Ensemble Model for the Prediction of Hypertension using KNN and SVM Algorithms

Saadatu Ali Jijji
Department of Computer Science
Federal University of Kashere,
Gombe State, Nigeria

Asabe Ahmadu Sandra
Department of Computer science
Modibbo Adama University of
Technology, Yola, Nigeria

Malgwi Yusuf Musa
Department of Computer Science
Modibbo Adama University of
Technology, Yola, Nigeria

ABSTRACT
Hypertension also known as high blood pressure is a dangerous illness because it can lead to strokes, heart disease, heart failure, kidney problem and many more ailment, but when hypertension is detected early it can be prevented or controlled. Thus an intelligent and accurate system is in need for early prediction. Data mining applied to medical field provide innovative results and when two data mining techniques are combined a better performance and more accurate model was developed. A model for the prediction of hypertension in patient using Hybrid data mining technique was developed using hyper-parameter tuning and ensemble method. The model was based on hypertension data set collected from Federal teaching hospital and specialist hospital Gombe state Nigeria. The dataset was further preprocessed and standardized by scaling, fixing missing values and fixing imbalanced data using SMOTE. Grid search technique was using for hyper-parameter tuning. KNN, SVM and Naïve Bayes was used in the model before applying the ensemble technique on KNN and SVM which Gradient Boosting has the accuracy of 0.9985.

Keywords
Hypertension, Data mining, Ensemble technique, Hyper-parameter tuning, SMOTE, Accuracy

1. INTRODUCTION
High blood pressure, commonly known as hypertension, can cause serious health problems and raise the risk of heart disease, stroke, and even death. The force exerted on the walls of a person's blood vessels is known as blood pressure. The resistance of the blood arteries and how hard the heart needs to work determine the pressure [1]. In Nigeria, the majority of adults are at risk. The majority of adults in Nigeria are at risk of hypertension, which many are unaware of. Hypertension is also a major risk factor for cardiovascular disease. Keeping blood pressure under control lowers the risk of most hypertension-related illnesses. Hypertension, also known as high blood pressure, is a common type of illness defined by the American Heart Association's new hypertension guidelines as any systolic blood pressure (BP) measurement of 130mm Hg or higher—or any diastolic BP measurement of 80mm Hg or higher. Age, sex, ethnicity, alcohol and cigarette use, family history, existing health condition, individual lifestyle, high fat diet, salt rich, high potassium are all risk factors for hypertension. High blood pressure, or hypertension, is harmful because it can cause strokes, heart attacks, heart failure, kidney illness, and a variety of other diseases [2].

Data mining is a procedure or process for displaying meaningful patterns from massive amounts of data; it is also known as the Knowledge Discovery process, data mining, knowledge extraction, or pattern analysis. The purpose of data mining is to locate data that is typically unknown; once the undiscovered pattern is discovered, it can be used in decision-making in a variety of fields, including medicine, marketing, and business [3]. Classification, clustering, regression, association rules, outer detection, sequential pattern, and prediction are some of the data mining approaches.

Marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining, and mobile computing have all used data mining to aid in decision making. Data mining has recently been used to successfully discover healthcare fraud and abuse incidents. According to [4], data mining is synonymous with knowledge discovery from data, and it is one of the most important processes in knowledge discovery. Data cleansing (to remove noise and inconsistent data), data integration (to merge numerous data sources), and data selection (where data relevant to the analysis task are retrieved from the database), Data transformation (the process of transforming and consolidating data into forms suitable for mining through summary or aggregate procedures), and data mining (an essential process where intelligent methods are applied to extract data patterns), Pattern evaluation (using interestingness measurements to find the truly fascinating patterns that represent knowledge), Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users). The preceding illustration of the knowledge discovery process demonstrates that data mining is only one phase in the overall process.

Machine learning is a branch of artificial intelligence, just as artificial intelligence is a part of computer science. Machine learning is defined as a program's ability to learn on its own through experience, converting experience into expertise or detecting meaningful patterns in complex data. Machine learning does not aim to create an automated version of intelligent behavior; rather, it aims to use computers' special abilities to supplement human intelligence, often by performing tasks that are beyond human capabilities. For example, the ability to scan and process large data sets allows machine programs to detect patterns that are beyond human perception. Speech recognition, pattern recognition, text mining, knowledge and data mining, face recognition, and other applications of machine learning [5].

