A Comparative Study of Traditional vs Hybrid AI models with A Medical Analytics Usecase

Komal Barge*[1], Nidhi Patel*[2], Mostafa M. Fouda†[3], and Zubair Md Fadlullah*‡[4].
*Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada.
†Department of Electrical and Computer Engineering, Idaho State University, Pocatello, ID, USA.
‡Thunder Bay Regional Health Research Institute (TBRHRI), Thunder Bay, Ontario, Canada.
Emails: [1]bargek@lakeheadu.ca, [2]pateln@lakeheadu.ca, [3]mfouda@ieee.org, [4]zubair.fadlullah@lakeheadu.ca.

***Abstract*—The National Health and Nutrition Examination Survey (NHANES) is a cross sectional database that have been used to examine the prevalence of sedentary lifestyle related chronic diseases. Inadequate dietary, nutritional, and behavioral habits, household environment, whole body, and bone measurements etc. which might risk of causing such severe illnesses. This project is aimed to investigate and analyze various data mining techniques for prediction of frequent chronic diseases in U.S. adults who participated in 2017-2018 NHANES dataset. The main focus is on using Hybrid AI Model to estimate the association among all the features of NHANES dataset and classify hypertension diseased participants. This paper utilized a highly imbalanced dataset of 8366 with 81.20% participants with no hypertension, and 18.80% participants with hypertension. Our results shows the best accuracy of 94% with Hybrid AI model which can be contributed to the health domain in identifying the person with risk of having hypertension by taking all the health prospects into the consideration.**

***Index Terms*—NHANES, Prediction, Hypertension, ANN, LDA, Logistic Regression, Hybrid AI Model.**

I. Introduction

Data Analytics plays a crucial role in decision making in different fields, as it provides insights from large datasets with multiple disciplines. Healthcare predictive analytics uses historical data to forecast the potential risk, personalizing care to each patient. The past medical history, demographic information, and behaviors of a person can be used together with the expertise and experience of healthcare professionals to predict the future.

The National Health and Nutrition Examination Survey (NHANES) database contains a complete health survey of adults and children in the United States [1]. The survey conducted by Centers for Disease Control and Prevention (CDC) is a combination of physical interviews and health examinations. It uses a unique and convoluted multistage probability design to sample the individuals living in 50 states and D.C. This dataset gives a glimpse of a relationship between the diet, health, nutrition and how it affects the overall health and potential risks for an individual. Because of its coverage of a broad range of health-related issues, and its combination of self-reported questionnaire data with lab results and examinations, NHANES has been a rich source of data for the investigation of specific health questions. [2]

In the past research, NHANES dataset have been used by researchers to find out the relationship between lifestyle, nutrition levels, medical conditions, and mortality. But there has not been a research to predict and analyze the chronic diseases using different data mining techniques using all the sections of NHANES dataset. This cross-sectional dataset consists of hundreds of features, among which several can be used for supervised learning purposes. The most difficult part is to find a subset of key features to combine them in an optimal way.

The purpose of this project is to study the relationship between the several health and behavioral features causing the common illnesses like Diabetes, Hypertension, Hypercholesterolemia, Depressive disorder, and Gastro-Esophageal Reflux disease. These diseases are characterized by high levels of blood glucose, blood pressure, cholesterol, sleep deprivation, stress, and stomach acid. It is often accompanied by other serious health complications and may lead to premature death. [2] Hence, exploring and analyzing the threat of these diseases would be a key contribution to the public health domain.

Despite of prominent research in prevention of chronic diseases, it has always been a leading cause of overall mortality all over the world. NHANES is a significant United States data sample for discovering associations in between behaviors and diseases. However, to the best of our knowledge, the NHANES data has not been strategically investigated for predicting severe disorders. The main objective of the study is 1) To explore the influence of diet, nutrition, behaviors, and environmental exposures on the risk of causing severe diseases, 2) To examine data mining techniques that could be used to inspect diseases on highly imbalanced NHANES dataset, 3) To develop and test different machine learning models for prediction of most frequent disorders discretely, here we have considered Hypertension disorder, as it is one of the top diseases in the dataset. The motivation of studying this cross sectional NHANES dataset is to assess the missingness of the data values, imbalanced class labels and nonlinearity of the input features.

