International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v10i10.5745
***Article Received**: 28 July 2022 **Revised**: 05 September 2022 **Accepted**: 12 September 2022*

Cardiovascular Disease Prediction Using ML and DL Approaches

[a,b] **Avvaru R V Naga Suneetha,** [c] **Dr. T. Mahalngam**
[a] Research Scholar, Department of Computer Science and Engineering,
Vignan's Foundation for Science, Technology & Research (Deemed to be University),
Vadlamudi, Guntur, Andhra Pradesh,India, 522213.
[b]Assistant Professor, Department of Computer Science and Engineering,
Vignan Institute of Technology and Science,
Deshmukhi(V),Pochampally(M), Yadadri-Bhuvanagiri District, Telangana, India, 508284.
suneethaavvaru@gmail.com
[c] Associate professor, Department of Computer Science and Engineering,
Vignan's Foundation for Science, Technology & Research (Deemed to be University),
Vadlamudi, Guntur, Andhra Pradesh,India, 522213.
arulmurugan1982@gmail.com

Abstract— Healthcare is very important aspects of human life. Cardiovascular disease, also known as the coronary artery disease, is one of the many deadly infections that kill people in India and around the world. Accurate predictions can prevent heart disease, but incorrect predictions can be fatal. Therefore, here this paper describes a method for predicting cardiovascular disease that makes use of Machine Learning (ML) and Deep Learning (DL). In this paper, SMOTE-ENN (Synthetic Minority Oversampling Technique Edited Nearest Neighbor) was used to equalize the distribution of training data. The K-Nearest Neighbor method (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), XGBoost (Extreme Gradient Boosting), Artificial Neutral Network (ANN), and Convolutional Neutral Network (CNN) are among the classifiers used in this paper. From Public Health Dataset required data is collected and focused on recognizing the best approach for predicting the disease in preliminary phase. This experiment end results show that the use of Artificial Neural Networks can be of much useful in prediction with better accuracy (95.7%) than compared to any other ML approaches.

Keywords- Cardiovascular Disease, Artificial Intelligence, Machine Learning, Deep Learning, Convolutional Neural Networks.

I. INTRODUCTION

A significant number of deaths are occurring around the world, primarily as a result of heart attacks. Due to the delay in determining the severity of the attack, developing countries, particularly Asia and Africa, face a significant number of failures to save lives. Early detection of the heart attack can significantly decrease the attacks risk. [1]. Medical practitioners will generate a wealth of datasets on a daily basis that can analyzed to determine the important characteristics to look for when diagnosing a heart attack.

Because it circulates blood to all other organs, the heart is the major important appendage in the human body. If heart fails to function properly, the mind and other organs will fail, and the person will die in a matter of seconds. As a result, maintaining proper heart function is critical. Cardiovascular disease (CVD), also known as heart disease, has risen to become one of the world's leading causes of death [2]. According to the World Health Organization, heart disease kills 17.7 million people each year are accounting for 31% of all deaths worldwide (WHO). Sometimes Heart diseases have also exceeded cancer as the leading cause of death in India.

Machine learning and deep learning methods have aided research in a variety of fields including medicine [3]. Large-scale medical diagnosis data has aided in the training of these algorithms. These algorithms can be used to create a clinical support system that saves money and improves accuracy. Machine learning algorithms can use a variety of medical features to classify a patient's danger profile [4]. Certain characteristics, such as age, sex and heredity are beyond the patient's control, where as other methods are, such as blood pressure, smoking and drinking habits. This algorithm divides patients into well and unhealthful groups based on a mixture of these characteristics.

Users can construct intelligent algorithms to acquire more accurate results and predict output within a fair range using artificial intelligence apps. There are two methods in machine learning: supervised and unsupervised learning. The algorithm be given input data as well as target values to prepare on and predict the output values with a

***Article Received**: 28 July 2022 **Revised**: 05 September 2022 **Accepted**: 12 September 2022*

certain level of correctness in supervised learning. The algorithm can learn a model from the specified data apply it to a new dataset and examine the dataset to find the pattern. Predictive modeling and data mining techniques are similar to this. On the other hand, unsupervised learning is used for more difficult tasks and does not require the provision of required resulting data. Unsupervised learning aims to categorize these datasets into meaningful groups [5].

