Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 8
August 2023

Lebesgue Prevosti Class Balanced Neural Turing Machine Data Classification For Healcathre Data Analytics

S. Sathish Kumar[1], Dr. P. Parameswari[2]

[1]Research Scholar, Department of Computer Science, Karuppannan Mariappan College, Muthur, Thirupur (Dt), Tamilnadu, India.
[2]Principal, Palanisamy College of Arts, No. 17, Erode Road, Perundurai, Erode (Dt), Tamilnadu, India.

Abstract

Medical data refers to the health-related information associated with routine patient care and clinical trial program. Data mining has a phenomenal perspective for healthcare services owing to the mushrooming increase in electronic health records. Digitalisation and innovation of new mechanisms minimise human efforts and make data straightforwardly significant. Machine Learning (ML) technique is used in the healthcare domain to diagnose different diseases. Also, with the aid of classification techniques and healthcare data, significant disease diagnoses are said to be ensured. Therefore, feature selection as a key element is very significant in ML-assisted diagnosis. So far has designed several feature selection methods. However, inspecting the accuracy has developed hardly any observations and time complexity with rational feature selection. To address class-imbalanced data without fine-tuning, utilising traditional feature selection models for classification can only be done with a smooth process. Lebesgue Prevosti Class-balanced Neural Turning Machine Data Classification (LPC-NTMDC) proposed for disease diagnosis to address this issue. The LPC-NTMDC method has split into two parts, namely, feature selection and classification. First, the Lebesgue Prevosti Rationality-based Feature selection algorithm proposed to carry out rational sparse feature selection applied to complex class-imbalanced data. Following this, to evaluate the rationality of feature selection has employed a recurrence-based scale represents a pipeline to select inherent features considering both rationality and class imbalance. Second, with the inherent selected features, Neural Turing Machine Data Classification-based disease diagnosis is designed. Here with Fuzzy feature matching capabilities, significant medical data classification is made for disease diagnosis. Experimental results on the cardiovascular disease dataset have demonstrated the feasibility of our method. Furthermore, the experimental results stipulate that the proposed LPC-NTMDC method outperforms the state-of-the-art methods regarding classification time, accuracy, false positive rate and space complexity.

Keywords: Machine Learning, Lebesgue Prevosti Rationality, Feature selection, Neural Turing Machine, Data Classification

1. Introduction

A straightforward clinical data encoding into fixed-length feature vector representation was conducted [1]. As a result, a new model has presented with efficient feature selection from representation. In addition, COVID-19 patients underwent machine learning algorithms for classification purposes. However, with the aid of the encoding model, the accuracy level was not improved.

A highly-scalable and robust machine learning framework was presented in [2] to forecast the adversity represented through mortality and ICU admission with readmission from vital sign time series. An unsupervised LSTM Autoencoder studied the optimal time-series representation to differentiate less frequent patterns with the adverse event from majority patterns. The means of static features improve the prediction performance has constructed the representation with the aid of a gradient-boosting model. However, computational complexity has not diminished by the designed framework.

Self-Rule to Multi-Adapt (SRMA) was presented in [3] with self-supervised learning for domain adaptation and eliminated the fully-labelled

source dataset needs. SRMA reassigned the discriminative knowledge attained from labelled source domain data to a target domain without any tissue annotation—the designed method harness domains structures by collecting visual similarity between intra-domain and cross-domain self-supervision. The designed method learnt the data from multiple source domains. However, using SRMA has not minimised the time complexity.

A deep spatio-temporal meta-learning model was introduced in [4] to predict the traffic revitalisation index (DeepMeta-TRI) with COVID-19 data. Meta-learning and external auxiliary information were combined to predict the urban traffic revitalisation index. In addition, disease prediction has combined the meta-graph convolution network and meta-temporal convolution modules. However, the deep spatio-temporal meta-learning model did not improve prediction accuracy.

Heart disease prediction is considered one of the significant issues as far as clinical data analysis is concerned. However, due to increasing the data size, many complications arise for analysing and are specifically found to be laborious in maintaining e-healthcare data.

A new heart disease prediction method involving Feature Extraction, minimisation of attributes and in [5] has proposed classification. To begin with, to extract the statistical and higher-order statistical features, using Component Analysis has performed the attribute minimisation. Finally, using Neural Network (NN) has conducted the actual prediction process to improve accuracy. As a result, the utilisation of ML derives as a solution to minimise heart disease symptoms.

In [6], applying a feature selection technique, a dimensionality reduction method has deigned to identify heart disease features. It results in dimensionality reduction and improved accuracy to a greater extent. However, another vaccine prediction model in the wakening of COVID-19 was designed in [7] employing neural molecular dynamics. The role played by ML in disease diagnosis was studied [8].

Early diagnosis of heart disease and classification play a significant role in clinical data analysis. Regarding e-healthcare, a novel feature selection in [9] has proposed the application of deep learning for detecting and classifying disease. To resolve disadvantages like information loss [10] has designed a dual-stp hybrid feature selection method was used. First, constructing the diagnostic model has identified the most relevant contributing features. Next, a timely manner makes the actual classification for disease diagnosis.

Over the past few years, medical technology has progressed. For example, to identify and classify heart disease in healthcare data analytics has designed a successful Lebesgue Prevosti Class-balanced Neural Turing Machine Data Classification (LPC-NTMDC) method. Because cardiovascular or heart disease is deadly, early detection is crucial to boosting patient survival rates.

1.1 Contribution remarks

The contributions of Lebesgue Prevosti Class-balanced Neural Turing Machine Data Classification (LPC-NTMDC) for disease diagnosis are listed below:

1. To propose a class-balanced and rational healthcare data analytics for cardiovascular disease diagnosis method.
2. To select the inherent features for cardiovascular disease detection via Lebesgue Prevosti distance has designed a Lebesgue Prevosti Rationality-based Feature selection and the linear combinatory matrix that checks with the corresponding threshold to selected relevant features.
3. Employing a Fuzzy feature matching with the selected feature for detecting either the presence or absence of cardiovascular disease has designed a Neural Turing Machine Data Classification-based diagnosis.
4. The proposed method has been implemented and tested with 50000 samples. The classification accuracy, classification time, false positive rate and space complexity for different numbers of samples while using the proposed method is determined, and theoretical analysis presents that the proposed method outperforms the state-of-the-art methods.

2. Related works

In recent decades, chronic disease has been one of the biggest threats to human life, resulting in the diagnosis and prediction of chronic disease

before minimising the mortality rate. Some leading chronic diseases are Parkinson's, lung cancer, heart disease, lung cancer, chronic kidney disease, Hepatitis and so on.

In [11] has designed a machine-learning-based prediction method. Here, as a predictive mechanism for Parkinson's disease (PD), prediction utilised a support vector machine (SVM). In addition, selected inherent features have utilised L1-norm SVM of features selection for precise classification between PD and healthy people. However, another machine learning technique-based method, fast conditional mutual information, in [12] has presented SVM for accurate prediction.
A review of feature selection in [13] has reviewed dimensionality reduction and classification methods. A literature survey on several methods for diagnosing several diseases and in [14] applied an overview of domain areas in which Artificial Intelligence (AI).