II. Related Work

The myriad of studies on NHANES and other similar topics have been carried out to gain richer insights on current health behaviors and trends. Alejandro et al. presented a combination of decision trees and a Genetic Algorithm (GA) to optimize the selection of a set of predictors from NHANES dataset that best predicts the presence of diabetes. In this paper, the authors have provided a relationship between diabetes and factors like age, ethnicity, socioeconomic and cholesterol data [3]. They exclaim that their approach is efficient for mining a dataset and identifying important set of predictors for diabetes without having to obtain inputs from experts or start from well defined hypothesis.

Zhengzheng et al. proposes a direct disease pattern mining method and an interactive disease pattern mining method to explore the NHANES data [4]. Their study provides summary of the data set via a disease influence graph and a hierarchical tree. Madhav V. Rao et al. have presented their multivariate logistic analysis by combining Kidney Early Evaluation Program (KEEP) and NHANES data from 1999-2004. Authors have mentioned that estimated glomerular filtration rate (eGFR) assessed as a continuous variable has a linear relationship to hypertension prevalence. Their results support the use of screening programs to improve public kidney and cardiovascular health [5].

Jun won Lee and other authors aims to deliver informative relations and association rules that are not trivial but have potential to provide valuable insights to clinical psychologists and people in medicine domain [6]. The authors of this paper have claimed that their experimental results with decision tree (C4.5) and association rule miner (Apriori Algorithm) provides several implicit relations such as association between high blood pressure and hearing problem as well as breathing problems and diabetes. Another research by Jun won lee with Christophe Giraud-Carrier is based on adapting and extending association rule mining and clustering algorithm to extract useful knowledge regarding diabetes and high blood pressure from the 1999-2008 NHANES data [2]. In this paper, they have focused on simple correlations between health conditions and issues, and then considered more global view in which MSapriori algorithm is used to apply association rule mining effectively.

Fernando et. al proposed a neural network classification model to estimate the associations in between different predictors in hypertensive patients. This paper uses 7 separate predictors to classify hypertensive patients with the help of cross validation experiments in a highly imbalanced NHANES data set [7]. This paper aspired us to achieve lower error rate using the whole NHANES dataset from year 2017 - 2018 by using hybrid ANN model. In our paper instead of using those 7 predictors, we have applied dimension reduction technique on all the NHANES sections to obtain the list of inputs to pass it to the ANN model which will be more flexible than just using a ANN model with raw features and traditional ML approaches like Logistic Regression, Random Forest etc.

III. Data

We downloaded NHANES datasets from NHANES 2017 - 2018. These datasets are published by the National Center for Health Statistics (NCHS), Division of Health and Nutrition Examination Surveys (DHANES), part of the Centers for Disease Control and Prevention (CDC) which conducts a series of health and nutrition surveys every year. Around 5000 individuals of all ages are interviewed to assess the health examination survey at home and mobile examination center (MEC). This data consists of 5 different sections that are Demographics, Dietary, Examination, Laboratory and Questionnaire data. Each data section has numerous data files depending on the variety of the examinations. These data files are in the SAS format with different number of instances and explanatory variables depending on the survey assessments.

IV. Data Preprocessing

Each data section is having numerous files. Hence, we tried to merge all the files in each section and come up with a final sectional dataset. It had many duplicates and multiple instances which we merged or dropped depending on the importance of the variable information. As from the observation, we got to see that the values missing in the dataset were at random. There were no relationship in between the observed responses or specific missing values which are expected to be obtained. The values for blank, period(.), refused, and don't know has been replaced after appropriate data analysis on the respective dataset. Each sectional datasets have been handled for missing values which can be seen in below data pre-processing section. Let's go through processing of each dataset.

