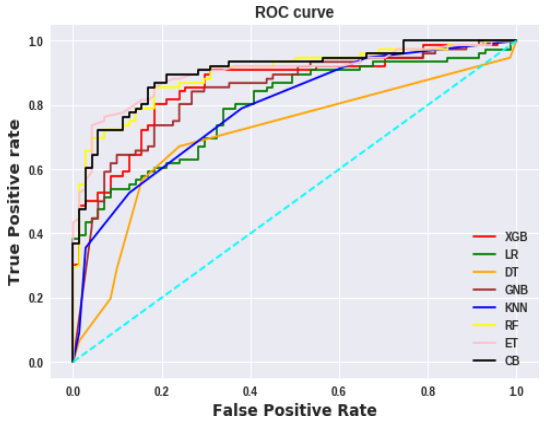

Fig. 5. Receiver operating characteristic curve of classifiers on Arrhythmia dataset.

Classification on the Z-Alizadeh Sani dataset with all features resulted in Catboost yielding a higher accuracy of 87.78%.

On performing feature selection with ModifiedBoostARoota (MBAR) on the imbalanced z-Alizadeh Sani dataset out of 55 features, 12 features were selected. On applying classifiers on these selected features, Catboost gave a better accuracy of 89.44%.

After balancing the dataset with SMOTE, the authors get 216 instances of both classes. The classification accuracy by Catboost applied on all features was 92.14%.

TABLE IV: COMPARISON OF CLASSIFIERS AFTER APPLYING SMOTE, FEATURE SELECTION BY MBAR AND REPEATED STRATIFIED K-FOLD CV APPLIED ON Z-ALIZADEH SANI DATASET

Classifier on Z-Alizadeh Sani dataset	Accuracy
XGBoost	91.21
Logistic Regression	91.98
Decision Tree	86.81
Support Vector Machine	78.32
Gaussian Naïve Bayes	89.28
K-Nearest Neighbors	74.93
Random Forest	92.52
Extra Trees	92.22
CatBoost	92.76

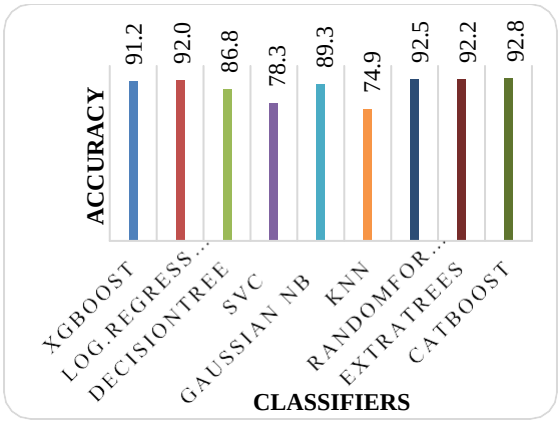

Fig. 6. Performance of classifiers on SMOTE-MBAR and repeated stratified k-fold CV applied Z-Alizadeh Sani dataset.

On performing feature selection by MBAR on the balanced dataset, 21 features were selected. Applying various classifiers using stratified ten-fold cross-validation with three repeats on the balanced dataset, the Catboost model outperformed others yielding an accuracy of 92.76%. Table VI displays the accuracy yielded by various classifiers by repeated stratified k-fold Cross Validation applied on the balanced Z-Alizadeh Sani dataset. Fig. 6 shows that most of the classifiers have performed equally well and Catboost leads by a small margin.

Table V displays the models proposed by various authors mentioned in Section II and exhibits the accuracy obtained by the models they used. Fig. 7 compares the performance of various models proposed by other authors. The proposed model MBAR and Catboost when applied on the balanced dataset outperform other authors' models by yielding an accuracy of 92.76%.

TABLE V: COMPARISON OF THE PROPOSED MODEL WITH EARLIER MODELS ON Z-ALIZADEH SANI DATASET

Authors	Models	Accuracy
dahal *et al.*,	SVM	89.47
koluisa *et al.*,	ensemble	88.38
koluisa *et al.*,	LDA-SVM	92.74
Cuvitoglu *et al.*,	ANN	85
R. Alizadehsani *et al.*,	SMO	92.09
Proposed-MBAR- Catboost	10-fold CV	89.44
Proposed-MBAR- Catboost (balanced with smote)	10-fold CV	92.76

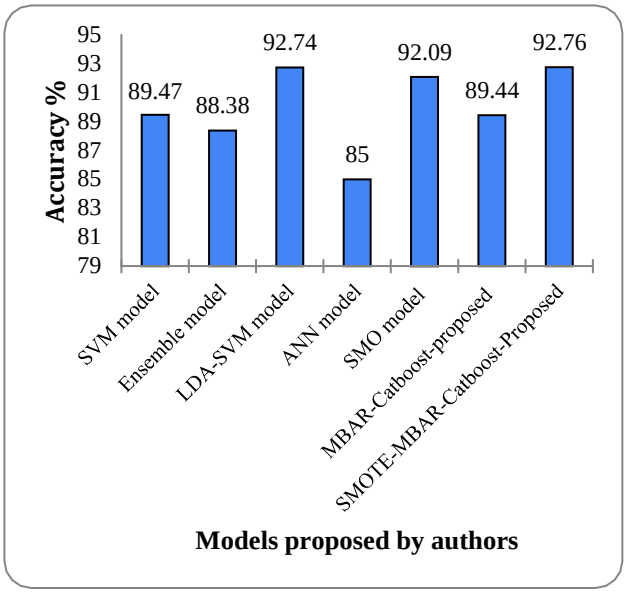

Fig. 7. Comparison of models by various authors on Z-Alizadeh Sani dataset.

Table VI shows the precision, recall and F1 score of the classifiers applied on 70-30 split of the balanced and selected features of Z-Alizadeh Sani dataset. CatBoost classifiers outperforms all other classifiers taken into comparison in this study. Fig. 8 showcases the ROC curve of the various classifiers considered in this study and finds CatBoost displaying better score that the others.

As CatBoost shows better performance compared to other classifiers, on modelling CatBoost on the 70-30 train-test split of the balanced Z-Alizadeh Sani dataset, we find as shown in Table VII that, MBAR-CatBoost combination

demonstrates better performance compared to the performance of the classifier applied on the dataset with no feature selection.

TABLE VI: PERFORMANCE OF CLASSIFIERS AFTER SMOTE AND FEATURE SELECTION BY MBAR ON THE TRAIN-TEST SPLIT Z-ALIZADEH SANI DATASET

Classifiers	class	Precision	Recall	F1-score	Accuracy
XGBoost	0	0.94	0.97	0.95	95.38%
	1	0.97	0.94	0.95	
Logistic Regression	0	0.92	0.94	0.93	93.08%
	1	0.94	0.92	0.93	
Decision Tree	0	0.86	0.89	0.88	87.69%
	1	0.89	0.86	0.88	
Gaussian NB	0	0.90	0.89	0.90	90%
	1	0.90	0.91	0.90	
K Nearest Neighbors	0	0.73	0.84	0.78	76.92%
	1	0.82	0.70	0.75	
Random Forest	0	0.94	0.95	0.95	94.62%
	1	0.95	0.94	0.95	
Extra Trees	0	0.93	0.97	0.95	94.62%
	1	0.97	0.92	0.95	
CatBoost	0	0.94	0.98	0.96	96.15%
	1	0.98	0.94	0.96	

TABLE VII: COMPARISON OF THE CLASSIFICATION PERFORMANCE WITH AND WITHOUT FEATURE SELECTION AND APPLYING TRAIN-TEST SPLIT ON Z-ALIZADEH SANI DATASET

Z-Alizadeh Sani dataset	class	Precision	Recall	F1-score	Accuracy
No features Selection and Catboost Classifier	0	0.93	0.98	0.95	95.38%
	1	0.98	0.92	0.95	
With feature selection by **MBAR** and **CatBoost** Classifier	0	0.94	0.98	0.96	**96.15%**
	0	0.98	0.94	0.96	

Table VIII shows the Area-Under-the-Curve (AUC) scores of the various classifiers applied on the feature-selected balanced datasets considered in this study. It shows that the Tree-Based models have better AUC scores compared to other models, and CatBoost also outperforms all models.

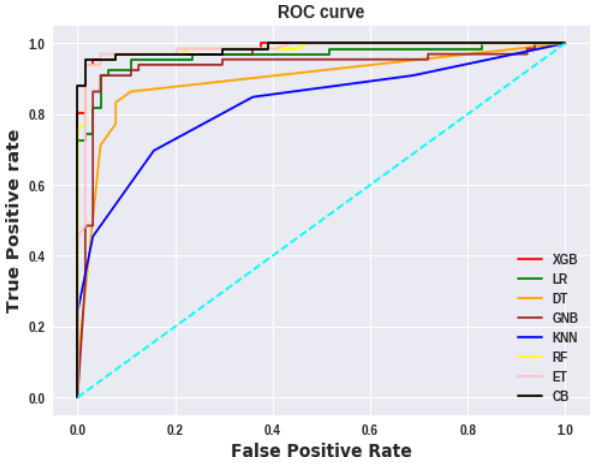

Fig. 8. Receiver operating characteristic curve of classifiers on Z-Alizadeh Sani dataset.

TABLE VIII: COMPARISON OF AUC SCORES OF VARIOUS CLASSIFIERS ON Z-ALIZADEH SANI DATASET AND ARRHYTHMIA DATASET

Classifier	Z-Alizadeh Sani dataset- AUC score	Arrhythmia Dataset- AUC score
XGBoost	0.985	0.859
Logistic Regression	0.965	0.795
Decision Tree	0.898	0.707
Gaussian NB	0.934	0.840
K Nearest Neighbors	0.826	0.788
Random Forest	0.985	0.896
Extra Trees	0.982	0.899
CatBoost	**0.987**	**0.904**

Table IX shows the comparison of the performance of the CatBoost Classifier on the Arrhythmia dataset with and without feature selection. The performance of MBAR with Catboost on the dataset is higher than that without feature selection.

TABLE IX: COMPARISON OF THE CLASSIFICATION PERFORMANCE WITH AND WITHOUT FEATURE SELECTION AND APPLYING TRAIN-TEST SPLIT ON ARRHYTHMIA DATASET

Arrhythmia DS	class	Precisio	Recall	F1-score	Accuracy
No features Selection and Catboost Classifier	0	0.80	0.83	0.81	81.63%
	1	0.84	0.80	0.82	
With feature selection by **MBAR** and **CatBoost** Classifier	0	0.82	0.85	0.83	**83.67%**
	0	0.85	0.83	0.84	

Therefore, both Table VII and Table IX evidence that classification done on feature-selected datasets yields high performance compared to datasets with all features considered.

V. Limitations

ModifiedBoostARoota (MBAR) can be tried on more high-dimensional datasets. Due to time constraints and the non-availability of high-dimensional datasets on heart disease, only two high-dimensional heart datasets were used in this research article.

VI. Conclusion

In this work, feature selection by ModifiedBoostARoota (MBAR) was applied on high dimensional datasets namely, Arrhythmia dataset and Z-Alizadeh Sani dataset. Various classifiers namely XGBoost, Logistic Regression, CatBoost, Decision Tree Classifier, Gaussian Naive Bayes, K Nearest Neighbors, Random Forest, Extra Trees and Support Vector Classifier were used on both the datasets. Their performances by repeated stratified k-fold cross-validation and by 70-30 train-test split were observed on both datasets. The accuracy yielded by classifiers when applied on features selected with MBAR was better than the accuracy obtained without feature selection. Moreover, on balancing both the datasets with SMOTE, the performance of the classifiers increased. Performing stratified 10-fold cross-validation with three repeats on the balanced Arrhythmia dataset with all the above-mentioned classifiers, the CatBoost model outperformed others by yielding an accuracy of 86.33%. Similarly, on the balanced Z-Alizadeh Sani dataset, the accuracy obtained by MBAR with Catboost was 92.76%. The precision, recall, and f1-score of the classifiers were compared and the highest performance was exhibited by CatBoost. The classification done by CatBoost on both the datasets with features selected by MBAR yielded a better performance as compared to datasets with no feature selection.

Thus, by selecting the prominent features and using a strong classifier, a correct prediction of heart diseases can be performed, thereby saving human lives and preventing death.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Anuradha.P had conducted the research, analyzed the performance of the models and wrote the paper.

Dr. Vasantha Kalyani David had guided towards the research work. Both the authors approve the final version.

References

[1] American Heart Association. [Online]. Available: https://www.heart.org/en /health-topics /arrhythmia/about-arrhythmia

[2] E. J. S. Luz, W. R. Schwartz, G. Cámara -Chávez, and D. Menotti, "ECG-based heartbeat classification for arrhythmia detection: A survey," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 144-164, 2016.

[3] National Heart, Lung, and Blood Institute (NHLBI). What Is Coronary Heart Disease? [Online]. Available: https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease

[4] Coronary Artery Disease - Coronary Heart Disease. (2015). [Online]. Available: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/coronary-artery-disease

[5] J. Brownlee. (2016). How machine learning algorithms work (they learn a mapping of input to output). *Machine Learning Algorithms*. [Online]. Available: https://machinelearningmastery.com/how-machine-learning-algorithms-work

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[7] Anuradha. P and V. K. David, "Feature selection using ModifiedBoostARoota and prediction of heart diseases using gradient boosting algorithms," in *Proc. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021, pp. 19-23.

[8] A. Mustaqeem, S. M. Anwar, M. Majid, and A. R. Khan, "Wrapper method for feature selection to classify cardiac arrhythmia," in *Proc. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC*), 2017, pp. 3656-3659.

[9] A. Mustaqeem, S. M. Anwar, and M. Majid, "Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants," *Computational and Mathematical Methods in Medicine*, vol. 1, pp. 1-10, 2018.

[10] S. Khare, A. Bhandari, S. Singh, and A. Arora, "ECG arrhythmia classification using Spearman rank correlation and support vector machine," in *Proc. the International Conference on Soft Computing for Problem Solving (SocProS 2011)*, K. Deep, A. Nagar, M. Pant, J. Bansal, Eds. *Advances in Intelligent and Soft Computing*, Springer, India, vol. 131, 2012.

[11] F. Yang, J. Du, J. Lang, W. Lu, L. Liu, C. Jin, and Q. Kang, "Missing value estimation methods research for arrhythmia classification using the modified kernel difference-weighted KNN algorithms," *BioMed Research International*, vol. 2020, p. 9, 2020.

[12] E. Yılmaz, "An expert system based on fisher score and LS-SVM for cardiac arrhythmia diagnosis," *Computational and Mathematical Methods in Medicine*, vol. 2013, p. 6, 2013.

[13] S. M. Jadhav, S. L. Nalbalwar, and A. A. Ghatol, "Modular neural network-based arrhythmia classification system using ECG signal data," *International Journal of Information Technology and Knowledge Management*, vol. 4, no.1, pp. 205-209, 2011.

[14] M. A. Khan and Y. Kim, "Cardiac arrhythmia disease classification using LSTM deep learning approach," *Computers, Materials & Continua*, vol. 67, no.1, pp. 427–443, 2021.