Deep Learning (DL) is also referred to as hierarchical or deep structured learning [6]. Unlike task-based methods, DL is a ML method based on the representation of supervised, unsupervised, or semi-supervised learning data. The functions of the biological nervous system, such as the processing and transmission of information, are vaguely influenced by the DL model. These DL technologies are structurally and functionally different from the human brain. They are incompatible with neuroscientific findings because of these differences. Convolutional neural networks are used in human speech recognition, Computer Vision (CV), speech recognition, natural language manipulation, machine translation, social website filtering, drug design, bioinformatics, medical imaging, and board game programs, CNN, DL-Network, Recurrent NN and Deep neural Network on paper [7].

The remaining of this paper is planned as follows. The literature review was summarized in Section II. Section III will look at how Machine Learning (ML) and Deep Learning (DL) is used for redict CVD (Cardiovascular Disease). The proposed model performance is evaluated in Section IV. Section V discuss in conclusion.

II. LITERATURE SURVEY

Himanshu Sharma, M A Rizvi et. al. [8] presents a range of classification techniques; it has developed successful data analysis models for forecasting serious cardiac syndrome. Noise features in data sets frequently distort valid data. As a result, efforts have been made to reduce noise in the dataset by cleaning and preprocessing it, as well as reducing its dimensions. They discovered that neural networks are capable of achieving high accuracy.

Verma et al. [9] presents a Correlation Feature Subset (CSF), A hybrid prediction model was built using Particle Swam Optimization (PSO), k-means cluster, and Multi Layer Perceptron (MLP). To build predictive models, the Cleveland heart disease dataset was used. According to the findings, the proposed hybrid model had an accurateness of up to 90.28 %.

Chen et al. [10] presents as a symptom, It is used to educate patients with group Heart Failure (HF) and related medication, as well as to determine the type of patients with unexplained heart failure based on prescription pharmaceuticals. After a ten-fold cross validation using the radial basis task of Support Vector Machine (SVM) with a cost of 0.075 and gamma of 0.5, they obtained an average precision rate 75.26%. The self learning module in LIBSVM (Library For Support Vector Machine) was used to automatically nominate these constraints. They concluded that the deceased patients had lower Exposure Factor (EF) differences than all patients with heart failure, based on the examination paper results of all compiled medical records.

Theresa Princy. R et.al, [11] describes the Naive Bayes, K-Near Neighborhood Algorithm, and the Neural Network were used in a survey of various machine learning methods to assess a person's risk of heart disease based on various characteristics such as epoch, sex, pulsate, and cholesterol. The accuracy with which the hazard is detected increases as the properties increase. It is possible to increase the accuracy with fewer features by using different methods.

Sana Bharti, Shailendra Narayan Singh et. al. [12] presents to predict health diseases; a researcher used an artificial neural network and a hereditary algorithm. The data mining strategy is incorporated into the association rules and classification methods in this reference. In this aspect, the author's methodology is quite helpful in detecting cardiac syndrome.

Long et al. [13] Presents In a medical decision sustain system for heart disease, the Chaos Firefly Algorithm and Rough Sets-Based Attribute Reduction (CFARS-AR) for Stat log datasets are used. After reducing the number of attribute with rough phrases, the Chaos Firefly algorithm was used to identify illnesses. The built representation was then compared to other models for example NB, SVM, and ANN. The described model had highest accuracy, sensitivity, and specificity of all models, with 88.3 %, 84.9 %, and 93.3 %, respectively.

Guidi et al. [14] presents a method for diagnosing heart failure as well as a clinical decision support system that can aid in early prevention. They compared deep learning and machine learning algorithms like SVM, Random Forest, and CART (Classification And Regression Trees). Random Forest and CART outperformed everybody in the category with an accurateness of 87.6%.

Binal A. Thakkar et.al [15] proposed a prototype that can forecast swine flu is being developed using the Naive-Bayes classification. Based on the numerous symptoms collected by the doctor, this algorithm classifies the likelihood of becoming unwell. An accuracy of the Naive Bayes classification is 63.33% and more work in this area can improve efficiency.

