ML-BASED DIAGNOSIS PREDICTION IN TELEMEDICINE APPLICATIONS

G. Vidyulatha[1*], K. Rakesh[2], B. Sai Kiran[2], S. Sai Kiran[2], J. Swapnil[2]

[1] Assistant Professor, [2]UG Student, [1,2] Department of Computer Science and Engineering (AIML)

[1,2]Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana

ABSTRACT

Telemedicine, which enables remote diagnosis and treatment through telecommunications technology, is revolutionizing healthcare delivery by improving accessibility and efficiency. However, traditional telemedicine systems often depend heavily on manual interpretation by healthcare professionals, making them time-consuming, subjective, and susceptible to human errors due to fatigue and cognitive bias. Additionally, these systems face scalability challenges when managing large volumes of patient data, limiting their utility in remote or underserved areas with constrained medical resources. This paper aims to overcome these limitations by integrating machine learning (ML) into telemedicine for automated diagnosis prediction. The goal is to enhance diagnostic accuracy, reduce the workload on healthcare professionals, and expand the reach of healthcare services. By automating routine diagnostic tasks, ML can free clinicians to focus on complex care, while also ensuring timely and consistent medical assessments in areas lacking adequate healthcare infrastructure. The proposed system employs supervised learning algorithms, including Logistic Regression and Decision Tree Classifier, to develop predictive models using patient data such as demographics, symptoms, and diagnostic test results. These models will be trained to identify patterns and correlations indicative of specific medical conditions, enabling accurate, data-driven diagnosis predictions. Ultimately, this ML-driven approach seeks to increase the effectiveness and equity of telemedicine services, making quality healthcare more accessible across geographic and socioeconomic boundaries.

Keywords: Telemedicine, Diagnosis Prediction, Machine Learning, Supervised Learning, Healthcare Accessibility

1. INTRODUCTION

The history of telemedicine traces back to the late 19th century when medical professionals first began experimenting with telegraphy to communicate with patients remotely. One of the earliest recorded instances occurred in 1879 when the Lancet medical journal published an article discussing the use of the telephone to transmit patient information between Paris and London. [1] However, it wasn't until the latter half of the 20th century that significant advancements in telecommunications technology laid the foundation for modern telemedicine practices. [2] In the 1950s and 1960s, pioneering initiatives such as the Nebraska Psychiatric Institute's video conferencing consultations and NASA's telemedicine experiments for space missions demonstrated the potential of remote healthcare delivery. [3] These early experiments highlighted the feasibility of using audiovisual communication technologies to facilitate medical consultations between healthcare providers and patients separated by distance. [4] The 1990s witnessed further progress with the widespread adoption of the internet, enabling real-time audiovisual communication and data exchange between remote locations. [5] Telemedicine expanded beyond consultations to include telemonitoring, tele-education, and tele-surgery, transforming healthcare delivery and improving access to medical services,

particularly in rural and underserved areas. Today, telemedicine encompasses a broad range of services, including remote consultations, telemonitoring of patients' vital signs and health metrics, and tele-education for medical professionals. [6] Technological advancements in digital health, wireless connectivity, and data analytics continue to drive innovation in telemedicine, making healthcare more accessible, efficient, and patient-centered.

2. LITERATURE SURVEY

Manoranjan Dash et al. [8] aimed at identifying the elements that will encourage patients in India to utilize telemedicine during the COVID-19 pandemic. In order to analyze the information gathered from 146 patients using a structured questionnaire, multiple regression and ANN techniques are applied. According to the experimental findings, the ANN model outperformed multiple regressions in terms of nonlinearity and linearity and predicted outcomes with a high degree of accuracy. Syed Thouheed Ahmed et al. [9] developed a dynamic user clustering method based on heterogeneous multi-input multi-output data. The suggested methodology employs networking nodes to add machine learning concepts for dynamic user grouping and classification, resulting in the construction of clusters reflecting similarity indexing ratios. The experimental findings revealed that the proposed method is effective for transmitting delicate medical datasets with pre-processed data. However, the proposed method cannot handle noisy data. Praveen Kumar Sadineni [10] presented on how big-data analytics and machine learning may be combined to improve the quality of healthcare services using techniques like decision trees, SVM, and KNN. The provision of individualized solutions to specific issues, such as the detection and treatment of epidemics, the enhancement of life value, the reduction of needless care, etc., is made possible by enhancing the quality of healthcare services.