[15] M. Mitra and R. K. Samanta, "Cardiac arrhythmia classification using neural networks with selected features," *Procedia Technology*, vol. 10, pp. 76-84, 2013.

[16] P. Shimpi, S. Shah, M. Shroff, and A. Godbole, "A machine learning approach for the classification of cardiac arrhythmia," in *Proc. 2017 International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 603-607,2017.

[17] M. Ayar and S. Sabamoniri, "An ECG-based feature selection and heartbeat classification model using a hybrid heuristic algorithm," *Informatics in Medicine Unlocked*, vol. 13, pp. 167-175, 2018.

[18] B. Kolukisa *et al*., "Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease," in *Proc. 2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2232-2238.

[19] B. Kolukisa *et al*., "Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm," *International Journal of Bioscience, Biochemistry, and Bioinformatics*, vol. 10, no. 1, 2020.

[20] A. Gupta, R. Kumar, H. S. Arora *et al*., "C-CADZ: Computational intelligence system for coronary artery disease detection using Z-Alizadeh Sani dataset," *Appl. Intell.*, 2021.

[21] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm," *Comput Methods Programs Biomed*, vol. 141, pp. 19-26, 2017.

[22] R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*, 9th ed. Elsevier Science, 2011.

[23] K. R. Dahal and Y. Gautam, "Argumentative comparative analysis of machine learning on coronary artery disease," *Open Journal of Statistics*, vol. 10, no. 4, pp. 694-705, 2020.

[24] A. Cüvitoğlu and Z. Işik, "Classification of CAD dataset by using principal component analysis and machine learning approaches," in *Proc. 2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)*, 2018, pp. 340-343.

[25] R. Alizadehsani, M. J. Hosseini, Z. A. Sani, A. Ghandeharioun, and R. Boghrati, "Diagnosis of coronary artery disease using cost-sensitive algorithms," in *Proc. 2012 IEEE 12th International Conference on Data Mining Workshops*, 2012, pp. 9-16.

[26] Arrhythmia Dataset. (1998). [Online]. Available: https://archive.ics.uci.edu/ml/datasets/arrhythmia

[27] Z-Alizadeh Sani Dataset. (2017). [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani

Anuradha. P is a research scholar at the Department of Computer Science at Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University, Coimbatore, India. She has done the MCA and is KSET and UGC NET qualified. She is a professor and the head of the Department of Computer Science at Indian Academy Degree College (Autonomous), Bangalore, Karnataka, India. Her areas of interest include object-oriented programming, design and analysis of algorithms, database management system, operating system, systems programming, advanced java programming and data science.

Vasantha Kalyani David is a professor and the head of the Department of Computer Science at Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University, Coimbatore, India. Earlier she is a mathematician with a master of philosophy in mathematics and later did research in computer science.

Dr. Vasantha Kalyani David has published many papers in areas of soft computing. Her research interests, include neural networks, artificial intelligence, fuzzy logic, genetic algorithms, cellular automata, theoretical computer science, and automata theory. She has authored a book on “Pattern Recognition Using Neural and Functional Networks”.

SUPER LEARNER MODEL IN PREDICTION OF HEART ATTACK BASED ON CARDIAC BIOMARKERS

Anuradha. P

Research Scholar, Dept. of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University,
Coimbatore, T.N, India
anujith72@gmail.com

Dr. Vasantha Kalyani David

Professor and HOD, Dept. of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University,
Coimbatore, T.N, India
vasanthadavid@gmail.com

Abstract

Unstable angina and/or a heart attack is caused when restricted flow of blood to the heart occurs due to the narrowed or blocked coronary arteries. On observing Electro cardiogram (ECG), ST segment Elevation Myocardial Infarction (STEMI) can be diagnosed but ECG might not show variation for Non-ST Segment Elevation Myocardial Infarction (NSTEMI). So, cardiac biomarkers could be tested in patients presenting chest pain to confirm whether heart attack or Acute Myocardial Infarction (AMI) is onset or not. Myoglobin, Troponin-I and CK-MB are sensitive biomarkers for diagnosing heart attack/ AMI within specific time frames. In this work, a novel real dataset from a hospital comprising cardiac biomarkers' values of patients was taken and Machine Learning (ML) classifiers namely Support Vector Machine, Logistic Regression (LR), XGBoost (XGB), CatBoost, Random Forest (RF), Decision Tree Classifier, Gaussian Naïve Bayes (GNB), Majority Vote Ensemble Classifier comprising of LR, XGB, GNB, RF were applied on the dataset. Then a Super Learner was designed by taking a novel combination of these classifiers. The comparison of these classifiers resulted in Super Learner outperforming the other ML classifiers. Subsequently, a graphical user interface prediction tool using the Super Learner model was designed which would guide those who have chest pain due to AMI, to undergo emergency medical care and thereby save lives.

Keywords: **Heart attack; Biomarkers; Machine Learning; Super Learner.**

1. Introduction

Coronary Artery Disease (CAD) is a foremost cause of death universally [Hanson *et al*. (2013)]. CAD occurs due to the accumulation of plaque (cholesterol and other deposits) in the coronary arteries which prevents oxygen rich blood from flowing to the heart. CAD includes Acute Myocardial Infarction (AMI), angina pectoris, silent cardiac ischemia and sudden cardiac death [Hanson *et al*., (2013)].

A common indicator of AMI/ Heart Attack is Angina/Chest pain. Angina can be of stable and unstable types. Stable angina may arise due to stress or strenuous activities and would subsides with rest or medication. When the blood clots in the arteries obstruct the blood supply to the heart, Unstable angina occurs. The symptoms occur while doing very little or resting. The pain often radiates to the left shoulder, neck, or arm. It aggravates over a period of a few minutes [Mythili and Malathi, (2015)].

Non-cardiac chest pain (NCCP) may occur due to gastroesophageal reflux disease, depression, anxiety, stress or stomach problems [Schey *et al*., (2007)]. As NCCP symptoms are similar to ischemic heart disease, a thorough check-up by a cardiologist would be required [Schey *et al*., (2007)].

ST segment Elevation Myocardial Infarction (STEMI) is caused by clogging in the coronary arteries. Changes on the Electrocardiogram (ECG) will be noticeable for this type of heart attack. STEMI causes damage to large area of the heart muscle [my.clevelandclinic.org, (2021)]. The levels of cardiac biomarkers in the blood will raise indicating an AMI.

In Non-ST Segment Elevation Myocardial Infarction (NSTEMI), changes may not be noticeable on Electrocardiogram (ECG) but there may be partial or temporary blockage and relatively small damage to the heart muscle compared to STEMI [my.clevelandclinic.org, (2021)]. Damage of the heart muscle raises the levels of the chemical markers in the blood with the help of which NSTEMI can be identified [my.clevelandclinic.org, (2021)].

Inappropriate diagnosis of patients with chest pain often leads to unfortunate admission of patients without AMI [Mythili and Malathi, (2015)]. This can be avoided by testing the Cardiac biomarkers' levels to confirm or rule out an AMI. Biomarkers have helped clinicians in rapid diagnosis of heart attack and saving the life of patients thereby reducing the mortality rate [Mythili and Malathi, (2015)].

1.1 *Cardiac Biomarkers*

1.1.1 Troponin-I

Cardiac troponin-I (cTnI) is a preferred biomarker for early diagnosis of AMI [Anderson *et al.,* (2007)]. cTnI is abnormal or may not rise for the first 4 to 8 hours after the onset of chest pain. It peaks at 12 to 16 hours so the test has to be repeated after 8hrs of the onset of chest pain. Also, cTnI remains high for 5 to 9 days [Larue *et al.,* (1993)] [Dolci and Panteghini, (2006)] [Jaffe *et al.,* (1996)]. Troponin-I (cTnI), when used together with other clinical information, provides a valuable diagnostic test for AMI, thereby assisting in informed clinical decision making [9].

1.1.2 Creatine Kinase- Myocardial Band

Creatine Kinase Myocardial Band (CK-MB), a cytosolic carrier protein for high-energy phosphates, is less sensitive and less specific for AMI than cardiac troponins. Healthy persons have low levels of CK-MB in the blood and if damage occurs to skeletal or cardiac muscle, CK-MB levels would elevate [Larue *et al.,* (1993)] [Dolci and Panteghini, (2006)] [Tsung, (1981)] [Surya *et al.*, (1999)].

1.1.3 Myoglobin

Myoglobin rises within 1 hour of the onset of infarction. As it also rises up in other conditions such as skeletal muscle injury and in case of renal impairment, its specificity for AMI is low [Hanson *et al*. (2013)] [Anderson *et al.,* (2007)] [Surya *et al*., (1999)] [Wong *et al.*, (2004)] [Kruger *et al.,* (2002)][Kontos *et al.*, (2007)] [Hamm and Katus, (1995)]. However, myoglobin can be used for diagnosis and risk stratification in the early hours after symptom onset [Dolci and Panteghini, (2006)] [Hamm and Katus, (1995)] [Apple, (1992)] [Ruzich, (1992)] [Aldous, (2013)].

1.1.4 Brain Natriuretic Peptide (BNP)

BNP testing is used to diagnose or rule out heart failure, including diastolic heart failure [Levin *et al*., (1998)]. BNP levels are greater than 100 pg/ml when there is cardiac damage [Cowie *et al.,* (2003)] [Valli *et al*. (2001)]. Heart failure is highly unlikely for values of Brain Natriuretic Peptide (BNP) less than 100 pg/ml [20]. BNP provides prognostic information following an MI [Levin *et al*., (1998)], Cowie *et al.,* (2003)] [de Lemos *et al.,* (2001)] [Arakawa *et al.,* (1996)]. Increased levels of BNP concentrations may also be due to cardiac ischemia [Cowie *et al.,* (2003)].

1.1.5 D-Dimer

D-Dimer and fibrinogen levels were found to be higher in patients with acute ischemic events (myocardial infarction and unstable angina) than non-ischemic patients [Bayes-Genis, A *et al.,* (2000)]. D-Dimer level > 500 microg/L had a self-determining diagnostic value for myocardial infarction [Bayes-Genis, A *et al.,* (2000)].

With reference to specific time frames, the biomarkers CK-MB, Troponin-I, and Myoglobin yield satisfactory diagnostic sensitivity [Chiu *et al.,* (2000)]. D-Dimer level > 500 microg/L had a self-determining diagnostic value for myocardial infarction [Bayes-Genis, A *et al.,* (2000)]. On combining these biomarkers, clinicians get valuable information to provide proper treatment to AMI patients [Chiu *et al.,* (2000)].
For a patient diagnosed with AMI, medication is given to reduce pain, prevent blood clots and to help blood flow through the coronary arteries. Oxygen may be given to make sure that the heart, lungs, organs and tissues stay healthy. The damage to the heart can be considerably reduced or prevented, if the treatment for heart attack is given as soon as Myocardial Infarction (MI) symptoms are first experienced.
In this work,

- A novel real-world dataset from Specialist hospital in Bengaluru, India, comprising of cardiac biomarkers was used.
- Machine Learning (ML) models, both base models and ensemble models were used to estimate the probability of Myocardial Infarction (MI) for an individual patient.
- A Super Learner consisting of novel combination of classifiers was also used in MI prediction.
- The performance of all the classifiers was evaluated on the real-world dataset comprising the biomarkers, Cleveland heart dataset from UCI ML repository and Cardiovascular Disease dataset from Mendeley data.

- A novel GUI prediction tool with the Super Learner as predictor was designed. This tool on providing the biomarker values, helps patients to know whether there is an onset of MI and take emergency treatment and thereby prevent deaths. It could be used as a support tool in making clinical decisions in patients with suspected MI.

The remaining part of this paper is structured as follows: Section 2 introduces the Machine Learning algorithms used in this work, Section 3 briefs the related work, Section 4 portrays the methodology adopted , Section 5 describes the dataset used, Section 6 discusses the results, Section 7 contains limitations and section 8 concludes the paper.

2. Machine Learning Algorithms

2.1. *Support Vector Machine (SVM)*

SVMs developed by Vladimir Vapnik, is a machine learning technique, used for both linear and non-linear classification. In the SVM algorithm, each data item is plotted as a point in n-dimensional space, where n is a number of features [Uddin *et al.,* (2019)]. By transforming low dimensional input space to a higher dimensional space, SVM kernel converts non-separable problem to separable problem. Then, the hyper-plane that separates the two classes is identified and classification is done.

2.2. *Logistic Regression*

Logistic Regression is used in classification of binary or multiple classes by employing the cost function called as sigmoid function,

$$f(x) = 1/(1 + e^{-x})). \quad (1)$$

It is a predictive analysis algorithm based on the concept of probability.

2.3. *Naive Bayes classifier*

Naive Bayes classifier, based on Bayes' theorem, is a probabilistic classifier. It assumes that each predictor variable/feature contributes independently and equally to the class/outcome/target variable [Uddin *et al.,* (2019)].

2.4. *Random Forest*

Random forest (RF) is a tree-based ensemble classifier, which consists of a group of decision trees. Various sub-samples of the dataset are given as input to the decision trees. When a new sample has to be classified, each DT gives a classification outcome by considering different parts of that input vector and the final output of RF is made based on majority voting or average of all the trees in the forest [Uddin *et al.,* (2019)][Breiman L. (2001)].

2.5. *XGBoost (Extreme Gradient Boosting)*

In boosting, a strong classifier is built from several weak classifiers. Extreme Gradient Boosting, in short XGBoost, developed by Tianqi Chen and Guestrin, is an implementation of gradient boosted decision trees which gives improved performance [Tianqi Chen and Carlos Guestrin, (2016)] [Jason Brownlee, (2016)]. In XGBoost, Objective function is the sum of Training Loss and Regularization, where Training Loss is the differentiable convex loss function which computes the difference between the target yi and the prediction ŷi. The regularization term is added in order to smooth the final learnt weights and thereby prevent over-fitting [Tianqi Chen and Carlos Guestrin, (2016)].

2.6. *CatBoost*

Yandex researchers and engineers developed CatBoost, a gradient boosting algorithm, which has both CPU and GPU implementations use binary decision trees as base predictors. As it does efficient vector representation of categorical data, CatBoost shows high performance when applied on categorical data [Liudmila Prokhorenkova *et al.,* (2018)] [Alexander Marz, (2020)]. With the default values of the hyper parameters, i.e., when hyperparameters are not tuned, Catboost classifier performs better than XGBoost.

2.7. Majority Vote Ensemble

Majority vote ensemble model combines the predictions of several other models considered in the ensemble. The predictions of each model in the ensemble for the respective class label is counted and the final prediction of the ensemble is the class label with the maximum votes.

3. Related Work

Lan Shou et al., determined that resting rate pressure product (rRPP), a measure of cardiac workload, can be predicted using combinations of biomarkers in the blood. XGBoost model outperformed other ML models to predict rRPP [Lan Shou et al., (2021)].