A. Demographics dataset

The demographics file provides individual, family, and household-level information on different topics. In this dataset, we first removed the correlation in between the variables, checked for null values and then read through NHANES demographics variables documentation and came up with important variables for the final analysis of the dataset.

B. Dietary dataset

The dietary dataset mainly consisted of nutrient intakes, food information, individual and total dietary supplements. After merging the whole dataset, it ended up around 8 GBs. So, we only considered important files for our analysis. This dataset had some columns with null values which needed to be imputed by zero. For ex., these variables indicated the reasons for taking each dietary supplement reported, so for some participants if the dietary supplement is not taken, that reason value is given as null. Therefore, we decided to encode it as zero and if the supplement is taken then by

default reason value will be given for a participant. After this, we dropped the attributes with null values more than 500.

C. Examination dataset

For examinations, the controlled environment of the MEC allowed physical measurements to be done under standardized conditions. In this dataset, we studied all the files descriptions to understand the explanatory variables and came up with the final set of variables which are necessary for the model. In this data, separate bodily measures were dropped, and instead final aggregated measurements were considered for prediction analysis. In this, we had 'BPXCHR' column which indicated heart rate value for participants with age 0-7 years old, whereas 'BPXPLS' column which represented heart pulse for 8-150 years old participants. Hence, instead of considering two different columns with mutual null values we combined it into one attribute to deal with missing values. For the remaining attribute's null value substitution, we considered the KNN imputation method with k value as 5.

D. Laboratory dataset

NHANES collected biological specimens (biospecimens) for laboratory analysis to provide detailed information about participants' health and nutritional status. This dataset is having an important information about the individual's health. So, we scaled the dataset for suppressing any outliers and the variable units. After normalization, we performed KNN imputation on top of it to substitute the null values.

E. Medication dataset

Medication dataset is not explicitly mentioned in the NHANES official dataset. We took it apart from the Questionnaire dataset as it contains the crucial information regarding disease analysis. This file mainly provides personal use of prescription medications and health problems.

The dataset contains duplicated entries for participants of the survey as some of them were prescribed more than one medication or having more than one disease. The attributes with string values were encoded in bytes which we decoded before applying any other operation on the dataset. First, we deleted entries with no description of diseases. Secondly, we dropped unimportant variables and figured out the most frequent unique keywords in disease description columns to come up with a list of top five diseases. Among all the instances, we took instances which were having diseases from top five diseases. This dataset had a column with different kinds of medications that a person is taking for a particular disease from which we made dummy columns for each medication. For the final dataset, we took dummy columns of the drugs and diseases.

Here, after getting the final dataset, we have tried to separate datasets by the target value which is different diseases. In these separate datasets, we had duplicate entries for some of the participants which we dealt by merging instances with same target values or dropping the instance. The separate datasets of each disease were then merged with other datasets such demographics, examination and laboratory. We still need to figure out how to proceed with the dietary dataset and questionnaire dataset.

F. Questionnaire dataset

Questionnaire dataset provides the information on participant's behavior, diet and nutritional habits. This dataset consists humongous amount of null values. Therefore, we went through every file's documentation to look at the attributes and check whether the attribute is useful or not to come up with a final list of necessary variables and considered only important explanatory variables which can be causing underlying issues to any kind of a health problem like smoking etc. In this some values are imputed with KNN imputation and others with 0 depending on the variable description.

V. Proposed Research Methodologies

We present and discuss traditional ML and different hybrid approaches using artificial neural network to classify hypertensive patients. In this section, we have described the model architecture that we followed to assess the NHANES unstructured data which can be seen in Fig. 1. This study has been solely carried out with the obtained hypertension dataset.