In [15] has presented a detailed and systematic literature review on the application of AI for disease diagnosis. However, feature selection becomes paramount, specifically with the inclusion of several variables and features. Unimportant variables would discard the reason, resulting in accurate classification performance. In [16], for handling feature selection issues, applied Random Forest and classified accurately even in the case of a higher number of variables.

For chronic kidney disease, machine learning has presented identification and detection [17]. However, another hybrid method was designed to employ feature selection and classification ensembles [18]. This type of ensemble model has eliminated weak selectors, and aggregating the strong selectors ensured accurate and error-free predictions.

A data-driven approach using ML is presented [19]. Incorporating the data-driven model resulted in accurate and timely cardiovascular disease prediction. However, the cost factor was not involved in the above design. Instead, the deep neural network is applied [20] for heart disease diagnosis, which improves the accuracy and sensitivity rate to a greater extent.

Most of the existing prediction mechanism tries to optimal a single objective, like reducing the classification time or enhancing the classification accuracy, or else if both objectives were arrived at, the falsification results were not analysed. However, single objective optimisation may reduce the overall performance and efficiency. So, to run such types of healthcare data analytics, i.e., prediction of cardiovascular disease diagnosis with minimum classification time, false positive rate and higher classification accuracy, remains challenging. To overcome the existing mechanisms' shortfalls, we have developed a method called Lebesgue Prevosti Class-balanced Neural Turing Machine Data Classification (LPC-NTMDC) for solving multiple objectives. In this paper, we have designed a Lebesgue Prevosti Rationality-based Feature selection. To has ensured a class-balanced and rational feature selection. In turn, it reduces the space complexity and false positive rate, improving the classification accuracy of identifying diseased patients.

3. Methodology

This study developed a successful Lebesgue Prevosti Class-balanced Neural Turing Machine Data Classification (LPC-NTMDC) for disease diagnosis in the context of the healthcare domain. Lebesgue Prevosti Rationality-based Feature selection and the Neural Turing Machine Data Classification-based disease diagnosis comprise the proposed LPC-NTMDC approach. Using LPC-NTMDC to identify the most inherent features improves classification accuracy while decreasing processing complexity.

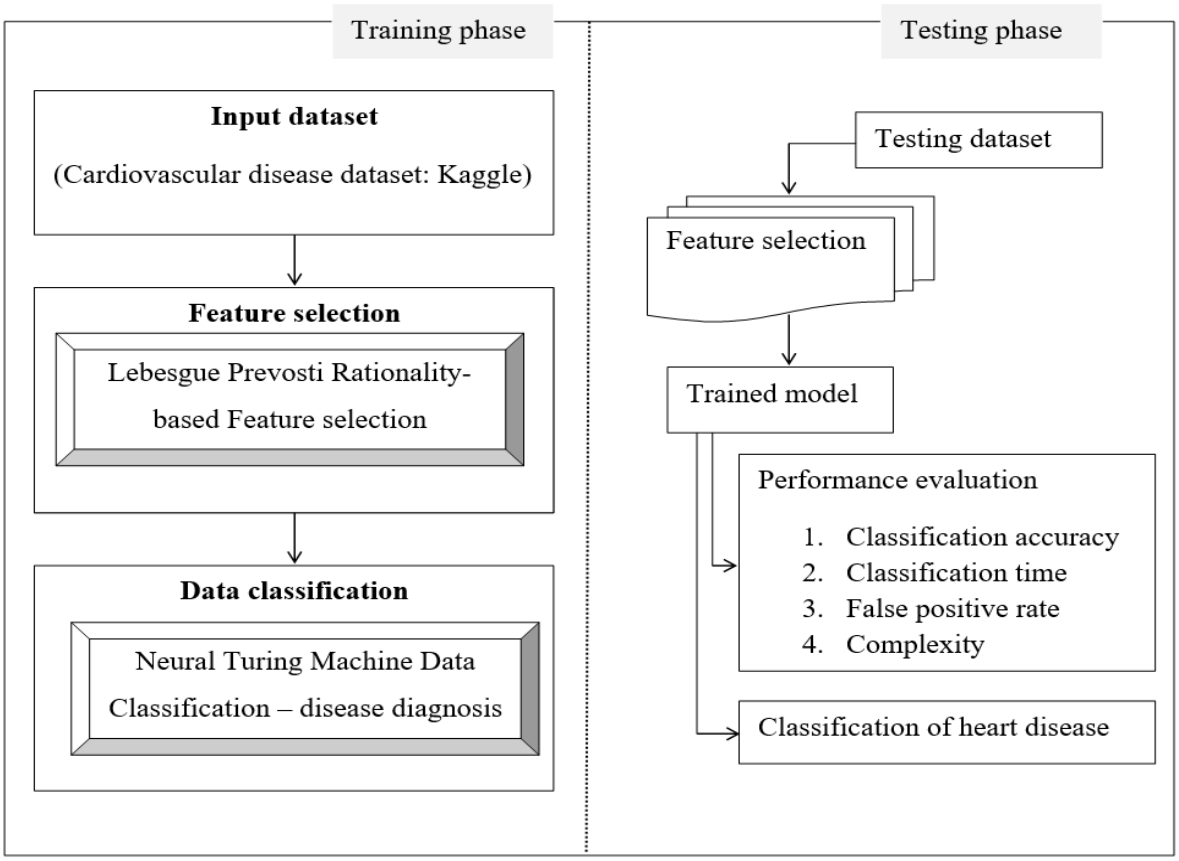

Figure 1 Block diagram of LPC-NTMDC

Into the training and testing phase has split the block diagram of the LPC-NTMDC method. First, in the training phase with the cardiovascular disease dataset as input, feature selection and classification are made using Lebesgue Prevosti Rationality-based Feature selection and Neural Turing Machine Data Classification-based disease diagnosis. Next, in the testing phase, with the trained model, the performance evaluation in classification accuracy, classification time, false positive rate and space complexity are measured to validate the LPC-NTMDC method. The subsequent sections following the dataset description provide a detailed description of the LPC-NTMDC method.