The outcomes of the experiment show that combining machine learning methods with Big-Data Analytics raises the excellence of healthcare services. However, in order to deliver accurate results, the suggested method needed high-quality data. M. Sornalakshmi et al. [11] proposed an approach that coupled the context ontology and enhanced apriori algorithm for mining and modelling physiological data utilizing the concepts and connections established by the rules that were generated. A growing number of rules are obtained by combining the EAA with the context ontology. According to the performance analysis, the proposed method produces better support and confidence. The comparison analysis shows that the suggested EAA-SMO technique achieves maximum accuracy and requires the least amount of time to execute than the semantic ontology. The scalability of the suggested approach is constrained. So-Young Choi and Kyungyong Chung [12] presented a big-data knowledge procedure for the health sector using association mining and Hadoop's MapReduce technology. By combining WebBot and the common data model to gather and process heterogeneous health information, the suggested solution offers effective health management knowledge services. Documents that are periodically generated by dynamic linking and distributed file processing are assembled into a corpus for the purpose of finding relationships between data. The processing of large amounts of health-related data using MapReduce-based association mining can aid in disease prevention, the detection of hazards, and post-management using a common data model. As a result, healthcare services that are more advanced can be provided, which helps to enhance people's health and quality of life. D.M. JeyaPriyadharsan et al. [13] presented machine learning techniques for keeping track of human health. The UCI dataset is used for the initial training and validation of ML algorithms. In the testing phase, anomalies in the health state are predicted using sensor data collected to use an IoT framework. IoT device data that has been stored in the cloud is statistically analyzed to determine the accuracy of the prediction percentage. Also, according to the results, the K-Nearest Neighbour beats other traditional classifiers. The major limitation of the study is that, when the training set is large, it takes a lot of space. R. Sandhiya and M. Sundarambal [14] created a clustering

model with enhanced semantic smoothing that is based on ontology and domain knowledge. The model used TF-IGM and modified n-grams to enhance the clustering process. Hierarchical and partitional clustering techniques are used to assess the model's performance. The proposed method outperformed the semantic smoothing model in almost 80% of the quality criteria, proving its efficacy. A drawback of using n-gram overlap to assess document similarity is that it performs poorly when the original document has been updated. T. K. Anusuya and P. Maharajothi [15] designed a method to manage various multimedia medical databases in the telemedicine system. The primary objective of this work is to convey the medical services to the patient, instead of transporting the patient to the medical care services.

This is accomplished through the use of web-based solutions, such as Modern Medical Informatics Services, which are simpler, quicker, and less expensive. The fragmentation of databases, clustering of network locations, and allocation of fragments were three enhanced services that were added into this method. In order to calculate the cost of communications, an estimating model was also put forth, which aids in the search for efficient data allocation strategies. The outcome demonstrated that the suggested technique considerably raises the level of satisfaction with services requirements in web systems. The main shortcomings of this proposed study are the lack of standardization and privacy issues. Syed Thouheed Ahmed and M. Sandhya [16] provided a cutting-edge method and presented about recursive image reduction in the cloud/server. The method depends on pixel value density matching with edge extraction for the suggested Real-Time Biomedical Imaging Recursion Detection. The suggested method reduced initial processing by 60% while achieving time optimization. According to time and space optimization, the suggested system has a 97.8% efficiency rate. It is difficult to schedule the suggested system's total synchronization. P. Sukumar et al. [17] proposed an ontology-oriented architecture that utilized a knowledge base (KB), enabling the integration of data from several heterogeneous sources. The proposed strategy has been used in the area of personalized medicine. In order to find knowledge concealed in diverse data sources, the AI approach is also utilized to mine data in the healthcare industry. The suggested system was subjected to three ontology phases. According to the findings, the textual documents were successfully grouped using the suggested system. The suggested system has many drawbacks, including limited language support and an inability to manage unstructured data.

