www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Machine Learning-Based Diagnosis Prediction in Telemedicine Applications

Sayyed Hasanoddin[1*], Shivani Kolipaka[2], Mohd Adnan[2], Nomula Akshay Kumar[2], Mohammad Abdur Rahman[2]

[1,2]Department of Computer Science and Engineering (Data Science), Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana.

*Corresponding Email: hasanoddin.sayyed@gmail.com

ABSTRACT

Telemedicine, the practice of remotely diagnosing and treating patients through telecommunications technology, is transforming modern healthcare delivery. With its rising adoption, the demand for efficient and accurate diagnostic prediction systems has become increasingly apparent. In traditional telemedicine systems, diagnosis prediction often relies heavily on manual interpretation by healthcare professionals. This method is time-consuming, subjective, and prone to errors due to human factors like fatigue and cognitive biases. Furthermore, traditional systems may struggle with scalability and efficiency in handling large volumes of patient data, limiting their effectiveness in delivering remote healthcare, particularly in underserved areas. The motivation behind this work lies in addressing the shortcomings of traditional telemedicine systems and harnessing ML's potential to revolutionize remote healthcare delivery. By automating diagnosis prediction, the work aims to improve the efficiency, accuracy, and accessibility of telemedicine services, particularly in underserved communities and remote regions where access to healthcare resources is limited. Furthermore, automated diagnosis prediction can alleviate the burden on healthcare professionals, enabling them to focus on critical patient care tasks while leveraging technology for routine diagnostic assessments. The proposed system entails the development and deployment of ML-based diagnosis prediction models tailored for telemedicine applications. Leveraging MLP Classifier, the system will analyze patient data, including demographics, symptoms, and diagnostic tests, to predict the likelihood of various medical conditions.

Keywords: Machine Learning, Telemedicine, Logistic Regression, MLP Classifier, Disease diagnosis.

1. INTRODUCTION

The concept of telemedicine encompasses the use of information and telecommunication technology to render medical services irrespective of geographical separation between physicians and patients [1]. Telemedicine has been in practice as far back as the 1900s. It covers any form of electronic communication between health workers and patients from a remote location [1,2]. Recently, researchers have focused more on wireless communication technologies for telemedicine to provide effective and reliable health care service delivery from remote location especially during emergencies. Various communication technologies have been proposed and implemented for providing expert medical services to patients without the need for the conventional face-to-face encounters with patients. This has greatly reduced the cost of medical diagnosis and the need to travel long distances in search of professional consultations. Available studies on telemedicine implementations suggest the need for continuous research to address several issues and challenges [3,4]. There is a need to compare relevant studies in the field in order to provide a broad overview of available communication technologies suitable for modern designs as well as to identify the most viable means of practical implementation. This is not to say that telemedicine should completely replace the conventional practice of physical diagnostic and other medical processes, as certain services require physical face-to-face contact.

Nonetheless, the deployment of telemedicine could greatly reduce congestion in hospitals, and consequently limit the spread of infectious diseases. The advent of the COVID-19 pandemic has increased the need to leverage the benefits of telemedicine and eHealth. Telemedicine and eHealth research has seen an upsurge in recent works [5]. For instance, the authors of [6] examined variability in services between in-person and telemedicine.

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 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. The

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

primary goal of the suggested design is to obtain a higher order of signal optimization for secure and reliable telemedicine consultation of biological samples via low line transmission channels. The RTSRV method has considered feature extraction and layered decomposition of a signal. The features are grouped using the KNN method to categorize each attribute based on the frequency with which it occurs within a certain time span. The suggested algorithm has a higher rate of accuracy. However, the effectiveness of the device is dependent on the signal quality. R. Sandhiya and M. Sundarambal [21] created an effective clustering approach for biomedical documents and health data based on chicken swarm optimization with dynamic dimension reduction to support telemedicine applications. The data are initially preprocessed using concept mapping and semantic annotation, which increase document representation, frequency, and inverse gravity moment factor, and the modified n-gram, which rectifies for substitution and deletion errors. The outcomes of the experiments demonstrated that the proposed methodology can be very effective in telemedicine applications and remote monitoring of medical treatment. The benefits of the suggested model include a reduction in time complexity, accurate clustering findings regardless of dataset rescaling or normalization, and independence from document order. The proposed system is sensitive to localization.

