IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Healthcare Data Analytics Using Machine Learning Approaches

Mr. S.MANO VENKAT (MTECH)
Assistant Professor

N.Y.V.V.S.S. THANUJA
M-Tech Final Semester

Masters of Technology,
Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

ABSTRACT

Data has become an integral part of the digital world with the advancement in computing technologies. The collection of data is very crucial with regards to data analytics. Every industry makes use of data analytics ranging from financial to other commercial applications but it becomes even more important in healthcare domain for the analysis of healthcare data. The present research work is mainly focused on classification/prediction problems of healthcare data based on machine learning (supervised) approaches using data mining techniques.

There is a need to design an intelligent model (based on machine learning) which can classify the amount of data that is stored in our databases. Human data analytical capability rate is much smaller when compared to the amount of data that is stored. This (classification) becomes even more critical when it comes to healthcare data as it can help to detect, diagnose and treat the patients based on these classified data. These healthcare data are voluminous, high dimensional and diversified in nature which in turn demands more and more data mining applications for classification/prediction purposes. Machine learning has emerged as a significant tool in healthcare research to solve complex classification problems efficiently, effectively and quickly.

In general, treatment of diseases is done by physicians based on their clinical knowledge and personal experience. Since physician's experience varies from expert to expert, they may sometimes diagnose the cases inaccurately and may take more time to do manual treatment. In addition, demand of medical experts is also increasing every-day with population growth. Keeping in mind all these, computer-based Medical Disease Diagnosis Systems (MDDSs) are developed (on the basis of healthcare data) that can empower clinicians to make a timely and better-informed healthcare decisions. However, there are various issues and practical restrictions associated with health-related information among which the most important being, certain class of diseases may occur relatively in small number of patients, which in turn arises a data unbalancing

problem in the medical datasets. Besides, certain diseases have low prevalence, so obtaining large volume of patients representing those diseases can be consistently troublesome. Third, the clinical appearance of patients with the same medical condition differs substantially and it causes conflicting issues. As a result of this uncertainty, MDDSs are often expected to deal with a wide range of features.

The main goal of the thesis is to develop a machine learning-based model for classification tasks using several standard machine learning algorithms like decision tree, logistic regression, support vector machine, neural network, naïve bayes etc. in order to address the issues mentioned above.

There are several Medical Disease Diagnosis System (MDDSs) but unfortunately most of them suffer from one or more drawbacks such as disease-specific clinical model, black box diagnostic model, statistical approaches based clinical model etc. The major contribution of the presented study is to address the issues mentioned above and are briefly discussed below-

1. A case study on healthcare bigdata is conducted to show the way how the size of healthcare data is rapidly increasing, importance of healthcare big data and its challenges, the latest trends to manage and process such data to provide quality health services.
2. Healthcare data sets are usually complex in nature that causes degradation of overall performance of the developed diagnosis system. The thesis includes an innovative entropy-based work to improve the performance of the system, removing the irrelevant features from the medical datasets.
3. Attention is paid to draw awareness of the researchers on the neglected cases like Preterm Birth (PTB) which is a serious public health problem that adversely affects both families and the society. A machine learning based model is introduced to predict PTB cases in the earlier stages itself. For this purpose, three learner classifiers namely, Logistic Regression, Support Vector Machine (SVM) and Decision Tree (DT) are used along with Minimum Information Loss (MIL) discretizer.
4. Specific work related to cardiovaslar disease is introduced to identify the effect of behavioural data on each of the risk factors and predict the occurrence of cardiovascular diseases. A neural network (NN)-based prediction model is designed for cardiovascular diseases. This model is evaluated against four different models namely, Linear Regression (LR), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM).
5. The performances of the models are evaluated in terms of specificity, sensitivity and accuracy. All the experiments are carried out under WEKA tools and TensorFlow framework. The results obtained during the experiments suggest that the performance of the introduced learning models is better than its competent models. In fact, the introduced MDDS can be used in healthcare domain to improve the diagnostic speed, accuracy and reliability. Future studies can be explored by developing MDDS for healthcare big data using big data analytical tools and techniques. Security concerns should also be taken as future opportunities.

