INVESTIGATION STUDY ON HEART DISEASE PREDICTION WITH PATIENT HEALTHCARE DATA

P. MUTHULAKSHMI[1]
M. PARVEEN[2]

Department of Computer Science, Cauvery College for Women (Autonomous),
[Affiliated to Bharathidasan University], Tiruchirappalli, India
[1]muthulakshmi.cs@cauverycollege.ac.in
[2]parveen.it@cauverycollege.ac.in

Abstract. Big data comprises the structured, semi-structured and unstructured data collected by organization mined for the predictive analytics. Heart disease is the common disease that caused the peoples worldwide. Early heart disease prediction is an essential process for diverse healthcare providers to save their lives. Heart disease prediction is carried out with signs, symptoms and physical examination of patient. Data pre-processing, feature selection and classification process are performed for efficient heart disease prediction. Data pre-processing is carried out to refill the missing values in the input database. The feature selection process is performed to choose the relevant features from pre-processed data. The classification process is performed to classify the input data as normal or abnormal data for performing heart disease prediction. Many researchers carried out their research on the heart disease prediction. But, the accuracy level was not increased and time consumption was not minimized during the heart disease prediction. In order to address these problems, existing heart disease prediction method was reviewed.

Keywords: Big data, predictive analytics, heart disease prediction, feature selection, relevant features,classification process.

(Received July 19th, 2021 / Accepted September 1st, 2021)

1 Introduction

Healthcare professional like doctors take clinical decisions depending on experience and observations for treating diseases and ailments. It has probability of wrong diagnosis and judgmental errors that resulted in wrong treatment and redundant costs [14, 24]. When patient data was employed for removing the relevant analysis at cumulative and patient-level through decision support systems in clinical decisions, it resulted in safe healthcare diagnosis and treatment. Heart disease occurs in several forms like chest pain, stroke and heart attack.The heart disease type comprises the heart rhythm issues, congestive heart failure, congenital heart disease and cardiovascular disease (CVD). Heart disease is a deadly human disease that increases globally in developed and undeveloped countries. The heart not supplied adequate amount of blood to other parts of body in order to accomplish their normal functionalities. Early disease diagnosis is an essential one for preventing the patients from damage and save their lives.

This paper is organized as follows: Section 2 studies the review on different heart disease prediction methods. Section 3 reveals the study and analysis of the three existing heart disease prediction techniques. Section 4 describes the possible comparison of existing heart disease prediction methods at earlier stage. In section 5, the discussion and limitations of the existing heart disease prediction techniques are listed with future direction and Section 6 concludes the paper.

2 Literature Review

An intelligent computational predictive system was designed in [11] for cardiac disease diagnosis. Feature selection algorithms were employed to eliminate the irrelevant the noisy data from extracted feature space. However, the prediction accuracy was not increased by designed system. An imperialist competitive algorithm with meta-heuristic approach was designed in [12] to select the prominent heart disease features. K-nearest neighbor algorithm [1, 2, 5, 19] was introduced for performing the classification task [8, 9, 13, 17]. But, designed algorithm failed to perform feature selection [7, 15, 16, 18, 19] for partial and missed data.

A new feature selection approach was designed in [3] for supervised learning. The patient record was forecasted the existence of heart disease with minimal false alarming rate. But, an optimal method was not employed to manage the dimensionality issues by ensemble classification plan.

A new 0-1D coupled, personalized hemodynamic model was designed in [25] to forecast the pressure waveform and flow velocities in the arteries. The multi-scale CVS model was combined with Levenberg-Marquardt optimization for solving the inverse problem. But, the error rate was not reduced through designed model. The predictive value of pathological factors was computed in [26] for HF detection through social network approach. The similarity values were determined to construct the unweighted and weighted medical social network. Though accuracy level was improved, cost and time performance was not enhanced by designed approach.

The machine learning algorithm were designed in [21] for prediction with training data. The designed algorithm was obtained once the person entered information. The prediction model was designed over real-life hospital data. But, the prediction time was not minimized by designed machine learning algorithm.

An efficient neural network with convolutional layers was designed in [6] to classify class-imbalanced clinical data. A least absolute shrinkage and selection operator (LASSO) based feature weight assessment was introduced with the majority-voting based feature identification. However, CNN was not executed for prediction from similar clinical data sets where imbalanced number of positive and negative classification takes place.