3.1 Dataset description

The cardiovascular disease dataset (https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset) employed in our work comprises 70 000 patient data records, 11 features and a target feature. Also, the overall 11 features have split into three types of input features:

1. Objective – factual information
2. Examination – results of medical examination
3. Subjective – the information was given by the patient

S. No	Features	Description
1	Age	Objective feature
2	Height	Objective feature
3	Weight	Objective feature
4	Gender	Objective feature
5	Systolic blood pressure	Examination feature
6	Diastolic blood pressure	Examination feature
7	Cholesterol	Examination feature
8	Glucose	Examination feature

9	Smoking	Subjective feature
10	Alcohol intake	Subjective feature
11	Physical activity	Subjective feature
12	Presence or absence of cardiovascular disease	Target variable

Table 1 Cardiovascular disease dataset features

The input vector matrix has modelled the raw data obtained from the cardiovascular disease dataset as given below, with 'S' representing the sample set used for simulation concerning the feature set 'F' respectively.

$$IVM = \begin{bmatrix} S_1F_1 & S_1F_2 & S_1F_3 & \dots & S_1F_n \\ S_2F_1 & S_2F_2 & S_2F_3 & \dots & S_2F_n \\ \dots & \dots & \dots & \dots & \dots \\ S_mF_1 & S_mF_2 & S_mF_3 & \dots & S_mF_n \end{bmatrix} \quad (1)$$

Input vector matrix 'IVM' as given in (1), 'S' corresponds to the samples and 'F' represents the features of use. With the above 'IVM', a machine learning-based classification method for disease diagnosis is elaborated in detail.

3.1 Lebesgue Prevosti Rationality-based Feature selection model

Over the past few years, feature selection has gained a sizeable amount of interest in class imbalance and rationality learning owing to the high dimensionality of class-imbalanced data in healthcare data. To address high-dimensional data presented different feature selection models. However, to handle the rationality and class distribution in the class imbalance environment scientifically designed, only a negligible number of them. Hence, executing feature selection from imbalanced class data and rationality evaluation remains a demanding task owing to the intrinsic complex aspects of such medical data, and to significantly reconstruct extensive amounts of raw medical data into knowledge representation has necessitated a novel principle. Therefore, the Lebesgue Prevosti Rationality-based Feature Selection process is carried out in the proposed method to select the necessary features from the input database. Figure 2 shows the structure of the Lebesgue Prevosti Rationality-based Feature selection model.

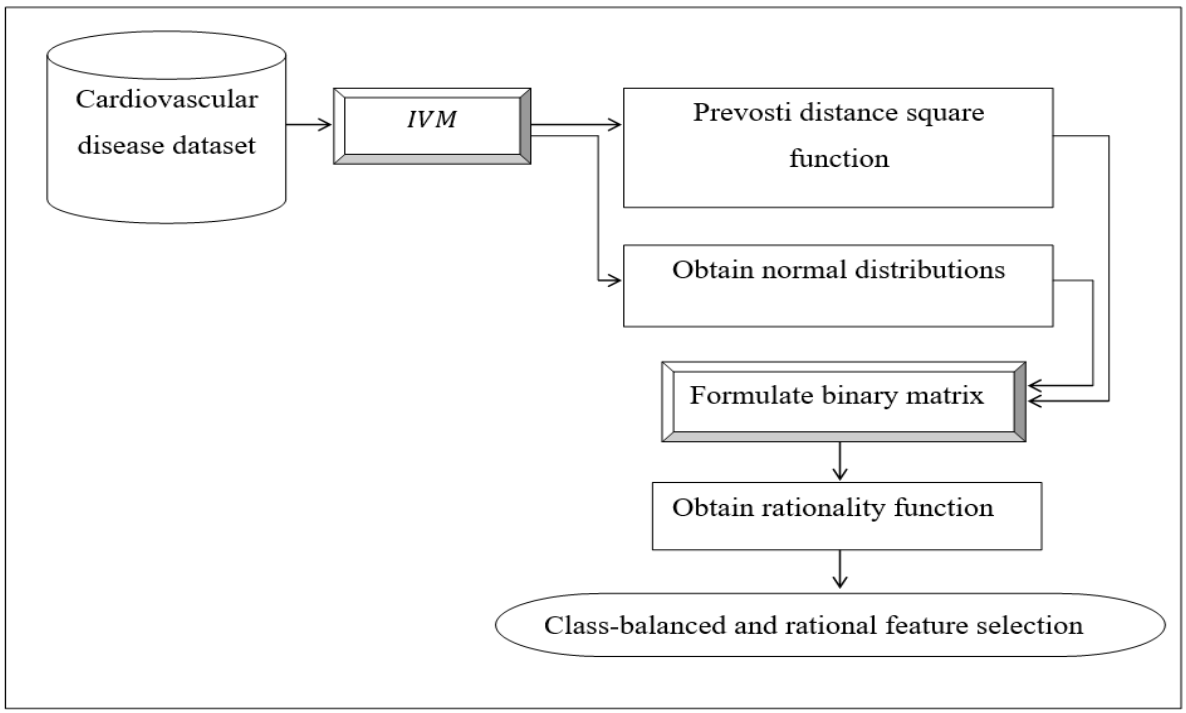

Figure 2 Structure of Lebesgue Prevosti Rationality-based Feature selection model

As illustrated in the above figure, with the cardiovascular disease dataset provided as input, to obtain class-balanced and rational features, in determining the variation between two variables for feature selection, has utilised Prevosti distance along with the rationality function. With the input vector matrix 'IVM', Lebesgue Prevosti distance is applied to reduce the space complexity involved in the identification or selection of features. The Prevosti distance is a measure of the distributional divergence. Let 'X' and 'Y' be two continuous probability measures concerning a third probability measure, 'α'.

Let us further consider 'A' and 'B' represent two populations '$(a_1, a_2, \dots, a_r)$', '$(b_1, b_2, \dots, b_r)$', then '$A_i = (a_{i1}, a_{i2}, \dots, a_{iki}), \sum_{j=1}^{ki} a_{ij}$' and '$B_i = (b_{i1}, b_{i2}, \dots, b_{iki}), \sum_{j=1}^{ki} b_{ij}$'. As given below has measured the square of the Prevosti distance.

$$Dis_{AB}^2 = \int \left(\sqrt{\frac{dA_{ij}}{d\alpha}} - \sqrt{\frac{dbB_{ij}}{d\alpha}} \right)^2 d\alpha \quad (2)$$

From the above formulate (2), 'Dis_{AB}^2' represents the Lebesgue measure '$\frac{dA_{ij}}{d\alpha}$' and '$\frac{dB_{ij}}{d\alpha}$' denoting the two probability density functions. Based on the bi-normal assumption, 'A' and 'B' represents the two normal distributions formulated as given below.

$$\frac{dA_{ij}}{d\alpha} = f_1(x) \sim S(\mu_1, \sigma_1^2) \quad (3)$$

$$\frac{dB_{ij}}{d\alpha} = f_0(x) \sim S(\mu_0, \sigma_0^2) \quad (4)$$

Then, the above square of the Prevosti distance with two normal distributions from equation (2). Below it was rewritten.

$$Dis_{AB}^2 = \int \left(\sqrt{f_1(x)} - \sqrt{f_0(x)} \right)^2 dx \quad (5)$$

With the above formulation, let 'x_i' represent the '$i-th$' feature '$i = (1,2,3,..,r)$' and '$X = (x_1, x_2, x_3, \dots, x_r)$' and '$\beta = (\beta_1, \beta_2, \dots, \beta_r)$' represents a coefficient vector. Then, below has formulated the class-balanced data concerning linear combinatory matrix 'X' rather than a single feature.

