Proceedings of the 5th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS-2025)
DVD Part Number: CFP25UG3-DVD; ISBN: 979-8-3315-5283-1

Multilayer Perceptron Neural Network-based Enhanced Big Data Analysis for Healthcare Data to Identify High-Risk Sensitive Data

[1st]R.Anitha
Research Scholar
Department of Computer Science
Vels Institute of Science, Technology and Advanced Studies (VISTAS),
Pallavaram, India.
anithanathan.tvm@gmail.com

[2nd]Dr.R. Bagavathi Lakshmi
Associate professor,
Department of Computer Applications,
Vels Institute of Science, Technology & Advanced Studies,
Pallavaram, India.
rbagavathi.scs@vistas.ac.in

***Abstract*— In today's era, data compression is crucial for managing big data in the healthcare sector, reducing dimensionality and simplifying registration complexity. The role of artificial intelligence (AI) models in big data analysis in healthcare is crucial for identifying patients with high-risk, disease-impacting features that are essential for diagnosis. Traditional machine learning (ML) models typically do not analyse feature edges to minimise false positives. As dimensionality increases, accuracy decreases proportionally, affecting the accuracy of sensitive data recognition in feature analysis. To address this issue, we propose a Multilayer Perceptron Neural Network (MLP-NN) to identify risks in healthcare data. Initially, Z-Score Normalisation (ZSN) is used to preprocess the dataset. Moreover, the Support Vector Machine (SVM) technique is employed to select the most essential features. After that, an MLPNN is used to accurately classify maternal health risk levels into low, medium, and high. The proposed system enhances high-dimensional feature reduction by leveraging auxiliary resources to improve detection accuracy. The experimental results show that the classification accuracy is 96.4%, which is higher than that of other methods.**

Keyword*— Big data, healthcare sector, Z-Score normalization, SVM, Multilayer Perceptron Neural Network (MLP-NN), High-Risk Sensitive Data

I. INTRODUCTION

Due to the increasing amount of big data being generated in the health industry, medicine has enhanced more explicit approaches to analyzing big data and predicting patient results. The use of big data analytics, especially Deep Learning (DL) and Electronic Health Records (EHR), has been found to have great potential in early warning systems for various critical conditions, including suicide attempts, heart failure, and chronic diseases, among others [1-3]. Nevertheless, the healthcare domain has specific issues related to data sensitivity, dataset shortage, especially with equal distribution, and identifying patterns often hidden in high-risk groups. As a result, these problems are known to challenge traditional ML techniques in terms of their ability to predict the results and their inaptness to deal with the heterogeneous data characteristic of healthcare [4-5]. To tackle these challenges, we suggest an improved big data analysis model for healthcare, using a MLPNN. This method combines the advantage of deep learning for capturing nonlinear relationships with feature optimization to guarantee sound classification of highly sensitive high-risk records [6-7-8].

This proposed method consists of a feature selection heuristic and an optimized small neural network that provides high-quality classification and alleviates the problem of data imbalance while increasing the model interpretability. MLPNN are especially suited for high dimensional data analysis this is so because of the multiple layers it is composed of to help identify key information for prevention and detecting potential complications, including renal failure, sepsis, and medication errors [9, 10]. This novel approach extends state-of-the-art DL models and big data analytics and solves the drawbacks. It offers a tailored package for healthcare systems to screen at-risk patients and enhance appropriate care, timely interventions, and clinical decision-making. Smart features are employed coupled with nonlinear optimization in MLPNN, hence bringing to bear the real possibility of unlocking healthcare big data to improve patient care.

A. The main contribution of the research

- The MLPNN improves model interpretability, reduces the issue of data imbalance, and offers high-quality classification.
- Blood pressure, glucose levels, and maternal age are among the most significant features explored by the SVM method, which also eliminates noise and less important information from the data input.
- MLPNNs provide medical professionals with a dependable tool for early diagnosis and intervention in high-risk maternity patients by reducing the classification error through iterative training.
- From forward propagation to weight updates and loss computation, the MLPNN framework is essential for converting unprocessed maternal health data (such as blood pressure, blood sugar, and age) into precise risk level forecasts (low, medium, or high).

II. LITERATURE SURVEY

The authors [11] investigated the Smart Healthcare Prediction and Evaluation (SHPE) model, which is based on DL. In order to maximize resource allocation and enhance patient outcomes, this method was used for predictive evaluation in healthcare scenarios. The study did, however, identify difficulties in integrating multimedia data with criteria unique to the healthcare industry. Data heterogeneity and real-time processing limitations require more investigation. A K-Nearest Neighbour (K-NN) approach was used to fix the problem. The approach's capacity to effectively identify high-risk patients was demonstrated by its application to a variety of healthcare datasets. However, there are still issues with the models' scalability and generalizability to different populations [12].

The author [13] investigated Federated Learning (FD), which permits collaborative model training without centralized data storage, for its potential in healthcare informatics. This method improved predicted performance while protecting patient privacy. However, issues with system scalability, data heterogeneity, and communication overhead were noted. To address these issues, an Explainable Artificial Intelligence (XAI) approach was used. Numerous XAI techniques have been used for risk assessment and disease prediction; however, combining explainability with high-performance models is still difficult [14].