An atherosclerotic cardiovascular disease (ASCVD) risk prediction model was introduced in [23] for patients with ASCVD. The statin-treated patients were used with ASCVD from AIM-HIGH trial cohort. But, the computational cost was not reduced by ASCVD risk prediction model. A hierarchical neighborhood component-based-learning (HNCL) and adaptive multilayer networks (AMLN) approach was designed in [20] for heart failure risk prediction. The global weight vector was employed in building the AMLN model for HFR prediction. However, the prediction time was not reduced through HNCL and AMLN.

A smart healthcare system was designed in [4] for performing the heart disease prediction through ensemble deep learning and feature fusion approach. The conditional probability approach computed feature weight for every class to improve the system performance. But, the feature fusion performance was not enhanced through data mining techniques for heart disease diagnosis.

A multi-task deep and wide neural network (MT-DWNN) was introduced in [22] for forecasting fatal difficulties during hospitalization. However, MT-DWNN model not incorporate additional information in EHR to increase the prediction performance. Cardio Help method was designed in [10] to predict the probability of cardiovascular disease presence in patient through deep convolution neural networks. However, the designed method failed to forecast the occurrence of major diseases like cancer and brain diseases.

3 Heart Disease Prediction Methods

Heart disease is one of the risky and life snatching chronic diseases all over world. The heart not supplied adequate blood to other parts of body to perform their normal functionality. Heart failure happens due to the blockage and narrowing the coronary arteries. Coronary arteries are responsible one for providing blood supply to the heart. In United States, the large amount of peoples gets affected by heart disease. The heart disease symptoms are physical body weakness, breath shortness, feet swollen and tiredness with related signs. The heart disease risk gets increased by person lifestyle like smoking, unhealthy diet, high cholesterol level, high blood pressure, deficiency of exercise and fitness.

3.1 Early and accurate detection and diagnosis of heart disease using intelligent computational model

An intelligent computational predictive system was introduced for the recognition and cardiac disease diagnosis. The machine learning classification algorithms were introduced for performing the accurate heart disease detection and diagnosis. The four feature selection techniques were used to remove the irrelevant and noisy data from the extracted feature space. P-value and Chi-

square were determined for Extra-Tree Classifier with every feature selection method. The designed system was helpful for the physician to diagnose the heart disease exactly and efficiently. In intelligent medical decision system, ten different machine learning classification techniques like Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), Artificial Neural Network (ANN) were used to choose the best technique for accurate heart disease detection at an early stage. Fast Correlation-Based Filter Solution (FCBF), minimal redundancy maximal relevance (mRMR), Least Absolute Shrinkage and Selection Operator (LASSO) and Relief were employed for choosing the essential and correlated features that reveal the motif of desired target. The developed system was trained and tested on Cleveland (S1) and Hungarian (S2) heart disease data sets from UCI machine learning repository.

3.2 New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection

Heart disease is one type of disease that includes the heart or vessels. Ten percent of death in early twentieth century was resulted from heart disease and death rate was caused due to disease increased by 25% in twentieth century. An imperialist competitive algorithm with meta-heuristic approach was introduced to choose the prominent heart disease features. The designed algorithm presented optimal response for feature selection. K-nearest neighbor algorithm was introduced for performing the classification task. An imperialist competitive algorithm presented the optimal response for feature selection toward the genetic and additional optimization methods. After feature extraction process, the features supplied to the K-nearest neighbor (KNN) for classification proposes. The combination of two techniques enhanced the heart disease diagnosis performance and their different features with higher classification accuracy. The designed algorithm achieved better result with two merits like reducing number of features and increasing classification accuracy.

Imperialist Competitive Algorithm (ICA) was employed to choose the features in heart disease diagnosis. The number of features was denoted to identify the best features to increase the heart disease diagnosis accuracy. The number of selectable features in implemented test was equal to different data sets. An imperialist competitive algorithm was started with initial population. Every population member was considered as country. The countries were divided into two groups, namely countries where colony subordinated to the country and colonialist countries. Every colonialist country controlled the colonies depending on their power. K-nearest neighbor algorithm was learning algorithm with observer employed by imperialist competitive algorithm to categorize the selected features. The designed algorithm include two goals, namely to determine the density function of data distribution learning and to categorize the data depending on the learning patterns.