Hypertension is dangerous because it can lead to strokes, heart disease, Heart failure,kidney problem and many more ailment, [2], Data mining applied to medical field provide innovative results and when two data mining techniques are combined together a better performance and more accurate model will be developed, [6].

Researches like [7], [8], [9] and more have showed that when

predicted early it can be prevented and diagnosed.

2. REVIEW OF RELAED WORK

The aim of this work is develop a model that will best predict hypertension using data mining techniques. So much work related to this study has been done using several techniques and tools among which are; according to [10] used artificial neural network which is a machine learning technique to design a hypertensive predictive model from clinical dataset, the performance of the model was measured in Matlab achieving 82% accuracy.

Design a hypertensive predicting system using naïve bayes algorithm, 10 fold cross validation and the performance analysis on WEKA environment, the model has the accuracy of 83.6% as compared to decision tree with 77% accuracy [11]

Application of a mobile chronic disease management system by [12] Hadoop, Spark, and data mining, as well as a C4.5 classification and support vector machine were used in this investigation

[13] Describe how they used data mining techniques to model hypertension and hyperlipidemia. They employed logistic regression analysis, C5.0, CHAID, Exhaustive CHAID, and discriminant analysis to find risk variables for hypertension and hyperlipidemia. Their work has made the most significant contribution in determining the common risk factors for these two illnesses. The generated models correctly classify participants into one of four physiological states with a 93.07% accuracy.

In their paper Hybrid prediction model for type 2 diabetes and hypertension utilizing DBSCAN-based outlier identification, synthetic minority over sample method (SMOTE), and Random forest by [14]. The HPM performed better, with recall and specificity of 70.270% and 82.203% respectively

[15] compares the efficacy of several machine learning algorithms in predicting persons at risk of developing hypertension and who are likely to gain the most from interventions. LogitBoost (LB), Bayesian Network classifier (BN), Locally Weighted Naive Bayes (LWB), Artificial Neural Network (ANN), Support Vector Machine (SVM), and Random Tree Forest (RTF) were the algorithms used.

The research focus on designing an expert system for diagnosing hypertension perform by range of age participant, the first step is fuzification the rule evaluation and aggregation of the third rule, finally defuzzification, FIS tool in MATLAB was used in designing the system, the study was conducted by[7].

Design of Fuzzy expert system and a multi-layer neural network system for hypertension detection. Fuzzy expert system with two input and one output and also three membership function, also multi-layer neural network 5 input, 5 hidden layer and 1 output for the prediction of hypertension [9].

This study uses two decision tree algorithm to predict hypertension, 10 fold cross validation was used to trained and test the model, The Waikato Environment for Knowledge Analysis (WEKA) was used to simulate the prediction model, With a precision of 100%, the ID3 surpassed the C4.5, which had a precision of 86.36% [8].

A study of predictive modeling of blood pressure during hemodialysis: a comparison of linear model, random forest, support vector, regression, XGBoost, LASSO regression and ensemble method. According to the study for future SBP prediction, the ensemble technique (R2=0.50, RMSE=16.01, MAE=11.97) performed the best[16].

Using three supervised machine learning algorithms: decision tree, random forest, and logistics regression with 5-fold cross-validation, predictive models of hypertension were built and compared, the study was carried based on 987 record from Qatar biobank. Stata and weka were utilized for the analysis of the model. Random forest happens to be the best base on their study [17].

This paper focuses on a neural network classification model to evaluate the relationship between gender, race, BMI, age, smoking, renal disease, and diabetes. It also demonstrates that artificial neural network techniques applied to large clinical data sets can provide a useful data-driven approach for categorizing patients for population health management, as well as support in the control and detection of hypertensive patients, which is one of the key risk factors for heart disease. This study used an unbalanced data set of 24,434 patients, with non-hypertensive patients accounting for 69.71 percent and hypertensive patients accounting for 30.29 percent. The sudy also has a good performance [18].

3. METHODOLOGY

3.1 Hardware and Operating System

Windows 10 pro was the operating system used with Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz 2.50 GHz processor, 64 bit operating system type.

3.2 Software tool

The hypertension prediction model was implemented using a Python application that runs on the Jupyter Integrated Development Environment (IDE), the software was used to design the model as well as measure the performance of the model.