An optimization-assisted Bidirectional Gated Recurrent Unit (Bi-GRU) model was presented by the author [15] for big data-based healthcare monitoring. Both forecast accuracy and monitoring efficiency were shown to have significantly increased with this technique. The Bi-GRU did, however, point out that the model's ability to adjust to novel data patterns was limited. It was advised to use reinforcement learning for dynamic updates in order to get over this limitation. A Chaotic Red Deer Optimizer (CRDO) approach was suggested as a solution to the problem. The model surpassed conventional techniques in terms of accuracy and computational efficiency and successfully diagnosed illnesses. However, in other datasets, the optimizer's chaotic nature caused stability problems. Hybrid optimization strategies were put up to lessen this [16].

In a big data context, the author [17] created the best DL model for classifying medical images. The method demonstrated improved scalability and classification performance. However, medical picture preprocessing requires a lot of resources, indicating that future implementations should use lightweight preprocessing methods. A DL-based Visualization Algorithm (VA) for medical big data was suggested as a solution to the problem. Although the system offered clear insights for healthcare decision-making, it had trouble effectively managing massive datasets. It was suggested that distributed computing and visualization approaches be improved further [18].

The integration of DL and big data in healthcare was covered by the author [19]. Numerous applications were investigated, including personalized treatment and illness prediction. Nonetheless, the study made clear that strong frameworks are required for the effective management and processing of large amounts of healthcare data. Integrating blockchain technology to improve data security and interoperability should be the main emphasis of future research. An Improved Dragonfly Algorithm (IDA) approach was used to achieve the goal. Although the models' categorization accuracy was excellent, their deployment in contexts with limited resources was limited [20].

DeepMist, a mist computing system for DL-based healthcare large data management, was presented by the author [21]. Although the framework showed improvements in latency and resource usage, real-time applications still needed more optimization. To improve performance, adaptive resource allocation techniques were suggested. The Long Short-Term Memory and CNN (LSTM-CNN) approach was suggested as a solution to the problem. Although there were issues with model interpretability, the suggested approach showed improved prediction accuracy and resilience [22].

ST-Med-Box is a DL-based intelligent medication recognition system that was implemented by the author [23]. In addition to offering additional medication-related features, the suggested system can help chronic patients take several medications appropriately and prevent taking the incorrect ones, which could result in drug interactions. Although this work was more robust to unforeseen cases and medical terms, it was not interpretable due to certain data limitations, which caused issues with the system. A fuzzy C-Means (FCM) approach was suggested as a solution to the problem. Although the models performed well in detecting software problems, they were not scalable for bigger datasets. Future improvements were proposed using parallel processing methods and ensemble procedures [24].

The author [25] introduced the research of diabetes risk data mining method based on electronic medical record analysis and intends to provide some ideas and directions for the research of diabetes risk data mining method. It includes data mining and classification rule mining based on electronic medical record analysis, which is used in the research experiment of diabetes risk data mining method based on electronic medical record analysis. The proposed method provided valuable insights but faced challenges in integrating diverse healthcare datasets.

III. PROPOSED METHODOLOGY

This section introduces the proposed MLP-NN method, designed to accurately categorize risk levels—low, medium, and high using a maternal health risk dataset. Initially, we apply ZSSD as a data preprocessing step. This ensures that all inputs and outputs are normalized, resulting in outcomes with a zero mean and unit standard deviation. An SVM for feature selection enhances the process by identifying the most significant predictors, like systolic blood pressure and BMI, while filtering out irrelevant attributes that could introduce noise. Following feature selection, we employ MLP-NN for classification due to its capability to address complex nonlinear relationships within the dataset. The proposed MLP-NN offers several advantages: it is highly scalable, enabling accurate risk level categorizations in maternal health, including low, medium, and high risk.

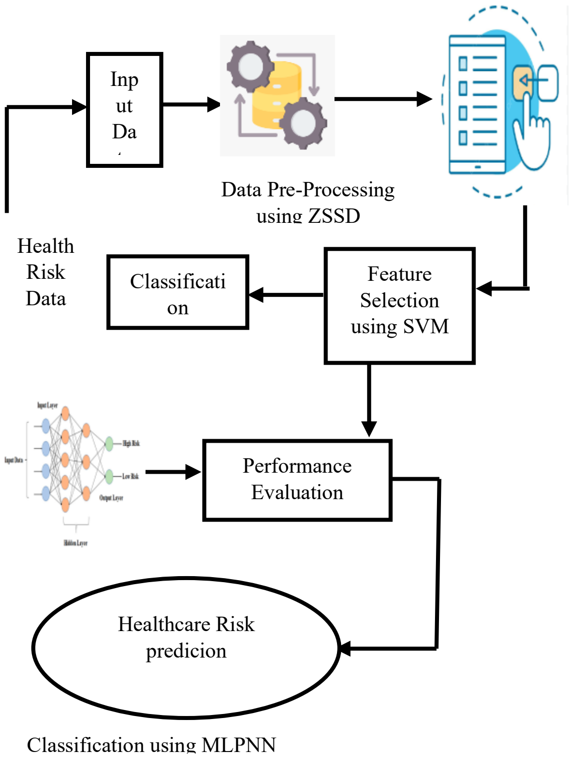

Fig. 1. The Proposed MLPNN Method Architecture Diagram

Figure 1 illustrates the proposed structure of the MLPNN method designed for predicting high and low health risk categories. This comprehensive approach begins with a detailed collection of health risk data as the basis for subsequent analysis. Preprocessing using the ZSSD technique for standardizing datasets and improving model performance. Feature selection is further refined with SVM, which ensures that only the most relevant variables contribute to the model's predictive power. Finally, the MLPNN classification process leverages big data to accurately assess health risks and make informed decisions in healthcare settings. This systems approach emphasizes the integration of advanced ML techniques to improve prognostic accuracy and support proactive health management.