3. PROPOSED METHODOLOGY

The proposed system of this research involves developing and comparing machine learning models to diagnose medical conditions using a telemedicine dataset. After preprocessing and encoding the data, both Logistic Regression and MLP Classifier models are trained and evaluated. The better-performing model is then used to make predictions on unseen test data.

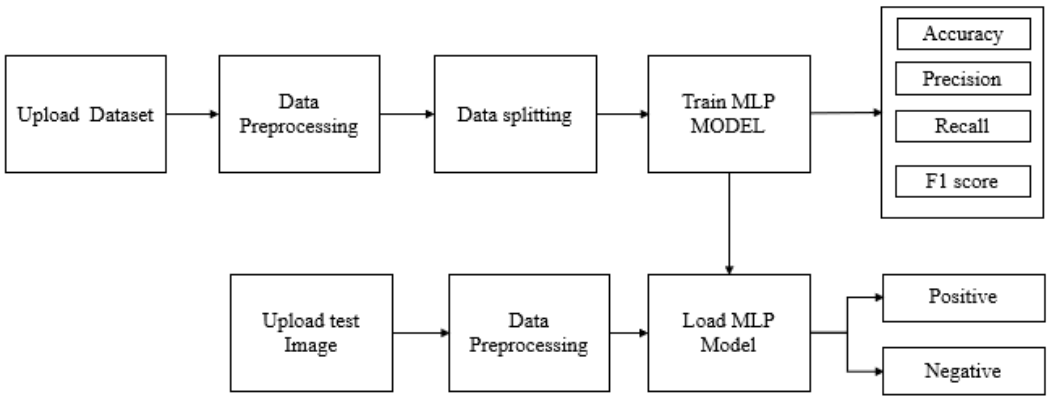


Fig. 1: Block Diagram of proposed system for diagnosis prediction in telemedicine.

Step 1: Telemedicine Application Dataset

The objective is to utilize a dataset containing patient information, including demographics, symptoms, and test results, to develop a model for diagnosing medical conditions. The dataset typically consists of features (independent variables) such as age, gender, symptoms, and diagnostic tests, along with a target variable (dependent variable) indicating the diagnosis, which is labeled as either 'positive' or 'negative'.

Step 2: Dataset Preprocessing

In this step, the dataset is checked for any missing values. These null values are either removed or appropriately filled to ensure the data is complete. This process is essential for maintaining the accuracy

and reliability of the model. After handling missing data, relevant features are selected for model training, excluding the target variable.

Step 3: Label Encoding

Categorical variables in the dataset are transformed into numerical values through label encoding. This transformation is crucial because most machine learning algorithms require numerical inputs. For instance, categorical labels such as 'positive' and 'negative' are encoded as 1 and 0, respectively.

Step 4: Logistic Regression

A Logistic Regression model is implemented as the baseline approach for predicting medical diagnoses. The model is trained using the preprocessed dataset, allowing it to learn the relationship between input features and the target variable. After training, the model's performance is evaluated using accuracy, precision, recall, F1 score, and a confusion matrix to understand its predictive capability.

Step 5: MLP Classifier

An MLP Classifier (Multilayer Perceptron) is proposed as an advanced model for diagnosis prediction. This model is trained on the same dataset used for Logistic Regression, with adjustments made to hyperparameters such as the number of hidden layers and the number of training iterations to optimize performance. The model is then evaluated using the same set of metrics for a fair comparison.

Step 6: Performance Comparison

The performance results of both models—Logistic Regression and MLP Classifier—are compared. This comparison is based on accuracy, precision, recall, and F1 score. The goal is to analyze which model performs better and under what conditions, thereby demonstrating the potential advantages of using a neural network approach like the MLP Classifier for this type of task.

Step 7: Prediction of Output from Test Data with MLP Classifier

In the final step, the trained MLP Classifier model is used to make predictions on a separate testing dataset, which the model has not previously encountered. For each prediction made, the model's output is interpreted to indicate whether the diagnosis is predicted to be 'positive' or 'negative'. These predictions are reviewed in the context of their real-world applicability, providing insight into the effectiveness of the model and offering a basis for potential improvements based on the observed results.