3.3 Research Model

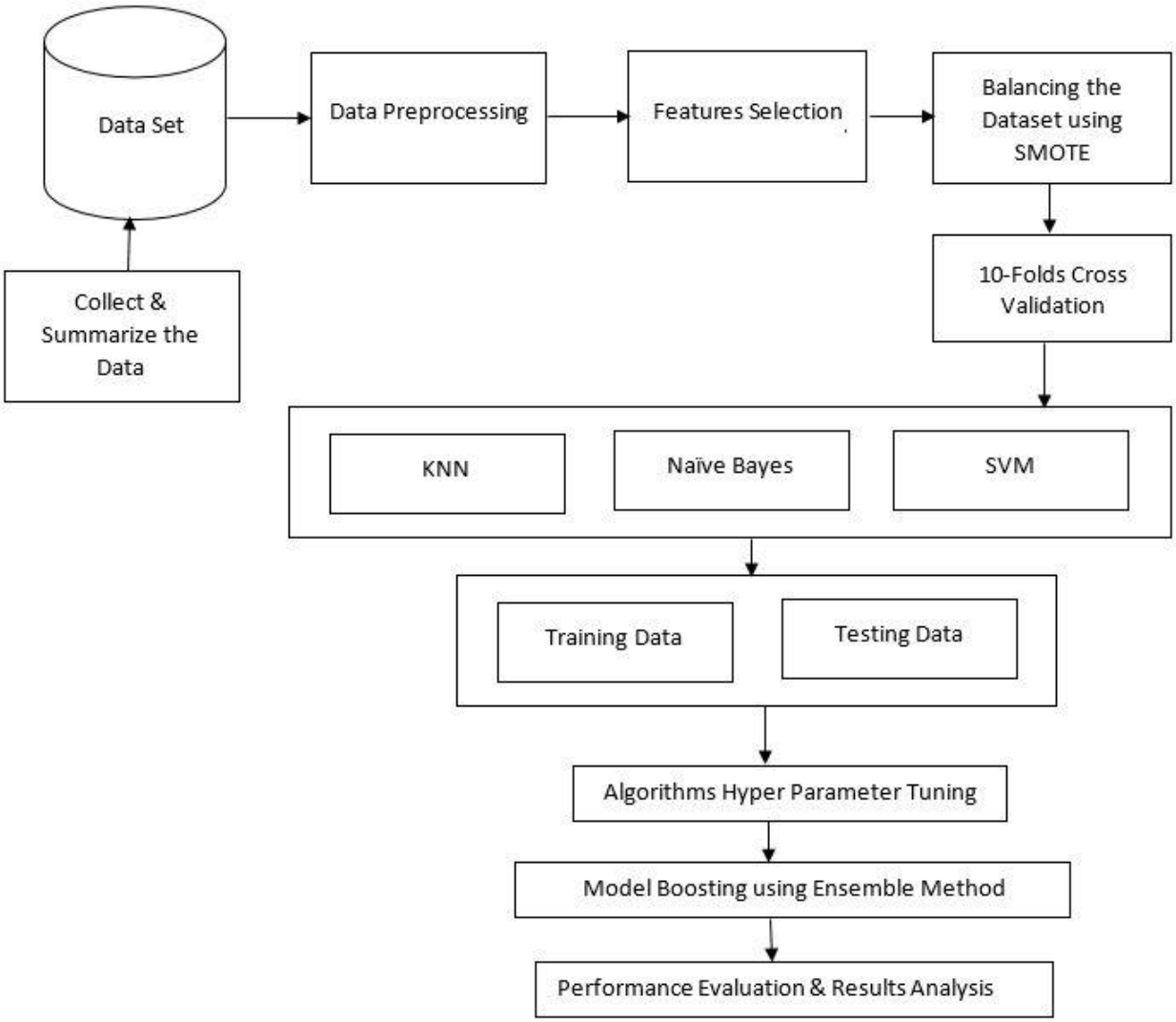

Fig 1: Block diagram for the ensemble algorithms process

3.4 Source of Data collection

A total of 499 data was collected for the purpose of the study which shows patients with elevated blood pressure and also patient with normal blood pressure. Medical history on hypertensive patient record from federal teaching Hospital Gombe was used in training the model and the record from specialist Hospital Gombe was used in testing the proposed model.

3.5 Dataset

The data has 499 instances with 20 attributes, all the attributes are of numerical types that is elevated has been mapped to 1 and normal has been mapped to 0, also gender male has been mapped to 1 for male and 0 for female, all the missing values problem has been fixed by replacing them with the mean values and also fixing data imbalance using the smote technique. The data was further divided into training and testing data.