Asan Agibetov et al., used the XGBoost model to diagnose Cardiac amyloidosis among patients having symptoms of Heart Failure based on laboratory parameters [Agibetov *et al.,* (2020)].
In order to detect undiagnosed Heart failure with reduced ejection fraction (HFrEF), Mathis et al., used random forest, XGBoost and L1 regularized logistic regression models based on perioperative data consisting of 628 preoperative and 1,195 intraoperative features. XGBoost was found to yield AUROC value better than other models. However, as the model has low positive predictive results because of the disease being less prevalent, the authors are of the opinion that the model prediction needs to be followed by confirmatory testing with echocardiography or cardiac biomarkers [Mathis et al., (2020)].
Weng et al., applied gradient boosting machines, neural networks, random forest and logistic regression, on routine medical data of patients from UK family practices. They found that neural networks performed better in identifying patients who could benefit from preventive treatment [Weng *et al.,* (2017)].
Meeshanthini V dogan et al., used machine learning techniques on datasets from Intermountain Healthcare (IM) and Framingham Heart Study (FHS) in order to propose an integrated genetic-epigenetic model for predicting 3-year incident CHD. They showed that their proposed model performed better in identifying patients at high risk for a heart attack, compared to Framingham Risk Score (FRS) and atherosclerotic cardiovascular disease Pooled Cohort Equation (PCE) [Meeshanthini *et al.*, (2021)].
C. Beulah Christalin Latha et al., had experimented on Multilayer Perceptron, SVM, PART, Bayes Net, C4.5 and Naive Bayes as individual classifiers and also used a combination of these classifiers in ensemble techniques namely stacking, boosting, majority voting and bagging. The dataset used was Cleveland heart dataset taken from the UCI machine learning repository. On comparing the accuracy yielded by all the techniques, it was observed that the Majority voting classifier yielded an improved accuracy of 7.26% [Latha *et al.,* (2019)].
Martin.P. Than et al., proposed myocardial-ischemic-injury-index, a machine learning model, which used gradient boosting on dataset which included paired high-sensitivity cardiac troponin-I concentrations of patients along with traditional factors. The proposed model was designed to diagnose type 1 myocardial infarction [Than *et al.,* (2019)].
Alaa AM et al., used Auto Prognosis, an automated ML tool for CVD risk prediction on the UK Biobank population. On comparing the tool with traditional techniques like Framingham score and Cox PH model, the automated tool resulted in yielding improved accuracy [Alaa AM et al., (2019)].
Van der Laan et al., had proposed a Super Learner algorithm in their work in [Van der Laan et al., (2007)], where selected classifiers were modelled on v-fold cross validation and the predicted value for each validation set was stored as meta-x values and the actual target values as meta-y. Then the classifiers were modelled on the whole dataset. The target values predicted by the selected classifiers when applied on the test data was saved as meta-x values. The meta learner or Super Learner then predicted the y-values for these meta-x values and the performance was evaluated [Van der Laan et al., (2007)].

4. Methodology

The methodology adopted in this work is depicted in the flowchart given in Figure 1.

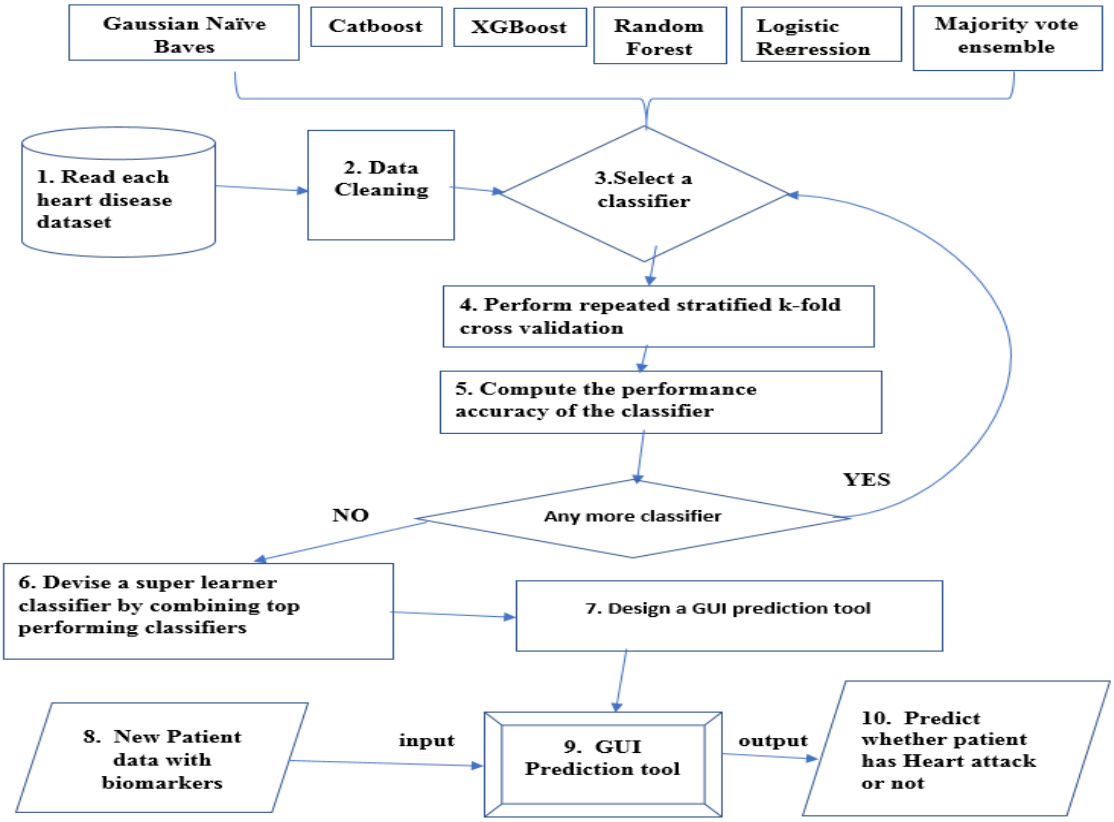

Fig. 1. Methodology

5. Dataset

On reviewing related work, it is clear that biomarkers play a vital role in predicting heart diseases. This work is aimed at developing ML model to predict heart attack/MI using a dataset containing cardiac biomarker values which would be useful to predict or rule out MI in patients presenting chest pain, especially in cases of NSTEMI. The first dataset used in this work is the data of patients who presented chest pain and had undergone lab tests, at Specialist hospital, Bangalore, in order to confirm whether they had an MI or not. The personal details of the patients were not disclosed. The data collected consists of Age, Gender, CKMB, Myoglobin, Troponin-I, BNP, D-Dimer, ACS_types, Disease, shown in Table 1, consists of 192 instances.

The normal and abnormal range for each of these biomarkers are as follows: Normal range of CKMB is 0.0 - 4.3 ng/ML and abnormal CKMB is a value >4.3 ng/ML. Normal range of Troponin-I is 0.0 - 0.4 ng/ML and abnormal Troponin-I is a value >0.4 ng/ML. Normal range of BNP is 0.0 - 100 ng/ML and abnormal BNP is a value>100 ng/ML. Normal range of D-Dimer is 0.0 - 400 ng/ML, abnormal D-Dimer is a value>400 ng/ML. Normal range of myoglobin 20-80 ng/ML and abnormal myoglobin is a value>80 ng/ML.

Two other datasets which were also used for evaluating the performance of the Super Learner are: (i) Cardiovascular disease dataset taken from Mendeley data [Doppala *et al.,* (2021)]. Table 2 exhibits the features of this dataset which has 1000 instances. (ii) Cleveland heart disease dataset at UCI Machine Learning Repository, created by Robert Detrano, M.D., Ph.D., V.A. Medical Center, Long Beach and Cleveland Clinic. This dataset has 303 instances. Table 3 displays the features of Cleveland heart dataset.

Features
1. Age
2. Gender: 0-female, 1-Male
3. CKMB- Creatine Kinase Myocardial Band
4. Myoglobin
5. Troponin-i
6. BNP- Brain Natriuretic Peptide
7. D-Dimer
8. ACS_types- heart disease types
9. Target: Disease: 0- no AMI, 1- AMI, 2- heart problems

Table 1. Cardiac biomarkers-Heart Dataset

Features
1. Age
2. Gender
3. Chest pain
4. Resting BP
5. Serum cholestrol
6. Fasting blood sugar
7. Resting relectro
8. Max heart rate
9. Exercise angia
10. Old peak
11. Slope
12. No of major vessels
13. Target

Table 2. Cardiovascular Disease Dataset (Mendeley Data)

Features
1. Age
2. Sex
3. Fbs-fasting blood sugar
4. Cp-chest pain type
5. Chol-serum cholesterol
6.
7. Trestbps- resting blood pressure
8. Restecg -ecg at rest
9. Thalach-maximum heart rate
10. Exang-exercise induced angina
11. Ca-number of major vessels colored
12. Slope- slope of the peak exercise st segment
13. Thal -defect type
14. Oldpeak-st depression induced by exercise
15. Target: num

Table 3. Cleveland Heart Dataset (UCI)

6. Results and Discussion

On exploring the cardiac biomarkers dataset using the WEKA tool, we observe in Figure 2 that the dataset consists of 81 females and 111 males. From figure 3, we can visualize that the age of the patients lie between 21 and 90, and more data is concentrated around the median, 55 years of age. Also, both male and female patients are found in all age groups. More people have undergone lab tests for MI between the age groups of 45 to 75.

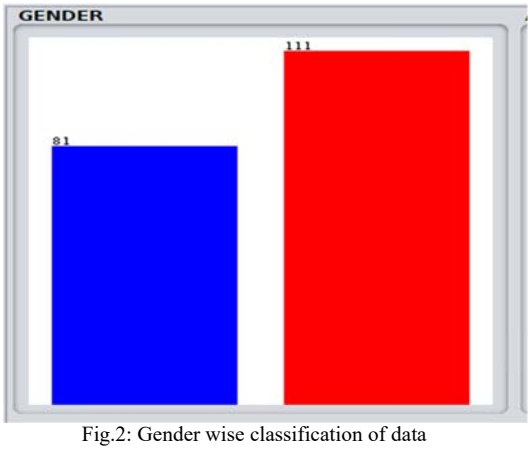

Fig.2: Gender wise classification of data

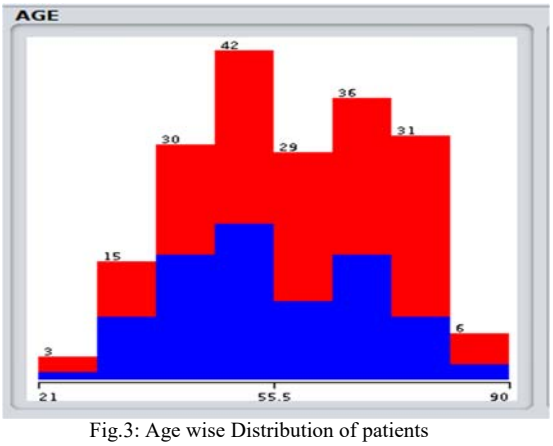

Fig.3: Age wise Distribution of patients

Figure 4 shows the count of patients with no MI(Myocardial Infarction) as 61, patients diagnosed with MI and Heart failure and in critical condition as 3, number of patients who had MI with heart muscle damage is 51, only cardiac muscle damage is 42, early MI was diagnosed in 18 patients, 7 patients had MI diagnosed early and also after 8 hrs, 1 patient was diagnosed of MI after 8 hrs, muscle/skeletal damage for 5 patients, higher level blood clots was diagnosed in 4 patients. Summing up the above data we get in figure 5, that the number of patients with no MI are 61 and belong to class 0, patients who had MI are 80 and classified as disease class 1, and those with cardiac muscle damage and clots are 51 in number and belong to class 2. The disease classes 0 and 2 have male and female patients of almost the same ratio but class 1 consisting of MI patients have more males compared to females. In the gender field, males were coded as 0 and females as 1.

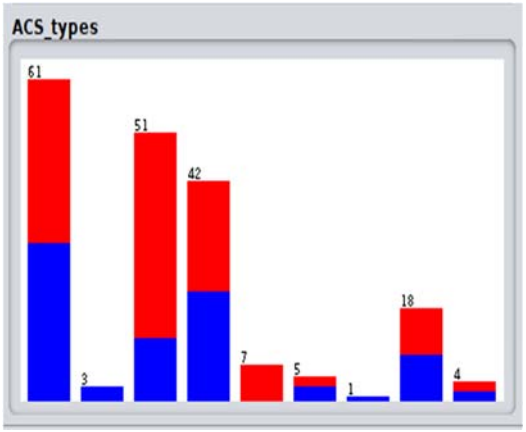

Fig.4: Heart disease types based on biomarkers' levels

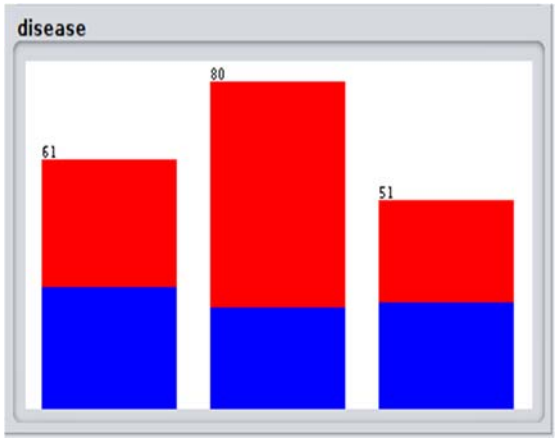

Fig.5: Heart disease types

The training and evaluation of the classifiers were done using Python on Jupyter notebook. The classifiers used were Support Vector Machine (SVM), Logistic Regression (LR), XGBoost (XGB), CatBoost (CB), Random Forest (RF), Decision Tree Classifier (DT), Gaussian Naïve Bayes (GNB), Majority Vote Ensemble Classifier comprising of LR, XGB, GNB, RF. The details of all these classifiers are found in section II. The Repeated Stratified K-Fold method was used to train and test the classifiers. The Stratified K-Fold Cross validator is repeated 3 times with different randomisation in each repetition. In this work, the Stratified K-Fold divides the entire data set into 10 splits. For each iteration it approximately maintains the same percentage of samples of each target class. In cross validation, for each iteration, 9 folds are considered as the training set and 1-fold is taken as the validation set. The number of iterations is 10, which is equal to the number of folds. The mean accuracy across all folds and all repeats is calculated for every classifier and a comparison of their performance is made. The training and testing process are repeated on Cleveland Heart dataset and Cardiovascular disease dataset. Table 4 shows the accuracy obtained by each classifier on performing repeated stratified K-fold cross validation on the three selected datasets.

