A Systematic Comparison of Traditional and Hybrid AI Models for NHANES Big Data Analytics

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 big health dataset, which has recently raised much research interest to use data mining and analytics techniques to examine the prevalence and risks of chronic diseases related to sedentary lifestyle, inadequate dietary, nutritional and behavioral habits, household environment, whole body measurement, bone measurements, and so forth. In this paper, we carry out a systematic investigation of comparative analytics based on several traditional machine learning techniques and hybrid AI models to estimate the association among all the features of the 2017-2018 NHANES dataset and classify hypertension diseased participants as a use-case. Based on the use-case, our research work utilizes a highly imbalanced dataset of 8,366 people with 81.20% participants with no hypertension and 18.80% participants with hypertension. Empirical results demonstrate that our proposed AI model, with an accuracy of 94%, significantly outperforms the other machine learning techniques and two hybrid model variants in identifying patients with a risk of having hypertension by considering all the health conditions.**

***Index Terms*—NHANES, prediction, hypertension, Artificial Neural Network (ANN), Linear Discriminant Analysis (LDA), logistic regression, hybrid AI model.**

I. Introduction

Big data analytics plays a crucial role in decision making in different fields since it provides insights from large datasets with multiple disciplines. Healthcare predictive analytics uses historical data to forecast the potential risk that is critical for conceptualizing personalized care for every patient. Thus, the past medical history, demographic information, and behavior of an individual can be jointly exploited with the expertise and experience of domain experts (e.g., healthcare professionals) to predict the risk of developing chronic diseases. The National Health and Nutrition Examination Survey (NHANES) database contains a complete health survey of adults and children in the USA [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 multi-stage probability design to sample the individuals living in all the states in the USA. A systematic data mining and analytics on this robust dataset could, therefore, provide intricate relationships between the diet, health, nutrition, and so forth; and how they impact the overall health and drive potential risks in case of an individual patient. Because of its coverage of a broad range of health-related issues and its combination of self-reported questionnaire data with pathological lab results and health examinations, NHANES has been a rich source of data for the investigation of specific health questions [2]. Traditionally, the NHANES dataset was utilized by researchers to find out the relationship between lifestyle, nutrition levels, medical conditions, and mortality in a specific section of the NHANES dataset. However, to the best of our knowledge, the existing research work overlooked the importance of predicting and analyzing the chronic diseases using different data mining techniques using all the sections of NHANES dataset. Therefore, in this work, we address how to take into consideration this huge, cross-sectional dataset consisting of hundreds of features, and determine the most useful features for supervised learning purposes. In other words, we identify the research challenge of finding a subset of key features and combining them in an optimal way for effectively training Artificial Intelligence (AI) models with high accuracies.

Indeed, the diseases considered in NHANES 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. While studying the relationship between the health and behavioral features causing the common illnesses like diabetes, hypertension, hypercholesterolemia, depressive disorder, and gastro-esophageal reflux, we identify hypertension to be the top health priority in the USA based on the NHANES dataset. Despite prominent research in prevention of chronic cardiac diseases, it has always been a leading cause of overall mortality all over the world. Therefore, we use the hypertension as our considered use-case in this paper to provide the following research contributions to perform comparative data analytics using traditional and hybrid AI models to inspect diseases on highly imbalanced NHANES dataset. Our contribution also includes studying the cross sectional big dataset of

NHANES to assess and treat the missing data values, imbalanced class labels, and non-linearity of the input features.

The remainder of the paper is structured as follows. Section II surveys relevant research work. The data preparation steps are described in section III. Our proposed AI model for the big health data classification in NHANES is presented in section IV. Empirical results and analyses are provided in section V. Finally, the paper is concluded in section VII.