BikashKanti Sarkar and ShibSankar Sana [18] created a disease decision support system in which the initial stage deals with determining the best training set in parallel with the best data-partition for each illness data set. The second stage investigates a general predictive model over the learned data for a precise disease diagnosis. The suggested method performs admirably on all of the selected medical data sets and can be a useful alternative for the well-known ML techniques. The findings demonstrated that, for the initial identification of the disorders, the suggested hybrid model consistently outperformed the basic learners. However, the quality of the data employed for training the model affects the accuracy of the model. Atta- Ur- Rahman and Mohammed Imran Basheer Ahmed [19] examined a telemedicine plan for a virtual clinic that would provide medical care in remote locations of developing nations. The suggested approach combines a fuzzy rule-based approach to rank the top doctors with a clinical decision support system that aids in selecting the best physician for a certain patient based on his prior prescriptions. The apriori algorithm and the inductive learning algorithm serve as the foundation for the clinical decision support system. The evaluation findings demonstrated that inductive learning performed better than the Apriori algorithm. Syed Thouheed Ahmd et al. [20] suggested a Real-Time Signal Re-Generator and Validator method based on neural networking and machine learning.

3.PROPOSED SYSTEM

Telemedicine, the remote diagnosis and treatment of patients using telecommunications technology, has gained significant traction in recent years, particularly in scenarios where physical access to healthcare facilities is limited. ML algorithms play a crucial role in automating diagnosis prediction, enabling timely and accurate assessments of medical conditions.

Data Handling: Importing necessary libraries and the dataset containing diagnostic information. The dataset is then subjected to data analysis to gain insights into its structure, including its dimensions, summary statistics, correlation between variables, and identification of missing values. Visualization techniques like count plots are employed to understand the distribution of diagnostic labels, crucial for assessing class imbalances.

Preprocessing: Preprocessing steps involve separating the independent variables (features) from the dependent variable (diagnostic labels). The dataset is split into training and testing sets using the train_test_split function, facilitating model training and evaluation.

Model Building: Two classification algorithms are employed: Logistic Regression and Decision Tree Classifier. For each algorithm, the code checks if a pre-trained model exists; if not, it trains the model using the training data and saves it for future use. Model performance evaluation is conducted using various metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive assessment of predictive capabilities.

Performance Evaluation: The performance_metrics function calculates and reports the accuracy, precision, recall, and F1-score for each algorithm. Additionally, it generates a classification report and a confusion matrix, aiding in understanding the model's predictive behavior and potential misclassifications.

Results Presentation: Performance metrics are tabulated for easy comparison between algorithms, allowing stakeholders to identify the most suitable model for deployment. The tabular format presents a concise summary of each algorithm's performance across different metrics.

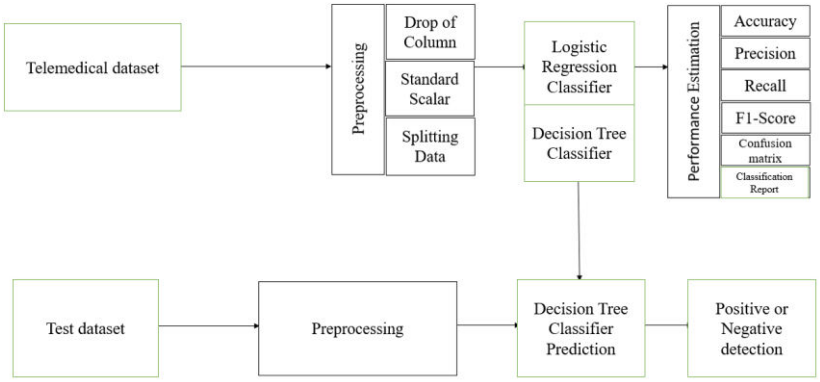

Figure.1: Block Diagram of Proposed system

3.1 Decision Tree Classifier

A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks. It builds a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is constructed by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met, such as the maximum depth of the tree or the minimum number of samples required to split a node.

During training, the Decision Tree algorithm selects the best attribute to split the data based on a metric such as entropy or Gini impurity, which measures the level of impurity or randomness in the subsets. The goal is to find the attribute that maximizes the information gain or the reduction in impurity after the split.