$$A = x\alpha|(y = 1) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r|(y = 1) \quad (6)$$

$$B = x\alpha|(y = -1) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r|(y = -1) \quad (7)$$

Finally, a recurrence-based scale is employed to evaluate the rationality of a feature-select method. The feature selection model is performed on the 'r' dataset 'DS' for selecting inherent features. A binary matrix 'X' is given below to indicate the feature selection results of 'm' probabilities.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1r} \\ x_{21} & x_{22} & \dots & x_{2r} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mr} \end{bmatrix} \quad (8)$$

From the above equation (8), each row in the dataset represents one feature selection try. Based on the binary matrix 'X', the rationality of feature selection regarding the recurrence-based criterion is defined below.

$$R(X) = 1 - \frac{\frac{1}{r}\sum_{j=1}^{r}\left[\frac{m}{m-1}\left(\frac{1}{m}\sum_{i=1}^{m} x_{ij}\right)\left(1-\frac{1}{m}\sum_{i=1}^{m} x_{ij}\right)\right]}{\frac{\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij}}{r}\left(1-\frac{\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij}}{r}\right)} \quad (9)$$

From the above equation (9) results, the value of '$R(X)$' ranges from '0 to 1', the greater the value better the rationality and higher probability of the inherent feature selection. Below has given the pseudo-code representation of Lebesgue Prevosti Rationality-based Feature selection.

Input: Dataset 'DS', Samples '$S = \{S_1, S_2, \dots, S_n\}$', Features '$F = \{F_1, F_2, \dots, F_n\}$'
Output: class-based and rational feature selection
Step 1: **Initialize** 'm', 'n' Step 2: **Begin** Step 3: **For** each Dataset 'DS' with Samples 'S' and Features 'F' Step 4: Formulate the input vector matrix as given in equation (1) Step 5: Evaluate the square of the Prevosti distance as given in equation (2) Step 6: Based on the bi-normal assumption, formulate two normal distributions as given in equations (3) and (4) Step 7: Rewrite the square of the Prevosti distance based on two normal distributions as given in equation (5) Step 8: Obtain class-balanced data concerning linear combinatory matrix 'X' as given in equations (6) and

(7)
Step 9: Obtain binary matrix as given in equation (8)
Step 10: Evaluate rationality-based features as given in equation (9)
Step 11: **If** '$R(X) \geq 0.5$ *and* $R(X) = 1$'
Step 12: **Then** feature selected
Step 13: **Return** feature selected 'FS'
Step 14: **End if**
Step 15: **If** '$R(X) \geq 0$ *and* $R(X) \leq 0.5$'
Step 16: **Then** the feature not selected
Step 17: **End if**
Step 18: **End for**
Step 19: **End**

Algorithm 1 Lebesgue Prevosti Rationality-based Feature Selection

As given in the Lebesgue Prevosti Rationality-based Feature selection algorithm, with the sample and features listed in the cardiovascular dataset as input, initially, the square of the Prevosti distance is evaluated for each input vector matrix. Second, employing a linear combinatory matrix, class-balanced data are obtained. Finally, with rationality-based features in recurrence-based criterion, relevant and inherent features are selected computationally efficiently.

3.2 Neural Turing Machine Data Classification-based disease diagnosis

After the relevant and inherent feature selection process, the proposed method to categorise medical data into two classes (i.e., presence or absence of cardiovascular disease) has carried out the Neural Turing Machine Data Classification process. The Neural Turing Machine Data Classification process joins the fuzzy feature matching capabilities with medical data mining objectives for efficient data classification. Figure 3 shows the structure of Neural Turing Machine Data Classification for efficient healthcare data analytics.

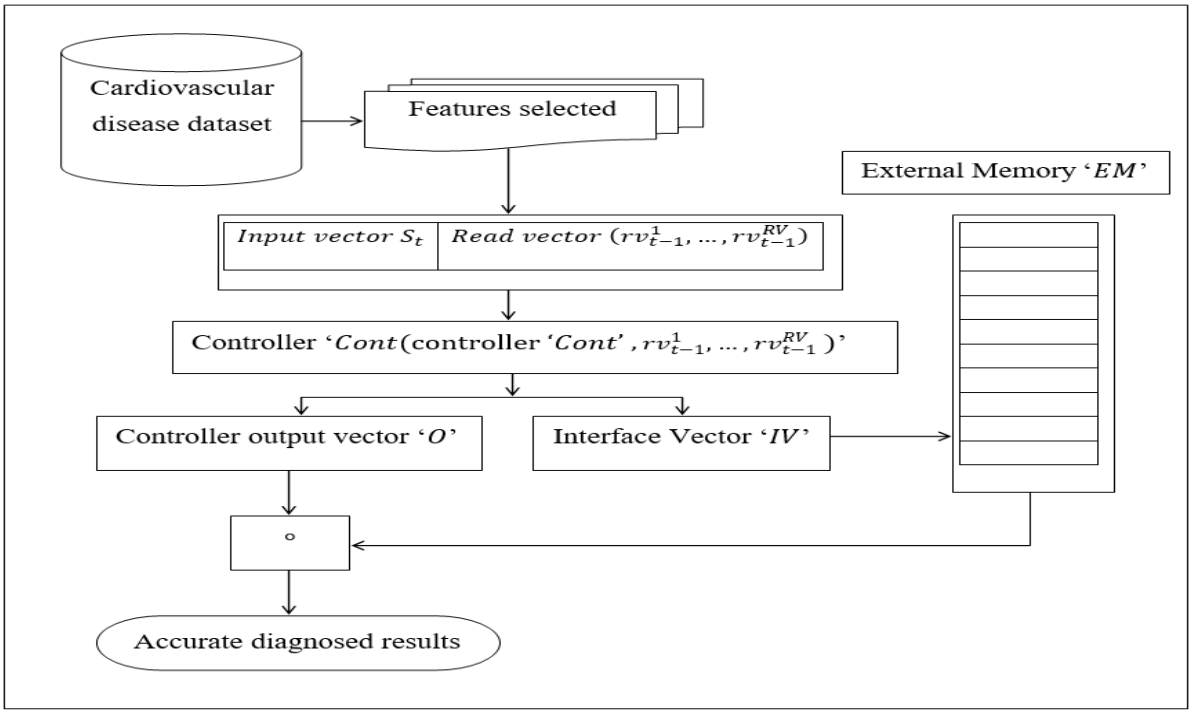

Figure 3 Structure of Neural Turing Machine Data Classification for efficient healthcare data analytics