3.1 MLP Classifier

A Multi-Layer Perceptron (MLP) Classifier is a type of artificial neural network used for supervised learning tasks, particularly for classification problems.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

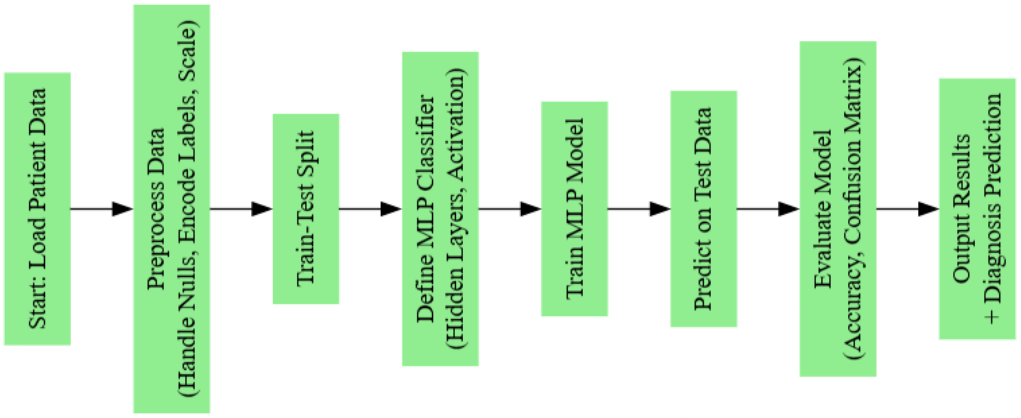


Fig. 2: Block diagram of working model of MLP Classifier.

It consists of multiple layers of neurons, including an input layer, one or more hidden layers, and an output layer. MLPs can capture complex patterns in data, making them suitable for a wide range of applications. The architecture consists of an input layer that takes feature inputs, one or more hidden layers that learn intermediate representations, and an output layer that produces the final class predictions. Each layer is composed of multiple neurons, with connections defined by weights that are adjusted during training.

The MLP Classifier operates by passing input features through multiple layers of interconnected neurons. Each neuron applies a linear transformation followed by a non-linear activation function (like ReLU or sigmoid) to introduce non-linearity. The output layer generates predictions based on the transformed data. The model is trained using backpropagation, where the error from the output is propagated backward through the network to update weights and biases, minimizing the loss function (such as cross-entropy).

4. RESULTS AND DISCUSSION

4.1 Dataset description

The dataset contains medical records with various health metrics and attributes for individuals, used to predict the presence or absence of a specific medical condition, indicated by the "Label" attribute. Each row represents an individual, with columns detailing their health status and demographics.

The Age column represents the individual's age in years, a key factor in diagnosis due to age-related health risks, with values ranging from 62 to 85 years. Gender is a categorical variable where '0' stands for female and '1' for male, acknowledging that disease patterns can differ between sexes. Systolic_BP captures the systolic blood pressure, an important cardiovascular indicator taken when the heart beats. Diastolic_BP reflects the pressure in arteries when the heart rests between beats, also essential in assessing cardiovascular risk.

Glucose_Level records blood sugar levels, used to evaluate diabetic conditions, with sample values like 113, 154, and 261 mg/dL. BMI (Body Mass Index) indicates the individual's weight category based on height and weight, aiding in assessing obesity or underweight conditions. Cholesterol_Level shows the cholesterol in the blood, a key factor in cardiovascular health, with typical values such as 141, 185, and 261 mg/dL.

Family_History is a binary variable denoting the presence ('1') or absence ('0') of disease history in the family, which is critical for predicting genetically influenced conditions. Finally, the Label column is

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

the target variable, representing whether the individual has ('1') or does not have ('0') the medical condition of interest, supporting binary classification in predictive modeling.

4.2 Results description

Fig. 3 displays a count plot of the Family_History attribute, which indicates whether individuals in the dataset have a family history of the medical condition being studied. The x-axis represents the two categories: '0' for individuals without a family history and '1' for those with a family history, while the y-axis shows the count of individuals in each category. The plot reveals that the number of individuals with a family history (label '1') is slightly higher than those without it (label '0'), with both categories having nearly equal representation—each just above 490 instances. This balanced distribution suggests that the dataset provides a fair representation of both groups, which is beneficial for training machine learning models without introducing bias toward one class.