III. CARDIOVASCULAR DISEASE PREDICTION

The block diagram of cardiovascular disease prediction using Machine Learning (ML) and Deep Learning (DL) approaches as shown below in fig. 1. The Cleveland Heart Disease dataset from the UCI repository was used for this work and it contains 303 records and 14 features in this dataset. The data is split into two sections. The remaining 20% is used for validation and the remaining 80% is used for training. They have no values for some of the examples have missing values at some of attributes. For the purpose of training in this architecture, those values have been replaced with the attribute's mean value. The majority of traditional classification architectures demand that all attributes fall within the same range. Because the attributes in this dataset are in different ranges, a standardization technique is used to convert them all to the same value.

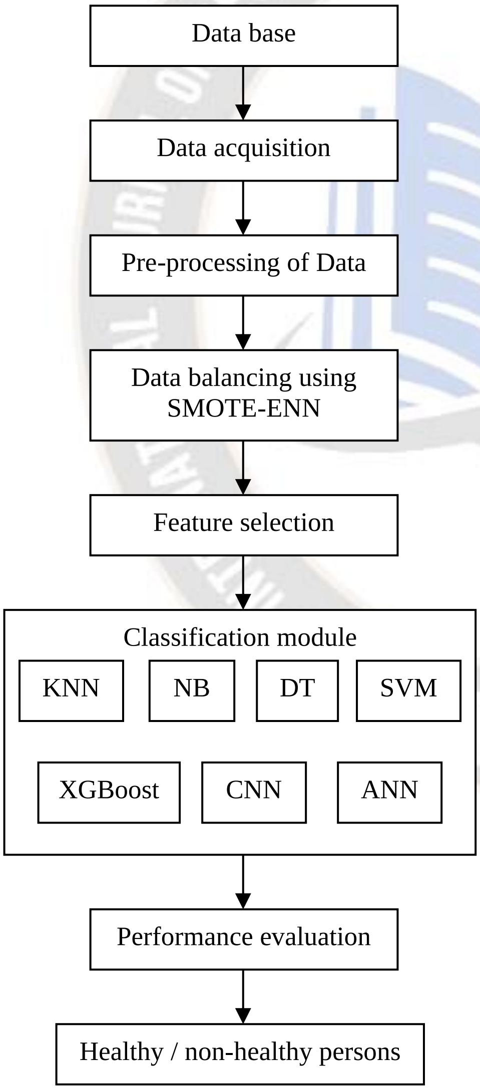

Fig. 1: Block Diagram of Cardiovascular Disease Detection

There are no missing values in the dataset. However, there were a lot of outliers to deal with, and the dataset wasn't distributed properly. Considering the data directly to the machine learning algorithm without using outliers or the feature selection procedure using two methods produced unfavorable results. However, the results obtained after using the dataset's normal distribution to overcome over fitting and applying isolated forests to detect outliers are very promising. All of these pretreatment techniques are critical in preparing data for classification or prediction.

Data sampling is also known as data balancing and it is a common method which deals the imbalanced data. In machine learning, there are three subcategories: oversampling, under sampling, and hybrid methods. The hybrid SMOTE-ENN method was used to correct an imbalanced heart disease training dataset. SMOTE (Synthetic Minority Oversampling Technique) is commonly used to oversample minority classes until the training set is balanced. Then, while maintaining a balanced distribution, ENN (Edited Nearest Neighbor) used to generate an unwanted double sample between the two classes[16]. Using the SMOTE technique, the number of members of the minority class through randomly generate new samples from the minority class samples' Nearest Neighbors (NN) has increased. ENN used to remove unnecessary duplicate samples. The total number of minority classes increases after SMOTE-ENN is implemented and the updated percentage of minority classes in the dataset becomes more balanced.

When performing feature selection, the Lasso algorithm, which is a part of embedded methods, is used to select the features and only choose the important ones. It outperforms filter methods in terms of predictive accuracy. This produces a functional subset suitable for existing algorithms. Then, to select the selected feature, choose from the models in the Skit-learn library that are part of the feature selection.

This paper employs a total of seven machine learning algorithms, including deep learning and machine learning. The K-Neighbors classifier was used to focus on neighbor selection, followed by the random forest classification of tree-based technology such as the decision tree classifier and finally the most popular technology of collective methods. The support vector machine was also used to test and handle the data's high dimensionality.

