**USE OF ML MODELS TO PREDICT THE TYPE OF ARRYTHMIA OCCURRING**

Vagadia Anand K.1

1 VIT Chennai, Vandalur-Kelambakkam road, India

**Abstract.** Arrythmia is an interruption or disarray in the rhythmic heartbeat of a patient. For context, the heartbeat of an average human being is highly rhythmic, as in a difference of tenth of a second could be classified as an occurrence of arrythmia. However, there are several types of arrythmia. And not all types are dangerous or require defibrillators. Here, a cardio log is used to determine if a person has arrythmia and to which type it belongs. We will be exploring the different accuracies observed in different models that we will be testing out. To have a ML program determine the type of arrythmia can be dangerous however, since such programs can make inaccuracies and this could lead to a wrong treatment being administered to the patient. Here, we will be exploring the following algorithms: KNN Classifier, multiple linear regression, support vector machine, Linear SVC, Kernelized SVC, Gradient boosting classifier. We will also be using Principal Component analysis (PCA) to reduce the dimensionality of the data to increase its accuracy.

**Keywords:** Arrythmia, KNN Classifier, multiple linear regression, support vector machine, Linear SVC, Kernelized SVC, Gradient boosting classifier, Principal Component analysis (PCA), scikit-learn.

1. Introduction
   1. Arrythmia

Arrhythmia refers to an abnormal heart rhythm, where the heart beats too quickly (tachycardia), too slowly (bradycardia), or irregularly. This disruption in the heart's natural rhythm can interfere with its ability to pump blood effectively throughout the body. Arrhythmias can occur in various forms, ranging from harmless to life-threatening. Irregular heart rhythms, known as atrial fibrillation or atrial flutter, can cause chaotic and ineffective contractions of the heart's upper chambers (atria). This can lead to blood pooling, increasing the risk of blood clots, stroke, and other complications. Arrhythmias can result from various factors, including heart disease, electrolyte imbalances, congenital heart defects, medication side effects, and lifestyle choices such as smoking or excessive alcohol consumption. Treatment for arrhythmias depends on their severity and underlying cause, ranging from lifestyle changes and medication to medical procedures like cardioversion, catheter ablation, or implantation of pacemakers or defibrillators. Regular monitoring and management are essential to prevent complications and maintain heart health.

* 1. ECG and other factors in a cardiolog:

In the sample dataset that we have taken for this project, the cardiolog contains 279 attributes about each patient. This is a huge number of attributes. Hence it causes the problem of high dimensionality in the data. Among these attributes, there may be data related to the patient's medical history, such as previous heart conditions, comorbidities, medication usage, lifestyle factors, and results from various diagnostic tests like electrocardiograms (ECGs), echocardiograms, and blood tests. Additionally, the dataset might include detailed information about the specific characteristics of the arrhythmias observed in each patient, such as the type, duration, frequency, and severity of the irregular heart rhythms. Having such a comprehensive dataset allows cardiologists and researchers to perform in-depth analyses to identify patterns, risk factors, and predictive models for different types of arrhythmias. This extensive information can ultimately enhance our understanding of arrhythmia pathophysiology, improve risk stratification, and guide personalized treatment approaches for patients with these cardiac conditions.

* 1. ALGORITHMS USED:

KNN Classifier. K-Nearest Neighbors (KNN) Classifier is a simple and intuitive algorithm used for classification tasks. It assigns a class label to a data point based on the majority class of its k nearest neighbors in the feature space. KNN is non-parametric and does not make any assumptions about the underlying data distribution, making it versatile for various datasets.

Multiple Linear Regression. Multiple Linear Regression is a statistical technique used to model the relationship between multiple independent variables and a single dependent variable. It assumes a linear relationship between the predictors and the response variable and estimates the coefficients of the regression equation using the method of least squares

Support Vector Machine (SVM). Support Vector Machine is a powerful supervised learning algorithm used for classification and regression tasks. SVM aims to find the optimal hyperplane that separates the data points of different classes with the maximum margin. It is effective in high-dimensional spaces and can handle non-linear decision boundaries through the use of kernel functions.

Linear SVC. Linear Support Vector Classifier (Linear SVC) is a variant of the SVM algorithm specifically designed for linearly separable datasets. It seeks to find the optimal hyperplane that separates the classes with a maximum margin while allowing for some misclassification.

