Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 19(3): 166-174, Sep. 2021 Regular paper

Enhance Health Risks Prediction Mechanism in the Cloud Using RT-TKRIBC Technique

Venkateswara Raju Konduru[1*] and Manjula R Bharamgoudra[2], *Member, KIICE*

[1]Department of School of ECE, REVA University, Bangalore, 560064 India
[2]Department of School of ECE, REVA University, Bangalore, 560064 India

Abstract

A large volume of patient data is generated from various devices used in healthcare applications. With increase in the volume of data generated in the healthcare industry, more wellness monitoring is required. A cloud-enabled analysis of healthcare data that predicts patient risk factors is required. Machine learning techniques have been developed to address these medical care problems. A novel technique called the radix-trie-based Tanimoto kernel regressive infomax boost classification (RT-TKRIBC) technique is introduced to analyze the heterogeneous health data in the cloud to predict the health risks and send alerts. The infomax boost ensemble technique improves the prediction accuracy by finding the maximum mutual information, thereby minimizing the mean square error. The performance evaluation of the proposed RT-TKRIBC technique is realized through extensive simulations in the cloud environment, which provides better prediction accuracy and less prediction time than those provided by the state-of-the-art methods.

Index Terms: Cloud, Healthcare heterogeneous data, Internet of things, Predictive analytics, Radix trie

I. INTRODUCTION

The cloud has recently been offering a wide range of data analytics services via the Internet. Data analysis is the process of collecting, storing, processing, and retrieving data. Cloud computing has several applications in the fields of education, social networking, and medicine. However, the benefit of the cloud for medical purposes is the seamless connectivity it offers to handle the large volume of data generated by the healthcare industry. To provide better services to patients over online healthcare applications, machine learning algorithms play a significant role in handling a very large volume of patient data to improve accurate disease risk prediction.

A fuzzy rule-based neural classifier (FRNC) was developed in [1] to predict disease and severity. The proposed method developed a cloud and internet-of-things based mobile healthcare application to monitor serious diseases. However, it consumes more time for disease prediction which minimizes accuracy. A parallel semi-naïve Bayes (PSNB) was introduced in [2] to predict future health using healthcare big data. The accuracy of PSNB was improved using the modified conjunctive attribute (MCA). However, the error rate was not minimized.

A Hadoop cluster architecture was developed in [3] for processing and analyzing healthcare big data on cloud computing. However, disease prediction was not performed. A novel medical cloud multi-agent system (MCMAS) was introduced in [4] to provide various services to patients. The designed system did not use any machine-learning technique to analyze patient data.

An intelligent healthcare system was introduced in [5] for

Received 07 May 2021, **Revised** 29 July 2021, **Accepted** 09 August 2021
***Corresponding Author** Venkateswara Raju Konduru (E-mail: vkonduru@nextgen.com, kvrajukonduru@gmail.com Tel:+91 080 49072400, extn:27879)
Department of School of ECE, REVA University, Bangalore 560064, India.

Open Access https://doi.org/10.6109/jicce.2021.19.3.166 print ISSN: 2234-8255 online ISSN: 2234-8883

data analytics to improve healthcare services However, any advanced deep learning technique was not employed to further enhance the quality of the cloud-based service. A regularized stacked denoising auto-encoder (SDAE) method was developed in [6] to predict clinical risks using large volumes of electronic health records (EHRs). However, the designed method failed to conduct experiments using a large-scale dataset with heterogeneous data.

A multimodal data-based recurrent convolutional neural network (MD-RCNN) was developed in [7] to predict disease risk, but it failed to predict other diseases in minimal time. A nonlinear support vector classifier (SVC) with the radial basis function kernel algorithm was introduced in [8] for cardiovascular disease prediction. However, the designed algorithm failed to evaluate the predictions of various diseases. A deep risk based on the attention method and deep neural networks was introduced in [9] to predict the risk of cardiovascular disease. However, the designed method failed to improve the accuracy of risk prediction. A fuzzy-based reinforcement learning method using a neural network was developed in [10] for healthcare IoT in a fog-computing environment.

An ensemble multi-label classification technique was introduced in [11] to predict chronic disease risk, but this technique failed to provide a more accurate health risk prediction. An ensemble classification method was developed in [12] to achieve a higher accuracy for heart disease risk prediction. However, the designed methods failed to minimize risk prediction time. A deep learning paradigm was introduced in [13] to improve health prediction using big data, but it failed to improve prediction performance. An ant colony optimization (ACO)-based method was developed in [14] for processing mobile cloud-based large volumes of healthcare data.

A machine learning-based predictive model was introduced in [15] to predict type 2 diabetes with higher accuracy, but it failed to solve other disease-predictive problems in healthcare. A hybrid intelligent machine-learning-based predictive system was introduced in [16] for the diagnosis of heart disease. However, the designed system did not minimize the error rate of the disease prediction. Several machine learning methods have been developed in [17] for predicting chronic diseases using health data. However, these methods failed to minimize the false-positive rate of disease prediction.