II. Related Work

A myriad of studies on NHANES and other similar datasets have been carried out in the literature to gain richer insights on current health behaviors and trends. Heredia-Langner *et al.* [3] presented a combination of decision trees and a Genetic Algorithm (GA) to optimize the selection of a set of predictors from the NHANES dataset that best predicts the presence of diabetes. In particular, a relationship between diabetes and factors, such as age, ethnicity, socioeconomic and cholesterol data was revealed. On the other hand, Xing *et al.* [4] proposed a direct disease pattern mining method and an interactive disease pattern mining scheme to explore the NHANES data. Their study provided a summary of the dataset via a disease influence graph and a hierarchical tree. Next, Rao *et al.* [5] presented a multivariate logistic analysis by combining Kidney Early Evaluation Program (KEEP) and NHANES data from 1999-2004. Their estimated Glomerular Filtration Rate (eGFR), considered as a continuous variable, was found to have a linear relationship to hypertension prevalence. Their results, thus, supported the use of screening programs to improve public kidney and cardiovascular health.

Moreover, López-Martínez *et al.* [6] proposed a neural network classification model to estimate the associations between different predictors in hypertensive patients. They employed seven predictors to classify hypertensive patients with the help of cross-validation experiments in a highly imbalanced NHANES data set. However, no further work exists in the literature to achieve lower error rate using the whole NHANES dataset (instead of only a limited number of predictors) over the recent years (i.e., 2017 onward) by using hybridized AI models.

Lee *et al.* [7] aimed 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. 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.

III. Data Preparation

We obtained the recent, 2017-2018 NHANES dataset [1], published by the National Center for Health Statistics (NCHS), Division of Health and Nutrition Examination Surveys (DHANES), part of the CDC through a series of health and nutrition surveys every year. Around 5,000 individuals of all ages 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 Statistical Analysis System (SAS) format with different numbers of instances and explanatory variables depending on the survey assessments. The description of the 5 sections is as follows:

- *Demographics dataset*: provides individual, family, and household-level information on different topics.
- *Dietary dataset*: The dietary dataset mainly consists of nutrient intakes, food information, individual and total dietary supplements.
- *Examination dataset*: For examinations, the controlled environment of the MEC allowed physical measurements to be done under standardized conditions.
- *Laboratory dataset*: NHANES collected biological specimens (biospecimens) for laboratory analysis to provide detailed information about participants' health and nutritional status.
- *Questionnaire dataset*: Questionnaire dataset provides the information on participant's behavior, diet and nutritional habits. From this Questionnaire dataset, we generated another dataset called *Medication dataset* which contains the crucial information regarding disease analysis. The Medication dataset mainly provides personal use of prescription medications and health problems.

Each section in the obtained NHANES dataset has numerous files. Hence, we attemped to merge all the files in each section and derive a final sectional dataset. However, the resultant dataset contains many duplicates and multiple instances that we have merged or dropped depending on the importance of the variable information. We observed that the missing values in the NHANES dataset were at random. In other words, there were no relationship between the observed responses or specific missing values. The values for blank, period (.), refused, and "don't know" have been replaced after appropriate data analysis on the respective dataset. Thus, each of the sectional datasets has been treated for missing values. To remove the outliers in these datasets, we used two methods, namely, Interquartile Range (IQR) and Isolation Forest, then we removed the instances detected by both of the methods as outliers. Since the datasets are not balanced, we used Synthetic Minority Oversampling Technique (SMOTE) and random upsampling techniques for balancing it.

IV. Proposed AI Model for Big Health Data Classification

In this section, we present our AI model in terms of a combination of dimension reduction method and a simple Artificial Neural Network (ANN) model comprising input, hidden, and output layers as shown in Fig. 1.

First, we employ 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 use LDA component size of 1. After investigating through all the high weight features from LDA vector, we notice that LDA constructs its vector based on the most important features from each section in the NHANES dataset. Hence, we obtained five vectors for the five sections of the dataset.

Next, we utilize these five vectors as input nodes to our ANN model shown in Fig. 1. The motivation behind developing the flexible hybridized AI model is to filter the non-linearity of the features involved in the NHANES dataset. This non-linearity can be seen from the pair plot of input variables in Fig. 2.

To train our ANN model, we use a batch size of 8 samples and train our model on 66% of the data. The hyper-parameters used to train our ANN model are listed in Table I. The Rectified Linear Unit (ReLU) is used as the activation function in the hidden layer and the Sigmoid is used as the activation function in the output layer to generate the classification labels in the form of probabilities. Dropout layers are added during the training phase in order to avoid overfitting.