Kernelized SVC. Kernelized Support Vector Classifier (Kernelized SVC) extends the capabilities of the traditional SVM by employing kernel tricks to map the input data into a higher-dimensional space where it becomes linearly separable. This allows for the classification of non-linearly separable datasets by finding complex decision boundaries.

Gradient Boosting Classifier. is a powerful ensemble learning technique used for classification tasks. It builds a series of weak learners, typically decision trees, sequentially, with each subsequent tree learning to correct the errors of the previous ones. It works by optimizing a loss function, such as deviance or exponential loss, using gradient descent.

1. NEED OF THE STUDY:
   1. CHALLENGE OF HANDLING THE HIGH DIMENSIONALITY OF THE DATA:

The challenge of handling the high dimensionality of the data in the context of arrhythmia classification is significant due to the vast number of attributes associated with each patient. With 279 attributes per patient, the dataset presents a formidable task in terms of computational complexity, memory usage, and the curse of dimensionality. This high dimensionality not only increases the computational burden but also introduces noise and redundancy in the data, which can hinder the performance of machine learning algorithms.

Principal Component Analysis (PCA) offers a promising solution to address this challenge. PCA is a dimensionality reduction technique that identifies the underlying structure in the data by transforming it into a new coordinate system composed of orthogonal components, called principal components. PCA achieves this transformation by identifying the eigenvectors and eigenvalues of the covariance matrix of the dataset. Eigenvectors represent the directions of maximum variance in the data, while eigenvalues represent the magnitude of variance along those directions. The eigenvectors are used as the new coordinate system (principal components), and the eigenvalues indicate the amount of variance explained by each principal component.

* 1. The Dataset:

There are a total of 3 datasets used in this paper. Each dataset has its own characteristics. Each dataset can be used to make a model that predicts if a person has arrythmia or not. This is quite important as datasets 1 and 3 both produce results which show if a person might have that particular arrythmia, whereas the dataset 2 produces results based on signal analysis and pattern finding.

Dataset1. Link to the dataset: <https://archive.ics.uci.edu/dataset/5/arrhythmia> .The dataset has: 279 attributes, 206 of which are linear valued and the rest are nominal. Columns 11-15 have missing values. Originally used in a study by: H. Altay Guvenir. Dataset creators: H. Guvenir, Burak Acar, Haldun Muderrisoglu, R. Quinlan.It concerns 16 classes, which follow a class label system mentioned on the site. It encapsulates several ECG features and also includes the physical condition of the patient.

**Dataset2**. Link <https://physionet.org/files/mitdb/1.0.0/mitdbdir/mitdbdir.htm#toc> the dataset is a collection ECG signals of 48 patients. It contains annotations of where the arrythmias are occurring and for what duration as well as their types. It can be easily imported and operated upon with the help of WFDB library. It was a contribution by the Harvard-MIT Division of Health Sciences and Technology biomedical engineering center.

### Dataset3. Link to the dataset: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease This dataset is similar to dataset1, wherein it accounts for the physical condition of the patient rather than ECG features. It was originally contributed by CDC and has the physical info of more than 400,000 individual adults. This dataset can be easily used for risk assessment rather than risk identification and classification.

1. Literature review.
   1. KNN algorithm:

(Chen et al., 2023) This research proposes a model for detecting cervical cancer using electronic health records. It introduces a stacked ensemble voting classifier made up of XGBoost, random forest and extra tree classifier integrated with a KNN imputer to handle missing values. The model achieves 99.41% accuracy, 97.63% precision, 95.96% recall and 96.76% F1 score. It significantly outperforms other machine learning algorithms and existing studies.

(Mohanty et al., 2020) This document discusses using machine learning for medical data analysis and diagnosis. It analyzes a dataset containing medical information like age, gender, region, blood group of individuals collected through an online form. Various data visualization and machine learning algorithms like KNN are used to study relationships between different parameters, identify patterns and classify data. The study aims to help develop better diagnostic tools by analyzing existing medical data.

(Turki Aljrees ,2024) reports that cervical cancer develops in the cervix cells and is usually caused by HPV. Research focuses on detecting cervical cancer using machine learning approaches, addressing missing data handling. Various machine learning models and ensemble methods are used for prediction, with high accuracy rates reported. Imputing missing values with KNN improves model performance. SMOTE combined with RF achieved a 96.06% accuracy score.