An ensemble machine learning algorithm (EMLA) was designed in [18] to predict coronary artery disease risk, but the designed algorithms failed to achieve a higher risk prediction accuracy. A data-driven approach uses supervised machine learning techniques that were introduced in [19] to predict a patient's disease risk. However, the performance of the disease prediction time has not been solved. An intelligent decision support method was introduced in [20] to predict heart disease using the historical data of patients. The model failed to implement the ensemble model in a cloud environment to achieve a higher accuracy of disease prediction.

A. A Major Contribution of the Work

The major issues discussed in the preceding literature are addressed by introducing a new technique called the radix-trie-based Tanimoto kernel regressive infomax boost classification (RT-TKRIBC) technique. The main contributions of the proposed RT-TKRIBC technique are summarized as follows:

- The cloud-based healthcare data analytic architecture RT-TKRIBC is designed to process heterogeneous patient health data and predict the risk with higher accuracy. Tanimoto kernel regression is applied to analyze the input training data with the testing data. The regression function measures the similarity between these two datasets and predicts a higher risk of a particular disease. The infomax boosting technique improved the regression results and achieved a higher prediction accuracy.
- To minimize the false-positive rate of risk prediction, the infomax boosting technique finds mutual information between patient data and their predictive classes using a gradient ascent function. Additionally, the ensemble boosting technique finds a better weak learner with a smaller mean square error. This helps to improve the accuracy of prediction and minimize incorrect predictions.
- To reduce the time of risk prediction, RT-TKRIBC uses the radix trie to store heterogeneous patient health data on a cloud data center to easily access the data instantly.

B. Outline of Paper

This paper is structured as follows. In Section 2, a brief explanation of the proposed RT-TKRIBC technique with a detailed diagram is presented. The experimental evaluation and dataset description are presented in Section 3. In Section 4, various evaluation metrics are described. The results and discussion of the proposed technique and state-of-the-art methods are presented in Section 5. Finally, Section 6 presents the conclusions of the study.

II. METHODOLOGY

With the recent developments in healthcare systems, the amount of health data in various formats is rapidly increasing. These types of data are collected from various sources, including digital records, mobile devices, and wearable health devices. Big health data provide more opportunities for health risk analysis and improvement of health services

via cloud-based architecture. This research aims to enhance health risk prediction using machine learning paradigms. In this work, the RT-TKRIBC technique was introduced for a cloud-based architecture.

As shown in Fig. 1, the flow process of the proposed RT-TKRIBC technique is illustrated to perform risk prediction through data classification. Consider a cloud-based healthcare environment where heterogeneous data (D_1, D_2, D_3,…. D_m) are collected from a large number of patients (p_1, p_2, p_3,….p_n). Heterogeneous data contain various types of patient data. With the rapid growth of healthcare applications, many devices used in healthcare create various types of patient data such as heart disease-related data, diabetes data, kidney data. These types of data were collected and analyzed. The collected data are stored in data centers C_1, C_2, C_3,…C_s for performing predictive analytics. Here, the datacenters are related to hospitals where the various departments (i.e., servers) are available to predict patient health status. The proposed RT-TKRIBC technique is designed using the above system model, and the different processes are explained in the subsequent sections.

A. Radix Trie Based Data Storage

The inputs of the heterogeneous patient data were collected and stored in the cloud datacenter. The data center uses a radix trie to store heterogeneous patient data at the server. The radix trie is a data structuring technique in each child node is connected to its parent node. The number of child nodes of each internal node is created based on radix 'R' of the tree, where 'R' indicates a positive integer. For example, when radix is set to 2 each node can have at most two children.

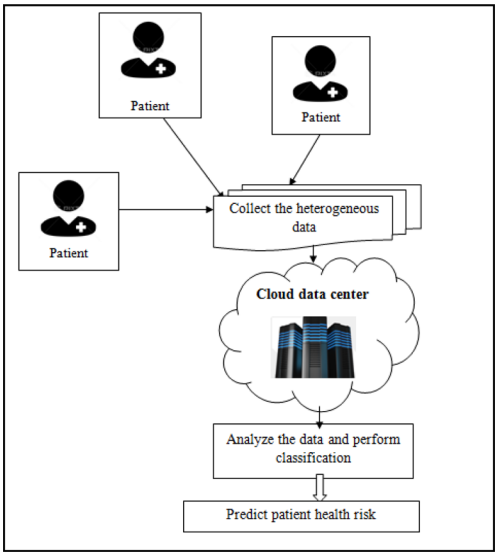

Fig. 1. Flow process of RT-TKRIBC technique.