I. Target attribute

Recommendation is the target attributes, based on the input supplied by the patient the model will decide if the blood pressure is Elevated or Normal. The Target variable is dependent on the input variable

II. Input attribute

The input variable consist of the following;

1. Age – in years
2. Gender
3. Weight
4. Systolic blood pressure (SBP)
5. Diastolic Blood Pressure (DBP)
6. History – if the patient has a family history of hypertension
7. Physical changes
8. Mouth and voice – that is if the patient speech is impaired or the mouth has change shape
9. Fever
10. Lifestyle – exercise, diet, etc.
11. Shortness of breath
12. Body pain
13. Nose bleed

14. Flushing
15. Chest pain
16. Visual pain
17. Blood in urine
18. Epigastric pain
19. Other diseases – like diabetes, heart disease etc.

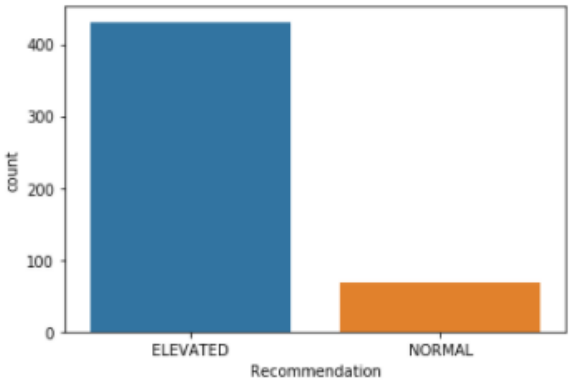

Fig 2: Distribution of the data in bar plot

3.6 Techniques Used

The Hypertension data set was tested and trained with K Nearest Neighbor Algorithm, Supper vector Machine and Naïve Bayes algorithm were used in this study, Also Hyper-parameter tuning and ensemble techniques were also used to boost the classification algorithm employed.

3.7 Performance Matrix

Confusion Matrix was used to measure the performance of the model using the following parameter Accuracy, Specificity, Sensitivity, error rate and precision.

4. RESULT AND DISCUSSION

The dataset is divided into validation and testing split, 80% of the data is for validation (training) the model and 20% of the data is for testing the model. For the validation its uses 10 fold cross validation method on all the algorithms that is Naïve Bayes, K-Nearest Neighbor and Support Vector Machine. Table 1 shows the accuracy of the model after testing the algorithms.

Table 1. Initial accuracy of the algorithm before tuning

Algorithm	Accuracy Score %
KNN	81
Naïve Bayes	90
Support Vector Machine	**97**

4.1 Standardizing the data set

Due to the fact that our dataset has different dimensions for many of the attributes, standardizing the dataset is required so that the features can be on the same scale, and standard scalar was used because it is best suits our dataset as we are using distance based algorithms. Table 2 shows the accuracy of the model after scaling using pipelining.

Table 2. Scaled algorithm accuracy

Algorithm	Accuracy Score %
Scaled KNN	86
Scaled Naïve Bayes	85
Scaled Support Vector Machine	**98**

4.2 Algorithm parameters tuning

Algorithm Tuning is a technique used to improve machine learning model accuracy, it is applied to the most promising algorithms from table 4.4 that is KNN and SVM .fig 4.10 shows the tuning parameters for KNN, utilizing 1,3,5,7,9,11,13,15,17,19,21 as a random neighbors. The best accuracy was obtained at n_neighbors equal to 1 which is 0.91715.

4.3 Ensemble technique

Ensemble technique is technique whereby more than one algorithmic models are used together to improve the global accuracy of the model, for this research, Adaboost classifier, Gradient boosting classifier, Random forest classifier and Extra tree classifier were used in order to achieved the best accuracy. Table 3 shows that gradient boosting classifier has the highest accuracy of 0.9985 which is very promising one.