(Shee et al., 2014) The aim of designing a repository with efficiently classified data for easy data mining is to make it easier to retrieve, organize, update, and manipulate all of the data within its database. It also helps in managing the utilization and maintenance of the database, including a meta database to define application-specific views of the database.

* 1. Multiple Linear Regression algorithm:

(Mukhopadhyay et al., 2020) The research paper focuses on the application of multiple linear regression modeling in the analysis and interpretation of clinical data, utilizing simulation with computer-generated data. The study developed a multiple linear regression model using computer-generated data for a sample size of 40, with one dependent variable and four independent variables. The process involved bivariate correlation, identification of significant variables, testing for multicollinearity, and developing the best-fit model.

(Kim et al., 2021) The study aimed to develop a multiple linear regression model to estimate health-related physical fitness (HRPF) in Korean adults using easy-to-measure variables such as gender, age, body mass index, and percent body fatThe study found that these easy-to-measure dependent variables were able to predict hand grip strength, muscular endurance, and cardiorespiratory fitness in adults, with high explanatory power for the regression models.

(Sidey-Gibbons & Sidey-Gibbons, 2019) We demonstrate the use of machine learning techniques by developing three predictive models for cancer diagnosis using descriptions of nuclei sampled from breast masses. These algorithms include regularized General Linear Model regression (GLMs), Support Vector Machines (SVMs) with a radial basis function kernel, and single-layer Artificial Neural Networks. The trained algorithms were able to classify cell nuclei with high accuracy (.94 - .96), sensitivity (.97 - .99), and specificity (.85 - .94). Maximum accuracy (.96) and area under the curve (.97) was achieved using the SVM algorithm. Prediction performance increased marginally (accuracy=.97, sensitivity = .99, specificity = .95) when algorithms were arranged into a voting ensemble.

(Trunfo et al., 2022) the text discusses the analysis of factors influencing the length of hospital stay (LOS) for patients undergoing laparoscopic appendectomy. The study used multiple linear regression to predict LOS based on variables such as age, complications, and diagnosis. The results showed that age, pre-operative LOS, presence of complications, and complicated diagnosis were the most influential factors. The study suggests the need for a more robust predictive model and future validation with multicenter studies.

* 1. Support Vector Machine (SVM)

(Kampourakia et al., 2013) Automated diagnostics using advanced algorithms are becoming more prevalent in industries worldwide. This paper proposes a web-based architecture for implementing automatic diagnosis techniques transparently to users, focusing on Support Vector Machines (SVMs) for disease analysis, prediction, and classification. The system allows users to define new diseases, train the SVM, and diagnose patients automatically.

(Yu et al., 2010) studied application of support vector machine modeling for prediction of common diseases. The text discusses the use of support vector machine (SVM) techniques as an alternative approach to classify individuals with common diseases like diabetes. The SVM models showed comparable performance to traditional logistic regression methods. SVM allows for better classification in cases with small sample sizes and many variables. The SVM approach appears to perform as well as logistic regression in terms of discriminative abilities.

(Sandhya, 2020) The research paper focuses on using machine learning algorithms, specifically the Support Vector Machine, to predict heart diseases in early stages based on factors like age, sex, and blood pressure. The study compares the accuracy of the SVM algorithm with other machine learning algorithms and suggests that SVM provides better results. Controllable factors like smoking, drinking, weight, blood pressure, and cholesterol can help reduce the risk of heart diseases.

(Seth et al., 2020) Lung cancer is a major cause of cancer-related deaths worldwide, primarily linked to long-term cigarette use. Machine learning technology, such as Particle Swarm Optimization, is being used to classify and detect lung cancer early. Data mining techniques are also employed to analyze health data and improve predictive modeling. The integration of PSO with SVM shows promise in detecting lung cancer at an early stage.

* 1. Linear SVC:

(Hiremath & Patil, 2022) The process is carried out on a drug review data in order to determine whether the patient’s behaviour towards a medicine, product, treatment etc is positive, negative or neutral using NLP techniques. The prominent statistical sklearn models used are support vector machines (SVM), Random Forest Classification, LinearSVC, MultinomialNB. SVM algorithm is found to perform better compared to other in terms of accuracy.