4.RESULTS AND DISCUSSION

Fig 1 presents a sample of the dataset used for predictive modeling. This dataset includes key health metrics and demographic information for individuals, such as Age, Gender, Systolic_BP (Systolic Blood Pressure), Diastolic_BP (Diastolic Blood Pressure), Glucose_Level, BMI (Body Mass Index), Cholesterol_Level, Family_History, and Label (indicating the presence or absence of a certain medical condition). The displayed sample helps in understanding the type of data collected and its potential relevance to health predictions. Fig 2 shows the total count of each class in the dataset. This figure illustrates the distribution of the target variable (Label), which is crucial for understanding the balance between the two classes. An imbalanced dataset can affect the performance of machine learning models, making it essential to visualize and address any significant discrepancies between the class counts. Fig 3 presents the performance metrics of the Logistic Regression model. This figure includes key evaluation metrics such as Precision, Recall, F-Score, and Accuracy. These metrics provide a detailed view of how well the Logistic Regression model performs in predicting the target variable. High values in these metrics indicate a robust and reliable model.

	Age	Gender	Systolic_BP	Diastolic_BP	Glucose_Level	BMI	Cholesterol_Level	Family_History	Label
0	62	1	150	103	254	25.210068	141	0	1
1	65	1	178	93	222	24.690708	261	1	1
2	82	1	151	102	113	41.472272	185	0	0
3	85	1	149	65	154	18.220462	201	1	1
4	85	0	164	89	261	10.060127	196	1	1

Figure.2: Presents the Sample Dataset.

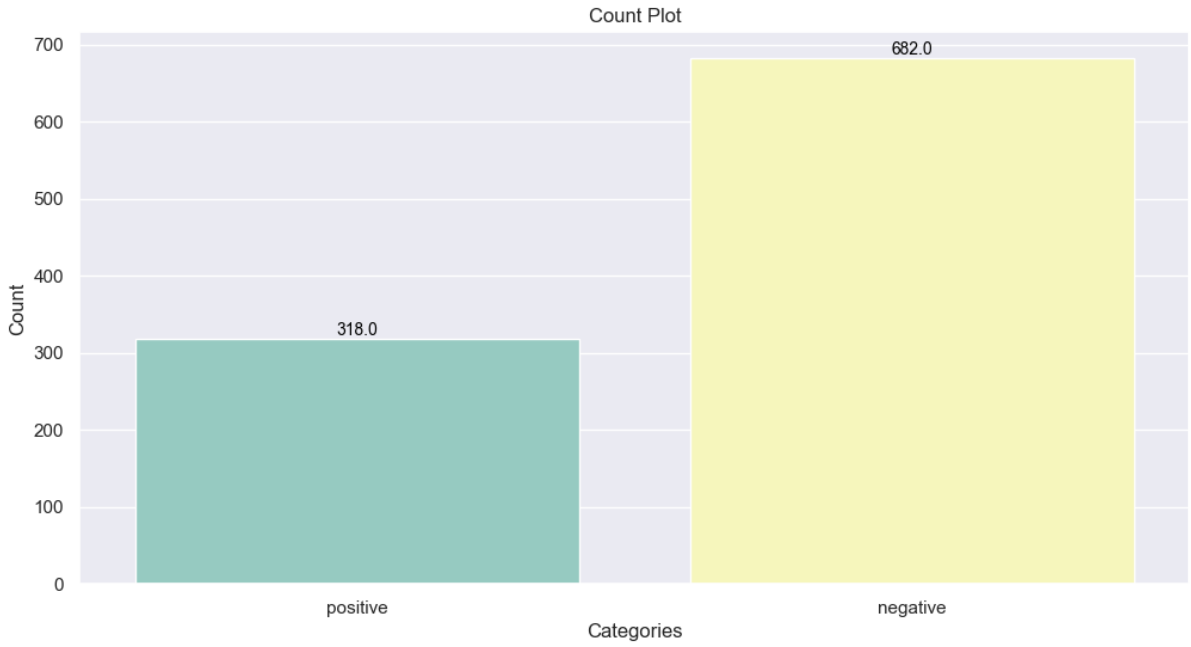

Figure.3: Shows the total count of each class in dataset.

```
LogisticRegression Accuracy    : 97.0
LogisticRegression Precision   : 96.13970588235294
LogisticRegression Recall      : 96.93782141187471
LogisticRegression FSCORE      : 96.52415710809872

 LogisticRegression classification report
              precision    recall  f1-score   support

    positive       0.97      0.94      0.95        32
    negative       0.97      0.99      0.98        68

    accuracy                           0.97       100
   macro avg       0.97      0.96      0.97       100
weighted avg       0.97      0.97      0.97       100
```