To evaluate our ANN classification model, we utilize the average test error and loss for the model. The evaluation is performed on 33% of the data.

V. Comparative Analytics Results and Discussion

In this section, a comparison is carried out to evaluate our proposed AI model in contrast with other traditional machine learning (ML) scenarios and two other hybrid variants. All the comparisons are performed by employing three different class imbalance techniques, namely stratified split, random upsampling with replacement, and SMOTE (Synthetic Minority Oversampling Technique); and 2 ML techniques, namely logistic regression and random forest with LDA and PCA (Principal Component Analysis) as dimension reduction techniques. These techniques helped in not only classifying the hypertensive patients but also in identifying the most important features from each section in the big NHANES data. The various scenarios considered in our performance evaluation are described next.

TABLE I: Optimal hyper-parameters.

Hyperparameter	Value
Optimizer	Adamax
No of Epochs	50
Batch Size	8
No of Layers	7
Loss Function	Binary cross-entropy

TABLE II: Comparative analytics of traditional AI models in ML scenario 2.

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 upsampling	Logistic regression	TN=1572	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	

1) ML scenario 1: As the NHANES data is highly imbalanced, in this scenario, we consider only the participants with top 5 diseases which were around 2423 instances and around 426 features. Here, the ratio of hypertension to no hypertensive participants was 2 to 1. After applying all the aforementioned approaches on this sample dataset, we obtained the highest accuracy of around 95%. Since PCA performs worse on the small dataset as depicted in Fig. 3, it is omitted in the remainder of this section.

2) ML scenario 2: In this scenario, we considered all the unique participants from NHANES dataset which are around 8,500 and around 1,064 feature variables. The respective class imbalance issue and reduction of dimensions has been handled by utilising similar techniques mentioned above. Fig. 4 demonstrates the dataset size increase from ML scenario 1 to ML scenario 2 significantly impacts the class separation by LDA. Furthermore, Table II lists the performances of the logistic regression and random forest models using the three class imbalance cases. These results demonstrate that the random forest-based model obtains the highest accuracy with 91.25% and 90.18% with SMOTE and stratified traint-test split, respectively.

3) Hybrid model variants based on Random Forest and ANN: From ML scenario 2, it can be observed that the random forest outperformed logistic regression with SMOTE. Inspired by this encouraging performance of random forest, we use this model as a feature extractor and construct a hybrid model variant. Some of the sections