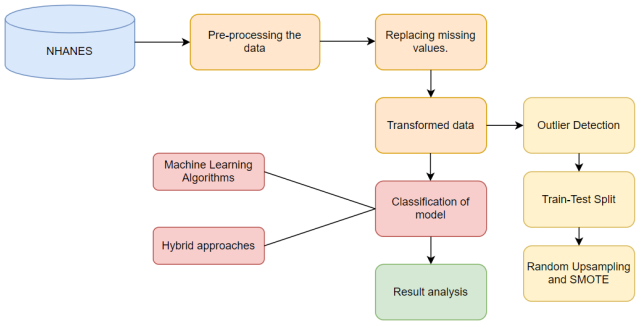

Fig. 1: Architecture of the model

First, We plotted pair-plots of feature correlation for each dataset with the more collinearity, and from these graphs we were able to observe that some data points were outliers. We tried using Z score, Interquartile range, and Isolation Forest for removing outliers from our datasets but Isolation Forest is more effective unsupervised technique for high dimensional data whereas other methods are less convenient for such sparse data. After evaluating the results from all the techniques, we figured that the outliers detected by Interquartile Range were somewhat similar to the instances detected by Isolation Forest. Among five instances detected as outliers from the Isolation forest, three same instances were also detected by IQR. Therefore, we removed those 3 exact instances as final outliers.

The obtained datasets after preprocessing has some class imbalance issue. After pondering upon class imbalance resolving options of adding artificial instances using Synthetic Minority Oversampling Technique (SMOTE), random upsampling and downsampling, we carried the comparative analysis with the help of SMOTE as well as random upsampling. Usually, if we perform random downsampling on the obtained datasets, we might lose potential important medical information required for the data analysis. For the train test data split, we used different techniques such train-test split with 2:1 ratio, stratified train-test split, and 5-Fold cross validation. The detailed analysis for our hybrid ANN methodology is discussed further in the sections below.

A. *Hybrid AI model*

This hybrid AI model is a combination of dimension reduction method and then passing those attributes to an ANN model that is made of input, hidden, and output layers. Applied hybrid AI model architecture can be seen in Fig. 2.

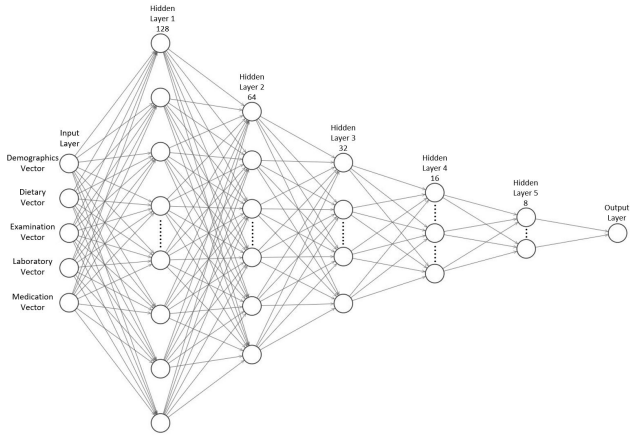

Fig. 2: Hybrid AI Model

Firstly, we have used Linear Discriminant Analysis (LDA) technique for reducing the dimensions of each section. LDA reduces the number of dimensions by taking the target variable into account and retaining as much information as possible from the features. LDA generally models the linear combination of features in such a way that it maximizes the separation between the classes which really helps in constructing the classification model afterwards.

For each section, as per the number of class labels and features constraints for LDA, we used LDA component size as 1. After investing through all the high weight features from LDA vector, we noticed that LDA constructed its vector based on most important features from the each section. Hence, we obtained five vectors for five sections in the dataset.

We used these five vectors as input nodes for our ANN model. The motivation behind developing the flexible hybrid AI model is to weed out the nonlinearity of the features involved in the NHANES dataset. This nonlinearity can be seen from the pair plot of input variables in Fig. 3.