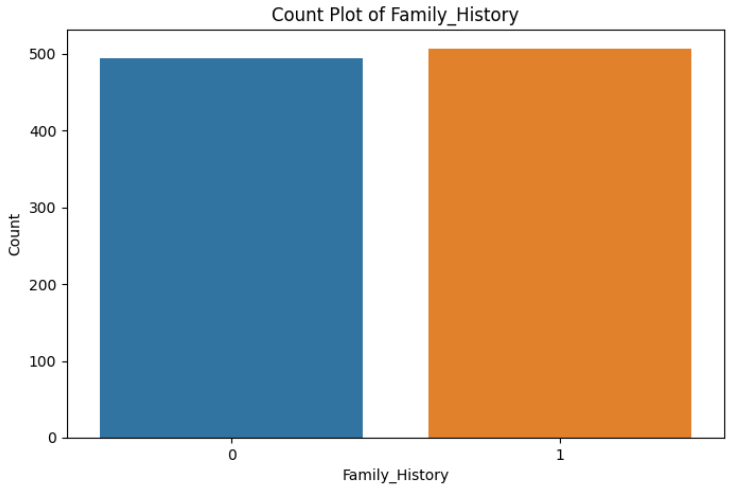
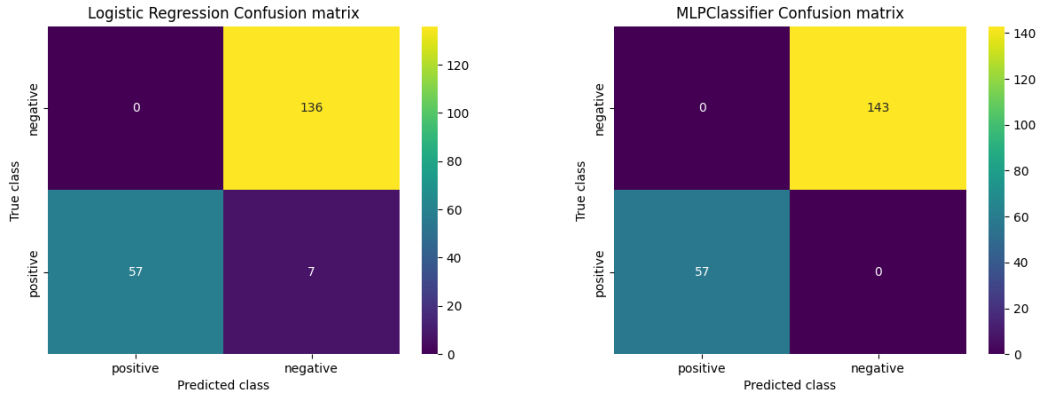


Fig. 3: Count plot of Family History.



(a) (b)

Fig. 4: Confusion matrices obtained using (a) Logistic Regression. (b) MLP Classifier model.

The confusion matrices compare the classification performance of Logistic Regression (Fig. 4 (a)) and MLPClassifier (Fig. 4 (b)). In the Logistic Regression confusion matrix, the model correctly predicts 136 negative cases but misclassifies 57 positive cases as negative and 7 negative cases as positive,

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

leading to some classification errors. This explains why the recall is slightly lower, as it fails to capture all positive instances. In contrast, the MLPClassifier confusion matrix shows perfect classification, correctly predicting 143 negative cases and 57 positive cases, with no misclassifications. This aligns with its 100% accuracy, precision, and recall, suggesting that it has perfectly fit the dataset. However, such results may indicate overfitting, meaning the model might not generalize well to unseen data.

Fig. 5 illustrates the graphical user interface (GUI) of the Telemedicine Application after it has successfully performed predictions on a test dataset. The interface displays individual patient data records including attributes such as Age, Gender, Systolic_BP, Diastolic_BP, Glucose_Level, BMI, Cholesterol_Level, and Family_History. For each record, the predicted medical outcome is shown as either "positive" or "negative.