(Garg, 2021) It also presents a study on sentiment analysis of drug reviews to build a recommender system using various machine learning classifiers. LinearSVC was found to be the best algorithm with an AUC score of 90.7%. The research evaluated different metrics and found that Linear SVC on TF-IDF outperformed other models with 93% accuracy.

(Khushi et al., 2021) The study focused on class-imbalanced datasets and found that using class-imbalance techniques improved performance compared to baseline models. Random forest with random over-sampling was identified as the best predictive model for lung cancer datasets. The research demonstrated that class imbalance techniques can aid in diagnosing lung cancer and that over-sampling was effective for imbalanced medical data

(Abdar et al., 2019) Recent advances in artificial intelligence have led to the development of new intelligent automatic systems. A new hybrid ensemble learning model called Nested Ensemble nu-SVC (NE-nu-SVC) was introduced for Clinical Decision Support Systems. This model combines nu-SVC with other machine learning techniques to achieve high accuracy rates for CAD datasets.

* 1. kernelized SVC:

(Subashini, A et al., 2024) This paper compares Support Vector Machines (SVM), Least Square SVM (LSSVM), Relevance Vector Machine (RVM), and Probability Classification Vector Machine (PCVM) on sixteen binary and multiclass medical datasets. Special emphasis is placed on comparing the commonly used Gaussian radial function (GRBF) kernel with the relatively new generalized min-max (GMM) kernel and exponentiated GMM (eGMM) kernel. SVM-GMM achieves an accuracy of 98.92% on the thyroid disease dataset of 7,200 samples.

* 1. Gradient Boosting Classifier:

(Theerthagiri, 2022) Gradient boosting is an efficient approach for illness prediction and disease classification. A study introduces a recursive feature elimination-based gradient boosting method for accurate cardiac disease prediction, achieving high accuracy. This method can assist in predicting and treating cardiovascular disease.

(Subashini & Raju, Year) The article discusses an IoT-based heart disease diagnosis system using Gradient Boosting and Deep Convolution Neural Network. The original article can be found online at the provided DOI link.

(Sai et al., 2023) Diabetes is a metabolic condition primarily caused by aberrant insulin secretion. Various ensemble models using k-NN, Naïve Bayes, LightGBM, Adaboost, and Random Forest algorithms were implemented for diabetes detection. The LightGBM + k-NN + Adaboost technique showed the highest detection accuracy of 90.76%.

(Hama Saeed, 2023) Diabetes is a widespread disease caused by a lack of insulin hormone. Machine learning techniques, particularly the extra trees classifier, have shown high accuracy in predicting diabetes. Various machine learning models have been proposed for classifying diabetes type 2 using datasets like PIMA and BRFSS. The extra trees classifier has outperformed other models with an accuracy of 96% for PIMA and 99% for BRFSS.

* 1. Principal Component Analysis (PCA):

(Pham et al., 2023) The proposed SAA is efficient for SCA detection with a small number of extracted features and relatively high diagnosis performance, such as accuracy of 99.52%, sensitivity of 97.69%, and specificity of 99.91%. Machine learning (ML) and deep learning (DL) in combination with advanced signal processing techniques have been widely proposed for SAA design in terms of AED performance improvement.

(Kong et al., Year) Healthcare quality assessment is essential for proper resource allocation and standardizing medical practice. A new hybrid method combining PCA and the ER approach is proposed for assessing healthcare quality based on patient experience. The method involves numerical scoring, PCA to identify item relationships, and ER to aggregate patient evaluations.

(Wanzala et al., 2019) Strong health systems are essential platforms for accessible, quality health services, and population health and attainment of the Sustainable Development Goal (SDGs). The Kenyan provides that essential health service delivery is assigned to county governments while the national government retains health policy, technical assistance, management of national referral health facilities and human resources.

(Jackson et al., 2015) Service Provision Assessment surveys have been conducted in East and South Asia and sub-Saharan Africa to assess primary health care and family planning clinical readiness. Principal Components Analysis is used to identify common indicators and create composite indices for comparative analysis. The Tanzania Service Provision Assessment survey evaluated healthcare capacity in Tanzania. The study aims to link indices with healthy facility survey data for further analysis.