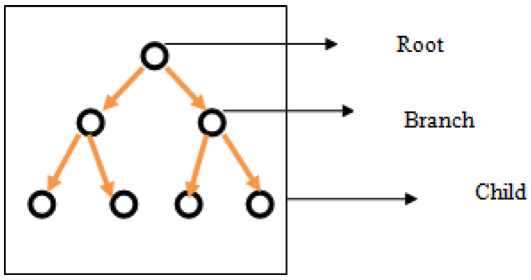

Fig. 2. Structure of a tree with radix two.

Fig. 2 shows the structure of the tree with a radix two. The radix trie performs insertion, deletion, and searching operations to add and remove data. Insertion adds a new string to the trie while attempting to minimize the amount of data stored. To insert the data, a new labeled outgoing edge added and subsequently split into two edges (i.e., children's). This splitting process ensures that no node has more than two children because radix is set to '2'. To remove data from a tree, first the leaf node is located and then the corresponding leaf node is removed from the tree. Following this procedure, all patient data were stored in the cloud server.

B. Tanimoto Kernel Regressive Infomax Boosting Classification Technique

After storing the data, the data center starts to perform predictive analytics by analyzing the input data and performing the risk prediction. The infomax boost technique was applied to predict patient health by analyzing the risk factors using the regression function. The infomax boost is a machine learning ensemble technique that provides accurate classification by analyzing patient data. The infomax boost uses the weak learner as a base classifier to provide the results with some training errors. In contrast, boosting is a strong classifier that provides accurate prediction results.

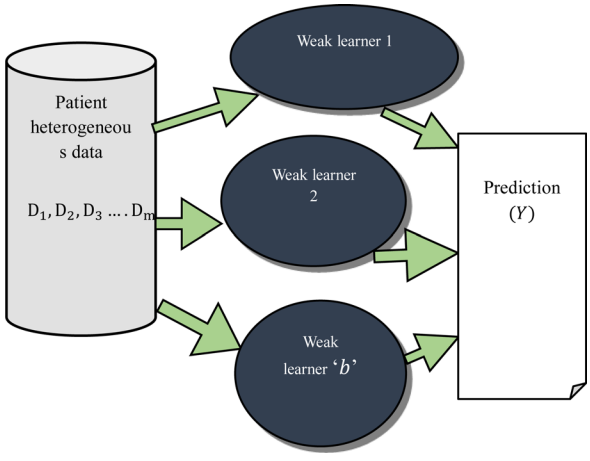

Fig. 3. Infomax boost ensemble learning classifier.

Therefore, the proposed RT-TKRIBC technique uses this machine-learning ensemble technique to improve the classification performance by summing all weak classifiers. The formation of the ensemble learning classifier is shown in Fig. 3.

Fig. 3 depicts a structure of the infomax boost ensemble-learning classifier with multiple weak learners. The ensemble-learning classifier considers the training sets (D_m, W), where D_m denotes the patient heterogeneous data and W denotes strong prediction results. The ensemble-learning classifier uses 'b' number of weak learners s_1, s_2, s_3,…s_h to classify the input data. The proposed infomax boost ensemble-learning classifier uses kernel regression as a weak learner to predict the patient health risks.

Tanimoto kernel regression is a statistical process for determining the relationships between a dependent variable (i.e., output) and one or more independent variables (i.e., input) based on the similarity measure. Here, the kernel indicates the similarity function between two variables.

As shown in Fig. 4, a block diagram of the Tanimoto kernel regression was developed using patient data. Consider D_1, D_2, D_3,….D_m as input training data and T_1, T_2, T_3,….T_r testing disease data. Then, the similarities between these two data sets are calculated using the Tanimoto kernel function, which is expressed as follows:

$$\beta = \frac{m*\sum D_m T_r}{\sqrt{\sum Dm^2}+\sqrt{\sum T_r^2-\sum D_m T_r}}, \quad (1)$$

where β represents the Tanimoto kernel coefficient, m represents the number of training data, D_m and T_r denote the training and testing disease data sets, respectively, ΣD_m^2 denotes the sum of the squared scores of the D_m, ΣT_r^2 denotes the sum of the squared scores of the T_r and $\Sigma D_m T_r$ denotes the sum of the product of the paired score of D_m and T_r. The Tanimoto kernel coefficient provides output ranges from 0 to +1. The threshold is set to the similarity value and predicts a patient with a higher risk of the disease. If the similarity value is higher than the threshold, the health risk to the disease is predicted. The output of the weak classifier had a mean square error. The error is minimized by combining the output of the weak classifiers into a strong classifier. The output of the final strong classification results was obtained as follows:

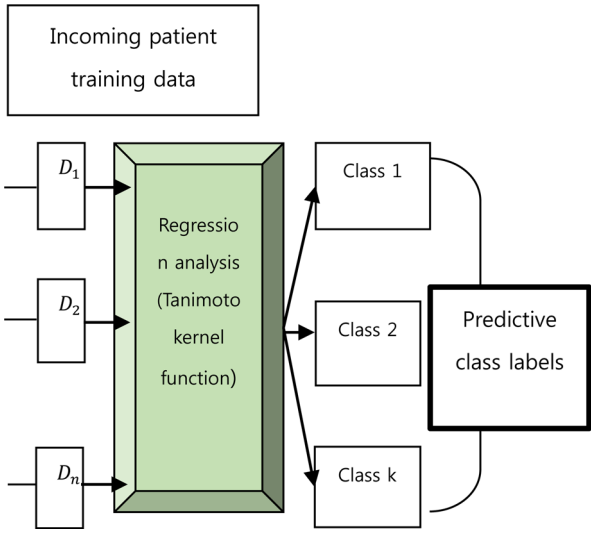

Fig. 4. Block diagram of Tanimoto kernel regression.

$$W = \sum_{i=1}^{b} F_i(D), \quad (2)$$

where the output of strong classification results is denoted by 'W' and $F_i(D)$ indicates the weak classification results. For each weak learner, the weight is initialized to accurately predict the results. After weight initialization, the mutual information is measured between the classified data and their predictive classes to minimize the mean square error. Therefore, mutual information was applied to validate the prediction results of each weak learner. Mutual information is a function that identifies the mutual dependence between the results and a particular class. The probabilities of mutual dependence were measured as follows:

$$M_{UI} = P(D_m, g_n) \log_2 \left(\frac{P(D_m, g_n)}{P(D_m)P(g_n)}\right), \quad (3)$$

where M_{UI} denotes the mutual information between the patient's data (D_m) and predictive classes (g_m), $P(D_m, g_n)$ denotes a joint probability distribution, $P(D_m)$, and $P(g_n)$ denotes a marginal probability between the patient's data and predictive classes. Then, the gradient ascent function is applied to determine the maximum dependence.

$$F(x) = argmax M_{UI}. \quad (4)$$

In (4), $F(x)$ denotes a gradient ascent function and *argman* denotes an argument of the maximum function. Based on the validation results, the mean square error was calculated based on the difference between the actual and predicted classification results. The error is calculated mathematically as follows:

$$E_{MS} = \frac{1}{m}\left(P_a - P_p\right)^2, \quad (5)$$

where E_{MS} denotes mean square error, P_a represents the actual results of a weak learner, P_p denotes predicted results of a weak learner, and 'm' denotes the amount of patient data. Based on the error value, the initial weight was updated. If the weak learner predicted correctly, the initial weight was minimized. Otherwise, the weight increases. Therefore, the ensemble-learning technique determines the weak learner that has a smaller mean square error. The output of the strong classification results was obtained as follows:

$$W = \sum_{i=1}^{b} F_i(D) * \Delta\vartheta, \quad (6)$$

where W represents the output of the ensemble classifier and $\Delta\vartheta$ represents the updated weight. Therefore, the ensemble-learning technique accurately predicts the patient's risk, and the cloud datacenter sends alerts to the doctors for informing them of the existence of higher-risk patients. This helps provide the exact treatment and minimize the mortality of patients in an emergency. The algorithmic process of proposed RT-TKRIBC technique is described as follows:

Algorithm 1. Radix Trie based Tanimoto Kernel Regressive Infomax Boost Classification

Input: Datasets '*B*'

Begin
  Collect patient heterogeneous data D_1, D_2, D_3..D_m from '*B*'
  Store the data to a data center using radix trie
  For each collected data D_m
  Constructs '*b*' weak learners
  Analyze the training patient data and testing disease data 'β'
  If ($\beta > th$) then
  Predict the patient with a higher risk of a particular disease
  else
  Check other possible classes
  end if
  Obtain weak classification results
  Combine all the weak classifiers $\sum_{i=1}^{b} F_i(D)$
  For each $F_i(D)$
  Initialize the weight 'ϑ'
  Measure mutual information 'M_{UI}'
  Compute mean square error E_{MS}
  End for
  Update the weight $\Delta\vartheta$
  Find weak learner with minimum error
  End for
  Return (strong classification $W = \sum_{i=1}^{b} F_i(D) * \Delta\vartheta$)
End
Output: Improve prediction accuracy

The step-by-step process of the health risk predictive analysis was performed using a machine learning ensemble classifier. Different heterogeneous data were collected from the dataset and stored in the cloud data center. Then, the data center performs a patient health risk predictive analysis using regression and classification. The Tanimoto kernel regression function analyzes patient training and testing disease data. When the training data are more similar to the testing disease data, it predicts the patient health risk. In other words, the disease-affected patient was correctly predicted. The output of weak classification results is summed to create a strong classifier by calculating the mean square error and measuring the mutual information. The strong classifier finds the weakest learner with the minimum error. This helps improve the prediction accuracy and minimizes the false-positive rate.