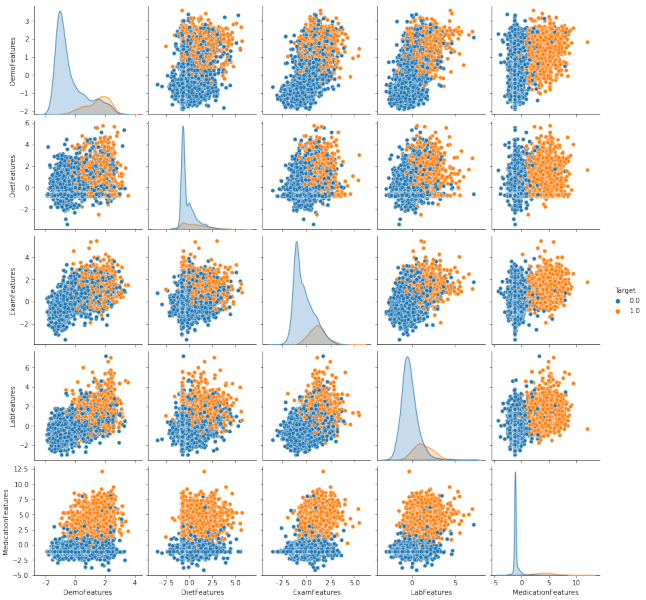

Fig. 3: Pair Plot of input features

Here each iteration works on 8 samples and the model is trained on 66% of the data. The optimal hyperparameters required for the hybrid model can be seen in Table. I

TABLE I: Optimum Hyperparameters

Hyperparameter	Value
Optimizer	Adamax
No of Epochs	50
Batch Size	8
No of Layers	7
Loss Function	Binary˙crossentropy

For each single layer, Rectified Linear Unit (ReLU) function has been applied to learn the complex relationships in between the input vectors which allows the models to perform better. To evaluate the classification model, We have utilized average test error and loss for the model. The evaluation is performed on 33% of the data. We trained the model by adding dropout layers to avoid overfitting. For each instance, the model uses sigmoid as the evaluation function that returns the classification labels in the form of probabilities as an output. Results are shown in Fig. 7. In the evaluation matrices, we can see that we obtained . True positive value (468), True Negative value (2134), False Negative (54) and False Positive value (102). Compared to the [7], The precision and recall of the model has definitely increased because of the better choice of the input components even in highly imbalanced dataset.

VI. Results and Analysis

In this paper, a comparison of our proposed hybrid AI model with other traditional ML scenarios and different hybrid models has been carried out.

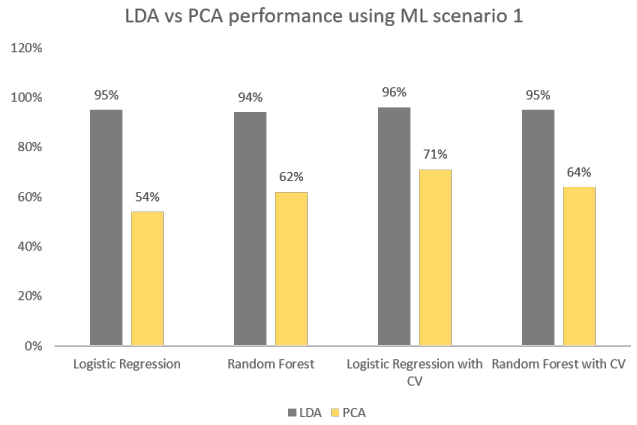

Fig. 4: Performance Comparison of LDA and PCA

All the comparisons have been performed by using 3 different class imbalance techniques (Stratified Split, Random Upsampling with replacement, SMOTE) and 2 machine learning algorithms (Logistic Regression, Random Forest) with the help of dimension reduction techniques that is LDA and PCA. These techniques helped in not only classifying the hypertensive patients but also in identifying the most important features from each section.

1) Machine Learning Scenario 1: As NHANES data is highly imbalanced, this analysis just had participants with top 5 diseases which were around 2423 instances and around 426 features. Here, the ratio of hypertension to no hypertensive participants is 2:1. After applying all those above explained approaches on this sample dataset, we obtained highest accuracy of around 95%. The respective comparison can be found in Fig. 4. Here you can see that PCA is performing worse on small dataset. Therefore, we eliminated PCA from rest of the data analysis.

2) Machine Learning Scenario 2: In this scenario, we have considered all the unique participants from NHANES dataset which are around 8500 and around 1064 feature variables. The respective class imbalance issue and reduction of dimensions has been handled by utilising similar techniques mentioned above. The confusion matrix for the same analysis can be seen in Fig. 5.