Classifiers	**Accuracy in the dataset**		
	Cardiac biomarkers DS	**Cleveland Heart DS**	**Cardiovascular disease DS**
CatBoost (CB)	96.5%	82.5%	98.37%
XGBoost (XGB)	96.3%	80.9%	98.0%
Random Forest (RF)	95.7%	82.6%	97.8%
Decision Tree Classifer (DT)	95.7%	74.5%	96.3%
Majority Voting Classifier (VC)	92.7%	84.3%	96.1%
Support Vector Machine (SVM)	91.5%	82.9%	96.5%
Logistic Regression (LR)	86.1%	83.7%	95%
Gaussian Naive Bayes (GNB)	78.1%	83.5%	94.6%
*** Super Learner (DT, CB)**	**97.9%**	**88.53%**	**98.8%**
*Super Learner combining DT and CB classifiers and LR as meta learner			

Table 4. Performance Accuracy of the Classifiers

Then Super Learner technique proposed by Van der Laan et al., described in section 3, was tried with all combinations of the classifiers mentioned in Table 4 and Logistic regression as meta learner. The Super Learner with a combination of Decision Tree and CatBoost classifier with Logistic Regression as meta learner gave an accuracy of 97.9% on cardiac biomarkers real world dataset, 88.53% accuracy on Cleveland dataset and 98.8% accuracy on Cardiovascular disease dataset. Table 5 displays the precision, recall and fi-score and Accuracy of the Super Learner model on the three heart disease datasets.

Results show that the Super learner with DT and CB combination and LR as meta learner has consistently performed better than other classifiers in all the three datasets. The graphical form of the comparison of classifiers on the three heart disease datasets is shown in fig.6.

Datasets	class	precision	recall	f1-score	Accuracy
Cleveland Heart DS	No Disease	0.85	0.97	0.91	88.53%
	Has Disease	0.95	0.77	0.85	
Cardiac Biomarkers DS	No Heart Attack	1.00	1.00	1.00	97.9%
	Has Heart Attack	0.95	1.00	0.97	
	Has Disease	1.00	0.94	0.97	
Cardiovascular disease DS	No Disease	0.98	0.99	0.99	99%
	Has Disease	0.99	0.99	0.99	

Table 5. Classification Matrix of the Performance of the Super Learner on Heart disease datasets

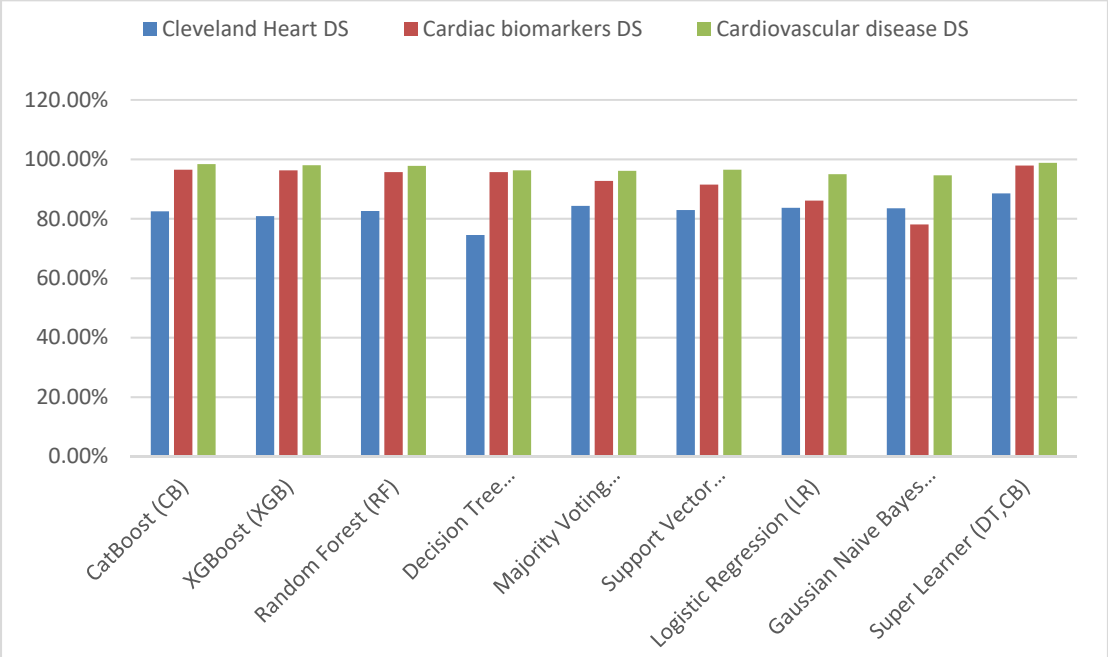

Fig.6: Comparison of Performance of the Classifiers on heart disease datasets

Using the Super Learner classifier, a GUI was created in python, using which, one can check whether he/she had chest pain due to heart attack or not, by entering the biomarkers' values. Fig.7 shows images of the prediction tool. As often people have lost lives by ignoring the chest pain, this tool can be used to check whether emergency medical care is required or not and thereby save lives. This tool can be useful especially in rural areas, where cardiac specialists are not available in the neighbourhood and the patient can rush for emergency treatment based on the prediction given in the tool. It is only a support tool which guides patients to approach the medical practitioners, who in turn would diagnose the illness and do the treatment required.

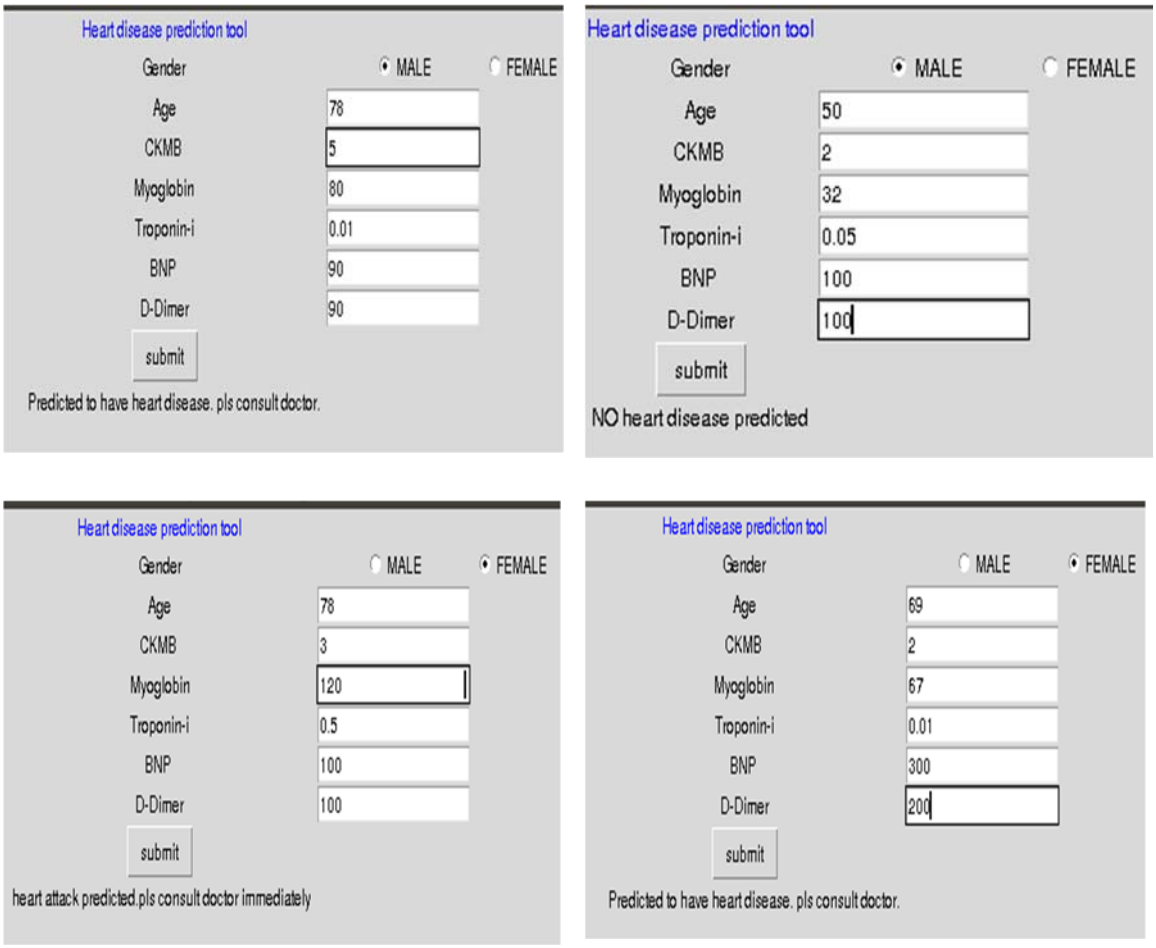

Fig. 7. Images of the GUI heart attack prediction tool

7. Limitations

The correlation of biomarkers with other parameters like family history, cholesterol, fasting blood sugar, blood pressure was not analysed in this work. In future work it can done inorder to predict heart diseases much earlier than the onset.

8. Conclusion

A common indicator of Acute Myocardial Infarction (AMI)/ Heart Attack is Angina/Chest pain. Angina can be of stable and unstable types. Stable angina may arise due to stress or strenuous activities and would subside with rest or medication. When the blood clots in the arteries obstruct the blood supply to the heart, unstable angina occurs. Troponin-I, myoglobin and CK-MB are sensitive biomarkers for diagnosing heart attack/ AMI within specific time frames. A patient presenting chest pain when checked for the above-mentioned biomarkers' levels will be given an emergency treatment if the tests confirm a heart attack. In this work, a novel real world dataset comprising cardiac biomarker values is used and Machine Learning models namely, Support Vector Machine (SVM), Logistic Regression (LR), XGBoost (XGB), CatBoost (CB), Random Forest (RF), Decision Tree Classifier (DT), Gaussian Naïve Bayes (GNB), Majority Vote Ensemble Classifier comprising of LR, XGB, GNB, RF were applied on the dataset by performing repeated stratified cross validation. Then a Super Learner was designed by taking combinations of these classifiers. The Super Learner with Decision Tree and CatBoost classifiers combination with Logistic Regression as Meta Learner gave 97.9% accuracy. Then the same set of classifiers and Super Learner model was applied on Cleveland Heart dataset and Cardiovascular disease dataset. The Super Learner with the novel combination outperformed other classifiers in all the three datasets. As many people by being reluctant to consult medical practitioners, had ignored the chest pain and have lost lives, a graphical user interface prediction tool was designed with the Super Learner classifier for predicting heart attack. The GUI prediction tool on entering the biomarkers' values would predict whether emergency medical attention is required or not. If patients take emergency treatment in case, they are predicted to have a heart attack, their lives can be saved.

References

[1] Hanson, MA*, et al*. (2013): Coronary artery disease. *Primary Care*, 40(1),1-16.
[2] Sabesan Mythili; Narasimhan Malathi. (2015): Diagnostic markers of acute myocardial infarction (Review). BIOMEDICAL REPORTS, 3: 743-748.
[3] Schey, R; Villarreal, A; Fass, R. (2007): Non Cardiac chest pain: current treatment. *Gastroenterology & hepatology*, *3*(4), 255–262.
[4] CAD: Acute Coronary Syndrome. Retrieved on September 20, 2021, from https://my.clevelandclinic.org/health/diseases/16713-cad-acute-coronary-syndrome
[5] Anderson, J.L, *et al.* (2007): ACC/AHA 2007 guidelines for the management of patients with unstable angina/non-ST-elevation myocardial infarction. *J Am Coll Cardiol.,* 250, 1-157.
[6] Larue, C, *et al.* (1993): Card iac specific immunoenzymometric assay of troponin I in the early phase of acute myocardial infarction, *Clinical Chemistry*, 39(6), 972–979.
[7] Dolci, A; Panteghini, M. (2006): The exciting story of cardiac biomarkers: From retrospective detection to gold diagnostic standard for acute myocardial infarction and more. *Clinica Chimica Acta,* 369, 179-187.
[8] Jaffe, AS, *et al.* (1996): Comparative sensitivity of cardiac troponin I and lactate dehydrogenase isoenzymes for diagnosing acute myocardial infarction. *Clin* Chem, 42(11), 1770-6.
[9] Collinson, P. O; Stubbs, P.J. (2003): Are troponins confusing?. *Heart (British Cardiac Society)*, *89*(11), 1285–1287.
[10] Tsung, SH. (1981): Several conditions causing elevation of serum CK-MB and CK-BB. *Am J Clin Pathol,* 75:711–5.
[11] Surya, P Rao, *et al.* (1999): Cardiac troponin I and cardiac enzymes after electrophysiologic studies, ablations, and defibrillator implantations, *The American Journal of Cardiology*, 84(4), 470.
[12] Wong, WM, *et al.* (2004): Population based study of noncardiac chest pain in southern Chinese: prevalence, psychosocial factors and health care utilization. *World J Gastroenterol*., 10, 707–712.
[13] Kruger, S; Graf, J, *et al.* (2002): Brain natriuretic peptide levels predict functional capacity in patients with chronic heart failure. *J Am Coll Cardiol*, 40, 718–22.
[14] Kontos, MC, *et al.* (2007): Ability of myoglobin to predict mortality in patients admitted for exclusion of myocardial infarction. *Am J Emerg Med*, 25, 873–9.
[15] Hamm, CW; Katus, HA. (1995): New biochemical markers for myocardial cell injury. *Curr Opin Cardiol*.10, 355–60.
[16] Apple, FS. (1992): Acute myocardial infarction and coronary reperfusion. Serum cardiac markers for the 1990s. *Am J Clin Pathol,* 97, 217–26.
[17] Ruzich, RS. (1992): Cardiac enzymes. How to use serial determinations to confirm acute myocardial infarction. *Postgrad Med*, 85–9.
[18] Aldous SJ. (2013): Cardiac biomarkers in acute myocardial infarction. *Int J Cardiol.*, 164(3), 282-94.
[19] Levin, ER; Gardner, DG; Samson, WK. (1998): Natriuretic peptides. *N Engl J Med*., 339: 321-328.
[20] Cowie, M.R, *et al.* (2003). Clinical applications of B-type natriuretic peptide (BNP) testing, *European Heart Journal*. 24, 1710-1718.
[21] Valli, N, *et al*. (2001): Assessment of brain natriuretic peptide in patients with suspected heart failure: comparison with radionuclide ventriculography data. *Clin Chim Acta,* 306, 19–26.
[22] de Lemos, JA, *et al.* (2001): The prognostic value of B-type natriuretic peptide in patients with acute coronary syndromes. *N Engl J Med.*, 345: 1014-1021.
[23] Arakawa, N; Nakamura, M; Aoki, H; Hiramori, K. (1996): Plasma brain natriuretic peptide concentrations predict survival after acute myocardial infarction. *J Am Coll Cardiol*., 27: 1656-1661.
[24] Bayes-Genis, A *et al.* (2000): D-Dimer is an early diagnostic marker of coronary ischemia in patients with chest pain. *Am Heart J*, 140(3), 379-84.
[25] Chiu, A; Chan, W.K; Cheng, S.H; Leung, C.K; Choi, C.H. (1999): Troponin-I, myoglobin, and mass concentration of creatine kinase-MB in acute myocardial infarction. *Q J Med*. 92, 711–718.
[26] Uddin, S, *et al.* (2019): Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak,* **19,** 281.
[27] Breiman L. (2001): Random forests. Mach Learn, 45(1):5–32.
[28] Tianqi Chen; Carlos Guestrin. (2016): "XGBoost: A Scalable Tree Boosting System," KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.
[29] Jason Brownlee. (2016). XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn. Machine Learning Mastery.
[30] Liudmila Prokhorenkova; Gleb Gusev; Aleksandr Vorobev; Anna Veronika Dorogush; Andrey Gulin. (2018): CatBoost: unbiased boosting with categorical features. *32nd Conference on Neural Information Processing Systems.* Montreal, Canada.
[31] Alexander Marz. (2020): Catboost Lss an Extension of Catboost to Probabilistic Forecasting. *arXiv*:2001.02121.
[32] Lan Shou, *et al.* (2021): Blood Biomarkers Predict Cardiac Workload Using Machine Learning. *BioMed Research International*, 6172815.
[33] Agibetov, A, *et al.* (2020): Machine Learning Enables Prediction of Cardiac Amyloidosis by Routine Laboratory Parameters: A Proof-of-Concept Study, *Journal of Clinical Medicine,* MDPI.
[34] Mathis, M.R, *et al.* (2020): Early Detection of Heart Failure With Reduced Ejection Fraction Using Perioperative Data Among Noncardiac Surgical Patients: A Machine-Learning Approach. *Anesthesia and analgesia*, *130*(5), 1188–1200.
[35] Weng, SF, *et al.* (2017): Can machine-learning improve cardiovascular risk prediction using routine clinical data. *PLoS ONE,* 12(4), e0174944.
[36] Meeshanthini, V Dogan, *et al.* (2021): External validation of integrated genetic-epigenetic biomarkers for predicting incident coronary heart disease. *Epigenomics, Future Medicine*, 13(14).
[37] Latha CBC; Jeeva SC. (2019): Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*,16,100203.
[38] Than, M.P, *et al.* (2019). Machine Learning to Predict the Likelihood of Acute Myocardial Infarction. *Circulation*, 140(11), 899–909.
[39] Alaa, AM, *et al.* (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS ONE*,14(5), e0213653.
[40] Van der Laan, Mark J; Polley ,Eric C; Hubbard Alan E. (2007): Statistical Applications in Genetics and Molecular Biology. *The Berkeley Electronic Press*, 6(1).
[41] Doppala, Bhanu Prakash; Bhattacharyya, Debnath. (2021). Cardiovascular_Disease_Dataset, Mendeley Data, V1, doi: 10.17632/dzz48mvjht.1.