III. EXPERIMENTAL SETTINGS

The experimental evaluation of the proposed RT-TKRIBC technique and existing methods, namely FRNC [1] and PSNB [2], are conducted using the Java language with various input heterogeneous data taken from the heart disease, diabetes, and kidney disease datasets.

A. Dataset's Description

The heart disease dataset was obtained from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/heart+Disease). This dataset contains 76 attributes, but only 14 of them are used for healthcare big data analytics in a cloud environment. The patient identification number, age, sex, patient name, resting blood pressure, serum cholesterol in mg/dl, chest pain type, and so on are among the attributes. This dataset comprises 303 instances. The attribute categorists are categorical, integer, real, and the dataset characteristic is a classification. The associated task performed by the dataset was classified.

The other dataset is a Pima Indian diabetes dataset, taken from https://www.kaggle.com/uciml/pima-indians-diabetes-database/data. All attributes are numeric values. The dataset provided data from diabetes cases in which all patients were women at a minimum age of 21 years. The dataset consists of 768 samples (i.e., instances) and the following nine attributes: pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome.

The Chronic_Kidney_Disease Data Set is taken from the UCI machine learning repository https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. The dataset comprises 25 attributes and 400 instances used for predicting kidney disease. The dataset characteristics were multivariate, and the characteristics of the attributes were real. The associated task was classified. As a result, these three datasets were used to perform experiments to collect patient data and perform predictive analytics.

IV. EVALUATION METRICS

The performance analyses of the RT-TKRIBC technique, FRNC [1], and PSNB [2] are discussed using different performance metrics, such as prediction accuracy, false-positive rate, and prediction time. The obtained results are discussed with the help of both tables and graphical illustrations. The performance of the various metrics is given below.

A. Prediction Accuracy

Prediction accuracy (A_p) is measured as the ratio of the number of patient data correctly predicted to be at risk to the

total number of patient data taken as input. The prediction accuracy was mathematically calculated as follows:

$$A_p = \left(\frac{m_{CP}}{m}\right) * 100, \quad (7)$$

where A_p denotes a prediction accuracy, m_{CP} represents the number of data correctly predicted by the cloud data center, and 'm' denotes a total number of patient data taken as input. Therefore, A_p is measured in terms of percentage (%).

B. False-positive Rate

The false-positive rate (F_p) is measured as the ratio of the number of patient data points predicted to the total number of patient data taken as input. Therefore, the formula for calculating the false positive rate is as follows:

$$\left(\frac{m_{ICP}}{m}\right) * 100, \quad (8)$$

where, F_p indicates a false-positive rate, m_{ICP} represents the number of patient data incorrectly predicted as a risk, and 'm' denotes a total number of patient data taken as input. The F_p is measured in terms of percentage (%).

C. Prediction Time

Prediction time (P_{time}) is measured as the amount of time taken by an algorithm to predict the risk of the patient using heterogeneous health data taken as input. Therefore, the overall prediction time was calculated as follows:

$$P_{time} = m * t\ (predict\ one\ patient\ data), \quad (9)$$

where P_{time} indicates the prediction time, t represents the time taken to predict single patient data. The prediction time is measured in terms of milliseconds (ms).

V. COMPARISON WITH STATE-OF-THE-ART MACHINE LEARNING ALGORITHMS

In this section, the clinical risk prediction performance of the proposed RT-TKRIBC technique and two state-of-the-art classification algorithms, that is, FRNC [1] and PSNB [2], using various numbers of instances (i.e., patient data) are compared.

The experimental results of the prediction accuracy and false-positive rate are shown in Table 1. For a fair comparison of the results and discussion, the numbers of input patient data were taken in the range from 50 to 500. To evaluate both accuracy and false-positive rate, similar input training heterogeneous data were taken as input. The heterogeneous data, that is, various types of patient data collected from the three disease datasets, namely, heart disease, diabetes, and kidney disease datasets. These data are stored in the cloud datacenters. The performance of the RT-TKRIBC technique showed considerable improvements in terms of accuracy and false-positive rate in predicting the health risks of patients. This is because of the use of an effective and efficient machine learning technique called the infomax boost ensemble classification (IBEC) technique. The ensemble classifier performs regression analysis between the training and testing disease patient data. If these two are correctly matched, patients are classified according to their risk of a particular disease. In other words, the training patient data are correctly matched to the testing heart disease data, and then classified based on their risk of heart disease. Similarly, the RT-TKRIBC technique was used to predict the variety of all patient health. This, in turn, improved prediction accuracy. Additionally, the IBEC technique uses mutual information to validate the prediction results. Furthermore, the boosting technique selects a weak classifier with the minimum mean square error. As a result, the RT-TKRIBC technique had a higher prediction accuracy and a lower false-positive rate.