1. METHODOLOGY
   1. KNN Classifier:

K-Nearest Neighbors (KNN) Classifier is a simple and intuitive algorithm used for classification tasks. It assigns a class label to a data point based on the majority class of its k nearest neighbors in the feature space. KNN is non-parametric and does not make any assumptions about the underlying data distribution, making it versatile for various datasets. The choice of distance metric should be based on the characteristics of the dataset and the problem at hand. The formula for computing the Euclidean distance between two data points xi and xj n-dimensional space is given by:

* 1. Multiple Linear Regression:

Multiple Linear Regression is a statistical technique used to model the relationship between multiple independent variables and a single dependent variable. It assumes a linear relationship between the predictors and the response variable and estimates the coefficients of the regression equation using the method of least squares.

In Multiple Linear Regression, the model assumes that the relationship between the dependent variable Y and p independent variables. X1, X2 … Xp is linear and can be represented the equation:

Here, we are assuming that the relationship between the variables is linear, and the variance of error is constant across all levels of the independent variables.

* 1. Support Vector Machine (SVM):

Support Vector Machine is a powerful supervised learning algorithm used for classification and regression tasks. SVM aims to find the optimal hyperplane that separates the data points of different classes with the maximum margin. It is effective in high-dimensional spaces and can handle non-linear decision boundaries through the use of kernel functions.

One of the key strengths of SVM lies in its ability to handle high-dimensional feature spaces efficiently, making it particularly suitable for datasets with a large number of features. Moreover, SVM can accommodate non-linear decision boundaries by transforming the input features into a higher-dimensional space through the use of kernel functions. These kernel functions allow SVM to implicitly map the input data into a higher-dimensional space, where linear separation may be feasible, even if the original data is not linearly separable in the input feature space. Here is the formula for the decision function:

Here, we will be using a soft margin SVM, as the data is non-linearly separable.

* 1. Linear SVC:

Linear Support Vector Classifier (Linear SVC) is a variant of the SVM algorithm specifically designed for linearly separable datasets. It seeks to find the optimal hyperplane that separates the classes with a maximum margin while allowing for some misclassification.

* 1. Kernelized SVC:

Kernelized Support Vector Classifier (Kernelized SVC) extends the capabilities of the traditional SVM by employing kernel tricks to map the input data into a higher-dimensional space where it becomes linearly separable. This allows for the classification of non-linearly separable datasets by finding complex decision boundaries. This optimization process involves determining the optimal coefficients (weights) for the linear decision function, as well as the bias term, through the method of least squares or other optimization techniques.

Linear SVC is particularly suitable for datasets with a large number of features and a linearly separable class distribution. Its simplicity, efficiency, and effectiveness make it a popular choice for various classification tasks, especially in scenarios where interpretability and computational efficiency are important considerations. Despite its focus on linear decision boundaries, Linear SVC remains a powerful and widely-used algorithm in the realm of supervised learning, providing a solid foundation for building classification models on linearly separable datasets.

* 1. Gradient Boosting Classifier:

It is a powerful ensemble learning technique used for classification tasks. It builds a series of weak learners, typically decision trees, sequentially, with each subsequent tree learning to correct the errors of the previous ones. It works by optimizing a loss function, such as deviance or exponential loss, using gradient descent.

One of the key strengths of Gradient Boosting Classifier lies in its ability to handle heterogeneous data types and nonlinear relationships effectively. Moreover, by leveraging the boosting technique, which focuses on learning from the mistakes of previous models, Gradient Boosting Classifier can achieve superior predictive performance compared to individual weak learners.

* 1. Pre-processing dataset3:

We used the wfdb library to import the data from its online repository. Once the data was imported, we converted it to a dataframe which had the signal info in it. This signal was in a raw, noisy format. To reduce the noise on these signals, we used nuerokit2 python library. Nuerokit2 is a powerful python tool to process, analyze and even simulate various types of bio-signals.

We used the nuerokit2 library to extract features of ECGs. However, before that, we extracted the relevant ECGs from the data, put each one in a separate dataframe and created a dictionary of signals. The signals that we extracted were in the duration of the annotations given in the data for e.g. patient0 has tachycardia at 25:13. So we extracted that signal for a duration of 5 seconds. Then we used the nuerokit library to extract the important features, such as r\_peaks, q\_intervals etc. Then we took the mean for each of the signals. This is done to have the data in a format that can be easily classified by a machine learning algorithm.