A. Dataset Collection

This section assists medics and healthcare professionals in minimizing deaths and complications through the proposed method. The open dataset utilized in this study comes from collected samples available on Kaggle. By analyzing the characteristics of class variables, we can predict the associated risk levels within the health data. Furthermore, classifying health data into low and high-risk categories allows us to estimate the risk levels of various health problems for patients. Additionally, maternal health risk datasets can be accessed via https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data.

Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
25	130	80	15	98	86	high risk
35	140	90	13	98	70	high risk
29	90	70	8	100	80	high risk
30	140	85	7	98	70	high risk
35	120	60	6.1	98	76	low risk
23	140	80	7.01	98	70	high risk
23	130	70	7.01	98	78	mid risk
35	85	60	11	102	86	high risk
32	120	90	6.9	98	70	mid risk
42	130	80	18	98	70	high risk
23	90	60	7.01	98	76	low risk
19	120	80	7	98	70	mid risk
25	110	89	7.01	98	77	low risk
20	120	75	7.01	100	70	mid risk
48	120	80	11	98	88	mid risk
15	120	80	7.01	98	70	low risk
50	140	90	15	98	90	high risk

Fig. 2. Dataset Feature Collection

As shown in Figure 2, health risk data can be used to collect dataset characteristics and analyze high and low blood pressure values and other vital factors. Furthermore, based on blood glucose level and mmol/L, the predicted risk severity of the preceding attributes can be identified.

B. Z-Score Standardized Data (ZSSD)

In this section, ZSSD technology can be used to process high-risk data from sensitive records proactively. Additionally, the ZSSD algorithm can be used to transform a data set to a zero-mean and unit-variance distribution. ZSSD is an analytical standardization process that converts the features of a health risk dataset into a specific range.

As shown in Equation 1, the minimum and maximum data can be normalized using the sensitive records of the health risk data in the decimal range. Let's assume y' −normalized value, y-attribute, c_a −mean value, std − standard deviation attribute, y_a −original value, M_{ax}, M_{in} −maximum and minimum value, $M_{ax}y_{new}$ − $M_{in}y_{new}$ −rescale value data

$$y' = \frac{y_1 - c_a}{std(y)} \quad (1)$$

$$y'_{a,n} = \frac{y_{i,n} - M_{in}(y_{a,n})}{M_{ax}(y_{a,n}) - M_{in}(y_{a,n})}(M_{ax}y_{new} - M_{in}y_{new}) + M_{in}y_{new} \quad (2)$$

A preprocessing technique can detect sensitive data using a health risk dataset based on large data sets normalized to the range of [-1, 1] and [0, 1].

C. Support Vector Machine (SVM)

In this section, we present an analytical approach to forecasting patient prognosis by utilizing data gathered from the database through the SVM algorithm, a method commonly employed in medicine. We also introduce an SVM technique to identify the hyperplane that separates the training data into two categories. Likewise, a selection of vectors from the training set can be employed to forecast high-risk medical conditions. Furthermore, nonlinear issues can be addressed by transforming hyperplane data into a high-dimensional space using the SVM algorithm and assessing the probability of linear separation. By applying kernel functions, bit data can facilitate the prediction of high-risk conditions in the medical field.

By maximizing the margins, we can identify the optimal hyperplane that separates the two shapes. SVM techniques can use kernel functions to map the inputs to data sets that are not linearly separable into a high-dimensional feature space. As shown in Equation 3, compute the possible separating hyperplanes in the transformed high-dimensional feature space1. Let's assume z^o −multidimensional vector, j- constant, u_a −feature instances.

$$z^o\varphi(u_a) + j = 0 \tag{3}$$

As illustrated in Equation 4, the linear function is computed by determining the optimum hyperplane and the inequalities of the input and output variables. Let's assume ξ_i −slack variable, v_a − set of variables.

$$v_a = |z^o\varphi(u_a) + j| \geq 1 - \xi_a \tag{4}$$

Equation 5 estimates the classification errors and the distance between the classification intervals. Compute the radial basis kernel functions described as eigenvectors in the input space, as illustrated in Equation 6. The Gaussian radial-based kernel function can be used to predict the highest risks in the sensitivity record based on the amplitudes based on the frequency. Let's assume e −relaxation variable, ξ_a −classification error, $\frac{1}{2}|z|^2$ − classification interval. Eigenvectors, $u - u'$ −eigenvectors input space, σ^2 −bandwidth Gaussian radial basis kernel function, γ − parameter, exp − exponential function, γ −general value.

$$\min = \left[\frac{1}{2}|z|^2\right] + e\sum_{a=1}^{h}\xi_a \tag{5}$$

$$p(u, u') = \exp\left(-\frac{\|u-u'\|^2}{2\sigma^2}\right) = \exp(-\gamma\|u - u'\|^2) \tag{6}$$