Keywords: Healthcare, Data Analytics, MDDS, Machine Learning, Classification, Feature Selection, PTB, Learning Classifiers, Data Mining

1. INTRODUCTION

The term healthcare refers to a system that entails the enhancement of health-related services in order to fulfill the clinical requests of the individuals. In healthcare services, patients, doctors, clinicians, researchers and medical industries are all making an effort to maintain and restore health records. In recent years, with the remarkable development of technologies, data is continuously increasing day by day in every sector including healthcare which in turn demands more and more data mining applications.

Fig 1. Healthcare data analytics using machine learning

However, due to digitization of healthcare system, the medical organizations are generating large amount of healthcare data. In general, healthcare data consists of all records related to health stored in a digital form. It may contain detailed information about patients' medical history, doctors prescribed notes, clinical reports etc. All these data are voluminous, high dimensional and diversified in nature. Healthy decision making is a challenge in the modern era due to this rising complexity of healthcare data.

Machine learning, data mining, and statistical techniques are the key fields of study that enhance the ability of individuals to make the right decisions in order to maximize the outcome of any working domain. Human data analytical capability rate is much smaller when compared to the amount of data that is stored. This becomes even more critical when it comes to healthcare domain as the number of available experts for healthcare data analysis is comparatively less.

Therefore, there is a requirement for computerized (semi-automatic) Medical Disease Diagnosis Systems (MDDSs) that can empower doctors and medical professionals to make informed (better) healthcare decisions for any individuals. This will enhance the quality of care, improve the diagnosis of diseases and ultimately reduce the healthcare costs. The present study is mainly focused on classification/prediction problems of healthcare data based on machine learning (supervised) approaches. There are several learning algorithms which can be used with data mining techniques to resolve classification problems in healthcare domain so as to increase the diagnostic speed, accuracy and reliability.

In recent time, cardiovascular disorders or cardiovascular diseases (CVDs) are considered one of the leading causes of death among men and women across the world. CVD is a very dangerous (deadly) disease that affect blood vessels and heart. In this disease, usually heart is unable to push the required amount of blood to other parts of the body to fulfil the normal functionalities, and due to this, ultimately the heart fail. Heart attacks, strokes and coronary heart disease are the major CVDs. The major risk factors of CVDs include high blood pressure, hypertension, high cholesterol, obesity, smoking and diabetes.

According to the recent survey by WHO, over 17 million deaths have been estimated annually due to cardiovascular diseases in the past few years, which represents about 31% of all deaths globally, including USA and out of these, 85% are caused by heart attacks and strokes. This number is further expected to be increased to 23.6 million by 2030.

CVDs are the major global health problem in the modern medicine. Nowadays, many researchers and clinicians have been using several data mining techniques by employing machine learning approaches in the diagnosis of cardiovascular disorder. Numerous computerized systems have been developed for the diagnosis of cardiovascular disease at an early stage but most of them take into account only a subset of risk factors. Moreover, the performance of the system remains a challenge in the case of complex interactions. In addition, the early detection of cardiovascular disease through traditional medical system is not efficient because of several contributing risk factors such as diabetes, hypertension, obesity and elevated cholesterol. Due to such limitations, medical practitioners depend upon machine learning-based medical systems.

For this reason, designing an automated-intelligent diagnosis system is very essential in order to predict cardiovascular disease accurately and in timely manner. The present study proposed a neural network-based prediction system (based on machine learning) for cardiovascular disease diagnosis. The introduced model deals with the patients' behavioural data in order to identify various risk factors responsible for cardiovascular disease. This computerized MDDS will help doctors to make informed decisions and predictions for diagnosis of CVDs more effectively.

2. LITERATURE STUDY

According to Kohli et al., electronic health records (EHRs) empower the clinicians by integrating patient health history for planning safe and proper treatment. Esfandiary et al. made a comprehensive study about application of data mining in medicine and healthcare, which is based on six medical tasks: screening, diagnosis, treatment, prognosis, monitoring and management.

Prema and Pushpalatha designed a preterm birth model based on the dataset of pregnant women having either diabetes mellitus or gestational diabetes mellitus. The logistic regression and Suppport Vector Machine are used for the prediction purpose.

In 2019, Tahmasseb et al. have developed a multi-parametric magnetic resonance imaging (MRI) machine learning model to predict the risk of breast cancer. In this regard, eight classifiers were used on MRI data of the breast in order to rank the features for early prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer patients.