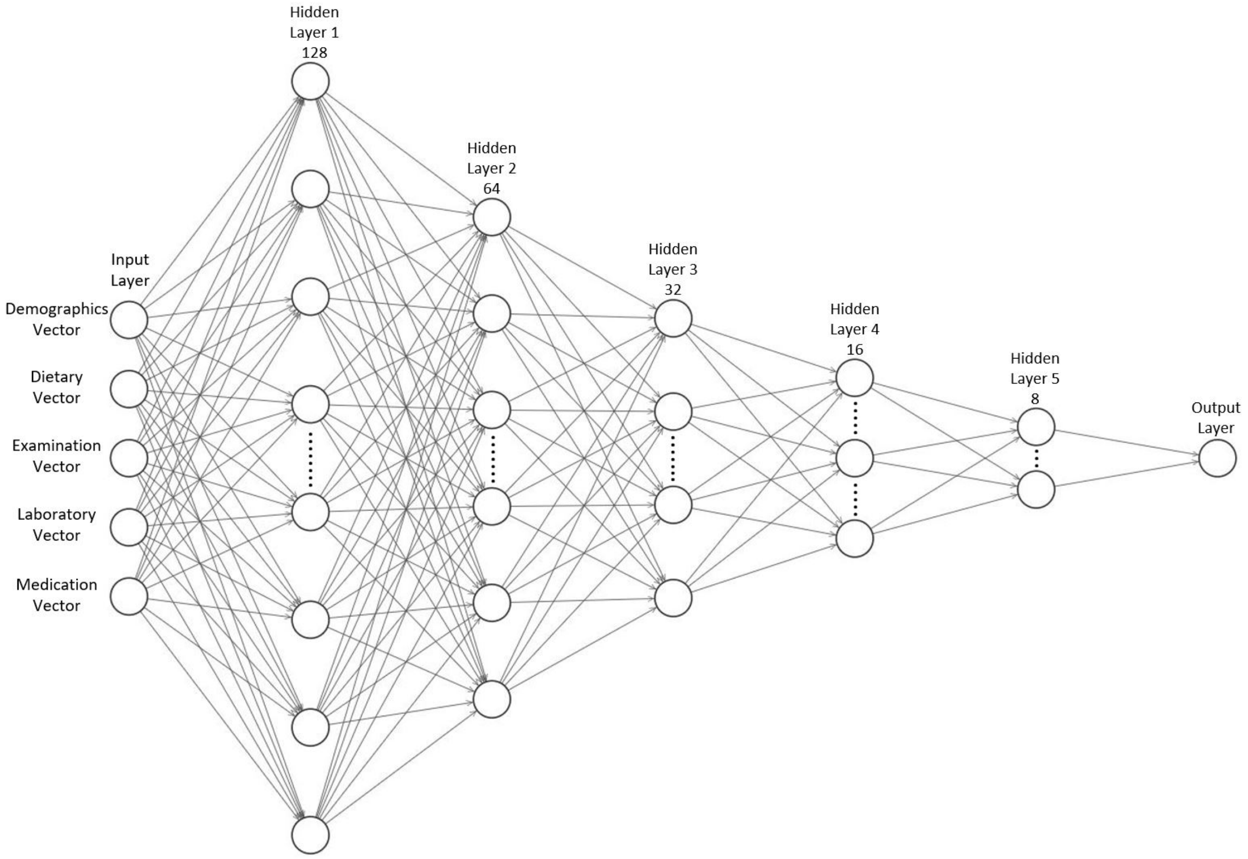


Fig. 1: Our trained ANN model.

of the NHANES dataset has missing values; but they also contain important feature information which was extracted through in the ANN layers of our proposed model in section IV. Therefore, in the first hybrid model variant, we consider extracting features from dietary and laboratory sections using ANN while for demographics, examination and medication sections, we employ random forest. These extracted features are concatenated and passed through the ANN model (the earlier proposed model in section IV) for classification. On the other hand, for the second hybrid model variant, instead of using ANN as a feature extractor, we extracted all the features using random forest from each section and concatenate them to pass through the ANN model. The architecture adopted in this scenario is depicted in Fig. 5.

Table III lists the performances of our proposed AI model and its two variants. Note that our proposal significantly outperforms both the variants utilizing the random forest/ANN combination and the random forest-only feature extractors concatenated with the ANN layer for both the class imbalance cases. Thus, the combination of LDA and ANN in our proposed AI model efficiently captures the nonlinearity of the input variables. To further analyze the performance, the test dataset consists of 2761 participants with 2242 non-hypertensive patients and 519 hypertensive participants. The model, as a whole, shows that 88.4% positives that are properly classified and 95.89% negatives that are properly classified. Hence, our proposed model outperforms in predicting the individuals who may develop hypertension than those who will not develop hypertension. With the sensitivity of 88.4% and specificity of 95.89%, our proposed 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 a critical contribution towards the healthcare diagnosis. Due to the consideration of most of the relevant predictor variables, our proposal definitely performs better compared to the ANN-based

TABLE III: Comparative analytics of proposed AI model and two hybrid variants.

Class imbalance	Classification	Confusion matrix		Accuracy
Statified 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	