Machine learning is the process of providing effective training to a dataset using effective learning algorithms. Algorithms are a system of rules and responsibilities based on assumptions about data. Using alternative datasets and the same training algorithms, the system can be utilized to construct several system models during training.

The KNN classifier assigns each class to the K-Nearest Neighbors (KNN) of a given data point, where the variation of neighbors in the class is taken into account. To calculate test scores, where test score range from one to twenty neighbors.

The Decision Tree (DT) classification builds a tree using the class values assigned to each data point. A number of possible features were obtained and considered in order to construct an effective data model. The features vary in number from 1 to 30.

Naive Bayes (NB) is a statistical classification. It classifies conditional independence, which assumes a trait value over a given class but is not affected by the values of other traits.

SVM is a supervised learning algorithm that, like the C4.5 algorithm, performs tasks without the use of decision trees. The Support Vector Machine (SVM) attempts to reduce misclassification [17].

Extreme Gradient Boosting (XGBoost) is a supervised ML technique intended for classification and regression modeling. XGBoost is an improved method that uses gradient boost DT with regularization, a loss function, and some column sampling optimizations.

An Artificial Neural Network (ANN), also called as a "Neural Network," (NN)[18]. It's a multi-level system that uses mimics of neurons found in human anatomy to perform computations and numerical models. It performs the same function as a single neuron in the human brain.

A recent paper found that neural networks with application-specific settings, such as several hidden layers, can increase performance dramatically in a range of domains. A Convolutional Neural Network (ConvNet/CNN) is deep learning method that is used recognize distinct elements / objects in an input image (via learnable weights and biases)[19].

A sequential representation with a fully linked opaque layer, as well as flatten and dropout layers, is used to avoid over fitting. Machine learning and deep learning results are compared, and learning variances such as computational actual time and accuracy are described and displayed in the statistics listed further down in the results section.

IV. RESULTS ANALYSIS

This method makes use of the Cleveland Heart Disease dataset from the UCI (University of California, Irvine) repository. Accuracy, Precision, Recall, Specificity and In the evaluation process, F1-score is used. The vast majority of the data is used for training, whereas just 20% is used for validation. These parameters are divided into four categories: The first is a True Positive (TP) i.e. the value is identified as true and is actually true. The second type is a False Positive (FP), which occurs when a false value turns out to be true. The third is False Negatives (FNs). This happens when the value is true but the negative is false identified. True Negative (TN) is the fourth option, in which the value was truly negative.

Accuracy can be obtained by dividing true positive and true negative by true positive and true negative, and false positive, false negative by true positive and false negative. The formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \dots (1)$$

Precision is a metric of exactness used to assess a classifier's performance. There are fewer false positives if the precision is high. There are more false positives in a model with lower precision means.

$$Precision = \frac{TP}{TP + FP} \dots (2)$$

Recall is a metric for determining a classifier's completeness. Higher recall equals fewer false negatives, while lower recall equals more false negatives. When recall improves, precision often suffers as a result.

$$Recall = \frac{TP}{TP + FN} \dots (3)$$

F1-score is a combination of accuracy and recall that can be calculated using the formula below:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \dots (4)$$

After accuracy, specificity is a measure of how well a classifier recognizes negative situations. This is the proportion of true negative cases classified. It is also referred to as the true negative rate. The formula is as follows:

$$Specificity = \frac{TN}{TN + FP} \dots (5)$$

Different classifiers performance parameters comparisons are described in below Table. 1 as:

***Article Received**: 28 July 2022 **Revised**: 05 September 2022 **Accepted**: 12 September 2022*

Table. 1: Performance Of Different Classifiers

Classification approach	Accuracy	Precision	Recall	F1-Score	Specificity
KNN	74.2	76.6	77.7	76.1	75.9
DT	76.8	75.4	74.8	77.9	76.4
NB	84.6	85.1	84.9	85.5	85.7
SVM	78.9	78.4	80.4	81.3	79.4
XGBoost	81.3	82.6	84.3	86.1	86.7
CNN	92.4	90.6	92.2	93.3	94.1
ANN	95.7	94.4	95.6	95.4	95.2

Performance parameters such as Accuracy, Precision, Recall, F1-Score and Specificity for different classifiers are represented in below Fig. 2, Fig. 3, Fig. 4, Fig. 5 and Fig. 6 respectively.