Dagli et al. have introduced a multilevel perceptron model for survival prediction. The introduced model predicts the survival of non-small cell lung cancer patients with multi-layer neural network. ReliefF and Recursive Feature Elimination algorithms were implemented to extract the features from clinical data in order to increase the accuracy of prediction model.

Zheng et al. suggested an intelligent-based expert system for identifying Type2 Diabetes Mellitus (T2DM) patients with the help of electronic health record (EHR) data. For this purpose, different

machine learning algorithms such as Random Forest (RF), logistic regression (LR), Decision Tree (DT), naïve bayes, K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) were applied.

Zacharaki at el. proposed a framework for brain tumour classification and prediction of degree of malignancy based on multi-parametric imaging profile data such as CT, conventional MRI and advanced MRI.

Rahimian et al. developed a machine learning based model for predicting the risk of emergency hospital admission using electronic health records (EHRs). The proposed model help healthcare practitioners to accurately predict whether a patient may require emergency care in the near future.

Raja et al. designed a machine learning model named as risk prediction conceptual model (RPCM) for the prediction of Preterm Birth (PTB) using three different classifiers, namely, decision tree (DT), logistic regression (LR), and support vector machine (SVM). In this model, feature selection approach is based on entropy-notion.

Sharma and Parmar proposed a deep learning neural network model using Talos in order to predict the heart diseases in human body. In this model, different learning classifiers namely, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB) and Random Forest (RF) were used.

Tan et al. designed a hybrid model named as GA-SVM for attribute selection from 5 medical datasets. The machine learning algorithms like Support Vector Machine (SVM), Naïve Bayes, C4.5, CART and Genetic Algorithms (GA) are used in this model.

Liu et al. [90] presented an efficient computer-aided diagnostic (CAD) base system consisted of fuzzy K-nearest neighbor (FKNN) classifier for diagnosis of thyroid diseases.

An architecture of big data analytics in healthcare have been introduced by Raghupathi and Raghupathi. The authors have discussed about the theoretical implementation of healthcare data analytics using big data. However, the application of data analytics in healthcare organizations are still not widely adopted.

Wang et al. [98] investigated the history of healthcare big data analytics, as well as the execution of architectural design and component functionality. The five main capabilities of healthcare data analysis that have been explored through case studies are namely, pattern recognition capability, analysing unstructured data capability, decision support system capability, predictive capability, and traceability. Unfortunately, no efficient data analytical methods were developed for analysing unstructured data.

Dewan and Sharma in have developed a computer-based system that is capable of predicting cardiovascular order from a past heart disease database record. For this purpose, a back propagation technique using genetic algorithm has been applied.

Pari et al. introduced a machine learning model for accurate prediction of premature birth in order to avoid unnecessary hospitalization for pregnant women having false labour pain. The suggested model is developed using multiple classifiers by analyzing the risk factors associated with preterm birth.

Elaveyini et al. have developed a neural network-based model to predict preterm delivery. A feed-forward backpropagation algorithm was applied to explore the major risk factors (responsible for PTB) present in women during their pregnancy period.

Parthiban and Srivatsa have designed an automated diagnosis system using machine learning methods to predict the chances of a diabetic patient getting a heart disease based on diabetes datasets.

Das et al. have designed an intelligent medical decision support system for diagnosing heart disease in order to improve the ability of the physicians. A neural network ensemble method is in the centre of the proposed system.

Jabbar et al. proposed a decision support system for the classification of cardiovascular disease using K-Nearest Neighbor (KNN) and genetic algorithm in order to improve the diagnostic decision for heart decision prediction.

Otoom et al. have developed an intelligent monitoring system that is capable of predicting a heart disease problem based on clinical data by utilizing three machine learning algorithms namely, Support Vector Machine, BayesNet and Functional Trees.

In worldwide, heart disease is the major issue in human life. The general term 'heart disease' actually comprises a range of conditions that affect the heart.

Some of the most critical ones are high blood pressure, hypertension, elevated cholesterol, lack of physical activity, obesity, smoking and diabetes. In 2016, approximately 17.6 million deaths globally were due to cardiovascular diseases. Cardiovascular diseases were leading among top 10 causes of deaths in United States. 9% (24.3 million in 2016) of the United States adults suffer from one of the cardiovascular diseases. People belonging to some age or race group tends to suffer more than others.