As illustrated in the above figure, the Neural Turing Machine Data Classification for efficient healthcare data analytics comprises a controller '$Cont$' and an external memory 'EM'. The controller controls the overall healthcare data analytics process in our work, whereas the external memory 'EM' is an element of a set 'R^{N*M}', where 'N' represents the number of '$M-dimensional$' vectors for analysis. Let us further assume that the read vectors are '$RV, rv_t^i = (i = 1,2, \dots, rv)$' are generated upon occurrence of the read operation at the time 't' with dimension 'M' respectively. The sample input vector 'S_t' (i.e., training data) and read vectors 'RV' (i.e., testing data) generated at the time '$t-1$' are merged therefore forming as the input of the controller '$Cont$'. The controller '$Cont$' issues an interface vector 'IV' and a controller output vector 'O'. Also, the interface vector 'IV' interact between controller '$Cont$' and external memory 'EM'. The read operation (i.e., acquiring the samples and features selected for healthcare data analytics of disease diagnosis) is written below with the above design.

$$rv_t^i = EM_t^T rw_t^{rv,i} \quad (10)$$

From the above equation (10), the '$i-th$' read vector at time 't' is acquired based on the transposed external memory values 'EM_t^T' at time 't' and the read weight vector 'rw_t^{rv}' at time 't'. Upon successful completion of the read operation, the generated output vector 'O_t' is finally utilised in obtaining the final output. The external memory in our work holds the values of the patient's cholesterol level, glucose value, smoking habit and alcohol intake with which the fuzzy feature matching capabilities perform medical data mining for efficient data classification. The below mathematically formulated the write operation (i.e., processing the data acquired based on the transposed external memory values and storing the results in the output vector).

$$EM_t = EM_{t-1} \circ (ww_t^{wv} EV_t) + ww_t^{wv} \quad (11)$$

From the above equation (11), with the resultant values of the external memory 'EM' at time '$t-1$', 't'and element-wise product '$\circ$', of the write weight vector 'ww_t^{rv}' at time 't', the erase vector 'EV_t^T', the final classified results are generated as output in the output vector. The pseudo-code representation of Neural Turing Machine Data Classification-based disease diagnosis is below.

Input: Dataset 'DS', Samples '$S = \{S_1, S_2, \dots, S_n\}$'
Output: Accurate and timely disease diagnosis
Step 1: **Initialize** 'm', 'n', feature selected 'FS', read vector '$rv_t^i = (i = 1,2, \dots, rv)$' Step 2: **Begin** Step 3: **For** each Dataset 'DS' with Samples 'S' and feature selected 'FS' Step 4: Perform the read operation as given in equation (10) Step 5: Perform fuzzy matching Step 6: **If** '$(Cholesterol = 3)$' and '$(Glucose = 3)$' and '$(Smoking = 1)$' and '$(Alcohol\ intake = 1)$' Step 7: **Then** '$O_t = Presence\ of\ cardiovascular\ disease$' and '$EV_t = 0$' Step 8: **Else** '$O_t = Absence\ of\ cardiovascular\ disease$' and '$EV_t = 1$' Step 9: **End if** Step 10: Perform the write operation as given in equation (11) Step 11: **End for** Step 12: **End**

Algorithm 2 Neural Turing Machine Data Classification-based disease diagnosis

As given above, early and timely disease diagnosis is the main objective of designing a Neural Turing Machine Data Classification model. The Neural Turing Machine Data Classification, as given above, includes an input vector (samples provided for simulation or training data), read vector (i.e., testing data) and external memory (i.e., the values of cholesterol, glucose, smoking and alcohol intake of each patient). For each sample and selected feature,

perform the read operation. Followed by which fuzzy matching has done to diagnose cardiovascular disease with the values collected at the time of medical examination. Finally, the results obtained via the interface vector that interacts between the controller and external memory has stored in the output vector.

4. Experimental setup

Python has implemented The proposed method Lebesgue Prevosti Class-balanced Neural Turing Machine Data Classification (LPC-NTMDC) for disease diagnosis, a high-level, general-purpose programming language. Using the benchmark cardiovascular disease dataset, https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset has evaluated the performance of the proposed method. In addition, parameters like classification accuracy, classification time, false positive rate and space complexity for varied samples collected at different time intervals carried out experimental evaluations.

4.1 Performance metrics

One of the significant parameters of analysis for healthcare data analytics concerning disease diagnosis is the rate of accuracy with which the prediction. Accuracy is directly associated with health sciences as early disease diagnosis can reduce further aggravation. Therefore, the classification accuracy towards disease diagnosis is measured.

$$Class_{acc} = \sum_{i=1}^{n} \frac{S_{CA}}{S_i} * 100$$

(12)

From the above equation (12), classification accuracy '$Class_{acc}$' is measured based on the samples involved in experimentation 'S_i' and the samples classified accurately 'S_{CA}'. In terms of percentage (%) has measured. The second factor required for disease diagnosis is the classification time. Early diagnosis is also said to be ensured and minimises the mortality rate. As given below evaluated, the classification time.

$$Class_{time} = \sum_{i=1}^{n} S_i * Time\ [Class]$$

(13)

From the above equation (13), the classification time '$Class_{time}$' is measured based on the samples involved in the classification process 'S_i' and the actual time consumed in the classification of target variables for the corresponding sample '$Time\ [Class]$' using fuzzy pattern matching. In terms of milliseconds (ms) measured. Third, that denotes the ratio between the number of negative events (nondisease patients classified as diseased patients) wrongly categorised as positive (false positive) has calculated a false positive rate and the total number of actual negative events (actual diseased patients). In other words, the false positive rate refers to the expectancy of a false positive ratio.

$$FPR = \frac{FP}{FP+TN}$$

(14)

From the above equation (14), the false positive rate 'FPR' is measured based on the number of false positives 'FP' (i.e., non-diseased patients identified as diseased) and the number of true negative rates 'TN' (i.e., number of non-diseased patients). Finally evaluated, the space complex involved in analysing the entire process. During disease diagnosis, the stack that has scored the intermediate evaluations, like features selected and classified results, consumes a certain amount of space. This space is referred to as the space complexity and is measured as given below.

$$SC = \sum_{i=1}^{n} S_i * Mem\left[rv_t^i + EM_t\right]$$

(15)

From the above equation (15), the space complexity 'SC' is measured based on the samples involved in the simulation process 'S_i' and the actual memory 'Mem' consumed in performing the read 'rv_t^i' and write 'EM_t' operations respectively.

4.2 Results and discussion

In this section, the analysis of the results for four different parameters, classification time, classification accuracy, false positive rate and health sciences, has discussed space complexity for disease diagnosis and a comparison is made with the proposed Lebesgue Prevosti Class-balanced Neural Turing Machine Data Classification (LPC-NTMDC) method and existing methods, Clinical data encoding using

machine learning [1] and Stacked ensemble [2] using the cardiovascular dataset from Kaggle.

4.2.1 Performance analysis of classification accuracy

Different performance metrics has used to determine the effectiveness of the proposed LPC-NTMDG method. Classification accuracy represents the overall classification ability of the proposed learning method. Table 2 below shows the performance evaluation of the classification accuracy using the three methods, LPC-NTMDC, Clinical data encoding using machine learning [1] and stacked ensemble [2], respectively.