1. RESULTS AND OBSERVATION:
   1. Random forest classifier on dataset1 and dataset3:

We applied the Random Forest algorithm on both dataset1 and dataset3. This is because both of the datasets had a huge number of attributes, and random forest is very useful in classifying such data. We also applied PCA on both the datasets and then evaluated the model. This showed some interesting results:

**Table 1.** Table of Test-accuracies fir Datasets 1 and 3

|  |  |  |
| --- | --- | --- |
|  | Without PCA | With PCA |
| Dataset1 | 0.74 | 0.68 |
| Dataset3 | 0.91 | 0.91 |

**Fig. 1.** Test accuracy of risk of Arrythmia.

* 1. Dataset2:

We applied all the previously discussed algorithms on this dataset. The dataset is a collection of ECG signals, and not a simple dataset such as dataset1 or dataset3. The results for the accuracy obtained even after extensive hyperparameter tuning were unacceptably low. We also applied PCA to the dataset, but to no avail:

**Table 2.** Table of accuracies for ML models on Signal Classification:

|  |  |  |
| --- | --- | --- |
| Model No. | Without PCA | With PCA |
| KNN | 0.18 | 0.18 |
| Linear SVC | 0.05 | 0.17 |
| Kernelized SVC | 0.03 | 0.17 |
| Gradient Boosting | 0.21 | 0.13 |
| Random Forest | 0.16 | 0.15 |

**Fig. 2.** Accuracies in classification processed signals

* 1. Conclusions:

In this project, we evaluated three datasets using machine learning techniques, with one dataset being signal-based. Surprisingly, the accuracies obtained for the signal-based dataset were notably low compared to the other datasets.

Upon careful analysis, it became apparent that the complexity and variability inherent in signal data pose unique challenges for traditional machine learning algorithms. Unlike structured or tabular data, signal data often exhibits intricate patterns and nuances that may not be effectively captured by conventional models. Consequently, despite our best efforts, the performance of the machine learning models on the signal-based dataset fell short of expectations.

To address this challenge, future projects should consider employing deep learning methodologies, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), which are better suited for analyzing signal data and can potentially yield higher accuracies. By leveraging neural network architectures such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), we can potentially unlock deeper insights and achieve higher accuracies on signal-based tasks.