A recent study indicated a relation between resting heart rate and cardiovascular disease. The study was made on diabetic and non-diabetic individuals. Due to higher heart rate in diabetic individuals, it was found that they suffer greater chance of cardiovascular diseases. Heart disease (Cardiovascular disorder) is one of the most common and dangerous disease in the world. Due to this, several medical analysis systems are developed rapidly to predict the risk of cardiovascular diseases.

3. METHODOLOGY

The methodology adopted to design the suggested model is described in this section. Here, the proposed model is formed based on neural network.

Notably, the designing process of the proposed model consists of four phases. The basic step of phase-I deals with collection of behavioural data from BRFSS data repository. In the phase-II, data pre-processing and data normalization is done to fill up the missing data values. Also, a special technique is applied to normalize the data values, particularly for the obesity problems. Then after, neural network based models have been proposed to analyse each of the behavioural risk factors in the next phase. Finally, a machine-learning based prediction model for cardiovascular disease is constructed in the last phase. The overall design of the proposed model is depicted in below figure.

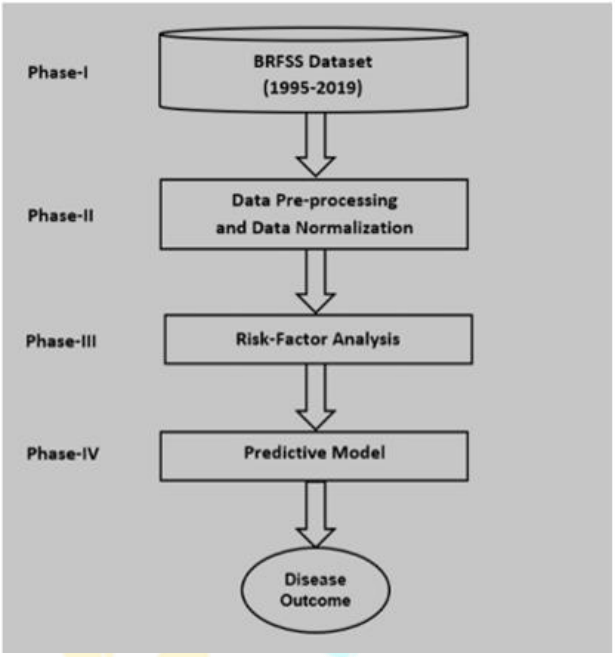

Fig 2. Phases in the Proposed Model

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Data Collection

BRFSS Datasets In this study, the data published by BRFSS from the year 1995-2015 was collected for analysis. The American Heart Association defines the following seven risk factors whose values determine the overall Cardiovascular Health.

- Hypertension (High BP)
- Physical Activity
- Cholesterol level
- Healthy Diet (or Nutrition)
- Healthy Weight (or Obesity)
- Smoking Status
- High Blood Glucose Level (or Diabetes)

4.2 Data Processing

The list of questions(s) that were used to extract data for a particular disease/risk factor is shown as follows.

The combination of Year, Locationabbr, Break_Out_Category and Break_Out formed a key. There was exactly one possible entry in the data set for each possible response for a particular key. When the data sets for two different questions were compared against each other, they were compared by Data values, after taking a join on the key columns.

4.3 Risk Factor Analysis

Each risk factor was analysed separately and then each disease was analysed by the combined effect of all the risk factors. The analysis of the entire data set was done in three stages –

- Raw Analysis – First stage analysis to understand the data and correlation between different variables.
- Subset Data Analysis – A subset of entire data (BRFSS 2011-2013) was analysed to get working models for prediction.
- Complete Analysis – The entire BRFSS data (1995-2015) was analysed.

The results of analysis for the first two stages are omitted as they were only meant for understanding the trend and designing a model for the analysis. The result of the last stage analysis is presented in the present study. After the pre-processing stage, each of the behavioural risk factor was analysed separately for cardiovascular diseases. All models for the risk factors and diseases (discussed in the present section) are built using neural network. Neural network was particularly chosen for its ability to learn complex curves and give continuous outputs. All models employed validation loss monitoring, i.e., the validation loss was continuously monitored during training and the model with the least validation loss was chosen.