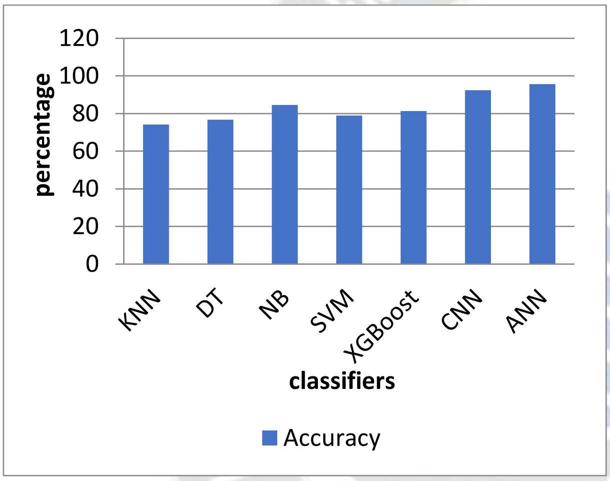

Fig. 2: Accuracy Performance of Different Classifiers

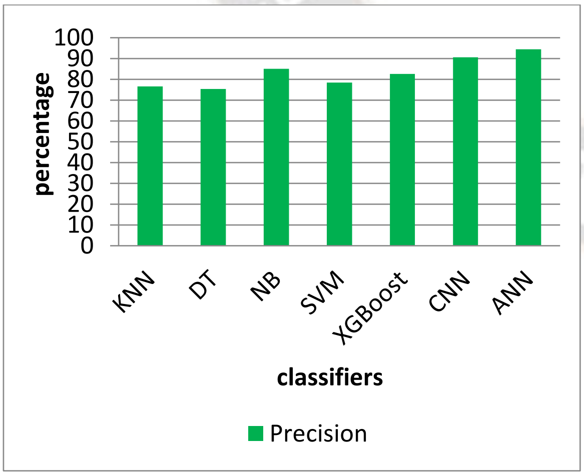

Fig. 3: Precision Performance Of Different Classifiers

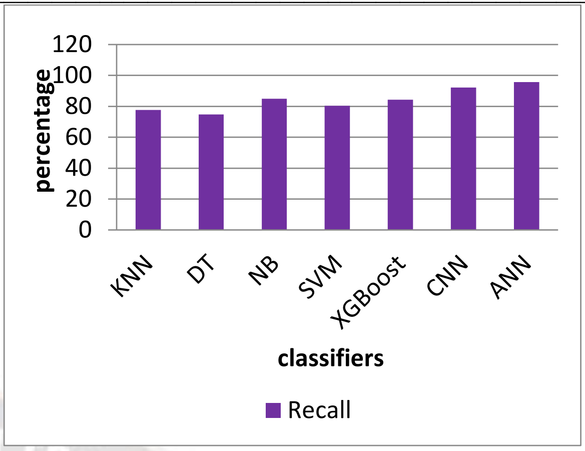

Fig. 4: Recall Performance Of Different Classifiers

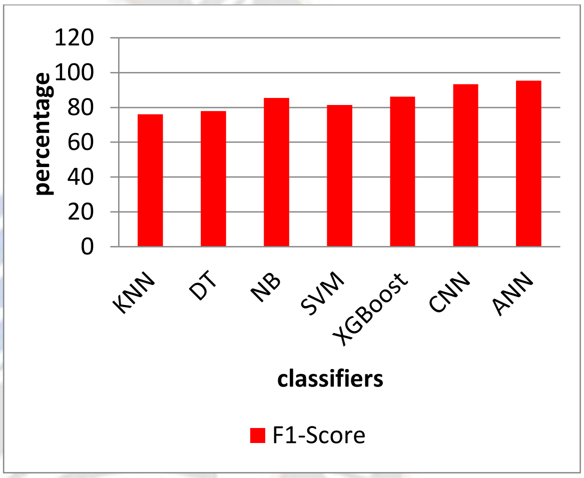

Fig. 5: F1-Score Performance of Different Classifiers

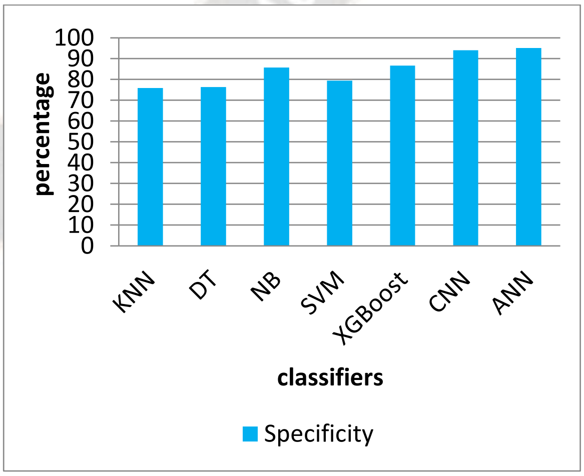

Fig. 6: Specificity Performance Of Different Classifiers

Therefore from above results it is clear that, performance of ANN and CNN are better than the machine learning models. In terms of Accuracy, Precision, Recall, Specificity, and F1-

***Article Received**: 28 July 2022 **Revised**: 05 September 2022 **Accepted**: 12 September 2022*

Score, ANN outperforms CNN in predicting cardiovascular disease.

V. CONCLUSION

Machine learning (ML) and Deep Learning (DL) algorithms are proposed in this paper for predicting cardiovascular disease. In this paper described techniques for Synthetic Minority Oversampling Edited Nearest Neighbors (SMOTE-ENN) were utilized to even out the distribution of training data. The seven different classifiers used in this method for classification and detection of Cardiovascular Disease are KNN, SVM, NB, DT, XGBoost, ANN and CNN. Some of the performance evaluation measures are accuracy, precision, recall, specificity, and F1-score. It increases the size of a dataset and then uses deep learning with a variety of other optimizations to achieve more promising results. ANN outperforms CNN in predicting cardiovascular disease in terms of accuracy, precision, recall, specificity, and F1-Score (accuracy as high as 95.7 %).

REFERNCES

[1]. Maria Sultana Keya, Muhammad Shamsojjaman, Faruq Hossain, Farzana Akter, Fakrul Islam, Minhaz Uddin Emon, "Measuring the Heart Attack Possibility using Different Types of Machine Learning Algorithms", 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Year: 2021.