3.3 Feature optimization by discrete weights for heart disease prediction using supervised learning

Healthcare information systems included the billing task, list and the purchase orders to aim on transactional statistics for managerial principles. A feature selection approach was introduced to perform the supervised learning. The designed approach identified that particular patient record was susceptible to the heart disease or not with lesser false alarming rate. The designed approach was dynamic n-gram features optimization with help of discrete weights of the feature correlation.The feature selection and optimization plan was introduced to distribute the variable size n-gram patterns of demographic features regarding the heart disease. The features were employed to train the classifier to perform the heart disease prediction with lesser false alarming as well as higher specificity and sensitivity rate. Naive Bayes classifier was trained with help of the optimal features chosen using the feature selection and optimization technique. The feature optimization method determined the optimal variable size n-gram features for performing the supervised learning termed discreet weights based n-gram feature selection. The designed method determined data structure with preprocessing, optimal attribute selection, attribute feature selection and classification strategy used to forecast the disease of patient medical records.

4 Performance Analysis Of Heart Disease Prediction Techniques

In order to determine the different heart disease prediction methods, number of patient data is considered as an input to conduct the experimentation. Different parameters are discussed for enhancing the heart disease prediction performance. For experimental evaluation, Cardiovascular Disease Data set from the UCI Machine Learning Repository is considered as input. The data set URL is given as `https://www.kaggle.com/sulianova/cardiovascular-disease-dataset`. The data set comprises 12 features and 70000 instances. The

data set comprises the patient information such as age, height, weight, gender, glucose, cholesterol,alcohol intake, etc. The quantitative analysis are compared with different parameters like,

- Prediction accuracy
- Prediction time and
- Error rate

4.1 Analysis on Prediction Time

Prediction time is described as the product of number of patient data and amount of time consumed for predicting the existence or absence of heart disease of one data. The prediction time is determined as,

$$P_{\text{Time}} = \text{Number of patient data} \times \text{time consumed for predicting one data} \quad (1)$$

From 1, the prediction time (P_{Time}) is calculated. The prediction time is measured in terms of milliseconds (ms). Table 1 illustrates the prediction time for different number of patient data varying from 100 to 1000. Prediction time comparison takes place on intelligent computational predictive system, imperialist competitive algorithm (ICA) with meta-heuristic approach and feature selection approach. The graphical representation of prediction time is revealed in figure 1.

Table 1: Tabulation for Prediction Time

Tabulation for Prediction Time	Prediction Time (%)		
	Intelligent computational predictive system	ICA with meta-heuristic approach	Feature selection approach
100	17	21	26
200	18	23	28
300	20	25	30
400	22	28	32
500	25	31	34
600	27	33	35
700	29	35	38
800	31	37	40
900	34	39	42
1000	36	41	45

Figure 1 illustrates the prediction time comparison for different number of patient data. From figure, the blue color cone denotes the prediction time of intelligent computational predictive system. The brown color cone and green color cone denotes the prediction time of imperialist competitive algorithm (ICA) with meta-heuristic approach and feature selection approach respectively. It is observed that prediction time consumption of intelligent computational predictive system is lesser when compared to imperialist competitive algorithm (ICA) with meta-heuristic approach and feature selection approach. This is because of using machine learning classification algorithms for performing the accurate heart disease detection and diagnosis. The feature selection technique removes the irrelevant and noisy data from the extracted feature space. This in turn helps to reduce the prediction time. Finally, the prediction time intelligent computational predictive system is 18% lesser than ICA with meta-heuristic approach and 27% lesser than feature selection approach.

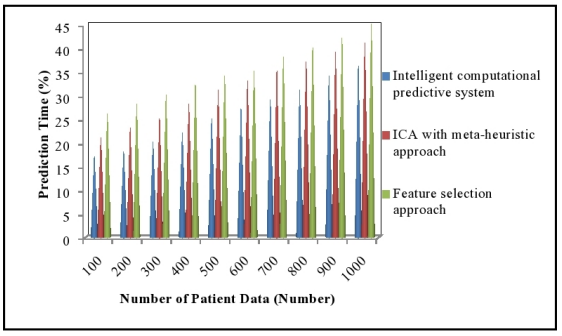

Figure 1: Measurement of Prediction Time

4.2 Analysis on Prediction Accuracy

Prediction accuracy is depicted as the ratio of number of patient data that are correctly predicted the existence or absence of heart disease to the total number of patient data considered as the input. Therefore, prediction accuracy is computed as,

$$P_{Acc} = \left(\frac{\text{Number of patient data that correctly predicted heart disease}}{\text{Number of patient data}}\right) \times 100 \quad (2)$$