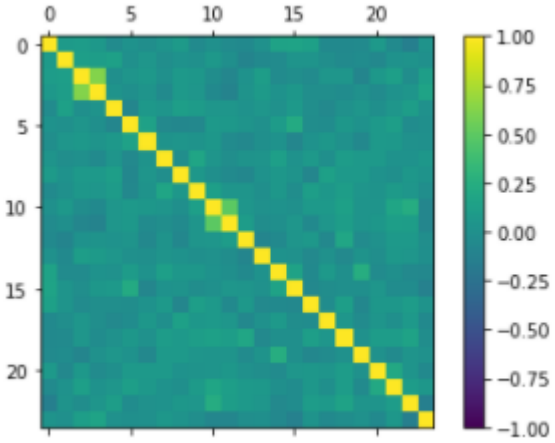

Fig 3: Heat maps diagram for the dataset

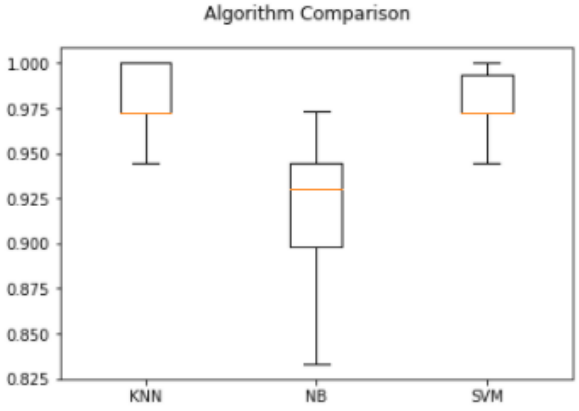

Fig 4: Accuracy scores comparison before feature scaling in box plot

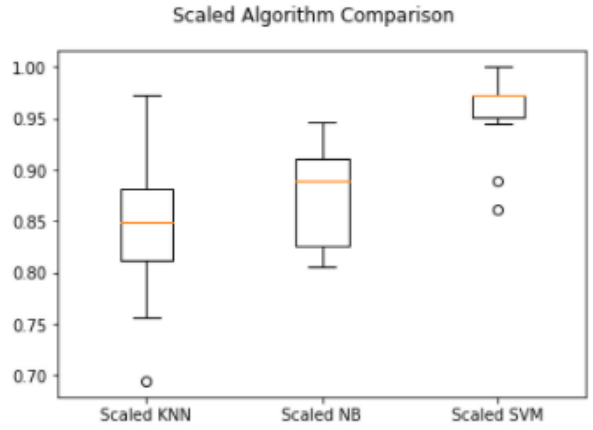

Fig 5: Accuracy scores comparison after feature scaling in box plot

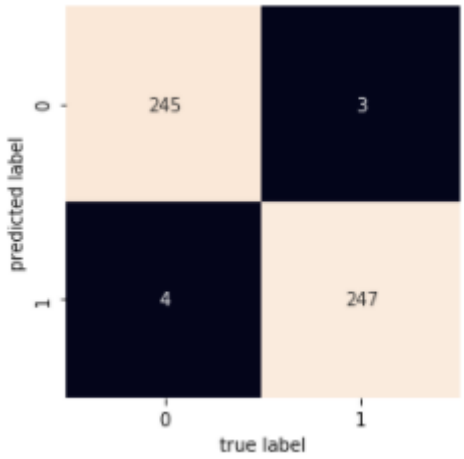

Fig 6: Confusion matrix for the final model

Table 3. Accuracy of ensemble technique

Algorithm	Accuracy Score
Adaboost	0.9942
Gradient boosting	**0.9985**
Random forest	0.9942
Extra tree classifier	0.9884

Table 3 shows that using tuning and ensemble technique KNN and SVM has the highest accuracy 0.9985.This part shows three machine learning algorithms, which were KNN, Nave Bayesian Classifier and SVM, on the hypertension dataset and obtaining the best fitted model amongst them. Finally, the most promising models were tuned and ensemble, which gave us the best accuracy score of 99%. The model was scalable as both the precision, recall and fi score of the target class were higher at 99%.

5. CONCLUSION

In the paper, hyper parameter tuning and ensemble techniques were used on our candidate models, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). Features selection and features scaling were performed to normalize the data as they were in different dimensions and also SMOTE technique was used to Balance the dataset. 86% and 98% accuracy scores were obtained for the KNN and SVM models, respectively after scaling as shown in table 2. Type I error and type II error were obtained as 3 and 4 respectively from the confusion matrix in Fig 6. The precision, recall, and f1-score of the ensemble model are all around 99%, respectively. The model has good performance compared to the ones studied in the reviewed papers. In Future scope an interface can be developed for the model in a web or mobile application and also advanced technique such as deep learning approach can further be applied to the data set.