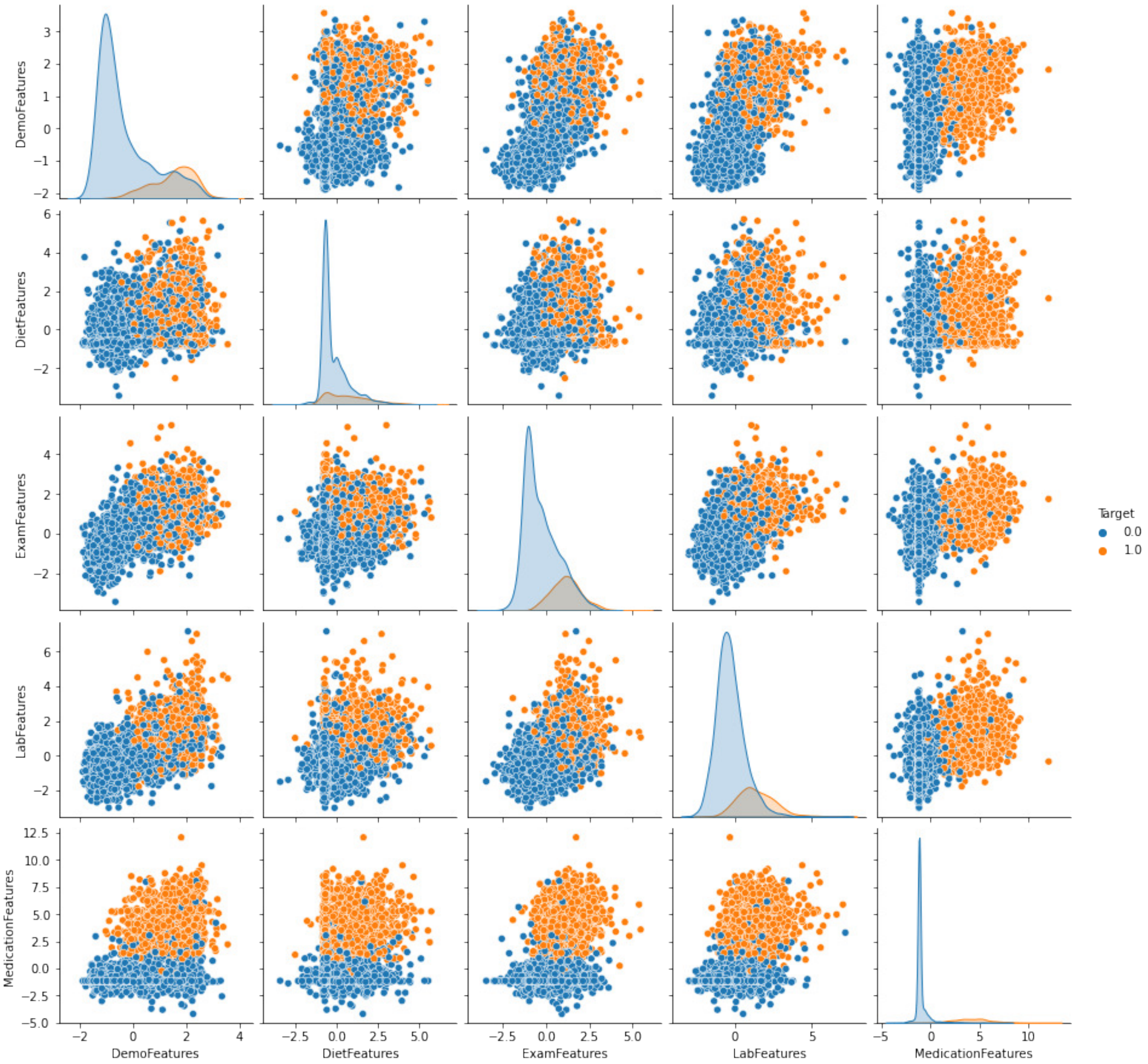


Fig. 2: Non-linearity in the NHANES dataset shown by the pair plot of the input features.