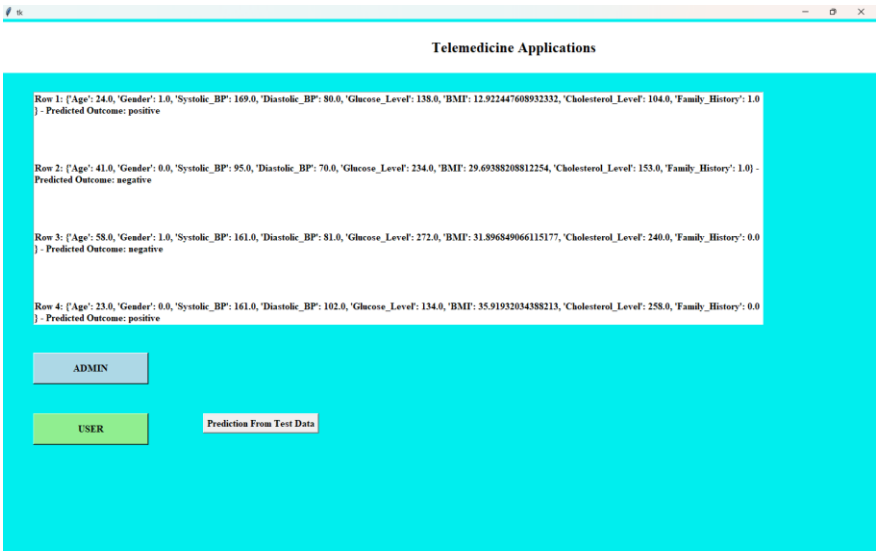


Fig. 5: Illustration of GUI application after successful prediction on test dataset.

Table. 1: Performance Comparison of Algorithms.

Metric	Logistic Regression	MLP Classifier
Accuracy	96.5%	100.0%
Precision	97.55%	100.0%
Recall	94.53%	100.0%
F1-Score	95.85%	100.0%

In Table.1 the performance metrics compare Logistic Regression and MLPClassifier on a given classification task. Logistic Regression shows strong results with an accuracy of 96.5%, meaning it correctly classifies most instances. Its precision (97.55%) suggests that when it predicts a positive class, it is correct 97.55% of the time. The recall (94.53%) indicates that it captures 94.53% of actual positive cases. The F1-score (95.85%) balances precision and recall, confirming it is a well-rounded model. On the other hand, MLPClassifier performs perfectly across all metrics (100% accuracy, precision, recall, and F1-score), suggesting it classifies every instance correctly.

5. CONCLUSION

The study demonstrates the effectiveness of machine learning models, particularly Logistic Regression and MLPClassifier, in a ML-based telemedicine diagnosis prediction system. Logistic Regression performed well with 96.5% accuracy, showing strong classification capabilities but with some misclassifications. In contrast, MLPClassifier achieved perfect scores (100%) in accuracy, precision, recall, and F1-score, suggesting flawless classification. The confusion matrices further highlight the strengths and weaknesses of both models, with MLPClassifier making no errors while Logistic Regression had minor misclassifications. The implemented system successfully integrates predictive analysis for telemedicine applications, offering valuable insights into patient diagnosis based on medical attributes.

REFERENCES

[1] Moore, M. The Evolution of Telemedicine. Future Gener. Comput. Syst. 1999, 15, 245–254.

[2] Pandian, P.S.; Safeer, K.P.; Shakunthala, D.T.I.; Gopal, P.; Padaki, V.C. Store and Forward Applications in Telemedicine for Wireless IP Based Networks. J. Netw. 2007, 2, 58–65.

[3] Devaraj, S.J.; Ezra, K. Current Trends and Future Challenges in Wireless Telemedicine System. In Proceedings of the 2011 3rd International Conference on Electronics Computer Technology, Kanyakumari, India, 8–10 April 2011; IEEE: Piscataway, NJ, USA, 2011.

[4] Chakraborty, C.; Gupta, B.; Ghosh, S.K. A Review on Telemedicine-Based WBAN Framework for Patient Monitoring. Telemed. e-Health 2013, 19, 619–626. [Google Scholar] [CrossRef]

[5] Alenoghena, C.O.; Onumanyi, A.J.; Ohize, H.O.O.; Adejo, A.O.; Oligbi, M.; Ali, S.; Okoh, S.A. eHealth: A Survey of Architectures, Developments in mHealth, Security Concerns and Solutions. Int. J. Environ. Res. Public Health 2022, 19, 13071.

[6] Campbell, I.; Crowley, T.; Keena, B.; Donoghue, S.; McManus, M.; Zackai, E. The experience of one pediatric geneticist with telemedicine-based clinical diagnosis. Am. J. Med. Genet. 2023, 1, 1–7.

[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

[21] Sandhiya R, Sundarambal M. Chicken swarm optimization-based clustering of biomedical documents and health records to improve telemedicine applications. Intern J Enterpr Netw Manage. 2019;10(3-4):305–328. doi:10.1504/IJENM.2019.103158.