From equation 2, prediction accuracy (P_{Acc}) is determined. The prediction accuracy is computed in terms of percentage (%). Table 2 explains the prediction accuracy for diverse number of patient data varying from 100 to 1000. Prediction accuracy comparison takes place on existing intelligent computational predictive system, imperialist competitive algorithm (ICA) with meta-heuristic approach and feature selection approach. The graphical representation of prediction accuracy is shown in figure 2.

Figure 2 explains the prediction accuracy comparison for diverse number of patient data. From the above figure, the blue color cone denotes the prediction accuracy of intelligent computational predictive system.

Table 2: Tabulation for Prediction Accuracy

Number of Patient Data (Number)	Prediction Accuracy (%)		
	Intelligent computational predictive system	ICA with meta-heuristic approach	Feature selection approach
100	75	84	80
200	78	86	83
300	76	82	78
400	79	85	81
500	81	87	84
600	82	90	86
700	84	91	88
800	86	93	90
900	87	94	91
1000	89	95	92

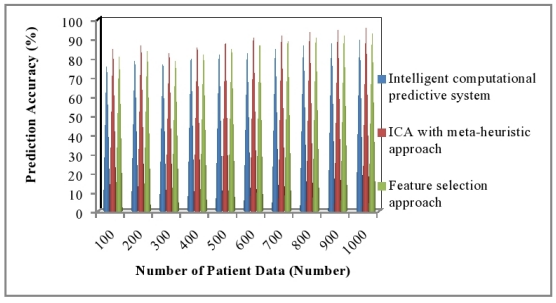

Figure 2: Measurement of Prediction Accuracy

The brown color cone and green color cone denotes the prediction accuracy of imperialist competitive algorithm (ICA) with meta-heuristic approach and feature selection approach correspondingly. It is clear that prediction accuracy of imperialist competitive algorithm (ICA) with meta-heuristic approach is higher when compared to intelligent computational predictive system and feature selection approach. This is due to the application of K-nearest neighbor (KNN) algorithm for performing the classification task. KNN was learning algorithm with observer by imperialist competitive algorithm to classify the selected features. The designed algorithm determined the density function and categorized the data depending on learning patterns. Finally, the prediction accuracy of ICA with meta-heuristic approach is 9% higher than intelligent computational predictive system and 4% higher than feature selection approach.

4.3 Analysis on Error rate

Error rate is computed as the ratio of number of patient data that are incorrectly predicted the existence or absence of heart disease to the total number of patient data taken. Therefore, the error rate is computed as,

$$Error_{Rate} = \left(\frac{\text{Number of patient data that are incorrectly predicted}}{\text{Number of patient data}} \right) \times 100 \qquad (3)$$

From 3, the error rate is computed. The error rate is determined in terms of percentage (%). Table 3 describes the error rate for different number of patient data varying from 100 to 1000. Error rate comparison takes place on existing intelligent computational predictive system, imperialist competitive algorithm (ICA) with meta-heuristic approach and feature selection approach. The graphical representation of error rate is illustrated in figure 3.

Table 3: Tabulation for Error Rate

Number of Patient Data (Number)	Prediction Accuracy (%)		
	Intelligent computational predictive system	ICA with meta-heuristic approach	Feature selection approach
100	20	18	12
200	22	20	14
300	23	23	15
400	25	25	18
500	27	22	16
600	30	21	15
700	31	23	17
800	33	25	18
900	35	27	20
1000	38	29	22