As illustrated in Equation 7, the general value is computed utilizing a grid search method to identify key records. Let's assume $f_{(e_1,e_2)}$ −distance measure, $g_1 - g_2$ −feature space centroid vectors, q- Length of space crosse's grid.

$$f_{(e_1,e_2)} = \|g_1 - g_2^2\| = \frac{1}{q_1^2}\sum_{a=1}^{q_1}\sum_{b=1}^{q_1}\exp\left(-\gamma\left\|u_a^{(1)} - u_b^{(1)}\right\|^2\right) + \frac{1}{q_1^2}\sum_{a=1}^{q_1}\sum_{b=1}^{q_1}\exp\left(-\gamma\left\|u_a^{(2)} - u_b^{(2)}\right\|^2\right) - \frac{1}{q_1 q_2}\sum_{a=1}^{q_1}\sum_{b=1}^{q_2}\exp\left(-\gamma\left\|u_a^{(1)} - u_b^{(1)}\right\|^2\right) \tag{7}$$

Calculate the average distance in the kernel interval, as described in Equations 8 and 9. The optimal kernel parameters can be determined as the feature space center vectors of the first and second-class data. Healthcare data can be analyzed to identify significant risks.

$$g_1 = \frac{1}{q_1}\sum_{a=1}^{q_1}\Phi\left(u_a^{(1)}\right) \tag{8}$$

$$g_2 = \frac{1}{q_1}\sum_{a=1}^{q_2}\Phi\left(u_a^{(2)}\right) \tag{9}$$

SVMs used for feature selection significantly improve the analytical process by efficiently identifying the most pertinent predictors, such as systolic blood pressure and BMI. This technique improves the accuracy and robustness of ensuing studies by methodically eliminating superfluous characteristics that could introduce noise into the data. By prioritizing only the salient features, an SVM makes a more targeted and valuable approach to predictive modeling possible.

D. Multi-Layer Perceptron Neural Network (MLPNN) method

An MLPNN is a class of artificial NN models that have been used frequently to solve several classification problems, such as analyzing the risk of maternal health. There is the input layer, one or more hidden layers, and the output layer, which are created from individual neurons corresponding to computers connecting each neuron of one layer to the neurons of the next layer. Where it serves as the model for an MLPNN in this case, its main aim is to map the dataset and reveal the relationships between the variables in order to determine and categorize maternal health risks as those of low, medium, or high risk. In MLPNN, endogenous factors, including maternal age, blood pressure, and glucose level, are utilized with supervised learning so that input features can be aligned to distinctive output classes. These neurons take inputs, multiply them by weights, add bias, and apply an activation function so as to help the network identify complicated nonlinearity existing in the data. The learned knowledge of the network is by the use of the backpropagation algorithm, which works by minimizing a loss function in order to get an optimum set of weights and biases in the network. In the case of maternal health risk classification, MLPNN has some advantages because it can handle a variety of inputs, and data can be of different natures and come in large sets. The additions of the hidden layers help the network to learn the implicit features that might not be obvious, such as the interaction between the medical parameters. Moreover, activation of functions such as ReLU, sigmoid, or softmax, which can be adapted to suit the model, also makes it suitable for addressing multi-classification type problems, thus ensuring high accuracy of the model. Experiments have shown that selecting the correct number of hidden layers, the size of each hidden layer or neuron, and the rate of learning for the neural network greatly enhances the accuracy of the MLPNN to provide reliable performance for important healthcare applications. In equation 10 we perform Forward propagation which is also used to denote the process by which the input generated and passes through the layers of the network to generate an output.

$$g_y^{(p)} = \sum_{x=1}^{n^{(p-1)}} u_{xy}^{(p)} c_x^{(p-1)} + b_y^{(p)} \tag{10}$$

Let assume, g as weighted sum of inputs, y as neuron, p as layer, x, y as inputs, $c_x^{(p-1)}$ as combined inputs, u as weight, $(p - 1)$ as previous layer, and b as bias term. This equation calculates the pre-activation value for each neuron. For maternal health data, the inputs could be features such as maternal age, blood pressure, glucose levels, and other health metrics. By following in equation 11 we perform activation function,

$$c_y^{(p)} = \frac{1}{1+e^{-g_y^{(p)}}} \tag{11}$$

Once the pre-activation value is computed, it is passed through an activation function to introduce non-linearity. The sigmoid function maps to a value between 0 and 1, making it suitable for binary or probabilistic outputs. It is used to processes maternal health features (e.g., age, blood pressure, glucose). In equation 12 we perform ReLU activation function,

$$c_y^{(p)} = \max\left(0, g_y^{(p)}\right) \quad (12)$$

In this equation 13 it is used to sets all negative values to 0, while retaining positive values, aiding in efficient gradient propagation. It is used to applies weighted sums, biases, and activation functions to learn non-linear relationships. Next, we perform Softmax function for multi-class classification,

$$c_y^{(P)} = \frac{e^{g_y^{(P)}}}{\sum_{k=1}^{D} e^{g_k^{(P)}}} \quad (13)$$

Let assume, P as last layer, and D as classes. This equation normalizes the outputs of the last layer to probabilities across classes, such as low, medium, and high maternal health risks. It is used to predict probabilities for risk classes. After perform the activation function, we perform loss function $\mathcal{L}$ through equation 14 to improve its predictions.