Samples	Classification accuracy (%)		
	LPC-NTMDC	Clinical data encoding using machine learning	Stacked ensemble
5000	97.7	96.1	95.1
10000	96.35	94.35	93.15
15000	94	93	90
20000	92.15	90	87.25
25000	92	88.25	85
30000	90.35	85.35	83.15
35000	90	83	80
40000	88.15	81.55	78
45000	86	80	75.25
50000	84.35	78.15	73

Table 2 Performance evaluation of classification accuracy

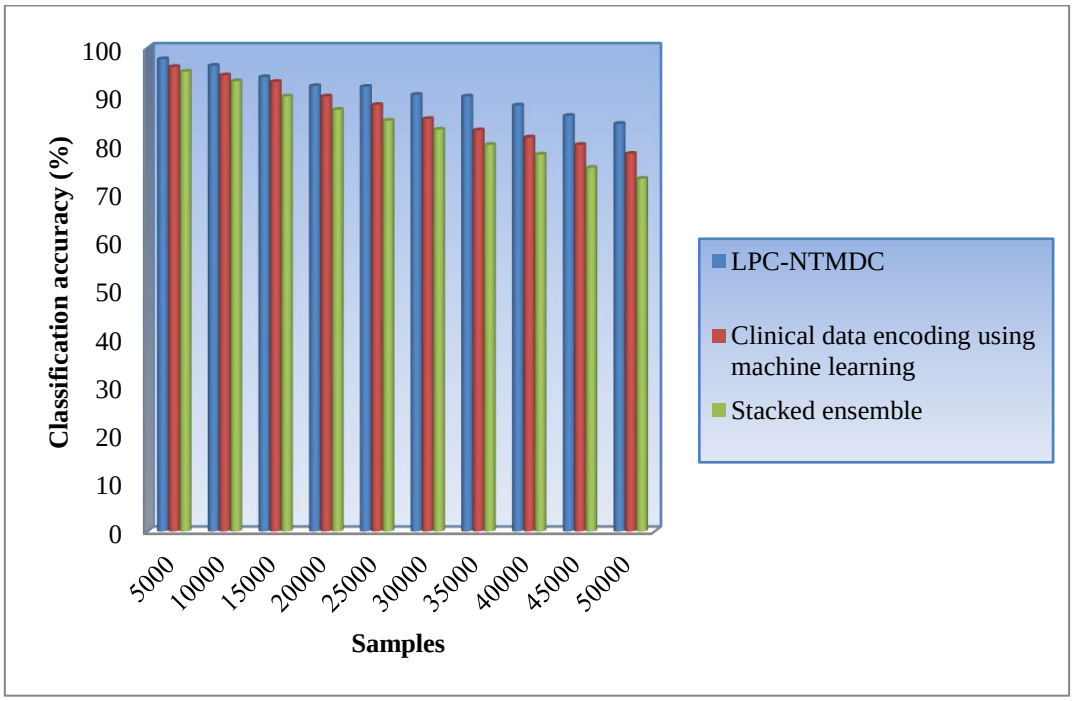

Figure 4 Graphical representation of classification accuracy

The classification accuracy concerning 50000 samples collected from males and females, from the figure, is inferred that the classification accuracy is inversely proportional to the number of samples given as input. Specifically, increasing the frequency of samples results in a proportionate decrease in the prediction accuracy. However, with '5000' numbers of samples considered for simulation and '4885' patient samples correctly classified using LPC-NTMDC, '4805' numbers of samples correctly classified using [1] and '4755' numbers of

samples correctly classified using [2], the overall classification accuracy was observed to be 97.7%, 96.10% and 95.1% using the three different methods. From this, the classification accuracy was better using the LPC-NTMDC method when compared to [1] and [2]. The improvement is due to applying the Lebesgue Prevosti Rationality-based Feature selection model. Only class-balanced and rational features involved in classification were selected using two normal distributions. Splitting the features into structured and categorical that, in turn, selects the inherent has achieved, therefore accurately classifying the cardiovascular disease using the LPC-NTMDC method by 5% compared to [1] and 9% compared to [2].

4.2.2 Performance analysis of classification time

This section presents the classification time involved in health sciences for cardiovascular disease prediction using the proposed LPC-NTMDC method. Table 3 below shows the performance evaluation of the classification time using the three methods, LPC-NTMDC, Clinical data encoding using machine learning [1] and stacked ensemble [2], respectively.

Samples	Classification time (ms)		
	LPC-NTMDC	Clinical data encoding using machine learning	Stacked ensemble
5000	65	90	105
10000	85	95	115
15000	90	115	140
20000	105	130	155
25000	115	145	170
30000	120	160	185
35000	135	185	215
40000	150	200	235
45000	165	215	250
50000	185	230	280