[2]. Nafis Mostafa, Muhammad Anwarul Azim, Md Rayhan Kabir, Rasif Ajwad, "Identifying the Risk of Cardiovascular Diseases From the Analysis of Physiological Attributes", 2020 IEEE Region 10 Symposium (TENSYMP), Year: 2020.

[3]. Saiful Islam, Nusrat Jahan, Mst. Eshita Khatun, "Cardiovascular Disease Forecast using Machine Learning Paradigms", 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Year: 2020.

[4]. Mariusz Filipowicz, & Waleed F. Faris. (2022). Recent Advancement in the Field of Analogue Layout Synthesis. Acta Energetica, (02), 01–07. Retrieved from http://actaenergetica.org/index.php/journal/article/view/462

[5]. Mauricio Rodríguez Segura, Orietta Nicolis, Billy Peralta Márquez, Juan Carrillo Azócar, "Predicting cardiovascular disease by combining optimal feature selection methods with machine learning", 2020 39th International Conference of the Chilean Computer Science Society (SCCC), Year: 2020.

[6]. Gudni Johannesson, & Nazzal Salem. (2022). Design Structure of Compound Semiconductor Devices and Its Applications. Acta Energetica, (02), 28–35. Retrieved from http://actaenergetica.org/index.php/journal/article/view/466

[7]. Sinkon Nayak, Mahendra Kumar Gourisaria, Manjusha Pandey, Siddharth Swarup Rautaray, "Prediction of Heart Disease by Mining Frequent Items and Classification Techniques", 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Year: 2019.

[8]. Kunal Rajput, Girija Chetty, Rachel Davey, "Risk Factors Identification for Heart Disease in Unstructured Dataset using Deep Learning Approach", 2019 International Conference on Data Mining Workshops (ICDMW), Year: 2019.

[9]. Yang Peili, Yin Xuezhen, Ye Jian, Yang Lingfeng, Zhao Hui, Liang Jimin, "Deep learning model management for coronary heart disease early warning research", 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Year: 2018.