Table 1. Comparative analysis of prediction accuracy false positive rate

Number of patient data	Prediction accuracy (%)			False-positive rate (%)		
	RT-TKRIBC	FRNC	PSNB	RT-TKRIBC	FRNC	PSNB
50	94	88	84	6	12	16
100	92	89	87	8	11	13
150	91	85	83	9	15	17
200	94	90	87	6	10	13
250	92	88	84	8	12	16
300	95	89	82	5	11	18
350	94	88	83	6	12	17
400	93	90	87	7	10	13
450	96	89	86	4	11	14
500	94	88	84	6	12	16

Ten results were obtained for each machine learning classifier. The RT-TKRIBC technique results were compared with those of the other two state-of-the-art classification algorithms. As shown in Table 1, the prediction accuracies of RT-TKRIBC, FRNC [1], and PSNB [2] were 94%, 88%, and 84%, respectively, when the number of patient data was 50. Similarly, various accuracy results were obtained using different patient data. The compared results prove that the average prediction accuracy is found to be higher using the RT-TKRIBC technique by 6% and 10% compared to the RT-TKRIBC technique, FRNC [1], and PSNB [2].

The average false-positive results of the RT-TKRIBC technique are significantly reduced by 44% compared to FRNC [1] and 57% compared to PSNB [2]. Therefore, the RT-TKRIBC technique showed considerable improvements in

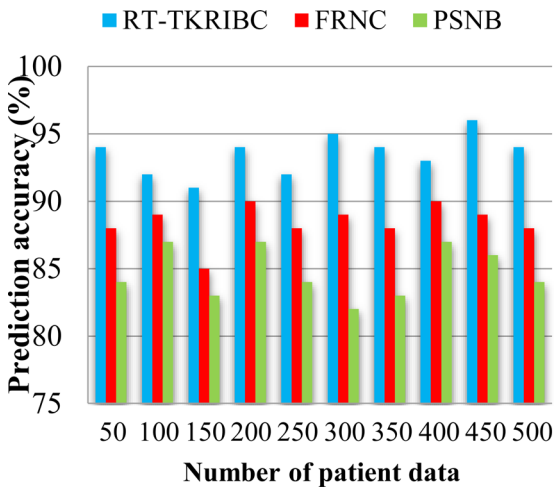

Fig. 5. Performance analysis of prediction accuracy.

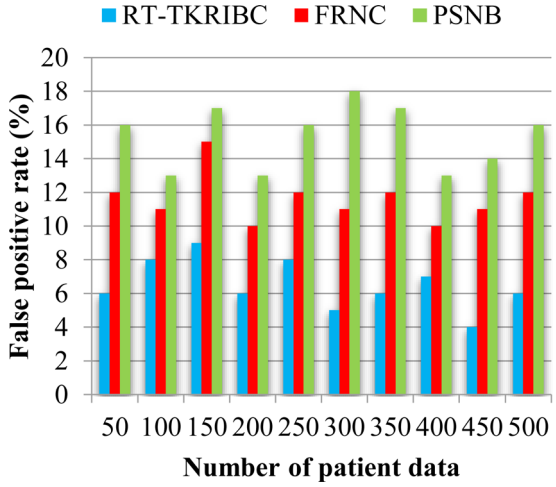

Fig. 6. Performance analysis of the false-positive rate.

Table 2. Comparative analysis of prediction time

Number of patient data	Prediction time (ms)		
	RT-TKRIBC	FRNC	PSNB
50	15	17	20
100	18	20	22
150	21	23	25
200	24	26	28
250	25	27	30
300	27	30	33
350	30	32	35
400	32	34	36
450	33	35	37
500	36	38	41

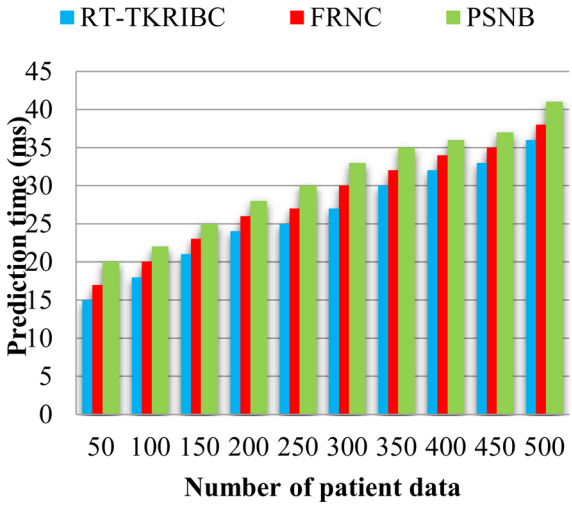

Fig. 7. Performance analysis of prediction time.

terms of both the prediction accuracy and false-positive rate. The graphical results of the prediction accuracy and false-positive rate are shown in Figs. 5 and 6.