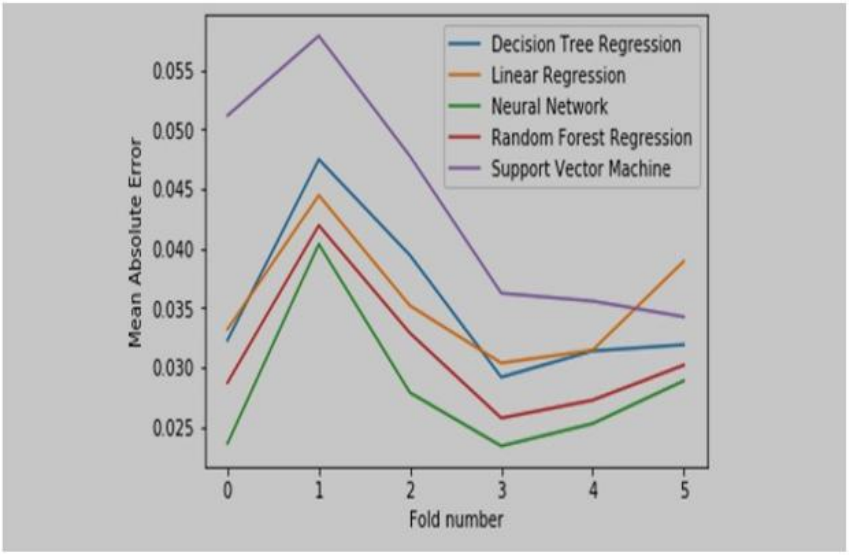

Fig 3. Plot showing the mean absolute error of each model for each fold of cross validation.

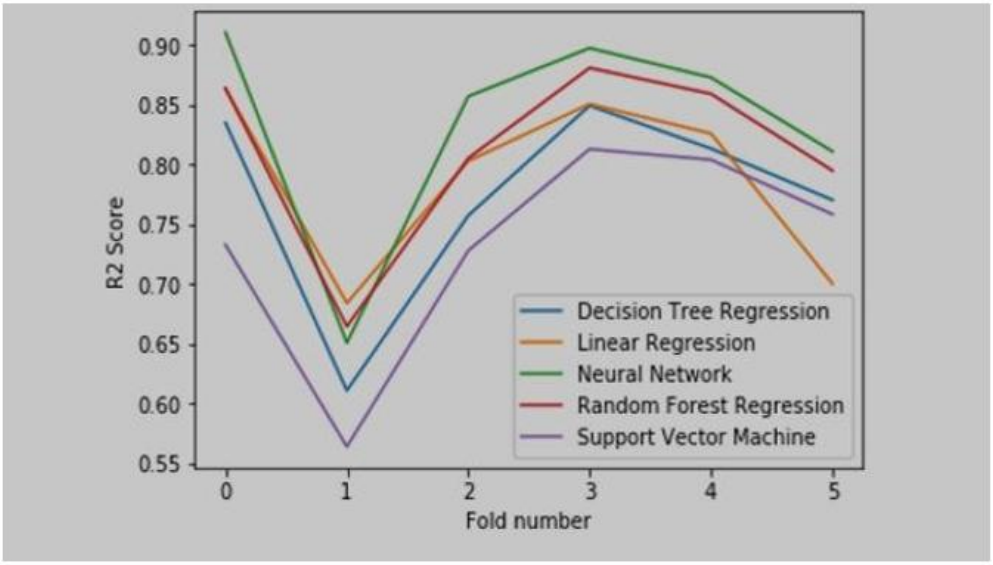

Fig 4. Plot showing the R2 score of each model for each fold of cross validation

Fold No.	Logistic Regression	Decision Tree	Random Forest	Support Vector Machine	Neural Network
1	0.0332	0.0323	0.0287	0.0512	**0.0236**
2	0.0445	0.0475	0.0419	0.0419	**0.0404**
3	0.0352	0.0394	0.0329	0.0329	**0.0279**
4	0.0304	0.0292	0.0258	0.0258	**0.0234**
5	0.0315	0.0314	0.0273	0.0273	**0.0253**
6	0.0390	0.0319	0.0302	0.0343	**0.0289**