Class imbalance	Classification	Confusion Matrix		Accuracy
Stratified Train-Test Split	Logistic Regression	TN = 2103	FP = 139	84.93%
		FN = 277	TP = 242	
	Random Forest	TN = 2145	FP = 97	90.18%
		FN = 174	TP = 345	
Random up sampling	Logistic Regression	TN = 1752	FP = 484	79.57%
		FN = 80	TP = 445	
	Random Forest	TN = 2097	FP = 139	89.75%
		FN = 144	TP = 381	
SMOTE	Logistic Regression	TN = 1788	FP = 448	80.62%
		FN = 87	TP = 438	
	Random Forest	TN = 2066	FP = 170	91.25%
		FN = 99	TP = 426	

Fig. 5: Results from ML Scenario 2

3) Hybrid models based on Random Forest and ANN: From previous ML scenario 2, it can be observed that random forest outperformed logistic regression with the SMOTE analysis. Hence for these fusion models, random forest was chosen to use as a feature extractor. Some of the sections of NHANES data has missing values, but they also had important feature information which was extracted through ANN layers. Hence, we ended up extracting features from dietary and laboratory sections using ANN whereas, for demographics, examination and medication sections, we used random forest. And, these extracted features were concatenated and passed through ANN model for classification. The architecture for this hybrid approach can be seen in Fig. 6.

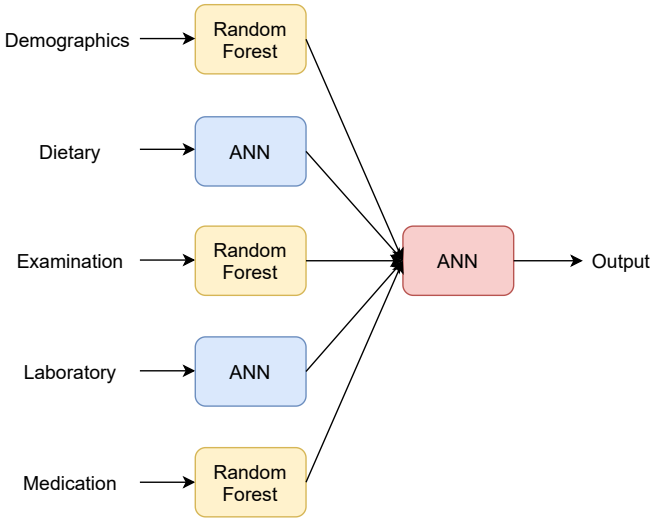

Fig. 6: Hybrid Approach 1 Architecture

On the other hand for hybrid approach 2, instead of using ANN as a feature extractor, we extracted all the features using random forest from each section and concatenated them to pass through ANN model. Same architecture was followed as Fig. 6.

The performance result for both these approaches and our proposed hybrid model can be seen in Fig.7.

Based on the performance evaluation measures shown in Fig. 5 and Fig. 7, it can be said that our proposed hybrid approach is outperforming than any other hybrid approaches as well as different traditional ML algorithms. Increasing the number of samples by using SMOTE has always produced better accuracy. We experimented around different dimension reduction methods and algorithms to find the best results for our proposed method. Table I shows optimal parameters and Fig. 7 shows the best confusion matrix.

In our paper, we used a combination of LDA and ANN model to explain the nonlinearity of the input variables. The test dataset consist of 2761 which includes 2242 nonhypertensive participants and 519 hypertensive participants. The model shows that 88.4% positives that are properly classified

Class imbalance	Classification	Confusion Matrix		Accuracy
Stratified Train-Test Split	Hybrid Approach 1	TN = 2030	FP = 212	86.64%
		FN = 157	TP = 362	
	Hybrid Approach 2	TN = 2175	FP = 67	84.00%
		FN = 354	TP = 165	
	Proposed Model	TN = 2134	FP = 102	94.24%
		FN = 57	TP = 468	
SMOTE	Hybrid Approach 1	TN = 1929	FP = 307	87.11%
		FN = 49	TP = 476	
	Hybrid Approach 2	TN = 1950	FP = 286	87.36%
		FN = 63	TP = 462	
	Proposed Model	TN = 2150	FP = 92	94.49%
		FN = 60	TP = 459	