$$\mathcal{L} = -\frac{1}{N}\sum_{x=1}^{N}\sum_{y=1}^{N} j_{xy}\log\left(c_{xy}^{(P)}\right) \quad (14)$$

Let assume, $\mathcal{L}$ as average cross-entropy loss, N as samples, j_{xy} as true label, (where 1 if class y is correct label, 0 otherwise), and $c_{xy}^{(P)}$ as predicted probability of class. This equation tends to punish the model heavily if the probability for the correct class is low, ensuring the network improves performance. In turning that to zero in order to minimize the loss function, the MLPNN uses the backpropagation algorithm, which computes the gradients of the loss with respect to the weights and the bias terms. The key equations include, in equation 15 we perform gradient of loss for output activation,

$$\delta_y^{(P)} = c_y^{(P)} - j_y \quad (15)$$

Let assume, δ as error term, y as neuron, and P as output layer. Then, we perform gradient of loss to u,

$$\frac{\partial\mathcal{L}}{\partial u_{xy}^{(p)}} = \delta_y^{(p)} c_x^{(p-1)} \quad (16)$$

Let assume, $\delta_y^{(p)}$ as error term, and $c_x^{(p-1)}$ as activation of the previous layer. This equation is used to compute the how the weights should be adjusted based on $\delta_y^{(p)}$ and $c_x^{(p-1)}$. By following equation 8 we perform gradient of loss to b ,

$$\frac{\partial\mathcal{L}}{\partial b_{xy}^{(p)}} = \delta_y^{(p)} \quad (17)$$

In this equation the gradient with respect to the biases depends solely on the error term $\delta_y^{(p)}$. At last, we perform error term for hidden layers through equation 18,

$$\delta_y^{(p)} = \left(\sum_{k=1}^{n^{(p+1)}} \delta_k^{(p+1)} u_{yk}^{(p+1)}\right) f'\left(g_y^{(l)}\right) \quad (18)$$

Let assume, $\delta_y^{(p)}$ as error term, y as neuron, p as layer, $(p+1)$ as propagated backward from the next layer, and $f'\left(g_y^{(l)}\right)$ as activation function. This equation computes prediction error using $\mathcal{L}$, and updates weights and biases to minimize the loss. If the real class for a record in maternal health is “medium risk,” the model predicts [0.1, 0.8, 0.1]. In that case, Cross-Entropy Loss discourages the wrong probabilities (for instance, if the model gives “high risk” instead of “medium risk”). For example, if the network assigns high blood pressure a low weight in the hidden layer, backpropagation re-adjusts this weight based on the contribution of this error to the output. By following in equation 19, we perform weight update rule through equation 10 by using an optimization algorithm,

$$u_{xy}^{(p)} \leftarrow u_{xy}^{(p)} - \eta\frac{\partial\mathcal{L}}{\partial u_{xy}^{(p)}}$$
$$b_y^{(p)} \leftarrow b_y^{(p)} - \eta\frac{\partial\mathcal{L}}{\partial b_y^{(p)}} \quad (19)$$

Let assume, $u_{xy}^{(p)}$ as weights, $b_y^{(p)}$ biases, and η as learning rate. In this equation, if maternal age consistently influences the "medium risk" class, the bias associated with this feature will be adjusted to improve predictions. By leveraging interconnected layers, weighted summations, activation functions, and optimization through backpropagation, the MLPNN captures complex, non-linear relationships within the dataset. Each equation in the MLPNN framework from forward propagation to loss calculation and weight updates plays a critical role in transforming raw maternal health features (e.g., blood pressure, glucose levels, age) into accurate predictions of risk levels (low, medium, or high). This method is particularly well-suited for maternal health data due to its ability to process diverse features, identify latent patterns, and generalize effectively to unseen data. By minimizing the classification error through iterative training, MLPNNs offer healthcare practitioners a reliable tool for early detection and intervention in high-risk maternal cases. This can lead to improved health outcomes for mothers and infants, highlighting the importance of advanced ML techniques in addressing real-world healthcare challenges.

IV. RESULTS AND DISCUSSION

This section presents the MLPNN algorithm, which identifies sensitive risk data using a health risk dataset sourced from Kaggle via a risk monitoring system. To enhance the accuracy of this algorithm, we employ KNN, BiGRU, and LSTM-RNN methods to evaluate the health risk dataset from the earlier stage against these classifier techniques. Furthermore, critical healthcare risk records can be extracted from large datasets based on performance measures, including precision, recall, F1 score, time complexity, and overall accuracy.