Authors Profile

	Ms. Anuradha P is a Research Scholar at Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University, Coimbatore, India. She has done MCA and is KSET and UGC NET qualified. She is Associate Professor and Head of the Department of Computer Science at Indian Academy Degree College (Autonomous), Bangalore, Karnataka, India. Her areas of interest include Object Oriented Programming, Analysis and design of algorithms, Database Management System, Operating System, Systems Programming, Advanced Java Programming and Data Science.
	Dr. Vasantha Kalyani David is Professor and Head in the Department of Computer Science at Avinashilingam Institute for Home Science and Higher Education for Women, deemed to be University, Coimbatore, India. Earlier a Mathematician with a Master of Philosophy in Mathematics and later did research in Computer Science. Dr. Vasantha Kalyani David has published many papers in areas of Soft Computing. Her research interests, include Neural Networks, Artificial Intelligence, Fuzzy Logic, Genetic Algorithms, Cellular Automata, Theoretical Computer Science, and Automata Theory. She has authored a book on “Pattern Recognition Using Neural and Functional Networks”.

Feature selection using ModifiedBoostARoota and prediction of heart diseases using Gradient Boosting algorithms

Anuradha.P
Research Scholar, Dept. of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Deemed to be University,
Coimbatore, T.N, India
anujith72@gmail.com

Dr.Vasantha Kalyani David
HOD and Professor, Dept. of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Deemed to be University,
Coimbatore, T.N, India

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397154

***Abstract*—Techniques in Machine Learning can be employed to detect cardiovascular disease at early stages thereby preventing deaths caused by the disease. Identifying the features that significantly participate in the target variable prediction using feature selection methods would help in achieving better accuracy and reducing the computational cost of a model. This paper portrays a new algorithm, ModifiedBoostARoota, developed similar to BoostARoota, differing in the feature elimination process. Also, by choosing XGBoost and catboost as base models in both BoostARoota and ModifiedBoostARoota, a comparison of both the algorithms' performances are done. ModifiedBoostARoota algorithm has faster performance compared to BoostARoota, when catboost is chosen as the base model. Also, the XGBoost and CatBoost classifiers modelled on features selected by ModifiedBoostARoota gave better accuracy than that of BoostARoota.**

Keywords—Feature Selection, BoostARoota, Modified BoostARoota, prediction, heart disease

I. INTRODUCTION

As heart disease is the number one killer disease in the world, early detection using techniques in Machine learning would help save lives. Feature selection (FS) methods identify the features that largely participate in the target variable prediction. FS techniques when applied on high dimensional datasets would be able to eliminate unimportant features without incurring much loss of data. D. C. Duro et al., used the Boruta FS algorithm on the full multi-source, multi-sensor data set consisting of 418 variables and 190 features were eliminated [1]. Na‟eem Hoosen Agjee et al., in [2], worked on multitemporal monitoring of infestation levels on water hyacinth plants and found that Random Forest coupled with the Boruta algorithm used for band-selection had lesser classification errors compared to the other methods [2]. Caraka RE et al., in [3], used the Boruta FS algorithm for feature selection. S.S. Kumar et al., used the Boruta FS algorithm with Random Forest classifier that yielded 98.8% accuracy [4]. Sanchez et al., formulated Incremental AFN-Feature Selection method (IAFN-FS) where the AFN algorithm carries out the sensitivity analysis over the functional constituents of the approximate function [5]. L. Ma et al., in [6], does first feature importance ranking by gain and then object-based image classification is carried out with a Support Vector Machine Classifier, by evaluating the subset of features using a tenfold CV which is polygon-based. [6].

Gradient Boosting builds an ensemble of tree-based models by training each of the trees in the ensemble on different labels and then combines the trees [7]. Boosting builds strong predictors by combining weaker models (base predictors) in a greedy manner [8]. XGBoost (eXtreme Gradient Boosting) is a gradient boosting algorithm where over-fitting is avoided by adding a regularization term to the loss function in the computation of the objective function [9][10]. Catboost, a gradient boosting algorithm, implements ordered boosting and is a well-suited algorithm for vector representation of categorical data [8][11]. Catboost outperforms other algorithms when applied on datasets consisting of many categorical features [8] [11]. A. Ogunleye and Q. Wang in [12], chose 12 most important features from each of the three FS techniques namely, RFE, Extra Tree Classifier and Univariate Selection. Then, features that were found in at least two of the above sets were selected [12]. The XGBoost model when applied on the reduced feature set gave 97.58% accuracy. In [13], CatBoost algorithm is used in predicting the production rate based on fracturing design parameters with an accuracy of 81%. N.S.Rajliwall et al., performed a comparative study of 7 classifiers namely, XGBoost, Nearest Neighbors (KNN), Random Forest, SVM, Naive Bayes, Logistic Regression, Ensemble (RPART), on NHANES dataset and Framingham Heart Study CHS dataset and concluded that XGBoost has faster performance compared to other classifiers [14].

BoostARoota, a wrapper-feature selection algorithm, developed by Chasedehan is modified version of Boruta algorithm. It uses Xgboost as the base model instead of RandomForest used in Boruta [15]. In BoostARoota, the

feature elimination process is modified and is computationally faster than Boruta. M. Zabihi et al., in their work on sepsis prediction, used BoostARoota to select five different sets of features. Then, an ensemble model consisting of five XGboost classifiers was used and the geometric mean of the outputs of the five classifiers was computed as the final output of the classifier [16].

This paper is organized as follows: Section II discusses the proposed ModifiedBoostARoota Algorithm. Section III discusses the methodology, Section IV describes the datasets used, section V discusses the results and compares the performance of modified BoostARoota with the original BoostARoota algorithm, Section VI concludes the paper.

II. PROPOSED MODIFIEDBOOSTAROOTA ALGORITHM

In this paper, the feature selection algorithm BoostARoota (BAR), developed by chasedehan (published in Python package Index (PyPI)), is modified and named as ModifiedBoostARoota (MBAR). In ModifiedBoostARoota, the feature elimination procedure is modified as compared to that of BoostARoota.

Algorithm ModifiedBoostARoota
{
1. Compute shadow feature (by shuffling original features at random) for each feature in the dataset and merge the shadow features with the dataset to form an extended dataset of „n" features.
2. Using any Tree based models, compute the Feature Importance (FI) of all features in the extended dataset.
3. Assign rank, r_i for all features $i = 1\ to\ n$.
4. If FI of original feature < FI of corresponding shadow feature then eliminate that original feature and its shadow feature.
5. If FI of any feature is insignificant then remove that feature.
6. Compute fscore for each feature in the extended dataset, $fs_i = \frac{r_i}{FI_i}, i = 1\ to\ n$
7. Compute weighted harmonic mean, $whm = \frac{\Sigma\, r_i}{\Sigma fs_i}, i = 1\ to\ n$.
8. For any feature i in the extended dataset, if $fs_i < whm$, eliminate the feature i
9. If *fs* of any original feature < *fs* of its corresponding shadow feature, then eliminate that original feature. Also, if *fs* of any feature is insignificant then remove that feature.
10. Repeat steps 1 to 9 until in each iteration at least 10% of the features are eliminated or if maximum iterations have not been completed. Else, return the remaining features and stop.
}

III. METHODOLOGY

In this paper, the BoostaARoota algorithm and the proposed ModifiedBoostARoota algorithm is used with base models XGBoost and Catboost for feature selection. These algorithms are applied on four different datasets namely Cleveland Heart dataset, Statlog heart dataset, SA heart dataset and Wisconsin Diagnostic Breast Cancer dataset which are described in section IV. Then, the classifiers XGBoost and Catboost are applied on the features selected by BoostARoota and ModifiedBoostARoota. A comparison of the performances of both the feature selection algorithms is done.

IV. DATASETS

In this paper, four datasets have been used, namely, 1) Cleveland heart disease dataset at UCI Machine Learning Repository, created by Robert Detrano, M.D., Ph.D, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. 2) Statlog heart disease dataset at UC Irvine Machine Learning Repository, created by Dua, Dheeru and Graff, Casey. 3) South African Heart dataset is taken from "Replication Data for: South African Heart Disease", Harvard Dataverse, contributed by Bartley, Christopher. 4) Wisconsin Diagnostic Breast Cancer dataset at UCI Machine Learning Repository, created by Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian, University of Wisconsin.

TABLE I. DATASETS

Datasets	**Cleveland Heart dataset**	**Statlog heart dataset**	**SA heart dataset**	**Wisconsin Breast Cancer dataset**
No. of instances	303	270	462	569
No. of predictor Attributes, Names of Attributes	13. age, sex, fbs, cp, exang, trestbps, ca, restecg, thal, thalach, slope, chol, oldpeak	13. age, sex, cp, cholestrol, maxheartrate, oldpeak, slope, coloredvessels, angina, thal, restbp, restecg, fbs	9. sbp, tobacco, ldl, adiposity, famhist, typea, obesity, alcohol, age	9. thickness, size, shape, epithelial, adhesion, nuclei, chromatin, nucleoli, mitoses
Target / class variable	Num (0-absent, 1-present)	disease (1-absent, 2 present)	Class label (-1 for negative, +1 for positive)	status *(2* for benign, 4 for malignant)

Table I describes the 4 datasets.

V. RESULTS AND DISCUSSION

The experiment was done using Python [17] in Jupyter notebook on Ubuntu OS run on a system with i5 processor and 4 GB RAM. On applying the BoostaARoota algorithm and the proposed ModifiedBoostARoota algorithm, with base models XGBoost and Catboost, the features selected are as shown in Table II.

From Table II it is obvious that in all the datasets, ModifiedBoostARoota (MBAR) selects reduced number of relevant features compared to BoostARoota (BAR), which is advantageous to reduce the training time for a model. From Table III, we find that BoostARoota with XGBoost as base model executes faster than ModifiedBoostARoota with XGBoost. But we observe that ModifiedBoostARoota with Catboost executes much faster than BoostARoota with Catboost.

TABLE II. FEATURE SELECTION ON ALL 4 DATASETS

Dataset (DS)	**Features selected by**			
	BoostARoota with XGBoost base model (BAR-XGB)	***ModifiedBoostARoota with XGBoost base model (MBAR-XGB)***	***BoostARoota with CatBoost base model (BAR-CB)***	***ModifiedBoostARoota with CatBoost base model (MBAR-CB)***
Cleveland Heart DS	sex, cp, trestbps, fbs, thal, chol, restecg, thalach, exang, ca, oldpeak, slope,	cp, sex, chol, thalach, oldpeak, slope, ca, trestbps, thal	sex, cp, trestbps, chol, restecg, thal, thalach, exang, ca, oldpeak, slope	cp, slope, ca, thal
Statlog heart DS	sex, cp, cholestrol, maxheartrate, oldpeak, slope, coloredvessels, angina, thal, restbp, restecg	sex, cp, cholestrol, maxheartrate, oldpeak, coloredvessels, angina, thal, restbp	sex, cp, cholestrol, maxheartrate, oldpeak, slope, coloredvessels, angina, thal, restbp, restecg	sex, cp, cholestrol, slope, coloredvessels, thal
SA heart DS	sbp, tobacco, ldl, adiposity, famhist, typea, obesity, alcohol, age	tobacco, ldl, famhist, typea, age	sbp, tobacco, ldl, adiposity, famhist, typea, obesity, alcohol, age	ldl, famhist, age
Breast cancer DS	thickness, size, shape, epithelial, adhesion, nuclei, chromatin, nucleoli, mitoses	thickness, size, shape, epithelial, adhesion, nuclei, chromatin, nucleoli	thickness, size, shape, epithelial, adhesion, nuclei, chromatin, nucleoli, mitoses	thickness, shape, nuclei, chromatin, nucleoli

TABLE III. COMPARISON OF ALGORITHMS' RUNNING TIME

Dataset (DS)	**Running time of BAR-XGB**	**Running time of MBAR-XGB**	**Running time of BAR-CB**	**Running time of MBAR-CB**
Breast cancer DS	0.373s	1.251s	25.097s	7.787s
Cleveland HDS	0.702 s	1.938 s	44.279 s	13.699 s
Statlog HDS	0.679 s	2.059 s	39.871 s	8.788 s
SA HDS	0.422 s	1.235 s	30.713 s	13.874 s

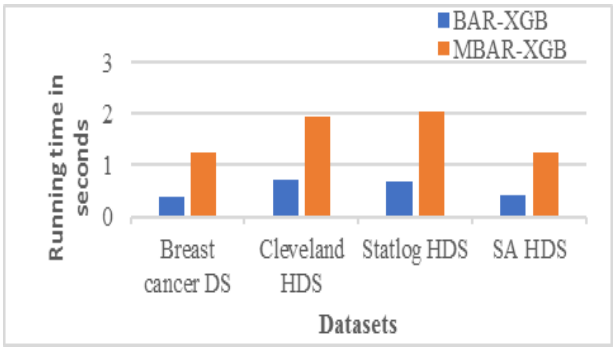

Fig. 1. Comparison of running time of BoostARoota(BAR) and ModifiedBoostARoota(MBAR) with XGBoost as basemodel

Fig.1 depicts the computational time of BoostARoota(BAR) with XGBoost base model is faster than ModifiedBoostARoota(MBAR) with XGBoost. Both algorithms terminate if less than 10% of the features are eliminated in an iteration. As features eliminated by MBAR are more than BAR, the iterations executed by BAR are lesser compared to MBAR, which leads to the difference of atmost one second in computation time between BAR and MBAR.