Table 3 Performance evaluation of classification time

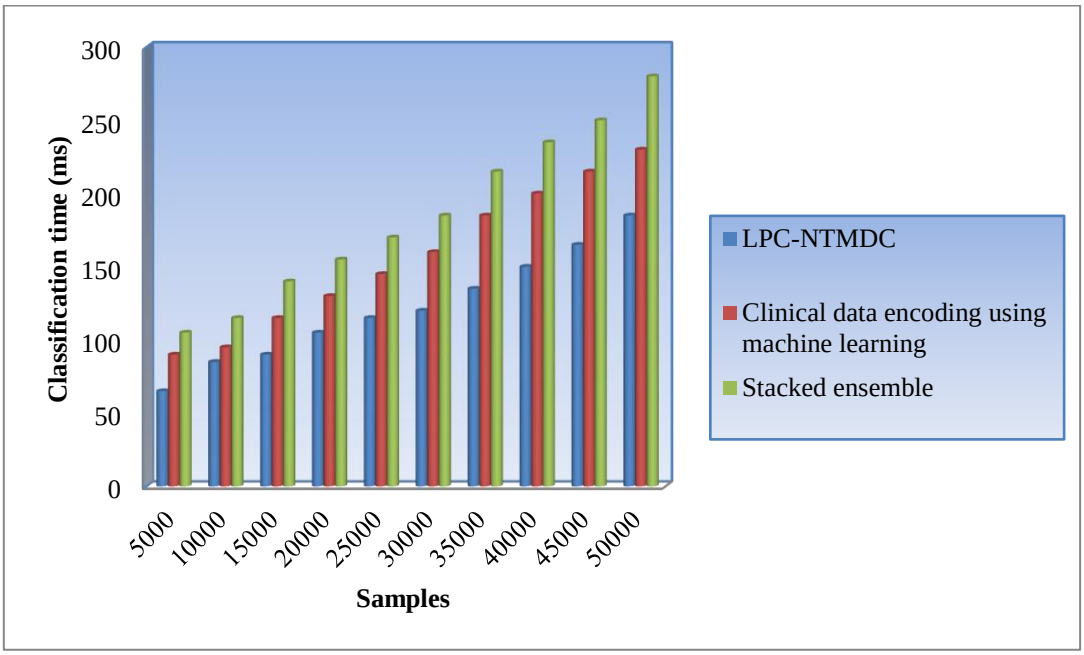

Figure 5 Graphical representation of classification time

The figure shows the classification time for 50000 samples for healthcare data analysis towards disease diagnosis. From the figure, the classification time is directly proportional to the number of samples given as input for diagnosing cardiovascular disease. Therefore, being performed increases the number of samples, testing, and classification time. With '5000' numbers of samples considered for experimentation and the time involved in classifying between the presence or absence of disease being '$0.013ms$' using LPC-NTMDC, '0.018ms' using [1] and time involved in classifying between the presence or absence of disease being '$0.021ms$' using [2], the overall classification time was observed to be 65ms, 90ms and 105ms using the three methods. From the results, the classification time is comparatively lesser using LPC-NTMDC when compared to [1] and [2] inferred. From the results, the classification time using LTC-NTMDC is comparatively better than [1] and [2] hypothesised. The improvement is due to the incorporation of read/write operations being separately performed by the controller via interface vector employing external memory separately to the input and read vector. With this, ensure rationality, reducing the classification time using LPC-NTMDC by 22% compared to [1] and 34% compared to [2].

4.2.3 Performance analysis of the false positive rate

This section measures the false positive rate involved in the classification process. Table 4 provides the results of the performance evaluation of the false positive rate using the three methods LPC-NTMDC, Clinical data encoding using machine learning [1] and stacked ensemble [2], respectively.

Samples	False positive rate		
	LPC-NTMDC	Clinical data encoding using machine learning	Stacked ensemble
5000	0.06	0.1	0.11
10000	0.09	0.11	0.14
15000	0.1	0.13	0.15
20000	0.11	0.14	0.17
25000	0.13	0.15	0.18
30000	0.15	0.17	0.2
35000	0.16	0.18	0.22
40000	0.18	0.2	0.23
45000	0.19	0.21	0.25
50000	0.21	0.23	0.28

Table 4 Performance evaluation of false positive rate

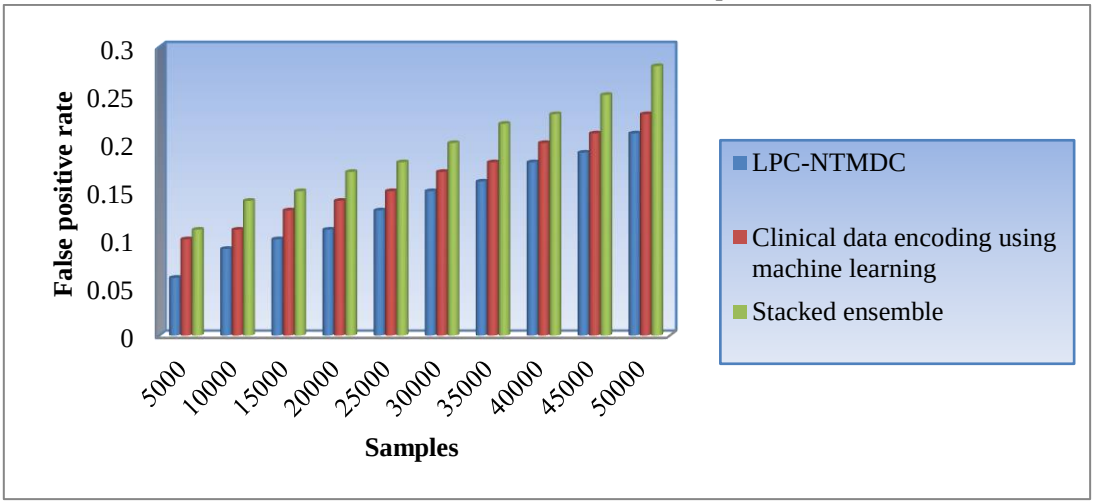

Figure 6 Graphical representation of the false positive rate

Figure 6 shows the false positive rate measure for three different methods. The false positive rate is one of the significant factors during the classification of diseased patients with non-diseased patients because not all the results are accurately classified, and specific results are evenly wrongly classified due to noise in the selected features. The figure shows a proportionate rise with the increase in samples (in all three methods) provided as input observed from the cardiovascular dataset. However, with simulations conducted with 5000 samples, 200 diseased and 4800 non-diseased patients, after evaluation of the proposed method, the diseased and non-diseased identified using LPC-NTMDC was 12, 188, 20, 180 using [1] and 22, 178 using [2]. This measure reduced the false positive rate using LPC-NTMDC upon comparison with the two existing methods. The improvement was due to applying the Lebesgue Prevosti Rationality-based Feature selection model. Following this, two normal distributions were made separately for the presence or absence of disease using a recurrence-based criterion. With this, the false positive rate was significantly reduced in LPC-NTMDC by 17% compared to [1] and 30% compared to [2].

4.2.4 Performance analysis of space complexity

Finally, this section discusses the space complexity in analysing healthcare data for disease diagnosis. Table 5 lists the results of space complexity.

Samples	Space complexity (KB)		
	LPC-NTMDC	Clinical data encoding using machine learning	Stacked ensemble
5000	300	450	500
10000	335	485	550
15000	350	515	585
20000	390	535	615
25000	415	580	635
30000	435	615	696
35000	480	635	715
40000	515	680	735
45000	555	715	780
50000	595	785	845

Table 5 Performance Evaluation of space complexity

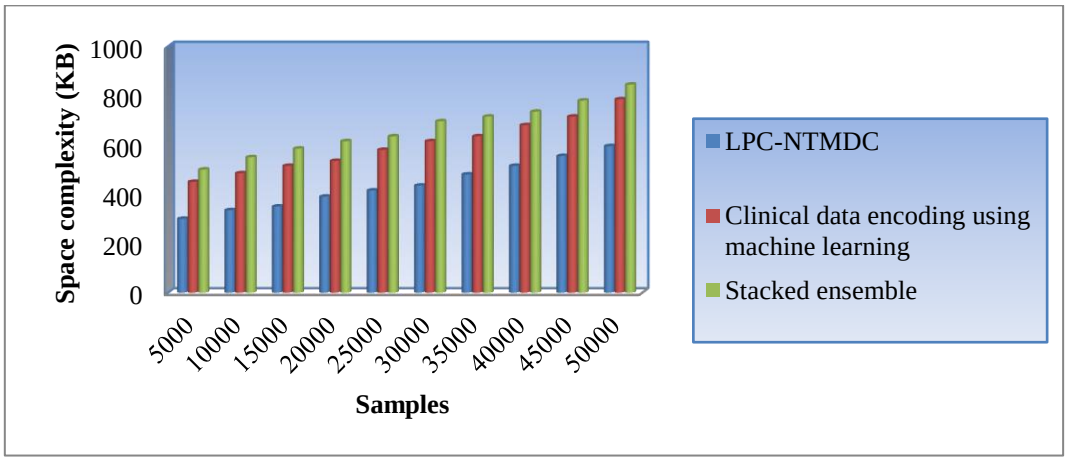

Figure 7 Graphical representation of space complexity

Figure 7 above portrays the graphical representation of exceptional complexity concerning 50000 samples obtained from different patients. Observing a proportionate amount of space complexity increases the sizeable amount of samples. Nevertheless, the

simulations performed with 500 samples showed 300KB memory consumption for reading and writing using the LPC-NTMDC method, 450KB memory consumption using [1] and 550KB memory consumption using [2], respectively. With this sample analysis, the space complexity involved in the LPC-NTMDC method for disease diagnosis was comparatively lesser than [1] and [2]. The minimum space complexity was due to applying the Lebesgue Prevosti Rationality-based Feature selection algorithm. Applying this algorithm, the given input vector matrix has initially formulated two normal distributions. Next, Lebesgue Prevosti distance was applied, which returned class-based and rational features for further processing. As a result, to be comparatively lesser by 27% compared to [1] and 35% compared to [2] found the LPC-NTMDC method's space complexity.