Table1 : Absolute Error for Different Models

Fold No.	Logistic Regression	Decision Tree	Random Forest	Support Vector Machine	Neural Network
1	0.8633	0.8350	0.8639	0.7329	**0.9106**
2	**0.6841**	0.6110	0.6646	0.5640	0.6509
3	0.8030	0.7574	0.8053	0.7280	**0.8570**
4	0.8506	0.8493	0.8809	0.8130	**0.8975**
5	0.8260	0.8134	0.8592	0.8041	**0.8723**
6	0.7005	0.7704	0.7951	0.7585	**0.8111**

Table2: shows the R2 score for Different Models

4.4 Diseases Analysis

Each disease was analyzed individually along with the effect of each risk factor on the disease. As discussed earlier, Year, Locationabbr, Break_Out_Category and Break_Out formed a key for a datum. For generating the dataset for disease analysis, the inner join of risk factor data and disease data were taken. All accuracies mentioned below are obtained through mean absolute error.

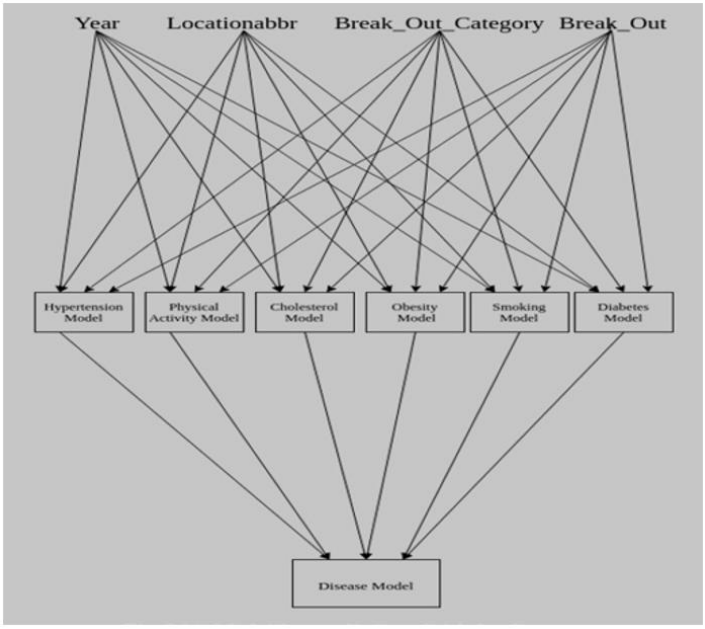

Fig 5. Model for prediction of risk for diseases

The final predictor of the cardiovascular disease model is a combination of 7 different predictors:

1. Hypertension risk predictor
2. Physical Activity level predictor
3. Cholesterol risk predictor
4. Obesity risk predictor
5. Smoking level predictor
6. Diabetes risk predictor
7. Disease risk predictor

Output:

```
✅ PDF Report generated successfully!
📄 brfss_thesis_outputs/Thesis_Results_Report.pdf
```

5. RESULTS AND DISCUSSION

The BRFSS dataset contains large-scale annual health survey responses collected across U.S. states, covering:

- Demographics (age, gender, income, education)
- Behavioral data (physical activity, smoking)
- Chronic conditions (obesity, hypertension, diabetes)

Data were cleaned and normalized; missing values were handled using mean and regression-based imputations.Separate regression equations were used for missing obesity records (as the dataset format changed after 2011).

5.1 Performance Results

Risk Factor	Neural Network MAE	R^2 Score (Approx.)	Remarks
Hypertension	~0.010–0.012	0.92–0.94	Very strong fit, model captures elevation and age effects accurately.
Physical Activity	~0.011	0.89	Slight underfitting due to behavioral variance.
Cholesterol	~0.009	0.91	Consistent across years; less noisy data.
Obesity	~0.012	0.88	Some noise from demographic crossover.
Smoking	~0.010	0.86	Good generalization despite categorical imbalance.
Diabetes	~0.009	0.93	Highly predictable due to stable medical indicators.

5.2 Discussion

- The models effectively captured the interrelationship between risk factors and cardiovascular diseases.
- Neural networks handled the non-linear correlations (e.g., obesity × diabetes) better than traditional methods.
- The generated predictive models can be integrated into early-warning systems or public-health dashboards.
- However, results depend on data accuracy and sampling bias in the BRFSS dataset.

6. CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

The present chapter summarizes the contributions made in this thesis towards the field of healthcare using machine learning algorithms. The main focus of this chapter is to highlight some of the important issues identified during the experiments and their possible future developments based on classification/prediction tasks. This research work deals with providing better healthcare infrastructure including analysis of diseases, addressing the limitations of the existing diseases and develop Medical Disease Diagnosis System (MDDS). It also focuses on the latest trends in the advancement of healthcare system, risk factor analysis of few specific heart-related diseases and the development of machine learning-based model for ignored preterm birth cases in

the field of obstetrics. Since data has become an integral part of the digital world with the advancement in computing technologies. A large volume of data is generating due to the digitization process. There is a need to design an intelligent model (based on machine learning) which can classify the amount of data that is stored in our databases. Human data analytical capability rate is much smaller when compared to the amount of data that is stored.

The facts really confirm that any medical data sets belong to the natural domain. Such data sets are voluminous and highly imbalanced in nature, conflict and vagueness along with high dimensionality. Consequently, developing an appropriate and generalized prediction model for disease diagnosis based on these data sets is a difficult task. Therefore, the key objective of the dissertation is to challenges raised above.

6.2 Future Scope

The ideas mentioned in this thesis can be enhanced in a number of interesting ways in future work. Some of these are outlined below.

- To improve clinical diagnosis decision, a number of experiments would be performed to develop better strategies.
- As much attention may be paid to strengthen the robustness of the medical disease diagnostic system for different types of health-related data in the medical field.
- Medical data sets are extremely complicated and diversified in nature, and our future study will focus to tackle them, as well as we are looking for these type of data sets that contain medical information.
- Future research can be explored by applying the models in certain applications where data are of random (imbalance and uncertain) nature.
- Some models for some specific cases (like Preterm Birth) may be beneficial for rural areas in most of the developing countries where there is a lack of healthcare facilities, and it can be extended in near future developing web-based applications.
- Security issues can also be covered while designing the extended version of Medical Disease Diagnosis System (MDDS).
- Last but not least, future studies can be explored by developing MDDS for healthcare big data using big data analytics tools and techniques.

7. ACKNOWLEDGEMENT

Mr.S. Mano Venkat working as an Assistant professor in Department of CSE in sanketika vidya parishad engineering college, Visakhapatnam, AP.

Nulu Yery V V S S Thanuja is pursuing her final semester M.Tech in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine Learning Nulu Yery V V S S Thanuja has taken up her PG project on **HEALTHCARE**

DATA ANALYTICS USING MACHINE LEARNING APPROACHES and published the paper in connection to the project under the guidance of Mr.S. Mano Venkat, Assistant Professor, SVPEC**.**

8. REFERENCES

1. K. Priyanka and N. Kulennavar, "A survey on big data analytics in health care", International Journal of Computer Science and Information Technologies, vol. 5, no. 4, pp. 5865–5868, 2014

2. R. Raja, I. Mukherjee and B. K. Sarkar, "A Systematic Review of Healthcare Big Data", Scientific Programming, vol. 2020, Article ID 5471849, 15 pages, 2020.

3. Rohan Bhardwaj, Ankitha R. Nambiar, "A Study of Machine Learning in Healthcare", IEEE 41st Annual Computer Software and Applications Conference, 2017.

4. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A, "Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute; 2011.

5. J. Sun and C. K. Reddy, "Big data analytics for healthcare", in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13, Chicago, IL, USA, 2013.

6. Dhillon A, Singh A, "Machine learning in healthcare data analysis: a survey", Journal of Biology and Today's World. 2019;8(6):1-0.

7. Jalpaiguri et al., "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology, vol.10, pp. 2137-2159, 2017.

8. M. Thirugnanam, P. Kumar, S.V. Srivatsan and C.R. Nerlesh, "Improving the prediction rate of diabetes diagnosis using fuzzy, neural network, case based approach (FNC)", Procedia Eng., vol. 38, pp. 1709-1718, 2012.

9. Rakesh Raja, Indrajit Mukherjee, Bikash Kanti Sarkar, "A Machine Learning-Based Prediction Model for Preterm Birth in Rural India", Journal of Healthcare Engineering, vol. 2021, Article Id 6665573, 11 pages, 2021.