Figure.4: Presents the performance metrics of the Logistic Regression.

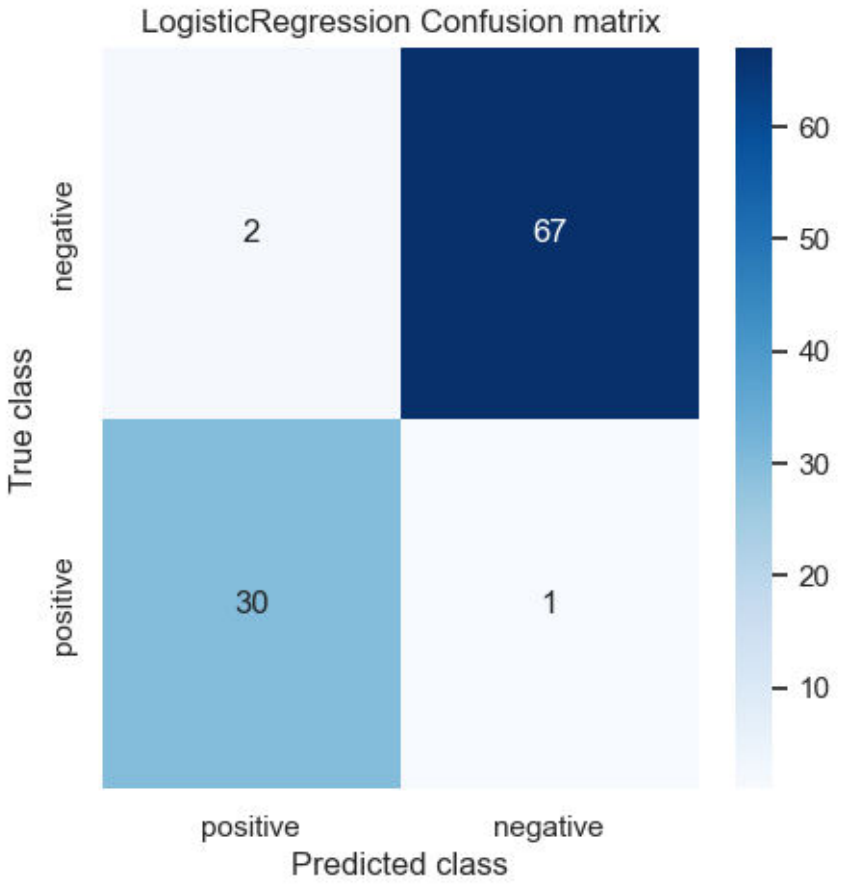

Figure.5: Presents the Confusion Matrix of the Logistic Regression.

Fig 5 displays the Confusion Matrix for the Logistic Regression model. The confusion matrix is a valuable tool for visualizing the performance of a classification model by showing the counts of true positive, true negative, false positive, and false negative predictions. This helps in understanding the model's performance in correctly and incorrectly classifying the instances. Fig 5 presents the performance metrics of the Decision Tree Classifier. Similar to Fig 3, this figure includes Precision, Recall, F-Score, and Accuracy, offering a detailed evaluation of the Decision Tree model's