6. REFERENCES

[1] Adam, F. (2019). Everything you need to know about hypertension. MedicalNewsToday. Retrieved july 22, 2019 from https://www.medicalnewstoday.com/articles/150109.php

[2] Lackland, D, T & Weber, M, A. (2015). Global burden of cardiovascular disease and stroke: hypertension at the core. The Canadian Journal of cardiology, 31(5):569 – 571. DOI:10.1016/j.cjca.2015.01.009

[3] Bharati, M., R. (2010). Data mining techniques and application. Indian Journal of computer science and Engineering, 1(4) pp 301-305. Retrieved October 16 from https://www.researchgate.net/publication/49616224_DATA_MINING_TECHNIQUES_AND_APPLICATIONS

[4] Han, J., & kamber , J., P. (2011) Data Mining Concepts and Techniques. New York : Morgan Kaufmann Publishers.

[5] Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From Theory to Algorithms. New York: Cambridge university press.

[6] Manimekalai, K. (2016).Prediction of heart disease using data mining techniques. International journal of innovative research in computer and communication, 4(2), 2320 – 9801. DOI: 10.15680/IJIRCCE.2016.0402013

[7] Azian, A, A., Zulkarnay, Z., & Nur, F, M. (2011). Design and Dwevelopment of fuzzy expert system for Diagnosis of Hypertension. Paper Presented at second International Conference on Intelligent system, Modeling and simulation. DOI 10.1109/ISMS.2011.27

[8] dowu, P, A. (2017). Predictive model for the classification of hypertension Risk using Decision Tree Algorithm. American Journal of Mathematical and Computer Modeling, 2 (2) : 48 – 59. DOI : 10.11648/j.ajmcm.20170202.12

[9] Abrishami, Z., & Tabatabaee, H. (2015 October). Design of a Fuzzy expert system and a multi layer Neural Network system for Diagnosis of hypertension. Bulletin of Environment, pharmacology and life science, 4 (11), 138-145.

[10] Daniel, L., Farhana, Z., David, B., & Ken, M. (2016). Using machine learning to predict hypertension from clinical data set. [Electronic version]. DOI: 10.1109/SSCI.2016.7849886

[11] Afeni, B. O., Aruleba, T. I., & Oloyeda, A. I (2017). Hypertension Prediction system using Naïve bayes classifier. Journal of Advances in mathematics and computer science, 24(2) 1-11

[12] Dingkun, L., Hyun, W.P., Erdenebileg, B., Lkhagvadorj, M., Ibrahim, M., Meijing, L. & Keun, H.R. (2018). Application of a mobile Health – care system for Hypertension Based on Big Data Platforms. Journal of Sensors. Arcticle ID 3265281, 13. Doi:10.1155/2018/3265281

[13] Chang, C., Wang, C. & Jiang, B. (2011). Using Data mining techniques for multi – disease prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert systems with Applications 38(2011), 5507 – 5513.

[14] Muhammad, F. I., Ganjar, A., Muhammad, S. & Jongtae, R. (2018). Hybrid Prediction model for type 2 diabetes and hypertension using DBSCAN – Based outlier Detection, synthetic minority over sample technique (SMOTE), and Random forest. Applied science, 8. Doi:10.3390/app8081325

[15] Sakr S, Elshawi R, Ahmed A, Qureshi W T, Brawner C, Keteyian S,Blaha M J., & Al mallah M H. (2018) Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project. PLoS ONE 13(4): e0195344. https://doi.org/10.1371/journal.pone.0195344

IComputer Modeling, 2 (2) : 48 – 59. DOI : 10.11648/j.ajmcm.20170202.12

[16] Huang , J., Tsai, Y., Wu, P., Lien, Y., Chien, C., Kuo, C., Hung, J., Chen, S & Kuo, C. (2020). Predictive Modelling of blod pressure during hemodialysis: a comparison of linear model,random forest,support vector,regression, XGBoost, LASSO regression and ensemble method. Computer methods and programs in biomedicine 195(0ctober 2020), 105536. Retrievied September 9 2020 from https://www.sciencedirect.com/science/article/abs/pii/S0169260720301206

[17] Alkaabi,L,A., Ahmed, L, S., Al attiyah, M, F & Abdel-Rahman, M, E.(2020).Predicting hypertension using machine learning: Findings from Qatar Biobank study. PLOS ONE 15 (10):e0240370.DOI: https://doi.org/10.1371/journal.pone.0240370. Retrieved September 9 2021 from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240370.

[18] López-Martínez, F., Núñez-Valdez, E.R., Crespo, R.G. & Giarcia-Diaz, V.(2020) An artificial neural network approach for predicting hypertension using NHANES data. Sci Rep 10, 10620. DOI :https://doi.org/10.1038/s41598-020-67640-z. retrieved on September 9 2021 from https://www.nature.com/articles/s41598-020-67640-z