TABLE I. SIMULATION PARAMETER

Simulation	Variable
Dataset Name	Health Risk Data
No of Records	1015
Training	789
Testing	226
Tool	Jupyter
Language	Python

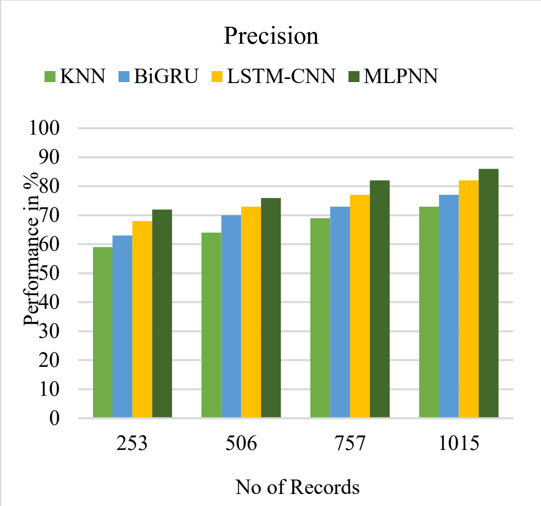

Fig. 3. Analysis of Precision

As shown in Figure 3, the health risk data obtained from the dataset can be used to predict the severity of risk in the health dataset, considering the previous attribute. Furthermore, the proposed method can be used to analyze the precision of performance measures in the sensitivity record to predict the high-risk factors in the health data compared to the previous KNN, BiGRU, and LSTM-CNN methods. In addition, the precision performance measure using the proposed MLPNN technique is shown to be 86%. Similarly, the precision performance measures are described in comparison with the previous KNN, BiGRU, and LSTM-CNN methods (73%, 77%, and 82%) alongside the presented technique.

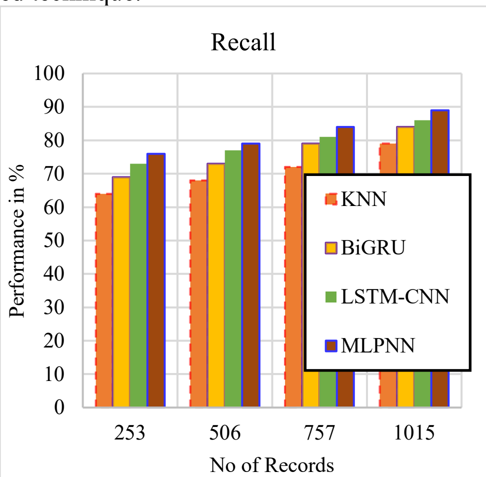

Fig. 4. Analysis of Recall

Figure 4 illustrates that the health risk data from the dataset is utilized to predict the severity of risks based on previous attributes. Additionally, in comparison to earlier KNN, BiGRU, and LSTM-CNN methods, the proposed approach effectively analyzes the recall of sensitive performance measures and predicts high-risk factors in healthcare data. The recall performance measure achieved with the proposed MLPNN technique is reported at 89%. Likewise, the recall performance metrics with this method were compared to those of the previous KNN, BiGRU, and LSTM-CNN techniques, yielding rates of 79%, 84%, and 86%, respectively.

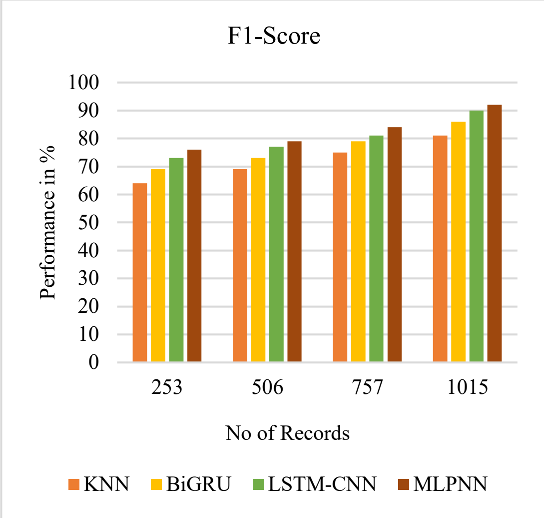

Fig. 5. Analysis of F1-Score

Figure 5 demonstrates that the dataset's health risk information is used to forecast risk severity based on past characteristics. Furthermore, the suggested method identifies high-risk indicators in healthcare data and efficiently analyses the F1-Score of sensitive performance measures when compared to previous KNN, BiGRU, and LSTM-CNN methods. The suggested MLPNN achieved an F1-Score performance metric of 92% algorithm. Comparing this method's F1-Score performance metrics to those of the earlier KNN, BiGRU, and LSTM-CNN approaches, the results showed that the previous method produced results of 81%, 86%, and 90%, respectively.

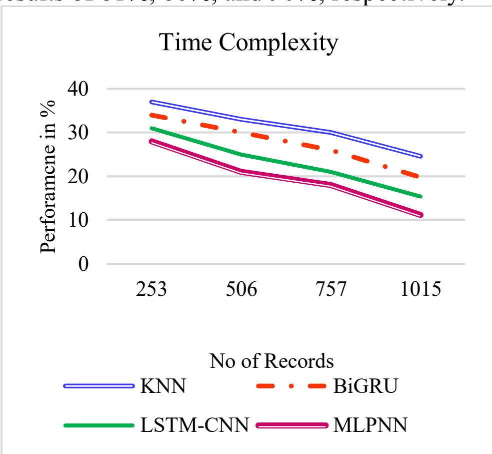

Fig. 6. Time Complexity

Figure 6 indicates the health risk data in the dataset used to predict the risk severity based on the preceding attributes. Furthermore, compared to the previous KNN, BiGRU, and LSTM-CNN methods, the method effectively analyzes the time complexity of the sensitivity performance metrics and predicts high-risk factors in clinical data. The time complexity performance measured using the proposed MLPNN method is reported as 11.2ms. Therefore, we compared the time complexity performance metrics of this method with the previous KNN, BiGRU, and LSTM-CNN methods, and the results were 24.6 ms, 19.8 ms, and 15.4 ms, respectively.