effectiveness. The metrics help compare the Decision Tree's performance against other models, such as Logistic Regression.

```
DecisionTreeClassifier Accuracy    : 100.0
DecisionTreeClassifier Precision   : 100.0
DecisionTreeClassifier Recall      : 100.0
DecisionTreeClassifier FSCORE      : 100.0

 DecisionTreeClassifier classification report
              precision    recall  f1-score   support

    positive       1.00      1.00      1.00        31
    negative       1.00      1.00      1.00        69

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```

Figure.6: Presents the Performance metrics of Decision Tree Classifier.

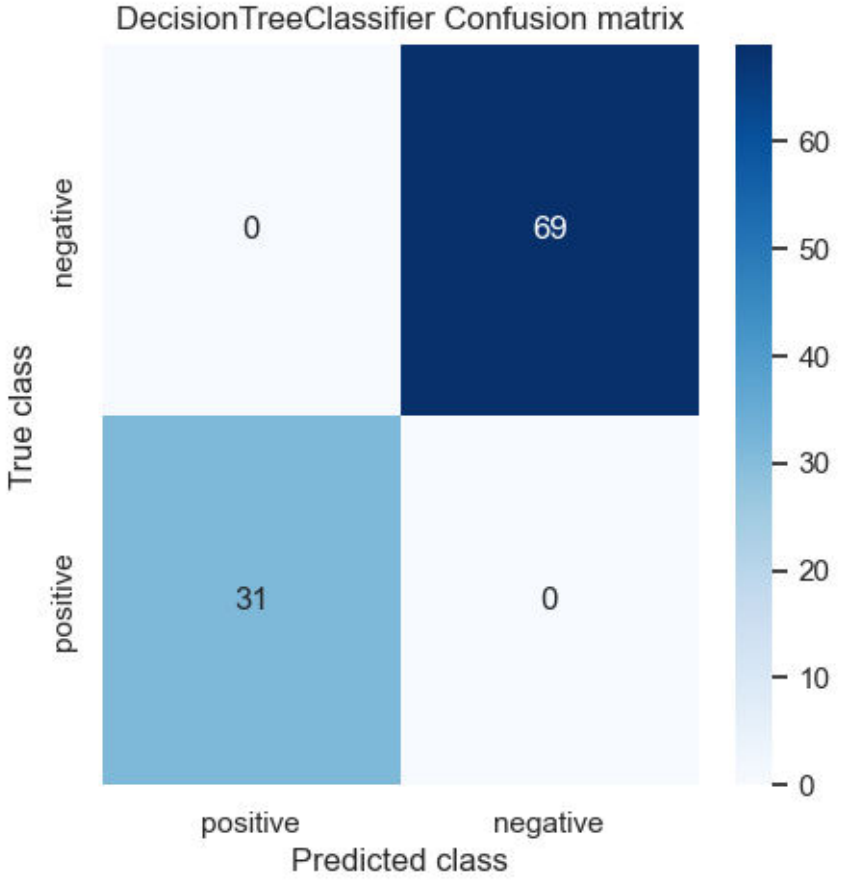

Figure.7: Shows the Confusion Matrix of Decision Tree Classifier.

Fig 7 shows the Confusion Matrix for the Decision Tree Classifier. Like the confusion matrix for Logistic Regression, this matrix helps visualize the classification results, providing insights into the Decision Tree model's accuracy and the types of errors it makes.

5. CONCLUSION

The project focuses on leveraging machine learning (ML) techniques to enhance diagnosis prediction in telemedicine applications. This approach aims to address the inherent limitations of traditional telemedicine systems, which heavily rely on manual interpretation by healthcare professionals. The integration of ML models, such as Logistic Regression and Decision Tree Classifiers, into telemedicine frameworks demonstrates significant improvements in efficiency, accuracy, and accessibility of healthcare services. These models analyze various patient data, including demographics, symptoms, and diagnostic tests, to predict the likelihood of various medical conditions.

The comparative analysis between Logistic Regression and Decision Tree Classifier models shows that the Decision Tree Classifier achieves perfect performance metrics with 100% precision, recall, F-score, and accuracy. On the other hand, Logistic Regression also performs exceptionally well with precision, recall, F-score, and accuracy rates around 97%, showcasing the robustness of both ML techniques in medical diagnosis prediction.

This project's outcomes indicate a promising potential for ML-based systems to transform remote healthcare delivery. By automating diagnosis prediction, telemedicine can become more scalable and efficient, reducing the burden on healthcare professionals and improving patient outcomes, particularly in underserved and remote regions.

REFERENCES

[1] Wootton R, Craig J, Patterson V. Introduction to telemedicine. United Kingdom: CRC Press; 2017.

[2] Lilly CM, Motzkus C, Rincon T, et al. ICU telemedicine program financial outcomes. Chest. 2017; 286–297. doi: 10.1016/j.chest.2016.11.029

[3] Mehrotra A, Jena AB, Busch AB, et al. Utilization of telemedicine among rural Medicare beneficiaries. Jama. 2016:2015–2016. doi:10.1001/jama.2016.2186

[4] Maheu M, Whitten P, Allen A. eHealth, Telemedicine & Telehealth: A comprehensive guide, New York; 2004.