5. Conclusion

This paper proposes machine learning-based healthcare data analytics for disease diagnosis. Consider two levels according to the theoretical model and the transposable formation of the proposed method. These levels include determining the relevant data or features for cardiovascular disease diagnosis according to the class-balanced and rational feature selection and disease classification via the Neural Turing Machine Data Classification algorithm. Different prediction methods have evaluated the proposed method. The applied predictors included machine learning and ensemble methods. The experimental results revealed that the classification algorithms using machine learning-based healthcare data analytics for disease diagnosis called the Lebesgue Prevosti Class-balanced Neural Turing Machine Data Classification (LPC-NTMDC) method performed well in terms of the classification accuracy, classification time, false positive rate and space complexity. Furthermore, the LPC-NTMDC method reached the highest performance for disease predicting in our scenario with 97.7% accuracy, 65ms classification time and a minimum false positive rate of 0.06%. The high accuracy of the LPC-NTMDC method in comparison with other applied predictors is a significant difference that makes it applicable in real-time cardiovascular disease diagnosis.

References

[1] Sarwan Ali, Yijing Zhou, and Murray Patterson1 "Efficient analysis of COVID-19 clinical data using machine learning models", Medical, Biological Engineering & Computing, Springer, Volume 60, 2022, Pages 1881–1896 [COVID-19 analysis using machine learning]

[2] Zina M. Ibrahim, Daniel Bean, Thomas Searle, Linglong Qian, Honghan Wu, Anthony Shek, Zeljko Kraljevic, James Galloway, Sam Norton, James T. H. Teo, and Richard JB Dobson, "A Knowledge Distillation Ensemble Framework for Predicting Short- and Long-Term Hospitalization Outcomes From Electronic Health Records Data", IEEE Journal of Biomedical and Health Informatics, Volume 26, Issue 1, January 2022, Pages 423-435 [Stacked ensemble]

[3] Christian Abbet, Linda Studer, Andreas Fischer, Heather Dawsonb, Inti Zlobec, Behzad Bozorgtabar, Jean-Philippe Thirana, "Self-rule to multi-adapt: Generalised multi-source feature learning using unsupervised domain adaptation for colorectal cancer tissue detection", Medical Image Analysis, Volume 79, 2022, Pages 1-20

[4] Yue Wang, Zhiqiang Lv, Zhaoyu Sheng, Haokai Sun and Aite Zhao, "A deep spatio-temporal meta-learning model for urban traffic revitalisation index prediction in the COVID-19 pandemic", Advanced Engineering Informatics, Elsevier, Volume 53, August 2022, Pages 1-12

[5] Renji P. Cherian, Noby Thomas, Sunder Venkitachalam, "Weight optimised neural network for heart disease prediction using hybrid lion plus particle swarm algorithm", Journal of Biomedical Informatics, Elsevier, Aug 2020

[6] Anna Karen, Garate-Escamila, Amir Hajjam E, Hassani, Emmanuel Andres, " Classification models for heart disease prediction using feature selection and PCA", Informatics in Medicine Unlocked, Elsevier, Apr 2020

[7] Amit Joshi, Bhuwan Chandra Joshi, M. Amin-ul Mannan, Vikas Kaushik, "Epitope-based vaccine prediction for SARS-COV-2 by deploying immuno-informatics approach", Informatics in Medicine Unlocked, Elsevier, Apr 2020

[8] Ibrahim Mahmood Ibrahim, Adnan Mohsin Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases", Journal of Applied Science and Technology Trends, Oct 2021

[9] Dwarakanath B., Latha M., Annamalai R., Jagadish S. Kallimani, Ranjan Wali, and Birhanu Belete, "A Novel Feature Selection with Hybrid Deep Learning Based Heart Disease Detection and Classification in the e-Healthcare Environment", Computational Intelligence and Neuroscience, Hindawi, Sep 2022

[10] Bikram Kar and Bikash Kanti Sarkar, "A Hybrid Feature Reduction Approach for Medical Decision Support System", Mathematical Problems in Engineering, Hindawi, Sep 2022

[11] Amin Ul Haq, Jian Ping Li, Muhammad Memon, Jalaluddin Khan, Asad Malik, Tanvir Ahmad, Amjad Ali, Shah Nazir, Ijaz Ahad, Mohammad Shahid, "Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings", IEEE Access, Apr 2019

[12] Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, Abdus Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare", IEEE Access, Jun 2020

[13] Afnan M. Alhassan, Wan Mohd Nazmee Wan Zainon, "Review of Feature Selection, Dimensionality Reduction and Classification for Chronic Disease Diagnosis", IEEE Access, Jun 2021

[14] Milad Mirbabaie, Stefan Stieglitz, Nicholas R. J. Frick, "Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction", Health and Technology, Springer, May 2021

[15] Yogesh Kumar, Apeksha Koul, Ruchi Singla, Muhammad Fazal Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesising framework and future research agenda", Journal of Ambient Intelligence and Humanized Computing, Springer, Nov 2021

[16] Rung-Ching Chen, Christine Dewi, Su-Wen Huang, Rezzy Eko Caraka, "Selecting critical features for data classification based on machine learning methods", Journal of Big Data, Springer, Oct 2020

[17] Dibaba Adeba Debal, Tilahun Melak Sitote, "Chronic kidney disease prediction using machine learning techniques", Journal of Big Data, Springer, Oct 2022

[18] Namrata Singh, Pradeep Singh, "A hybrid ensemble-filter wrapper feature selection approach for medical data classification", Chemometrics and Intelligent Laboratory Systems, Elsevier, Jul 2021

[19] Thoutireddy Shilpa, anal Paul, "CVDPF: A Hybrid Feature Selection Method with Data-Driven Approach for Cardiovascular Disease Prediction Framework using Machine Learning", Springer, Oct 2021

[20] Wiharto, Esti Suryani, Sigit Setyawan, Bintang Pe Putra, "The Cost-Based Feature Selection Model for Coronary Heart Disease Diagnosis System Using Deep Neural Network", IEEE Access, Mar 2022.