|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Citation | Dataset | Goal | Outlier Detection method | Outlier( more details) | ML Models |
| (Finnegan et al., 2021) | MIMIC III | Assess circadian rhythm of BP | Outliers were detected by comparing each PAT value to the median of a 30-second running window with a step size of 25 seconds | -patients are divided by age group  -focused on circadian rhythms  and BP dip status  -dip ratio (DR) of a vital sign computed as the  percentage increase from mean daytime,10 %  -accuracy of using vital sign DRs was assessed by a receiver operating characteristic  (ROC) curve | - |
| (Fitriyani et al., 2019) | type 2 diabetes, hypertension, prehypertension, Chronic Kidney Disease (CKD) | Prediction for type 2 diabetes and hypertension based on individual’s risk factors | isolation forest (iForest) based outlier detection method | - method is applied to the dataset as a whole  -an ensemble of isolation  trees (iTrees) for each dataset where outliers were defined as instances with short average length in the iTrees. The iTrees  is then recursively created by dividing the dataset until all  instances is isolated or specific tree height is achieved | MLP, SVM, Decision Tree, Logistic Regression, K-Means + logistic regression, DBSCAN+SMOTE+RF, proposed disease prediction model (DPM) |
| (Ijaz et al., 2018) | Uses diabetes, hypertension, and Chronic Kidney Disease (CKD) as dataset.  Hypertension dataset- Increased blood pressure by BMI, WC, and HC, and WHR on male subject | Hybrid prediction model for Diabetes and Hypertension based on risk factors. | Density-based spatial clustering of applications with noise (DBSCAN) | * Outlier is a distance based similar to k nearest neighbours. * Does not apply to patient level but as a dataset overall | Random Forest |
| (Kannan & Raj, 2019) | blood pressure data | Compare performance of outlier labelling methods | SD method, Median method, MADe method, Z−Score, Modified Z−Score, Tukey's method | - not specified whether data has multiple patient | - |
| (Kannan et al.,2015) | diabetes data from the primary health center in Tirunelveli | outlier labelling technique and issues of outlier identification | Z-Score, Modified  Z-Scores, Median Absolute Deviation (MADe) and Tukey Method (Boxplot). | -outlier is done on one patient with 50 records | - |
| (Liu & Hauskrecht, 2021) | MIMIC III dataset | detecting outliers in continuous-time  event sequences | CPPOD: point-process model using the history of both the target events and the context events as the context by using model adapted from the continuous-time LSTM | - uses a medication known as norepinephrine as a target arterial bp systolic and non invasive bp systolic blood pressure as vital signs  - all the measurements are from one patient admission in one event sequence with 500 sequences for above measurement  -50% training | - |
| (Nagavelli et al., 2022) | heart disease data | CDSS for cardiovascular diseases | DBSCAN | * Dbscan is applied to all not specifically to any feature * Bp pressure is using EKG instead of measurements | Naïve Bayes-weighted approach, 2SVM+XGBoost,SVM+DO,XGBoost |
| (Ranjeeth et al., 2020) | Pima  Indian diabetes dataset (PIDD) | Classify Type-I Diabetes, Type- II Diabetes | Radial basis function (RBF) | * Used as a node in nn to remove outlier * Classification with target, no patient level | Naïve Bayes, SVM, Random Forest, Decision Tree,MLP, Stochastic Gradient Descent |
| (Sarangi & Tripathy, 2022) | MIMIC dataset | Outlier detection in Wireless Sensor Network | Spearman's rank correlation and simple linear regression | - data of one patient, with 1000 records | - |
| (Zardkoohi & Fatemeh Molaeezadeh, 2022) | D12, D15, and DUNN  PhysioNet/CinC challenge 2009 used for blood pressure prediction | Event prediction | fuzzy clustering algorithms including Fuzzy C-Means  (FCM), Kernel-based FCM (KFCM), Density-oriented KFCM (DKFCM), and Gustafson–Kessel (GK)  [consists of outlier detection with threshold value and clustering] | * Not patient level but patients are separated into groups which already has targeted events * Use distance based nearest neighbour | Adaptive Neuro-Fuzzy Inference System (ANFIS) based on DKFCM clustering (DKFCM-ANFIS) |

MIMIC - 'Medical Information Mart for Intensive Care'

Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, *10*(2), 231-238.

Liu, S., & Hauskrecht, M. (2021, July). Event outlier detection in continuous time. In *International Conference on Machine Learning* (pp. 6793-6803). PMLR.

Dashdondov, K., & Kim, M. H. (2021). Mahalanobis Distance Based Multivariate Outlier Detection to Improve Performance of Hypertension Prediction. *Neural Processing Letters*. https://doi.org/10.1007/s11063-021-10663-y

Estiri, H., & Murphy, S. N. (2019). Semi-supervised encoding for outlier detection in clinical observation data. *Computer Methods and Programs in Biomedicine*, *181*. https://doi.org/10.1016/J.CMPB.2019.01.002

Finnegan, E., Davidson, S., Harford, M., Jorge, J., & Villarroel, M. (2021). The presence of a circadian rhythm in pulse arrival time and its application for classifying blood pressure night-time dip. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 488–491. https://doi.org/10.1109/EMBC46164.2021.9629883

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*. https://doi.org/10.1109/ACCESS.2019.2945129

Ijaz, M. F., Alfian, G., Syafrudin, M., & Rhee, J. (2018). Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest. *Applied Sciences (Switzerland)*, *8*(8). https://doi.org/10.3390/app8081325

Kannan, K. S., & Raj, S. S. (2019). *Outlier Labeling Methods for Medical Data*. 67–75. https://doi.org/10.1007/978-981-13-0872-7\_6

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/7351061

Ranjeeth, S., Kandimalla, V. A. K., & Reddy, G. (2020). Predicting Diabetes Using Outlier Detection and Multilayer Perceptron with Optimal Stochastic Gradient Descent. *Proceedings - 2020 IEEE India Council International Subsections Conference, INDISCON 2020*, 51–56. https://doi.org/10.1109/INDISCON50162.2020.00023

Sarangi, B., & Tripathy, B. (2022). Outlier detection in Wireless Sensor Network for Health Care: A statistical approach. *2022 International Conference for Advancement in Technology, ICONAT 2022*. https://doi.org/10.1109/ICONAT53423.2022.9725817

Zardkoohi, M., & Fatemeh Molaeezadeh, S. (2022). Long-term prediction of blood pressure time series using ANFIS system based on DKFCM clustering. *Biomedical Signal Processing and Control*, *74*, 103480. https://doi.org/10.1016/J.BSPC.2022.103480