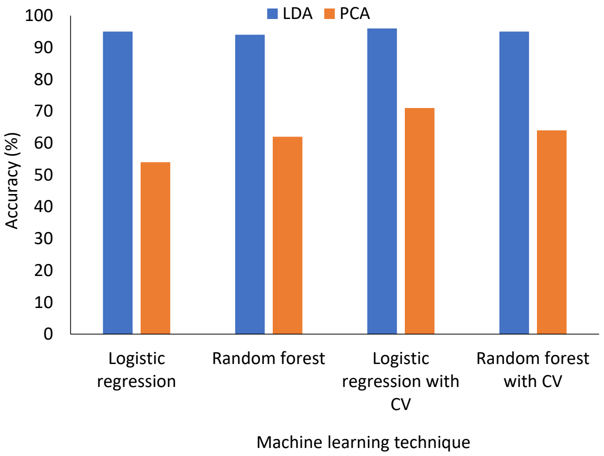
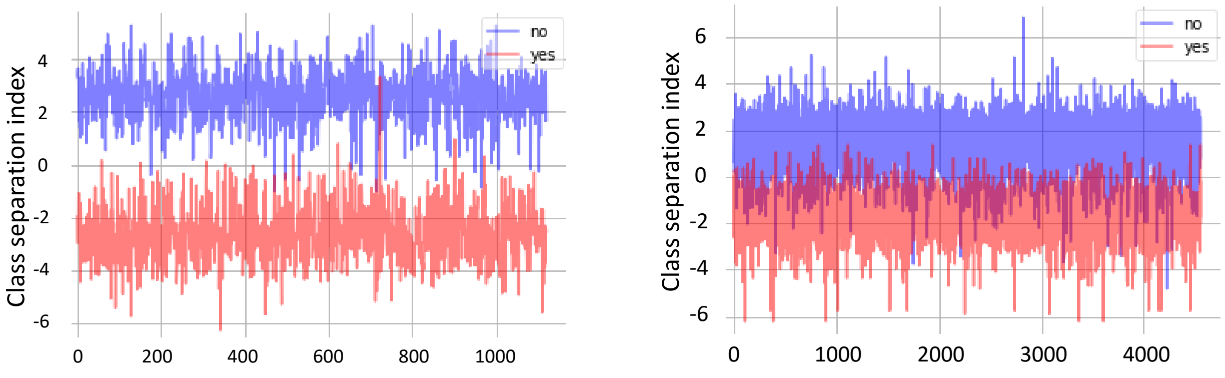


Fig. 3: Performance comparison of LDA and PCA.

model introduced in [6]. Therefore, this detailed analysis indicates that considering all the features yields the best results, and all the features or health conditions should be considered for an individual while examining them for an illness.

VI. 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 observed that the data is highly imbalanced and it involved humongous number of predictor variables. From the results, one can interpret 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.



(a) ML scenario 1 with 2423 participants.

(b) ML scenario 2 with 8500 participants.

Fig. 4: Maximizing the class separation by LDA in the two considered ML scenarios.

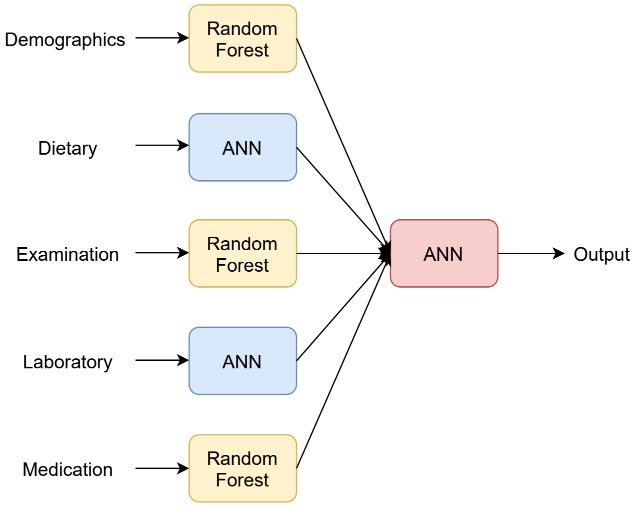


Fig. 5: Hybrid AI model variant.

VII. Conclusion

In this paper, we discussed the importance of developing analytic techniques for the entire large NHANES dataset with the hypertension use-case. A systematic, comparative analytics revealed that our proposed AI model combining LDA and a simple ANN structure outperformed other machine learning techniques and other hybrid model variants. In the health domain, it is always good to have less false negative values than false positive values, and our proposal effectively fulfilled this requirement. Overall, this paper demonstrated 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.

For the future work, we can analyse the remaining diseases from top 5 frequent ones as well as their diagnosis. But again, it might lead to highly class imbalanced data. To make our model more adaptable for training and testing purposes, we can expand this dataset by considering more data of previous years from NHANES website. 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.

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/.

[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 *9th International Conference on Data Mining (DMIN’13), a joint conference with 2013 World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP’13)*, 2013.

[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] 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.

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