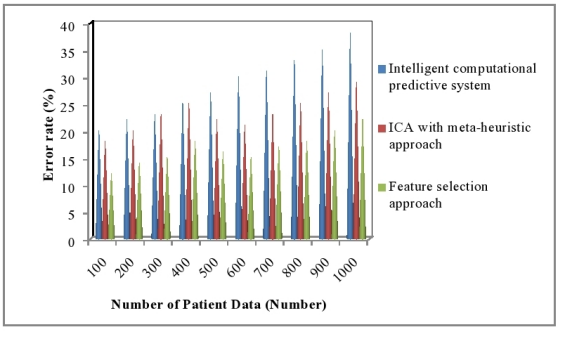

Figure 3: Measurement of Error Rate

Figure 3 illustrates the error rate comparison for different number of patient data. From the figure, the blue

color cone indicates the error rate of intelligent computational predictive system. The brown color cone and green color cone denotes the error rate of imperialist competitive algorithm (ICA) with meta-heuristic approach and feature selection approach correspondingly. It is apparent that error rate of feature selection approach is lesser when compared to intelligent computational predictive system and imperialist competitive algorithm (ICA) with meta-heuristic approach. This is due to the application of Naive Bayes classifier trained with optimal features selected through feature selection and optimization method. The feature optimization method computed the optimal variable size n-gram features for performing the supervised learning. By this way, the error rate gets minimized by feature selection approach. Finally, the error rate of feature selection approach is 41% lesser than intelligent computational predictive system and 29% lesser than ICA with meta-heuristic approach.

5 Discussion And Limiatation On Heart Disease Prediction Techniques

An intelligent computational predictive system was introduced for cardiac disease identification and diagnosis. The designed system performance was enhanced with the high variated optimal feature space. The designed system diagnosed the heart disease more accurately.However, the prediction accuracy was not enhanced through designed system. An ICA with meta-heuristic approach selected the prominent feature of heart disease. The designed algorithm provided the optimal response for feature selection with optimization algorithm. The feature selection accuracy was enhanced. However, the designed algorithm failed to perform feature selection method for imperfect and missed data.

The feature selection approach chose the relevant features. The n-gram sequence of features was considered as input for identifying the correlation between features and decision labels. The feature correlation was obtained by n-gram feature weight depending on the AHP method. The designed approach identified whether patient record was prone to heart disease or not with lesser false alarming. But, optimal method was not employed to handle the dimensionality issues by ensemble classification approach.

5.1 Future Direction

The future direction of heart disease prediction techniques with patient data can be carried out using machine learning techniques for reducing the time consumption and increasing the accuracy.

6 Conclusion

A comparison of different existing heart disease prediction techniques for patient data is studied. From the study, it is clear that the prediction accuracy was not enhanced by designed system. The review explains that the designed algorithm failed to perform feature selection technique for imperfect and missed data.In existing methods, the dimensionality issues were not addressed. The wide range of experiments on existing techniques describes the performance of many heart disease prediction techniques with its limitations. Finally, from the result, the research work can be carried out using machine learning and deep learning techniques for improving the performance of heart disease prediction techniques.

References

[1] Affonso, E. T., Nunes, R. D., Rosa, R. L., Pivaro, G. F., and Rodríguez, D. Z. Speech quality assessment in wireless voip communication using deep belief network. *IEEE Access*, 6:77022–77032, 2018.

[2] Affonso, E. T., Rosa, R. L., and Rodríguez, D. Z. Speech quality assessment over lossy transmission channels using deep belief networks. *IEEE Signal Processing Letters*, 25(1):70–74, 2018.

[3] Al-Yarimi, F. A. M., Munassar, N. M. A., Bamashmos, M. H. M., and Ali, M. Y. S. Feature optimization by discrete weights for heart disease prediction using supervised learning. *Soft Computing*, 25(3):1821–1831, 2021.

[4] Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., and Kwak, K.-S. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, 63:208–222, 2020.

[5] de Almeida, F. L., Rosa, R. L., and Rodríguez, D. Z. Voice quality assessment in communication services using deep learning. In *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pages 1–6, 2018.

[6] Dutta, A., Batabyal, T., Basu, M., and Acton, S. T. An efficient convolutional neural network for coronary heart disease prediction. *Expert Systems with Applications*, 159:113408, 2020.

[7] GuimarÃ£es, R., Rodríguez, D. Z., Rosa, R. L., and Bressan, G. Recommendation system using sentiment analysis considering the polarity of the adverb. In *2016 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 71–72, 2016.

[8] GuimarÃ£es, R. G., Rosa, R. L., De Gaetano, D., Rodríguez, D. Z., and Bressan, G. Age groups classification in social network using deep learning. *IEEE Access*, 5:10805–10816, 2017.

[9] Lasmar, E. L., de Paula, F. O., Rosa, R. L., AbrahÃ£o, J. I., and Rodríguez, D. Z. Rsrs: Ridesharing recommendation system based on social networks to improve the userâs qoe. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4728–4740, 2019.