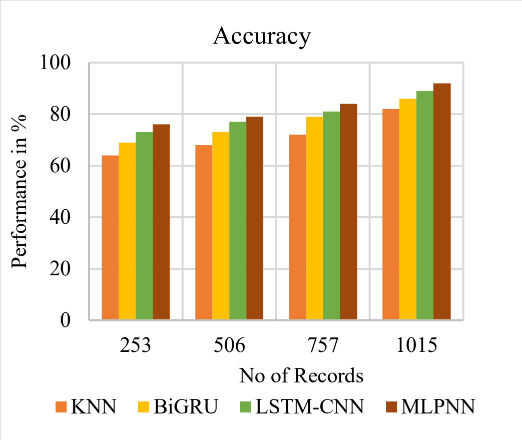

Fig. 7. Analysis of Accuracy

Figure 7 shows the health risk data in the dataset used to predict the risk severity based on the above attributes. Also, compared to the previous KNN, BiGRU, and LSTM-CNN methods, our method effectively analyzes the accuracy of the sensitivity performance metrics and predicts high-risk factors in clinical data. Therefore, we compared the accuracy performance metrics of the proposed method with the previous KNN, BiGRU, and LSTM-CNN methods, and the results were 82%, 86%, and 92%, respectively. The accuracy performance measured using the proposed MLPNN method is reported as 96.4%. Table 2 describes the comparison performance.

Table II. Comparison performance

Methods /Parameters	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
KNN	73	79	82	81
BiGRU	77	84	86	86
LSTM-CNN	82	86	92	90
MLPNN	86	89	96.4	92

A. Discussion

The findings of the current research indicate that MLPNN is more effective compared to the conventional techniques such as KNN, BiGRU, and LSTM-CNN in forecasting the level of health risk. MLPNN performed optimally on these measures, and its precision, recall, and accuracy were 86%, 89% and 96.4% which were higher in comparison to the rest of the models. Moreover, the MLPNN was also found to be less time-complex (11.2ms) compared to the various methods and hence can be used in real-time healthcare applications. The F1-score of 92% also reiterates the fact that MLPNN is efficient at balancing precision and recall. Such findings indicate that MLPNN is a more dependable, effective, and precise way of forecasting health risks, which would lead to a more promising direction of clinical decision support systems that would help to enhance patient outcomes and better distribution of healthcare resources.

V. CONCLUSION

In summary, we improved big data analysis for healthcare data analysis by employing the MLPNN technique. In this method, we use two phases: feature selection in the preparatory stage preprocessing and classification in the subsequent real-time processing. In the first phase, we utilized the ZSN, SVM method, which explores the most excellent influence features such as blood pressure, glucose level, and maternal age and excludes noise or less important information from the data input. This step reduces the dimensionality of the model and fine-tunes the interpretative capacity of the model as well as the model's efficacy. Second, the data set is passed through MLPNN; because of its multilayer structure, the network is able to identify the relationships of the features to classify the maternal health risk into low, medium, or high risk. This has been driven to a 96.4% accuracy by the MLPNN using the Maternal Health Risk Data dataset. Consequently, this approach can be leveraged to bring changes and enhancements when tested and implemented, enabling healthcare professionals and patients' outcomes, as well as promoting early interventions. Future work will use the Explainable AI (XAI) techniques could also be explored to provide more transparency into the model's decision-making process, which is crucial in healthcare settings to ensure trust and accountability. Finally, further validation of the model using real-world clinical datasets is necessary to assess its performance in diverse healthcare environments and on a broader patient population, ensuring its scalability and practical implementation in clinical decision support systems.

REFERENCES

[1] Zheng, L., Wang, O., Hao, S., Ye, C., Liu, M., Xia, M., Sabo, A. N., Markovic, L., Stearns, F., Kanov, L., Sylvester, K. G., Widen, E., McElhinney, D. B., Zhang, W., Liao, J., & Ling, X. B. (2020). Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. Translational Psychiatry, 10(1), 1-10. https://doi.org/10.1038/s41398-020-0684-2

[2] H. Sun, Z. Liu, G. Wang, W. Lian and J. Ma, "Intelligent Analysis of Medical Big Data Based on Deep Learning," in IEEE Access, vol. 7, pp. 142022-142037, 2019, doi: 10.1109/ACCESS.2019.2942937.

[3] Yang, Y., Li, Y., Chen, R., Zheng, J., Cai, Y., & Fortino, G. (2021). Risk Prediction of Renal Failure for Chronic Disease Population Based on Electronic Health Record Big Data. Big Data Research, 25, 100234. https://doi.org/10.1016/j.bdr.2021.100234

[4] Razzak, M.I., Imran, M. & Xu, G. Big data analytics for preventive medicine. Neural Comput & Applic 32, 4417–4451 (2020). https://doi.org/10.1007/s00521-019-04095-y

[5] Krishnamaneni, Ramesh, Ashwin Narasimha Murthy, and Souptik Sen. "A comparative study of big data mining algorithms for early detection of heart attack risk factors in electronic medical records." International Journal of Computer Engineering and Technology (IJCET) 10, no. 6 (2019): 139-154.