Fig. 5 shows the impact of prediction accuracy versus various patient data (i.e., instances). The plot uses the processed instances (i.e., number of patient data) as the x-axis and the prediction accuracy as the y-axis separately. Owing to space constraints, only ten iterations are performed for various input data and the results are obtained. The significant difference diagram shown in Fig. 5 shows that the RT-TKRIBC technique outperforms the conventional methods FRNC [1] and PSNB [2].

Fig. 6 shows the performance results of three machine learning techniques used to classify patients and predict health risks using different instances. The graphical plot illustrates the chart of three identical classification techniques RT-TKRIBC technique, FRNC [1], and PSNB [2], which are represented by three colors, blue, red, and green, respectively. Fig. 6 shows a continuous red color column, which indicates that the RT-TKRIBC technique minimizes the false-positives and is more robust in correctly predicting health risks compared to other conventional classifiers.

Table 2 and Fig. 7 depict the performance analysis of the prediction time based on the input patient data in the form of heterogeneity. Fig. 7 shows that the number of input patient data is directly proportional to the prediction time. As the number of data increases, so does the dimension of the extracted patient data. The time required for the prediction was also increased when all three classification methods were used. However, when compared to the two state-of-the-art methods, the RT-TKRIBC technique reduces prediction time. This is because the RT-TKRIBC technique uses the radix trie for data storage on the cloud datacenter. Subsequently, disease prediction is performed to minimize the time. Additionally, the Tanimoto kernel function is used to match the training and testing disease data, which reduce prediction time. The prediction time using the RT-TKRIBC

technique is reduced by 8% and 16% compared to FRNC [1] and PSNB [2], respectively.

The results demonstrate that the RT-TKRIBC technique improves the performance of health risk prediction with higher accuracy, minimum false-positive rate, and prediction time.

VI. CONCLUSION

A novel technique called RT-TKRIBC is designed for cloud-based healthcare risk prediction with higher accuracy, less time, and a lower false-positive rate. This is achieved by applying the IBEC technique for classifying patient data according to the risk of a particular disease. The Tanimoto kernel regression function was also applied to analyze heterogeneous patient data obtained from various databases. The regression function analyzes the input training data using the test data. The ensemble classification technique minimizes the mean square error of prediction through the mutual information measure. The performance of the RT-TKRIBC technique was evaluated with two existing machine learning techniques using different metrics, such as prediction accuracy, false-positive rate, and prediction time. The comparative analysis demonstrates that the RT-TKRIBC technique outperforms state-of-the-art methods in terms of health risk prediction in less time.

REFERENCES

[1] P. M. Kumar, S. Lokesh, R. Varatharajan, G. C. Babu, and P. Parthasarathy, "Cloud and IoT based disease prediction and diagnosis system for healthcare using Fuzzy neural classifier," *Future Generation Computer Systems*, Elsevier, vol. 86, pp. 527-534, 2018. DOI: 10.1016/j.future.2018.04.036.

[2] P. K. Sahoo, S. K. Mohapatra, and S. Wu, "SLA based healthcare big data analysis and computing in cloud network," *Journal of Parallel and Distributed Computing*, vol. 119, pp. 121-135, 2018. DOI: 10.1016/j.jpdc.2018.04.006.

[3] S. Rallapalli, R. P. Gondkar, and K. U. Pavan, "Impact of processing and analyzing healthcare big data on cloud computing environment by implementing hadoop cluster," *Procedia Computer Science*, vol. 85, pp. 16-22, 2016. DOI: 10.1016/j.procs.2016.05.171.

[4] J. Hanen, Z. Kechaou, and M. B. Ayed, "An enhanced healthcare system in mobile cloud computing environment," *Vietnam Journal of Computer Science*, vol. 3, pp. 267-277, 2016. DOI: 10.1007/s40595-016-0076-y.

[5] X. Ma, Z. Wang, S. Zhou, H. Wen, and Y. Zhang, "Intelligent healthcare systems assisted by data analytics and mobile computing," *Wireless Communications and Mobile Computing*, vol. 2018, pp. 1-16, 2018. DOI: 10.1155/2018/3928080.

[6] Z. Huang, W. Dong, H. Duan, and J. Liu, "A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 5, pp. 956-968, 2018. DOI: 10.1109/TBME.2017.2731158.

[7] Y. Hao, M. Usama, J. Yang, M. S. Hossain, and A. Ghoneim, "Recurrent convolutional neural network based multimodal disease risk prediction," *Future Generation Computer Systems*, Elsevier, vol. 92, pp. 76-83, 2019. DOI: 10.1016/j.future.2018.09.031.