[5] Chowdhury SM, Kabir MH, Ashrafuzzaman K, et al. A telecommunication network architecture for telemedicine in Bangladesh and its applicability. Intern J Digit Content Technol Applic. 2009;3(3):4), doi:10.4156/jdcta.vol3.issue3.20

[6] Huang EY, Knight S, Guetter CR, et al. Telemedicine and telementoring in the surgical specialties: A narrative re. Amer J Surg. 2019; 760–766. doi: 10.1016/j.amjsurg.2019.07.018

[7] Larose DT. Discovering knowledge in data. An introduction to data mining. New Jersey: John Wiley & Sons Publisher; 2005; ISBN 0-471-66657-2.

[8] Dash M, Shadangi PY, Muduli K, et al. Predicting the motivators of telemedicine acceptance in COVID-19 pandemic using multiple regression and ANN approach. J Stat Manage Syst. 2021; 319–339. doi:10.1080/09720510.2021.1875570

[9] Ahmed ST, Sandhya M, Sankar S. TelMED: dynamic user clustering resource allocation technique for MooM datasets under optimizing telemedicine networks. Wirel Person Commun. 2020; 1061–1077. doi:10.1007/s11277-020-07091-x

[10] Sadineni PK. Developing a model to enhance the quality of health informatics using big data. In 2020 fourth international conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC) (pp. 1267–1272). IEEE; 2020.

[11] Sornalakshmi M, Balamurali S, Venkatesulu M, …Muthu BA. Hybrid method for mining rules based on enhanced Apriori algorithm with sequential minimal optimization in the healthcare industry. Neural Comput Applic. 2020: 1–14.

[12] Choi SY, Chung K. Knowledge process of health big data using MapReduce-based associative mining. Pers Ubiquitous Comput. 2020;24:571–581. doi:10.1007/s00779-019-01230-3

[13] Priyadarshan DJ, Sanjay KK, Kathiresan S, et al. Patient health monitoring using IoT with machine learning. Intern Res J Eng Technol (IRJET). 2019;6(03.

[14] Sandhiya R, Sundarambal M. Clustering of biomedical documents using ontology-based TF-IGM enriched semantic smoothing model for telemedicine applications. Cluster Comput. 2019;22:3213–3230. doi:10.1007/s10586-018-2023-4

[15] Anusuya TK, Maharajothi P. A survey of telemedicine services using data mining. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN, 2456–3307; 2019.

[16] Thouheed Ahmed S, Sandhya M. Real-time biomedical recursive images detection algorithm for Indian telemedicine environment. In: Cognitive informatics and soft computing: proceeding of CISC 2017. Springer Singapore; 2019. p. 723–731.

[17] Sukumar P, Monika G, Gokila D, et al. An NLP based ontology architecture for dealing with Heterogeneous data to telemedicine systems. South Asian J Eng Technol. 2019;8(1):89–92.

[18] Sarkar BK, Sana SS. An e-healthcare system for disease prediction using hybrid data mining technique. J Model Manage. 2019;14(3):628–661. doi:10.1108/JM2-05-2018-0069

[19] Ahmed MIB. Virtual clinic: A CDSS assisted telemedicine framework. In: Telemedicine technologies. Academic Press; 2019. p. 227–238.

[20] Ahmed ST, Sandhya M, Sankar S. An optimized RTSRV machine learning algorithm for biomedical signal transmission and regeneration for a telemedicine environment. Procedia Comput Sci. 2019;152:140–149. doi:10.1016/j.procs.2019.05.036