[6] Potla, Ravi Teja. "Hybrid Deep Learning Models for Big Data: A Case Study in Predictive Healthcare Analytics." Distributed Learning and Broad Applications in Scientific Research 10 (2024): 319-325.

[7] Wang, Z., Chen, X., Tan, X., Yang, L., Kannapur, K., Vincent, J. L., Kessler, G. N., Ru, B., & Yang, M. (2021). Using Deep Learning to Identify High-Risk Patients with Heart Failure with Reduced Ejection Fraction. Journal of Health Economics and Outcomes Research, 8(2), 6. https://doi.org/10.36469/jheor.2021.25753

[8] Swinckels L, Bennis FC, Ziesemer KA, Scheerman JFM, Bijwaard H, de Keijzer A, Bruers JJ, The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review, J Med Internet Res 2024;26:e48320, doi: 10.2196/48320

[9] Corny, J., Rajkumar, A., Martin, O., Dode, X., Lajonchère, J., Billuart, O., Bézie, Y., & Buronfosse, A. (2020). A machine learning–based clinical decision support system to identify prescriptions with a high risk of medication error. Journal of the American Medical Informatics Association, 27(11), 1688-1694. https://doi.org/10.1093/jamia/ocaa154

[10] Kong, G., Lin, K. & Hu, Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Med Inform Decis Mak 20, 251 (2020). https://doi.org/10.1186/s12911-020-01271-2

[11] Zhihan Lv, Zengchen Yu, Shuxuan Xie, and Atif Alamri. 2022. Deep Learning-based Smart Predictive Evaluation for Interactive Multimedia-enabled Smart Healthcare. ACM Trans. Multimedia Comput. Commun. Appl. 18, 1s, Article 43 (February 2022), 20 pages. https://doi.org/10.1145/3468506

[12] Almustafa, K.M. Prediction of heart disease and classifiers' sensitivity analysis. BMC Bioinformatics 21, 278 (2020). https://doi.org/10.1186/s12859-020-03626-y

[13] Xu, J., Glicksberg, B.S., Su, C. et al. Federated Learning for Healthcare Informatics. J Healthc Inform Res 5, 1–19 (2021). https://doi.org/10.1007/s41666-020-00082-4

[14] Yang, C.C. Explainable Artificial Intelligence for Predictive Modeling in Healthcare. J Healthc Inform Res 6, 228–239 (2022). https://doi.org/10.1007/s41666-022-00114-1

[15] Shukla, Prashant Kumar, et al. "Optimization assisted bidirectional gated recurrent unit for healthcare monitoring system in big-data." Applied Soft Computing 138 (2023): 110178.

[16] Kumar, R.H., Sunitha, G. Big data analytics in healthcare environment using chaotic red deer optimizer with deep learning for disease classification model. Multimed Tools Appl 83, 77697–77715, (2024). https://doi.org/10.1007/s11042-024-18239-3.

[17] Alqahtani, T.M. (2023). Big data analytics with optimal deep learning model for medical image classification. Computer Systems Science and Engineering, 44(2), 1433-1449. https://doi.org/10.32604/csse.2023.025594.

[18] Qiu, Y., & Lu, J. (2021). A visualization algorithm for medical big data based on deep learning. Measurement, 183, 109808. https://doi.org/10.1016/j.measurement.2021.1098.

[19] Sidiq, S. Jahangeer, et al. "Big Data and deep learning in healthcare." Applications of Artificial Intelligence, Big Data and Internet of Things in Sustainable Development. CRC Press, 2022. 145-160.

[20] Lakshmanaprabu, S.K., Shankar, K., Ilayaraja, M. et al. Random forest for big data classification in the internet of things using optimal features. Int. J. Mach. Learn. & Cyber. 10, 2609–2618 (2019). https://doi.org/10.1007/s13042-018-00916-z

[21] Bebortta, Sujit, et al. "DeepMist: Toward deep learning assisted mist computing framework for managing healthcare big data." IEEE Access 11 (2023): 42485-42496.

[22] M. H. A. Banna et al., "A Hybrid Deep Learning Model to Predict the Impact of COVID-19 on Mental Health From Social Media Big Data," in IEEE Access, vol. 11, pp. 77009-77022, 2023, doi: 10.1109/ACCESS.2023.3293857.

[23] W. -J. Chang, L. -B. Chen, C. -H. Hsu, C. -P. Lin and T. -C. Yang, "A Deep Learning-Based Intelligent Medicine Recognition System for Chronic Patients," in IEEE Access, vol. 7, pp. 44441-44458, 2019, doi: 10.1109/ACCESS.2019.2908843.

[24] Li, N., Shepperd, M., & Guo, Y. (2020). A systematic review of unsupervised learning techniques for software defect prediction. Information and Software Technology, 122, 106287. https://doi.org/10.1016/j.infsof.2020.106287

[25] Liu, Y., Yu, Z., & Yang, Y. (2020). [Retracted] Diabetes Risk Data Mining Method Based on Electronic Medical Record Analysis. Journal of Healthcare Engineering, 2021(1), 6678526. https://doi.org/10.1155/2021/6678526.