[8] S. Mezzatesta, C. Torino, P. D. Meo, G. Fiumaraa, and A. Vilasi, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 9-15, 2019. DOI: 10.1016/j.cmpb.2019.05.005.

[9] Y. An, N. Huang, X. Chen, F. Wu, and J. Wang, "High-risk prediction of cardiovascular diseases via attention-based deep neural networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 21-29, 2019. DOI: 10.1109/TCBB.2019.2935059.

[10] S. Shukla, M. F. Hassanm, M. K. Khan, L. T. Jung, and A. Awang, "An analytical model to minimize the latency in healthcare internet-of-things in fog computing environment," *PLoS ONE*, vol. 14, no. 11, pp. 1-31, 2019. DOI: 10.1371/journal.pone.0224934.

[11] R. Li, W. Liu, Y. Lin, H. Zhao, and C. Zhang, "An Ensemble Multilabel Classification for Disease Risk Prediction," *Journal of Healthcare Engineering*, vol. 2017, pp. 1-10, 2017. DOI: 10.1155/2017/8051673.

[12] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, Elsevier, vol. 16, pp. 1-9, 2019. DOI: 10.1016/j.imu.2019.100203.

[13] H. Zhong and J. Xiao, "Enhancing health risk prediction with deep learning on big data and revised fusion node paradigm," *Scientific Programming*, vol. 2017, pp. 1-18, 2017. DOI: 10.1155/2017/1901876.

[14] MD. M. Islam, MD. A. Razzaque, M. M. Hassan, W. N. Ismail, and B. Song, "Mobile cloud-based big healthcare data processing in smart cities," *IEEE Access*, vol. 5, pp. 11887-11899, 2017. DOI: 10.1109/ACCESS.2017.2707439.

[15] G. Luo, "Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction," *Health Information Science and Systems*, Springer, vol. 4, pp. 1-9, 2016. DOI: 10.1186/s13755-016-0015-4.

[16] A. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, pp. 1-21, 2018. DOI: 10.1155/2018/3860146.

[17] T. S. Brisimi, T. Xu, T. Y. Wang, W. Dai, W G. Adams, and L. C. Paschalidis, "Predicting Chronic disease hospitalizations from electronic health records: An interpretable classification approach," *Proceedings of the IEEE*, vol. 106, no. 4, pp.690-707, 2018. DOI: 10.1109/JPROC.2017.2789319.

[18] S. M. Naushad, T. Hussain, B. Indumathi, K. Samreen, S. A. Alrokayan, and V. K. Kutala, "Machine learning algorithm-based risk prediction model of coronary artery disease," *Molecular Biology Reports*, vol. 45, pp. 901-910, 2018. DOI: 10.1007/s11033-018-4236-2.

[19] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Medical Informatics and Decision Making*, Springer, vol. 19, pp. 1-15, 2019. DOI: 10.1186/s12911-019-0918-5.

[20] N. Gupta, N. Ahuja, S. Malhotra, A. Bala, and G. Kaur, "Intelligent heart disease prediction in cloud environment through ensembling," *Expert System*, vol. 34, no. 3, pp. 1-14, 2017. DOI: 10.1111/exsy.12207.

Venkateswara Raju Konduru

Received a B. Tech in Electronics and Communication Engineering from J N T University, India in 2003, and an M.Tech from Satyabhama university, India in 2006. He has over 16 years of experience in the field of software Research and Development. He is currently employed as a senior engineering manager in Bangalore, India, where the company is developing a Healthcare mobile solution using IoT and AI. He developed several IoT products for home/building automation for Honeywell. Currently pursuing a Ph.D. at REVA University, Bangalore, India. He is a member of IEEE.

Manjula R Bharamgoudra

Received her Ph.D. in Electronics and Communication Engineering from Visvesvaraya Technological University. She is working as Associate Professor, Head Innovation Center, School of Electronics and Communication Engineering, REVA University. She has around 19 National/International Conference publications to her credit, 18 publications in reputed journals, and 02 publications as books/book-chapters. She has 16+ years of experience in teaching, research, and administration. She has 1 Indian, and 02 Australian Patent granted and copyrights awarded. She has filed 2 patents and published around 4 patents. Her areas of interest include wireless communication, underwater networks, Internet of Things and Artificial Intelligence. She is a reviewer for IEEE Journals, ACM/IEEE conferences and many others. She is an active researcher with her publications being cited by many other researchers with Scopus citation index of 5; Google Scholar citation h- index of 10 and i-index of 10 (as on Sept. 2021). She is active member of various professional bodies such as Member of IEEE (MIEEE) USA, Life Member ISTE (MISTE) India, Member of IEEE Educational Society. She is also a HAM Member with call sign of VU3UFS