References

1. Chen, X., Aljrees, T., Umer, M., Saidani, O., Almuqren, L., Mzoughi, O., Ishaq, A., & Ashraf, I. (2023). Cervical cancer detection using K nearest neighbor imputer and stacked ensemble learning model. Digital health, volume 9: 1-16
2. S. Mohanty, A. Mishra, A. Saxena, H. Shee, & J. Kenyatta (Eds.) (2020), Medical Data Analysis Using Machine Learning with KNN (No. 3580). EasyChair.
3. Aljrees T (2024) Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning. PLoS ONE 19(1): e0295632. <https://doi.org/10.1371/journal.pone.0295632>
4. Shee, Hassan & Cheruiyot, Wilson & Kimani, Stephen. (2014). Application of k-Nearest Neighbour Classification in Medical Data Mining.
5. Mukhopadhyay, K., Singh, R., Dalai, C. K., Ahmed, S. N., & Banerjee, K. (2020). Understanding the application of multiple linear regression model in health care research using simulation with computer-generated data. Journal of The West Bengal University of Health Sciences, Vol 1 Issue 2
6. Sidey-Gibbons, J., Sidey-Gibbons, C.(2019) Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19, 64 (2019). <https://doi.org/10.1186/s12874-019-0681-4>
7. Kim, Sung-Woo & Park, Hun-Young & Jung, Hoeryong & Lee, Jinkue & Lim, Kiwon. (2021). Estimation of Health-Related Physical Fitness Using Multiple Linear Regression in Korean Adults: National Fitness Award (NFA) 2015-2019. Frontiers in Physiology. 12. 668055. 10.3389/fphys.2021.668055.
8. Trunfo, T. A., Scala, A., Giglio, C., Rossi, G., Borrelli, A., Romano, M., & Improta, G. (2022). Multiple regression model to analyze the total LOS for patients undergoing laparoscopic appendectomy BMC Medical Informatics and Decision Making (2022) 22:141, <https://doi.org/10.1186/s12911-022-01884-9>
9. Kampourakia, A., Vassisa, D., Belsisb, P., & Skourlasa, C. (2013). e-Doctor: A Web Based Support Vector Machine for Automatic Medical Diagnosis.
10. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010 Mar 22;10:16. doi: 10.1186/1472-6947-10-16. PMID: 20307319; PMCID: PMC2850872.
11. Sandhya, Yamala. (2020). Prediction of Heart Diseases using Support Vector Machine. International Journal for Research in Applied Science and Engineering Technology. 8. 126-135. 10.22214/ijraset.2020.2021.
12. Jyotsna Seth, Dr. Parma Nand, Prabhdeep Singh, Dr. Rajbir Kaur. (2020). PARTICLE SWARM OPTIMIZATION ASSISTED SUPPORT VECTOR MACHINE BASED DIAGNOSTIC SYSTEM FOR LUNG CANCER PREDICTION AT THE EARLY STAGE. PalArch’s Journal of Archaeology of Egypt / Egyptology, 17(9), 6202 – 6219
13. Basavaraj N. Hiremath, Malini M. Patil, Enhancing Optimized Personalized Therapy in Clinical Decision Support System using Natural Language Processing, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 6, Part A,2022
14. Garg, Satvik. (2021). Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning.
15. Khushi, Matloob & Shaukat Dar, Kamran & Mahboob Alam, Talha & Hameed, Ibrahim & Uddin, Shahadat & Luo, Suhuai & Yang, Xiaoyan & Reyes, Maranatha. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access. 10.1109/ACCESS.2021.3102399.
16. Abdar, Moloud & Acharya, U Rajendra & Sarrafzadegan, Nizal & Makarenkov, Vladimir. (2019). NE-nu-SVC: A New Nested Ensemble Clinical Decision Support System for Effective Diagnosis of Coronary Artery Disease. IEEE Access. 7. 167605 - 167620. 10.1109/ACCESS.2019.2953920.
17. Qian, X., Zhou, Z., Hu, J., Zhu, J., Huang, H., & Dai, Y. (Year). A comparative study of kernel-based vector machines with probabilistic outputs for medical diagnosis. Biocybernetics and Biomedical Engineering, Volume(Issue), Page range.
18. Theerthagiri, Prasannavenkatesan. (2022). Predictive Analysis of Cardiovascular Disease using Gradient Boosting based Learning and Recursive Feature Elimination Technique. Intelligent Systems with Applications. 16. 200121. 10.1016/j.iswa.2022.200121.
19. Subashini, A., & Raju, P. K. (Year). Correction: An IoT Based Heart Disease Diagnosis System Using Gradient Boosting and Deep Convolution Neural Network. Publisher.
20. Sai, M. & Chettri, Pratiksha & Panigrahi, Ranjit & Garg, Amik & Bhoi, Akash Kumar & Barsocchi, Paolo. (2023). An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. International Journal of Computational Intelligence Systems. 16. 14. 10.1007/s44196-023-00184-y.
21. Hama Saeed, Mariwan. (2023). Diabetes type 2 classification using machine learning algorithms with up-sampling technique. Journal of Electrical Systems and Information Technology. 10. 10. 10.1186/s43067-023-00074-5.
22. Pham, VS., Nguyen, A., Dang, H.B. et al. Diagnosis of sudden cardiac arrest using principal component analysis in automated external defibrillators. Sci Rep 13, 8768 (2023). https://doi.org/10.1038/s41598-023-36011-9
23. Kong, G., Jiang, L., Yin, X., Wang, T., Xu, D.-L., Yang, J.-B., & Hu, Y. (Year). Combining principal component analysis and the evidential reasoning approach for healthcare quality assessment. Publisher.
24. Wanzala, Maximila & Oloo, James Aggrey & Ochieng, Nguka. (2019). Application of Principal Component Analysis to Assess Health Systems Capacity Using Cross Sectional Data in Rural Western Kenya. 7. 27-32. 10.12691/ajphr-7-1-5.
25. Jackson EF, Siddiqui A, Gutierrez H, Kanté AM, Austin J, Phillips JF. Estimation of indices of health service readiness with a principal component analysis of the Tanzania Service Provision Assessment Survey. BMC Health Serv Res. 2015 Dec 3;15:536. doi: 10.1186/s12913-015-1203-7. PMID: 26634449; PMCID: PMC4669619.