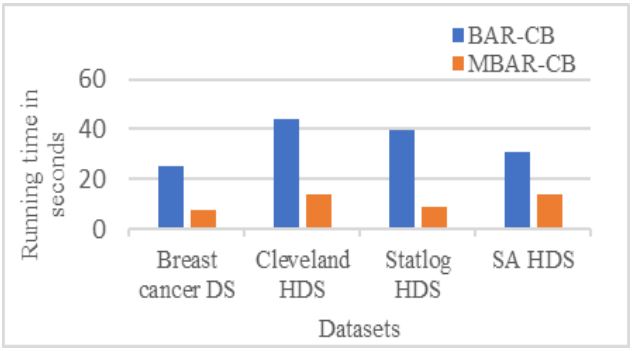

Fig. 2. Comparison of running time of BoostARoota(BAR) and ModifiedBoostARoota(MBAR) with Catboost as basemodel

Fig. 2 shows that there is significant difference between the computational time between MBAR with CatBoost (CB) as base model and BAR with CB. The performance of MBAR-CB is significantly better than BAR-CB.

Each of the 4 datasets with features selected by MBAR-CB and BAR-CB is classified using CatBoost Classifier (CBC) and features selected by MBAR-XGB and BAR-XGB is classified using XGBoost Classifier (XGBC). The performance of the classifiers based on the features selected by the respective algorithms is depicted in Table IV.

TABLE IV. PERFORMANCE OF THE CLASSIFIERS BASED ON THE FEATURES SELECTED

Datasets	FS algorithm, classifier	Training time	Prediction time	Accuracy	Precision	Recall
Cleveland HDS	MBAR-CB, CBC	1.736	0.001	**91.8**	0: 0.88 1: 0.96	0: 0.97, 1: 0.87
	BAR-CB, CBC	1.954	0.002	88.52	0: 0.88, 1: 0.90	0: 0.90, 1: 0.87
	MBAR-XGB, XGBC	0.034	0.004	**91.8**	0: 0.88, 1: 0.96	0: 0.97, 1: 0.87
	BAR-XGB, XGBC	0.044	0.002	85.25	0: 0.87, 1: 0.84	0: 0.84, 1: 0.87
Statlog HDS	MBAR-CB, CBC	2.494	0.025	**87.04**	0: 0.82, 1: 0.95	0: 0.96, 1: 0.77
	BAR-CB, CBC	2.626	0.003	87.04	0: 0.84, 1: 0.91	0: 0.93, 1: 0.81
	MBAR-XGB, XGBC	0.032	0.003	**85.19**	0: 0.79, 1: 0.95	0: 0.96, 1: 0.73
	BAR-XGB, XGBC	0.036	0.002	83.33	0: 0.77, 1: 0.95	0: 0.95, 1: 0.69
SA HDS	MBAR-CB, CBC	1.331	0.001	**74.19**	0: 0.77, 1: 0.65	0: 0.87, 1: 0.48
	BAR-CB, CBC	1.673	0.013	72.04	0: 0.78, 1: 0.59	0: 0.81, 1: 0.55
	MBAR-XGB, XGBC	0.051	0.003	**75.27**	0: 0.81, 1: 0.63	0: 082, 1: 0.61
	BAR-XGB, XGBC	0.084	0.002	69.89	0: 0.78, 1: 0.55	0: 0.76, 1: 0.58
Breast C.DS	MBAR-CB, CBC	2.115	0.006	**96.43**	0: 0.98, 1: 0.94	0: 0.97, 1: 0.96
	BAR-CB, CBC	2.038	0.004	95.71	0: 0.97, 1: 0.94	0: 0.97, 1: 0.94
	MBAR-XGB, XGBC	0.069	0.003	**95**	0: 0.96, 1: 0.94	0: 0.97, 1: 0.92
	BAR-XGB, XGBC	0.144	0.002	95	0: 0.96, 1: 0.95	0: 0.97, 1: 0.93

From Table IV, it is observed that using ModifiedBoostARoota with base model CatBoost for feature selection and Catboost classifier for training and prediction (MBAR-CB, CBC), gives more accuracy when compared to that of BoostARoota. Also, ModifiedBoostARoota with base model XGBoost for feature selection and XGBoost classifier for training and prediction (MBAR-XGB, XGBC) shows better performance than its BoostARoota counterpart.

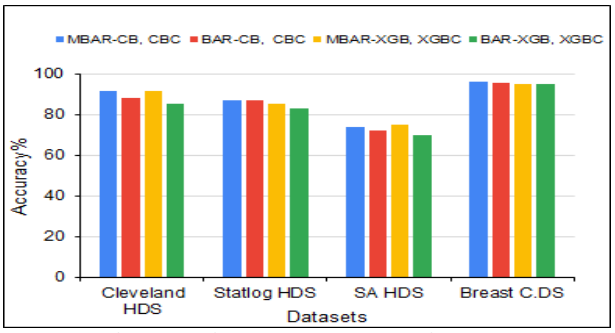

Fig.3. Performance of BoostARoota and ModifiedBoostARoota combined with XGBoost and CatBoost classifiers

Fig.3 shows that on all datasets, MBAR-CB feature selection with Catboost Classifier (CBC) combination gives more accuracy compared to other combinations. From Table IV, we find that XGBoost classifier requires lesser Training time and Catboost classifier takes lesser time to predict. Also, the features selected by proposed ModifiedBoostARoota algorithm is lesser and thereby consumes lesser training time compared to its counterpart.

VI. CONCLUSION

Early detection of heart diseases can be done with the help of algorithms in Machine learning. Features contributing to the prediction of the target variable (disease) are identified by using feature selection methods. The proposed feature selection algorithm, ModifiedBoostARoota, is a modified version of BoostARoota algorithm, where modifications are made to the feature elimination process. These algorithms are used with XGBoost and CatBoost base models on 4 different medical datasets to perform feature selection. On using classifiers XGBoost and Catboost on the selected features, the combination of ModifiedBoostARoota with CatBoost base model for feature selection and Catboost as classifier (MBAR-CB,CBC) gives higher accuracy on all 4 datasets. Also, the prediction time required by Catboost is comparatively lesser than that of XGBoost. Moreover, ModifiedBoostARoota with base model XGBoost for feature selection and XGBoost classifier for training and prediction (MBAR-XGB, XGBC) shows better performance than its BoostARoota counterpart. Future work must focus on reducing the algorithm execution time.

REFERENCES

[1] D. C. Duro, S. E. Franklin, and M. G. Dub, “Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests,” International Journal of Remote Sensing, vol. 33, no. 14, pp. 4502–4526, 2012.

[2] Na"eem Hoosen Agjee, Riyad Ismail, Onisimo Mutanga, "Identifying relevant hyperspectral bands using Boruta: a temporal analysis of water hyacinth biocontrol," J. Appl. Remote Sens.10(4),042002, 2016.

[3] Caraka RE et al., "Feature importance of the aortic anatomy on endovascular aneurysm repair (EVAR) using Boruta and Bayesian MCMC," Communications in Mathematical Biology and Neuroscience,2020.

[4] S. S. Kumar and T. Shaikh, "Empirical evaluation of the performance of feature selection approaches on random forest," 2017 IEEE International Conference on Computer and Applications (ICCA), Doha, pp. 227-231, 2017.

[5] Sánchez-Maroon N., Alonso-Benzos A., Calvo-Estevez R.M, "A wrapper method for feature selection in multiple classes datasets," In: Cabestany J., Sandoval F., Prieto A., Corchado J.M. (eds) Bio-Inspired Systems: Computational and Ambient Intelligence. IWANN 2009. Lecture Notes in Computer Science, vol 5517. Springer, Berlin, Heidelberg, 2009.

[6] L. Ma, M. Li, Y. Gao, T. Chen, X. Ma and L. Qu, "A novel wrapper approach for feature selection in object-based image classification using polygon-based cross-validation," in IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 3, pp. 409-413, March 2017.

[7] Michael Bowles, Machine Learning in Python: Essential Techniques for Predictive Analysis. Wiley India Pvt. ltd, ISBN 978-81-265-5592-5, 2019.

[8] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin, "CatBoost: unbiased boosting with categorical features," 32nd Conference on Neural Information Processing Systems, Montreal, Canada, 2018.

[9] Tianqi Chen, Carlos Guestrin, "XGBoost: A scalable tree boosting system," KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785-794, 2016.

[10] Jason Brownlee, XGBoost with Python: Gradient Boosted Trees with XGBoost and scikit-learn, MachineLearningMastery. 2016.

[11] Alexander März, "Catboost Lss--an Extension of Catboost to Probabilistic Forecasting," 2020.

[12] A. Ogunleye and Q. Wang, "Enhanced XGBoost-based automatic diagnosis system for chronic kidney disease," 2018 IEEE 14th International Conference on Control and Automation (ICCA), Anchorage, AK, pp. 805-810, 2018.

[13] Anton D. Morozov et al., "Data-driven model for hydraulic fracturing design optimization: focus on building digital database and production forecast", Journal of Petroleum Science and Engineering, Elsevier, volume 194, 2020.

[14] N. S. Rajliwall, R. Davey and G. Chetty, "Cardiovascular risk prediction based on XGBoost," 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp. 246-252, 2018.

[15] Miron B. Kursa, Witold R. Rudnicki, "Feature Selection with the Boruta Package," Journal of Statistical Software, Volume 36, Issue 11, 2010.

[16] M. Zabihi, S. Kiranyaz and M. Gabbouj, "Sepsis prediction in intensive care unit using ensemble of XGboost models," 2019 Computing in Cardiology (CinC), Singapore, Singapore, pp. 1-4, 2019.

[17] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, Sci kit-learn: Machine Learning in Python. JMLR 12: 2825-2830, 2011.

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

Feature Selection and Prediction of Heart diseases using Gradient Boosting Algorithms

Anuradha.P
Research Scholar, Dept. of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Deemed to be University,
Coimbatore, T.N, India
anujith72@gmail.com

Dr.Vasantha Kalyani David
HOD and Professor, Dept. of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Deemed to be University,
Coimbatore, T.N, India

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395819

***Abstract*— According to WHO, 31% of the global human mortality rate is due to Cardio Vascular Diseases and 85% of it is due to heart attacks and strokes. To prevent such deaths, several Machine Learning (ML) Algorithms are used in the early prediction of heart attacks. In order to reduce the computation time of the ML models, several feature selection techniques exist. In this paper, Feature Importance ranking of two gradient boosting algorithms XGBoost and CatBoost were computed on Cleveland, Statlog heart and SA heart data sets. With each feature importance rank as threshold, subsets of features were formed. Classifiers XGBoost, CatBoost and Majority voting ensemble were modelled on these subsets and the feature subset yielding highest accuracy was obtained. The range of feature importance ranking among which the feature subset with the highest accuracy would be obtained, was identified in this work. The classifiers exhibited improved performance on selected features when compared to their performance on all features. On comparing the classifiers, CatBoost outperformed the other classifiers.**

Keywords*— *XGBoost; Catboost; feature selection; feature importance ranking; heart disease; boosting algorithms; classification; Majority voting ensemble; prediction

I. INTRODUCTION

Cardiovascular diseases (CVD) which are diseases related to heart and blood vessels, are the number one killer disease worldwide. CVD includes myocardial infarction (commonly known as heart attack). According to WHO, in 2016, 17.9 million people died of CVDs, out of which 85% was due to heart attacks and strokes [1].

The risk factors causing heart attacks include high blood pressure, high cholestrol, smoking, diabetes, lack of exercise, obesity, poor diet and genetic factors. By identifying the risk factors that might lead to heart attack, medical practitioners would prescribe medicines to control those risk factors and prevent heart attacks. Machine Learning Algorithms can be used for early prediction of CVDs based on the risk factors and thereby prevent CVDs.

In Machine Learning, Feature Selection (FS) is performed to select features/ risk factors (predictor variables) which would predict the disease (target variable) with better accuracy. FS helps to simplify the model by eliminating the irrelevant and redundant features. It also reduces training and prediction time. Also, a feature that has correlation with the target variable and left unnoticed by the practitioners can be brought to light by the supporting ML tool which does the prediction of heart attacks. Table I lists few of the feature selection techniques (apart from directly using Feature Importance) applied on heart datasets in earlier research works.

TABLE I. EARLIER WORK DONE ON FEATURE SELECTION AND PREDICTION OF HEART DISEASES

Authors	Feature Selection technique	Classifier	Accuracy
R. Spencer et al.,[2]	Chi squared	BayesNet	85%
C. B. Gokulnath and S. P. Shantharajah [3]	Genetic Algorithm	Support Vector Machine (SVM)	88.34%
Swati Shilaskar, Ashok Ghatol [4]	Mann–Whitney test (Wilcoxon) distance criterion + forward feature subset selection	SVM	85%
Gazeloglu C [5]	1)No FS 2)Correlation-based Feature Selection (CFS) 3)Chi-Square and Fuzzy RoughSet	1)SVM (PolyKernel) 2)Naive Bayes 3)RBF Network	85.1% 84.8% 81.2%
Saba Bashir et al.,[6]	Minimum Redundancy Maximum Relevance Feature Selection (MRMR)	Logistic Regression SVM	84.85%
Anna Karen et al.,[7]	Chi-square and principal component analysis	Random Forest	98.70%
Debjani Panda et al., [8]	Least Absolute Shrinkage and Selection Operator	Gaussian Naive Bayes	94.92%
Jalil Nourmohammadi-Khiarak et al., [9]	Imperialist competitive algorithm with meta-heuristic approach	K nearest neighbours	94.43%

Table II in section III on related works shows the studies done on using feature importance for selection of features.

As tree-based models have inbuilt Feature Importance (FI) ranking in them, it is ideal to use FI ranking for feature selection. Also, as FI ranking consumes less time it would be advantageous to adopt it for feature selection when tree-based models are used for classification. Recent research works, mentioned in section III, had used XGBoost FI ranking for selecting features with random removal of low ranked features. Also, Catboost FI is untapped in the works on heart datasets.

This work aims first at utilizing the same algorithm for both feature selection and classification. Second aim is to reduce the number of feature subsets to be analyzed and thereby avoid the random elimination of low ranked features.