[10] Mehmood, A., Iqbal, M., Mehmood, Z., Irtaza, A., Nawaz, M., Nazir, T., and Masood, M. Prediction of heart disease using deep convolutional neural networks. *Arabian Journal for Science and Engineering*, 46(4):3409–3422, 2021.

[11] Muhammad, Y., Tahir, M., Hayat, M., and Chong, K. T. Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Scientific reports*, 10(1):1–17, 2020.

[12] Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M.-R., Behrouzi, K., Mazaheri, S., Zamani-Harghalani, Y., and Tayebi, R. M. New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. *Health and Technology*, 10(3):667–678, 2020.

[13] Rodríguez, D. Z., Abrahao, J., Begazo, D. C., Rosa, R. L., and Bressan, G. Quality metric to assess video streaming service over tcp considering temporal location of pauses. *IEEE Transactions on Consumer Electronics*, 58(3):985–992, 2012.

[14] Rodríguez, D. Z., Rosa, R. L., and Alfaia, E. C. A simple method to measure the image complexity on a fault tolerant cluster computing. In *2010 Sixth Advanced International Conference on Telecommunications*, pages 549–554, 2010.

[15] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., and Rodrigues, F. A. Clustering algorithms: A comparative approach. *PloS One*, 14(1), 2019.

[16] Rosa, R. L., Rodríguez, D. Z., and Bressan, G. Sentimeter-br: A social web analysis tool to discover consumers' sentiment. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 2, pages 122–124, 2013.

[17] Rosa, R. L., Rodríguez, D. Z., and Bressan, G. Music recommendation system based on user's sentiments extracted from social networks. *IEEE Transactions on Consumer Electronics*, 61(3):359–367, 2015.

[18] Rosa, R. L., Rodríguez, D. Z., Schwartz, G. M., de Campos Ribeiro, I., and Bressan, G. Monitoring system for potential users with depression using sentiment analysis. In *2016 IEEE International Conference on Consumer Electronics (ICCE)*, pages 381–382, 2016.

[19] Rosa, R. L., Schwartz, G. M., Ruggiero, W. V., and Rodríguez, D. Z. A knowledge-based recommendation system that includes sentiment analysis and deep learning. *IEEE Transactions on Industrial Informatics*, 15(4):2124–2135, 2019.

[20] Samuel, O. W., Yang, B., Geng, Y., Asogbon, M. G., Pirbhulal, S., Mzurikwao, D., Idowu, O. P., Ogundele, T. J., Li, X., Chen, S., et al. A new technique for the prediction of heart failure risk driven by hierarchical neighborhood component-based learning and adaptive multi-layer networks. *Future Generation Computer Systems*, 110:781–794, 2020.

[21] Shankar, V., Kumar, V., Devagade, U., Karanth, V., and Rohitaksha, K. Heart disease prediction using cnn algorithm. *SN Computer Science*, 1(3):1–8, 2020.

[22] Wang, B., Bai, Y., Yao, Z., Li, J., Dong, W., Tu, Y., Xue, W., Tian, Y., Wang, Y., and He, K. A multitask neural network architecture for renal dysfunction prediction in heart failure patients with electronic health records. *IEEE Access*, 7:178392–178400, 2019.

[23] Wong, N. D., Zhao, Y., Xiang, P., Coll, B., and López, J. A. G. Five-year residual atherosclerotic cardiovascular disease risk prediction model for statin treated patients with known cardiovascular disease. *The American Journal of Cardiology*, 137:7–11, 2020.

[24] Zegarra Rodríguez, D., Rosa, R. L., and Bressan, G. A proposed video complexity measurement

method to be used in a cluster computing. In *2013 IEEE Global High Tech Congress on Electronics*, pages 76–77, 2013.

[25] Zhang, X., Wu, D., Miao, F., Liu, H., and Li, Y. Personalized hemodynamic modeling of the human cardiovascular system: a reduced-order computing model. *IEEE Transactions on Biomedical Engineering*, 67(10):2754–2764, 2020.

[26] Zhou, C., Li, A., Hou, A., Zhang, Z., Zhang, Z., Dai, P., and Wang, F. Modeling methodology for early warning of chronic heart failure based on real medical big data. *Expert Systems with Applications*, 151:113361, 2020.