[10]. Himanshu Sharma, M A Rizvi Prediction of Heart Disease using Machine Learning Algorithms: A Survey (August 2017).

[11]. Kanna, D. R. K. ., Muda, I. ., & Ramachandran, D. S. . (2022). Handwritten Tamil Word Pre-Processing and Segmentation Based on NLP Using Deep Learning Techniques. Research Journal of Computer Systems and Engineering, 3(1), 35–42. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/39

[12]. L. Verma, S. Srivastava, and P. C. Negi, ''A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data,'' J. Med. Syst., vol. 40, no. 7, p. 178, Jul. 2016.

[13]. C.-J. Chen, Y.-T. Lo, J.-L. Huang, T.-W. Pai, M.-H. Liu, and C.-H. Wang, ''Feature analysis on heart failure classes and associated medications,'' in Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC), Oct. 2016, pp. 1382–1387.

[14]. Therasa Princy R, J. Thomas, " Human Heart Disease Prediction System Using Data Mining Techniques", International Conference on circuit, Power and Computing Technologies [ICCPCT],IEEE,2016.

[15]. Thota, D. S. ., Sangeetha, D. M., & Raj , R. . (2022). Breast Cancer Detection by Feature Extraction and Classification Using Deep Learning Architectures. Research Journal of Computer Systems and Engineering, 3(1), 90–94. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/48

[16]. Sana Bharti, Shailendra Narayan Singh, Amity university, Noida, India Analytical study of heart disease prediction comparing with different algorithms (May 2015).

[17]. N. C. Long, P. Meesad, and H. Unger, ''A highly accurate firefly based algorithm for heart disease prediction,'' Expert Syst. Nov. 2015.

[18]. Chiba, Z., El Kasmi Alaoui, M. S., Abghour, N., & Moussaid, K. (2022). Automatic Building of a Powerful IDS for The Cloud Based on Deep Neural Network by Using a Novel Combination of Simulated Annealing

Algorithm and Improved Self- Adaptive Genetic Algorithm. International Journal of Communication Networks and Information Security (IJCNIS), 14(1). https://doi.org/10.17762/ijcnis.v14i1.5264 (Original work published April 12, 2022)

[19]. G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 6, pp. 1750–1756, 2014.

[20]. Binal A. Thakkar, Mosin I. Hasan, Mansi A. Desai, "Health Care Decision Support System For Swine Flu Prediction Using Naïve Bayes Classifier", International Conference on Advances in Recent Technologies in Communication and Computing,2010.

[21]. Maloth, Bhav Singh. (2016). Privacy-Preserving Scalar Product Computation over Personal Health Records. International Journal of Computer Engineering In Research Trends. 3. 42-46.

[22]. (2022). Bug2 algorithm-based data fusion using mobile element for IoT-enabled wireless sensor networks. Measurement: Sensors. 100548. 10.1016/j.measen.2022.100548.

[23]. Roy, S. S., Mallik, A., Gulati, R., Obaidat, M. S., & Krishna, P. V. (2017, January). A deep learning based artificial neural network approach for intrusion detection. In *International Conference on Mathematics and Computing* (pp. 44-53). Springer, Singapore

[24]. Turovsky, O. L., Vlasenko, V., Rudenko, N., Golubenko, O., Kitura, O., & Drobyk, O. (2022). Two-Time Procedure for Calculation of Carrier Frequency of Phasomodulated in Communication Systems. International Journal of Communication Networks and Information Security (IJCNIS), 13(3). https://doi.org/10.17762/ijcnis.v13i3.5165 (Original work published December 25, 2021)

[25]. Viswanathan, P., & Krishna, P. V. (2013). A joint FED watermarking system using spatial fusion for verifying the security issues of teleradiology. *IEEE Journal of Biomedical and Health Informatics*, *18*(3), 753-764.

[26]. Maloth, Bhav Singh & Lakshmi, M & Kumar, Dr & Parashuram, N. (2017). International Journal on Recent and Innovation Trends in Computing and Communication Improved Trial Division Algorithm by Lagrange"s Interpolation Function. International Journal on Recent and Innovation Trends in Computing and Communication. 5. 1227-1231.