This work was experimented on Cleveland, Statlog and South African heart datasets. The Feature Importance (FI) ranking types gain and weight of XGBoost and FI types Prediction Values Change and Loss Function Change of CatBoost were explored. Forward selection of features was done by taking each FI value (sorted in descending order) as threshold. The classifiers XGBoost, Catboost and Hard Majority Voting Classifer (an ensemble of Logistic Regression (LR), Random Forest (RF), Gaussian Naïve Bayes (GNB), XGBoost, Catboost) were modelled on the selected subsets and a comparison of the classifiers was done.

This paper is organised as follows: Section II briefly describes the Machine Learning algorithms, Section III discusses the related work, Section IV introduces the methodology, Section V describes the datasets used, section VI discusses the results and compares the performance of the gradient boosting algorithms, Section VII concludes the paper.

II. METHODS

A. Gradient Boosting

By performing gradient descent in a functional space, an ensemble predictor is constructed by Gradient boosting. It builds strong predictors by iteratively combining weaker models to reduce the error [10].

B. XGBoost

XGBoost stands for eXtreme Gradient Boosting. It was developed by Tianqi Chen and Guestrin [11]. XGBoost is an implementation of gradient boosted decision trees which gives enhanced speed and performance [12]. The objective function in the XGBoost is defined as follows:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_t) \quad (1)$$

where $\hat{y}_i^{(t)}$ is the prediction at the t^{th} round, f_t denotes the structure of a decision tree, and the regularization item $\Omega(f_t)$ is given by:

$$\Omega(f_t) = \gamma T + \tfrac{1}{2} \lambda \sum_{j=1}^{t} \omega_j^2 \quad (2)$$

Objective function= Training Loss+ Regularization.
How well a model fits on training data is measured by Training Loss. Good predictive models are made by optimizing the training loss [11]. Complexity of the model is measured by Regularization. Optimizing regularization makes models simple [11].

XGBoost has feature importance types namely, gain, weight, cover, total_gain, total_cover. Weight is the default importance type. It is calculated as the number of times a feature was used to split the data across all trees [13]. Gain is computed as the average gain across all splits where the feature was used [13]. Cover is the average coverage across all splits where the feature was used [13]. Total_gain is computed as the total gain across all splits where the feature was used [13]. Total_cover is the total coverage across all splits where the feature was used [13].

C. CatBoost

CatBoost an implementation of gradient boosting was developed by Yandex researchers and engineers. Binary decision trees are used as base predictors in Catboost. Catboost algorithm is robust as only the learning rate and iterations need to be set. Rarely, other hyper parameters tuning is required. Random permutation-driven approach of ordered boosting is implemented in Catboost. As it does efficient vector representation of categorical data, CatBoost is well suited for handling datasets which have lot of categorical features [10][14]. If only default parameters are used and no tuning is done then CatBoost performs better than XGBoost. Catboost has both CPU and GPU implementations which are faster than other gradient boosting algorithms [14].

CatBoost has feature importance types: Prediction values change and Loss function change. Prediction Values Change computes the average changes in prediction when the feature value changes [15]. Loss Function Change of a feature calculates how much the loss value of the model with the feature differs from loss value of a model without it [15].

D. Hard Majority Vote Ensemble

Majority vote ensemble is a machine learning model that combines the predictions of several other models. Hard voting sums up the predictions of each model in the ensemble for each class label and predicts the class label with the maximum votes.

E. Evaluation Measure

Confusion matrix is used to evaluate the quality of a classifier's output. Accuracy is calculated as the ratio of correctly predicted observation to the total observations [16]. Precision gives a measure of how many observations are correctly classified as positive out of all those observations classified as positive [16]. Recall is the ability of the classifier to find all the positive observations [16]. F1 score is calculated as the weighted harmonic mean of precision and recall [16].

$$Precision = TP/(FP+TP) \quad (3)$$
$$Recall = TP/(FN+TP) \quad (4)$$
$$Accuracy = (TN+TP) / (TP+TN+FP+FN) \quad (5)$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

TABLE II. RELATED WORK USING FEATURE IMPORTANCE IN FEATURE SELECTION

Authors	FS	Classifier	Accuracy	Dataset	Remarks
Din.H et al.,[17]	XGBoost -FI ranking by gain	Weighted ensemble model (logistic regression, SVM, random forests, and gradient boosting)	AUC-ROC-83.1% using no laboratory results and 83.9% with laboratory results	NHANES	This work considered feature sets as those based on laboratory data and no laboratory data (by survey method). Top 24 features were selected and rest were eliminated due to low performance. This study shows models trained on both lab data and no lab data has almost equal performance [17].
Borak kolukisa et al.,[18]	Average of the FI done by Relief-F, Gain Ratio, Information Gain, and Chi-Square	Naive Bayes	85.47%	Cleveland	In addition to the mentioned FS, an experiment on selecting features based on doctors' recommendation was adopted but it resulted in lesser accuracy [18].
Divya Tomar and Sonali Agarwal [19]	F-score	LSTSVM	85.59%	Statlog	The limitation of this approach is that it works well only on binary classification [19].
Zahangir et al.,[20]	Infogain, Correlation and ReliefF	Random forest	a)83.13% b)77%	a) Statlog b) SA heart	a) The last feature was eliminated in the subset formed based on the ranking of features by ReliefF on statlog dataset. b) The last two features in the feature ranking by RF through ClassifierAttribute on SAheart dataset were eliminated in the selected subset [20].
Gao Liyuan, Yongmei Ding [21]	XGBoost – FI by average gain	XGBoost	73.5%	Kaggle's public dataset	Bayesian parameter optimization algorithm based on Gaussian processes was used in order to have stable hyperparameters [21].
Mohammad-Ashraf Ottom and Walaa Alshorman [22]	Accumulating features ranks by Information gain, reliefF, gain ratio, oneR, symmetric uncertainity	K nearest neighbours	78.90%	Cleveland	For each feature the ranks across all FS techniques were summed up to get the accumulated rank. Features with rank<1 were eliminated. KNN gave best enhancement accuracy compared to no Feature Selection [22].
A. Khemphila and V. Boonjing [23]	Information gain + backward elimination	Artificial Neural Networks	80.99%	Cleveland	Information gain was computed after discretization of continuous valued attributes by partitioning the range of values into a finite number of subsets using WEKA tool [23].
Rajliwall et al., [24]	Information gain	XGBoost	a) 97.7% b) 89.9%	a) NHANES dataset b) Framingham Heart Study CHS dataset	Comparison of XGBoost with Decision Tree (J48), K-Nearest Neighbours (KNN), Random Forest, SVM, Naive Bayes, Logistic Regression, Ensemble (RPART) was done and found that XGBoost required very less model building time [24].
Nalluri et al., [25]	-	a) Logistic Regression b) XGBoost	a)85.86% b)84.46%	Framingham	AS XGBoost requires very less computation time, for large datasets, XGBoost is preferred over logistic regression [25].
MarcoMamprin et al., [26]	FI using SHapley Additive exPlanations (SHAP)	CatBoost	90%	anonymized dataset obtained from Catharina Hospital, Netherlands	Transcatheter Aortic Valve Implantation (TAVI) mortality prediction model based on CatBoost, XGBoost, Random Forest Classifier, Logistic Regression, Support Vector Machine, and Gaussian Naive Bayes was analyzed and Catboost model outperformed others [26].

III. RELATED WORK

The related work on heart diseases using Feature importance ranking for selecting features are mentioned in Table II. From the review of literature, it is understood that gradient boosting models require less computing time and performs better than other models which were taken into comparison. Few of the works in Table II, where FI type of XGBoost is used for feature selection, had either randomly eliminated the low-ranking features or considered all combinations of the subsets.

IV. METHODOLOGY

In this paper, the methodology shown in Fig.1 was experimented on Cleveland, Statlog and SA heart datasets respectively. The threshold range of feature importance values

among which the feature subset with the highest accuracy would be obtained was identified. The performance of the models XGBoost, CatBoost and Hard Majority Voting

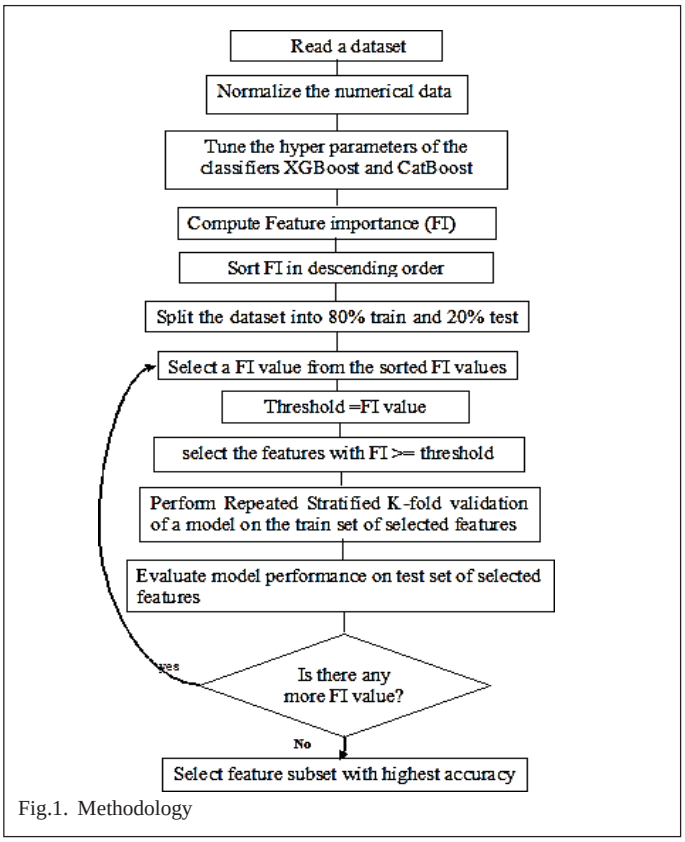

Fig.1. Methodology

Ensemble Classifer (an ensemble of LR, RF, GNB, XGBoost and CatBoost) were evaluated.

V. DATASETS

In this paper, the datasets used are: 1) South African Heart dataset taken from "Replication Data for: South African Heart Disease", Harvard Dataverse, contributed by Bartley, Christopher. 2) Statlog heart disease dataset at UC Irvine Machine Learning Repository, created by Dua, Dheeru and Graff, Casey. 3) Cleveland heart disease dataset at UCI Machine Learning Repository, created by Robert Detrano, M.D., Ph.D., V.A. Medical Center, Long Beach and Cleveland Clinic. Table III, IVand V shows the features of Cleveland, Statlog and SA heart dataset.

TABLE III. FEATURES OF SOUTH AFRICAN DATASET

adiposity
obesity
sbp: systolic blood pressure
tobacco: cumulative tobacco
ldl: low density cholesterol
famhist: family history
typea: type-A behavior
alcohol: alcohol consumption
age: age at onset
Target: class

TABLE IV. FEATURES OF STATLOG DATASET

Attributes/Features
Age
Sex
fbs-fasting blood sugar
cp-chest pain type
chol-serum cholesterol
restbp-resting blood pressure
restecg-ecg at rest
maxheartrate -maximum heart rate
angina-exercise induced angina
coloredvessels - number of major vessels colored
slope- slope of the peak exercise ST segment
thal -defect type
oldpeak- ST depression induced by exercise
Target: disease

TABLE V. FEATURES OF CLEVELAND DATASET

Attributes/Features
Age
Sex
fbs-fasting blood sugar
cp-chest pain type
chol-serum cholesterol
trestbps- resting blood pressure
restecg -ecg at rest
thalach-maximum heart rate
exang-exercise induced angina
ca-number of major vessels colored
slope- slope of the peak exercise ST segment
thal -defect type
oldpeak-ST depression induced by exercise
Target: Num

VI. RESULTS AND DISCUSSION

The experiment was done using Python in Jupyter notebook on Ubuntu OS run on a system with i5 processor and 4 GB RAM. The experiment as per the methodology described in section IV was performed on Cleveland Heart DataSet (CHDS), Statlog Heart DataSet (SHDS) and South African Heart Data Set (SAHDS).

The numerical features in each dataset were normalized using min-max scaler function. The Feature Importance (FI) ranking by FI types gain and weight of XGBoost classifier and FI ranking based on types Prediction Values Change (PVC) and Loss Function Change (LFC) of CatBoost classifier were done on Cleveland, Statlog and SA heart datasets. Fig.2. and Fig.3.depicts FI by gain of XGBoost and FI by PVC of Catboost, on SA heart DS. We can observe that the top 5 features selected by both methods are same. Fig.4 and Fig.5 displays the FI by Catboost and XGBoost on SHDS. Here 3 out of top 4 features of SHDS are same in both figures. Fig.6 and Fig.7 displays the FI by XGBoost and Catboost on CHDS. The top 3 features of CHDS are same in both FI types shown in figures 6 and 7.

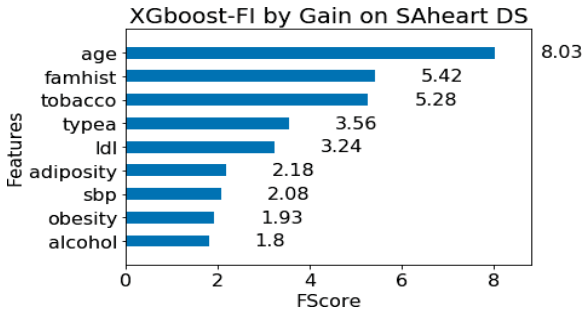

Fig.2. Feature Importance by XGBoost-gain on SA heart DS

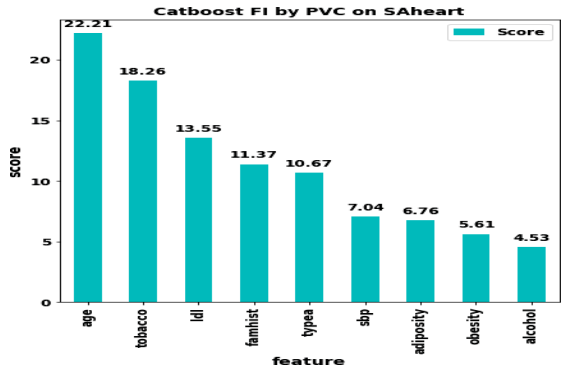

Fig.3.Feature Importance by Catboost-Prediction value change on SA heart DS

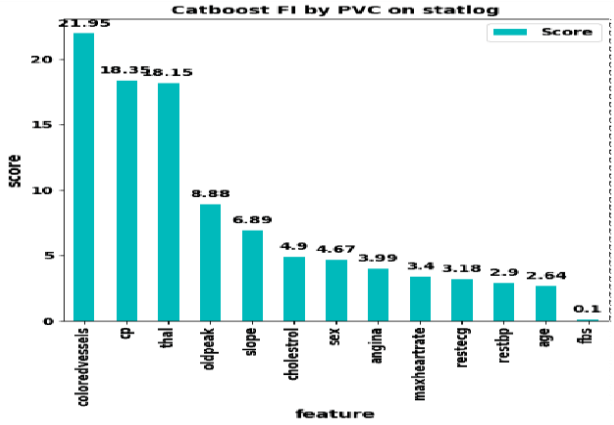

Fig.4. Feature Importance by Catboost -prediction value change on Statlog DS

Fig.5. Feature Importance by XGBoost-weight on Statlog DS

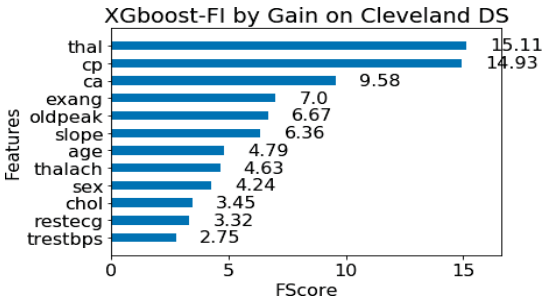

Fig.6. Feature Importance byXGBoost-gain on Cleveland DS

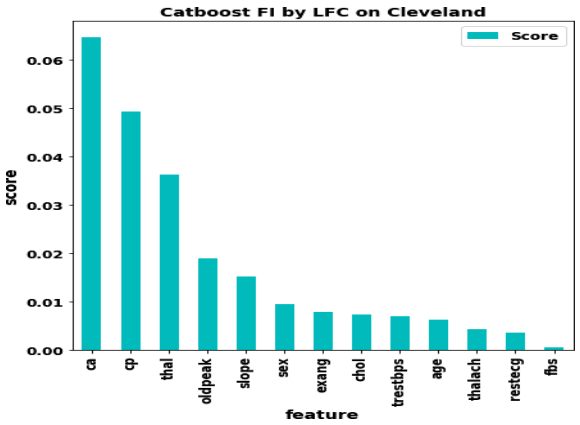

Fig.7. Feature Importance by Catboost-Loss Function change on Cleveland DS

Each dataset was split into 80% train and 20% test set. Feature subsets were formed by taking each Feature Importance (FI) value, starting from the highest value, as threshold. Repeated Stratified K-fold (K=5) validation of the models XGBoost and Catboost was performed on the train set containing the selected subset of features. Then prediction performance was tested using the test set. The feature subset which gave the highest accuracy was selected for each dataset.