Fig. 7: Results from Hybrid models

and 95.89% negatives that are properly classified. Whereas, a positive predicted value of 83.3% and a negative predicted value of 97.28%. This model outperforms in predicting the people who might have or develop hypertension than those who will not develop hypertension. With the sensitivity of 88.4% and specificity of 95.89%, our proposed hybrid technique shows that it might be effective in detecting the people who might develop or have hypertension than detecting non hypertensive people, which would be key contribution towards the healthcare diagnosis. Due to consideration of most of the relevant predictor variables, this proposed hybrid approach definitely performs better compared to the multiperceptron model proposed by the paper [7]. This detailed analysis gives an overview on how considering all the features will give you the best results and why all the features or health conditions should be considered for an individual while examining them for an illness.

VII. Limitations

The dataset that we have used for the analysis is a survey data provided by NHANES, there were some of the participants who refused to provide the information which led to the missingness in the data. The dataset was in sections and in each section, we had multiple files that we merged according to the unique sequence number assigned to the participants. Thus, it involved massive overhead of preprocessing. In addition, from the count of target variable, it can be seen that the data is highly imbalanced and it involved humongous number of predictor variables. From the results, it can be seen that, we were facing curse of dimensionality because as the number of features increased, there was an increase in error values. On the other hand, we will need medical expertise to regulate the predictor variables that has been used as a part of the investigation.

VIII. Conclusion and Future Work

In this paper, we have discussed about the analysis of only Hypertension disease. From the comparative analysis, it is observed that our proposed model is performing best out of all the models/approaches with accuracy of around 94% and by taking the False Negative values into the consideration, our proposed model outperforms among the other approaches with value as 60 as in health domain it is always good to have less False negative values than False Positive values.

For the future work, we can analyse the other diseases from top 5 frequent diseases as well as the diagnosis. But again, it might lead to high class imbalance in the data. Also, to make our model more adaptable and for testing or training purposes, we can expand the dataset by taking data from previous years from NHANES. Because of the time constraints, we did not consider other imputation techniques which can be tried on for training the model in a better way. Overall, this paper shows that the proposed model is giving best accuracy and performance among the traditional machine learning approaches by providing the same features as input to the models.

References

[1] "Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention," 2017-2018. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017

[2] J. won Lee and C. Giraud-Carrier, "Results on mining nhanes data: A case study in evidence-based medicine," *Computers in biology and medicine*, vol. 43, no. 5, pp. 493–503, 2013.

[3] A. Heredia-Langner, K. H. Jarman, B. G. Amidan, and J. G. Pounds, "Genetic algorithms and classification trees in feature discovery: diabetes and the nhanes database," in *Proceedings of the International Conference on Data Science (ICDATA)*. The Steering Committee of The World Congress in Computer Science, Computer ..., 2013, p. 1.

[4] Z. Xing and J. Pei, "Exploring disease association from the nhanes data: Data mining, pattern summarization, and visual analytics," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 6, no. 3, pp. 11–27, 2010.

[5] M. V. Rao, Y. Qiu, C. Wang, and G. Bakris, "Hypertension and ckd: Kidney early evaluation program (keep) and national health and nutrition examination survey (nhanes), 1999-2004," *American Journal of Kidney Diseases*, vol. 51, no. 4, pp. S30–S37, 2008.

[6] J. won Lee, Y. H. Lin, and M. Smith, "Dependency mining on the 2005-06 national health and nutrition examination survey data," 2008.

[7] F. López-Martínez, E. R. Núñez-Valdez, R. G. Crespo, and V. García-Díaz, "An artificial neural network approach for predicting hypertension using nhanes data," *Scientific Reports*, vol. 10, no. 1, pp. 1–14, 2020.