As forward selection of highest-ranking features is done, the feature subsets selected by this methodology are the best among all possible combinations of subsets available.

Table VI displays the accuracy score of each feature subset formed by the rankings done by respective FI types of XGBoost and Catboost. The highest accuracy achieved by the FI types PVC and LFC are same in all datasets. We can infer that both the Catboost FI types have equal performance. In the case of XGBoost FI types, feature subsets formed on the basis of gain type shows higher performance compared to that of the default weight type.

From Table VI it is observed that the highest accuracy in each category is close to the threshold with median FI value. Instead of taking all FI values as threshold and forming subsets, the threshold range can be 4 FI values above and below the median FI value, which holds good in Table VI. Therefore, for low dimensional datasets,

Threshold range = 4 FI values above Median to 4 FI values below median.

TABLE VI. ACCURACY SCORES OF CATBOOST , XGBOOST AND MAJORITY VOTE ENSEMBLE CLASSIFIERS ON FEATURE SUBSETS

Datasets	Cleveland DS						Statlog DS						SA Heart DS					
Feature subsets	MVE-gain	MVE-PVC	CB-PVC	CB-LFC	XGB-gain	XGB-weight	CB-PVC	CB-LFC	XGB-gain	XGB-weight	MVE-PVC	MVE-gain	CB-PVC	CB-LFC	XGB-gain	XGB-weight	MVE-gain	MVE-PVC
Top 1	78.69	75.41	75.1	75.41	78.69	75.41	72.22	72.22	74.07	72.22	72.22	77.78	77.74	72.04	67.74	62.37	66.67	68.82
Top 2	78.69	83.61	81.97	73.37	78.69	77.05	68.52	68.52	70.37	75.93	66.67	72.22	72.04	72.04	64.52	66.67	65.59	69.89
Top 3	86.89	85.25	86.89	86.89	86.69	86.89	85.19	85.19	72.22	74.07	**87.04**	77.78	72.04	69.89	72.04	60.22	70.97	74.19
Top 4	**88.52**	**90.16**	90.16	88.52	88.52	88.52	81.48	81.48	85.19	70.37	81.48	**85.19**	74.19	74.19	69.89	63.44	**73.12**	74.19
Top 5	85.25	88.52	91.8	91.8	83.61	86.89	79.63	79.63	83.33	70.37	83.33	79.63	77.42	77.42	72.04	66.67	69.89	**75.27**
Top 6	86.89	86.89	90.16	88.52	85.25	86.89	79.63	79.63	83.33	74.07	83.33	83.33	78.49	78.49	69.89	63.44	68.82	75.27
Top 7	86.89	86.89	88.52	88.52	85.25	81.97	79.63	79.63	83.33	81.48	81.48	85.19	76.34	76.34	72.04	64.52	69.89	75.27
Top 8	86.89	88.52	88.52	91.8	85.25	85.25	81.48	77.78	81.48	79.63	83.33	83.33	76.34	76.34	69.89	68.82	68.82	74.19
Top 9	86.89	88.52	88.52	88.52	83.61	83.61	83.33	79.63	83.33	79.63	85.19	83.33	75.27	77.42	68.82	68.82	69.89	75.27
Top 10	88.52	88.52	88.52	85.25	86.89	86.89	83.33	81.48	81.48	83.33	83.33	87.04						
Top 11	85.25	85.25	85.25	85.25	85.25	83.61	81.48	83.33	81.48	79.63	85.19	85.19						
Top 12	81.97	85.25	86.89	85.25	85.25	85.25	81.48	85.19	79.63	79.63	85.19	85.19						
Top 13	81.97	86.89	86.89	85.25	85.25	85.25	83.33	83.33	77.78	77.78	83.33	85.19						

Creating subsets only for the threshold range and modelling on the same would save a lot on computation time. If only FI types gain and PVC are considered, we can observe from Table VI that all highest accuracies are above median FI value. It shows that the best subsets are above the median and only the 4 FI values greater than the median FI need to be checked.

Table VII shows the classification report of the models. The f1 scores of both classes are same for the highest accuracy entries on Cleveland and Statlog datasets and it slightly differs in SA heart dataset.

It can be observed from the last entry of each column in the Tables VI that the performance of the models decreases if all features of the respective datasets are considered.

Fig.8 displays the highest accuracy scores of XGBoost, CatBoost and Majority vote ensemble (MVE) classifiers modelled on features selected by FI type PVC and gain. It also illustrates that CatBoost outperforms other classifiers on both Cleveland and SA heart datasets and the MVE classifier gives higher accuracy than others on Statlog dataset.

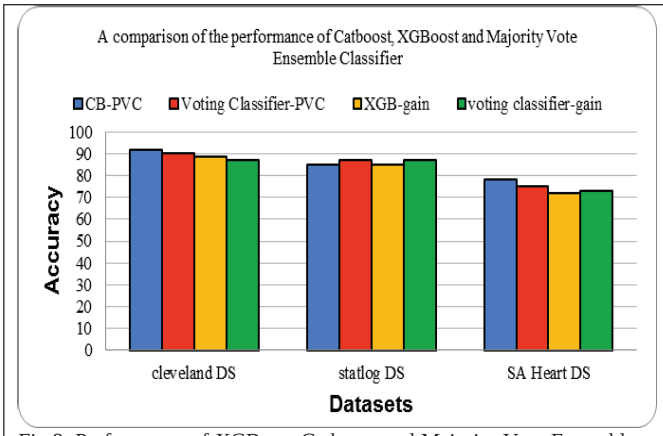

Fig.8. Performance of XGBoost,Catboost and Majority Vote Ensemble Classifier

Table VIII displays the features subsets which yielded highest accuracy in each dataset.

TABLE VII. CLASSIFICATION REPORT

		Cleveland Dataset				Statlog Dataset				SA Heart Dataset			
		precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score	accuracy
FS by cb-pvc /Catboost	0	0.89	0.97	0.93	**91.80%**	0.82	0.88	0.85	85.19%	0.78	0.91	0.84	**78.49%**
	1	0.96	0.85	0.9		0.88	0.82	0.85		0.8	0.57	0.67	
FS by cb-LFC /Catboost	0	0.89	0.97	0.93	91.80%	0.82	0.88	0.85	85.19%	0.78	0.9	0.83	77.42%
	1	0.96	0.85	0.9		0.88	0.82	0.85		0.77	0.57	0.66	
FS by xgb-gain/XGBoost	0	0.85	0.97	0.91	88.52%	0.85	0.85	0.85	85.19%	0.71	0.89	0.79	72.04%
	1	0.95	0.77	0.85		0.86	0.86	0.86		0.75	0.47	0.58	
FS by xgb-weight/XGBoost	0	0.89	0.91	0.9	88.52%	0.77	0.92	0.84	83.33%	0.69	0.87	0.77	68.82%
	1	0.88	0.85	0.86		0.91	0.75	0.82		0.7	0.42	0.52	
FS by cb-pvc /Maj. vote ensemble	0	0.89	0.94	0.92	90.16%	0.85	0.88	0.87	**87.04%**	0.76	0.88	0.82	75.27%
	1	0.92	0.85	0.88		0.89	0.86	0.87		0.73	0.54	0.62	
FS by xgb-gain /Maj. vote ensemble	0	0.85	0.97	0.91	88.52%	0.85	0.88	0.87	87.04%	0.72	0.89	0.8	73.12%
	1	0.95	0.77	0.85		0.89	0.86	0.87		0.76	0.5	0.6	

TABLE VIII. SELECTED FEATURE SUBSETS

Dataset	Features selected	Model	Accuracy
Cleveland HDS	ca, thal, cp, oldpeak, slope	Catboost	91.80
Statlog HDS	coloredvessels, cp, thal	Maj.vote ensemble	87.04
SA HDS	age, tobacco, ldl, famhist, typea, sbp	Catboost	78.49

The threshold range method proposed to reduce the feature subsets to be considered for modelling, was not applied on high dimensional datasets, which is a limitation of this work.

This study infers that for both feature selection and classification same gradient boosting algorithm can be used which would save time and memory.

VII. CONCLUSION

Selecting features that are highly correlated with the target variable is at most important as it improves the performance of the classifiers. In this work, XGBoost Feature Importance (FI) ranking based on gain/weight and CatBoost FI ranking based on Prediction Values Change (PVC) and Loss Function Change were explored. Feature subsets were formed by taking each FI rank as threshold. XGBoost, CatBoost and Hard Majority Vote Ensemble classifiers were modelled on these feature subsets and the subset which gave the highest accuracy score was selected. The threshold range of feature importance values for the feature subset selection was identified. Feature Selection by Catboost-PVC FI type with Catboost classifier achieved highest accuracy on Cleveland and SA heart datasets while Majority Vote Ensemble classifier gave highest accuracy score on Statlog dataset.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases -(cvds).

[2] Robinson Spencer, Fadi Thabtah, Neda Abdelhamid, Michael Thompson, “Exploring feature selection and classification methods for predicting heart disease,” Digital health, vol. 6, 2055207620914777, 2020.

[3] Gokulnath C.B., Shantharajah S.P, “An optimized feature selection based on genetic approach and support vector machine for heart disease,” Cluster Comput,vol.22, pp.14777–14787, 2019.

[4] Swati Shilaskar, Ashok Ghatol, “Feature selection for medical diagnosis : Evaluation for cardiovascular diseases,” Expert Systems with Applications, Vol.40, Issue 10, pp.4146-4153, 2013.

[5] Gazeloglu C, “Prediction of heart disease by classifying with feature selection and machine learning methods,” Progr Nutr [Internet], 22(2), pp.660-7, 2020.

[6] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, pp. 619-623, 2019.

[7] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès, “Classification models for heart disease prediction using feature selection and PCA,” Informatics in Medicine Unlocked, Volume 19, 100330, 2020.

[8] Debjani Panda, Ratula Ray, Azian Azamimi Abdullah, Satya Ranjan Dash, "Predictive Systems: Role of Feature Selection in Prediction of Heart Disease," Journal of Physics: Conference Series International Conference on Biomedical Engineering, Volume 1372, 2019.

[9] Jalil Nourmohammadi Khiarak et al., "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," Health and Technology,), vol.10, pp. 667–678, 2020.

[10] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin, “CatBoost: unbiased boosting with categorical features,” 32nd Conference on Neural Information Processing Systems, Montreal, Canada, 2018.

[11] Tianqi Chen, Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

[12] Jason Brownlee, XGBoost with Python, Gradient Boosted Trees with XGBoost and scikit-learn, MachineLearningMastery, 2017.

[13] https://xgboost.readthedocs.io / en/ latest / python / python_api.html.

[14] Alexander März, “Catboost Lss an Extension of Catboost to Probabilistic Forecasting,” Arxiv Preprint.

[15] https:/ /catboost.ai /docs /concepts /fstr.html

[16] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, Sci kit-learn: Machine Learning in Python. JMLR 12: 2825-2830, 2011.

[17] Dinh, A., Miertschin, S., Young, A. et al., “A data-driven approach to predicting diabetes and cardiovascular disease with machine learning,” BMC Med. Inform. Decis. Mak., vol.19, pp. 211, 2019.

[18] B. Kolukisa et al., "Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, pp. 2232-2238, 2018.

[19] Divya Tomar, Sonali Agarwal, "Feature Selection based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease," International Journal of Bio-Science and Bio-Technology ,Vol.6(2), pp.69-82, 2014.

[20] Md. ZahangirAlam,Saifur Rahman M.,SohelRahman M, “A Random Forest based predictor for medical data classification using feature ranking,” Informatics in Medicine Unlocked, Vol. 15, pp. 100180, 2019.

[21] Gao Liyuan, Yongmei Ding, “Disease prediction via Bayesian hyperparameter optimization and ensemble learning,” BMC research notes, vol. 13,1 205, 2020.

[22] Mohammad-Ashraf Ottom and Walaa Alshorman, "Heart Diseases Prediction Using Accumulated Rank Features Selection Technique," Journal of Engineering and Applied Sciences, vol. 14, pp. 2249-2257, 2019.

[23] A. Khemphila and V. Boonjing, "Heart Disease Classification Using Neural Network and Feature Selection," 2011 21st International Conference on Systems Engineering, Las Vegas, NV, USA, 2011, pp. 406-409.

[24] N.S. Rajliwall, R. Davey and G. Chetty, "Cardiovascular Risk Prediction Based on XGBoost," *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, Nadi, Fiji, pp. 246-252, 2018.

[25] Nalluri S., Vijaya Saraswathi R., Ramasubbareddy S., Govinda K., Swetha E., “Chronic Heart Disease Prediction Using Data Mining Techniques,” In: Raju K., Senkerik R., Lanka S., Rajagopal V. (eds) Data Engineering and Communication Technology. Advances in Intelligent Systems and Computing, vol 1079, 2020.

[26] MarcoMamprin, Svitlana Zinger, Peter H.N. de With, Jo M. Zelis,Pim A.L. Tonino, “Gradient Boosting on Decision Trees for Mortality Prediction in Transcatheter Aortic Valve Implantation,” arxiv.org, 2020.