A Deep Learning Model for Decision Making in Healthcare Systems

N.S. L Kumar Kurumeti[1*], Dr K.Gangadar Rao[2]

[1*]Research Scholar, Department of CSE, Acharya Nagarjuna University,
[2]Principal, University College of Science, Acharya Nagarjuna University, HOD, Department of CSE, Acharya Nagarjuna University, Guntur, AP, India. kancherla123@gmail.com
Corresponding Author: N.S. L Kumar Kurumeti
*Email: kurumeti.kumar@gmail.com.

Abstract:
Decision Making (DM) is one of the domains that provide decisions based on improvement in business without the involvement of humans. This is mainly focused on giving suggestions to various companies to increase their online product, disease prediction, and also in many applications. Decision-making is used in many research applications such as the medical domain, online marketing, E-commerce, Filtering of Email, and other types of domains that help experts to get better decisions. It is identified that machine learning (ML) algorithms have several disadvantages in finding accurate decisions on various applications. Huge research is done on ML algorithms to find accurate patterns in decision-making algorithms to take automated decisions. ML algorithms are very weak in the processing of large datasets and also in statistical analysis. In this domain, deep learning (DL) plays a significant role in finding accurate patterns in detecting decision-making in multiple domains. In this paper, a deep learning model is introduced to find the disease patterns from the medical data to take the correct decisions and also to predict the risk based on the status of the disease. The proposed deep learning model is the combination of several models. Thus the decision-making is applied to know the status of the disease. Experiments are conducted on Heart-2 Dataset, Haberman's Survival dataset, and Pima Indians Diabetes datasets and show the comparative performance.

Keywords: Decision making, machine learning, deep learning, risk prediction.

Introduction

Decision-making system (DMS) is a domain which is used helps experts to take accurate decision by analyzing present data and analytical data by utilizing real-time applications and external sources. Merging a large amount of data gives enlightened analytical models and tools that make the system easy to access, these provide better online sources of data to utilize in taking accurate decisions making. DSS is the class of intelligent systems that supports decision-making activities. Several domains are used in the DMS to communicate with several components such as information, documents, knowledge, and techniques that are more successful to complete the decision-making tasks.

Recently, people have been able to see the bright future of artificial intelligence (AI) in the healthcare industry thanks to the development of machine intelligence and AI. Promising outcomes have been observed in early illness prediction with deep learning technologies. The advancement of machine learning and deep learning technologies has significantly enhanced the performance of multi-disease prediction. Predicting diseases accurately and early on is the most challenging endeavour. It seeks to illustrate the important connection— primarily for early disease prediction—between deep learning and the healthcare sector.

Medical image classification and segmentation is a common use of deep learning techniques, with a focus on convolutional neural networks (CNN). I'll give you a few noteworthy recent examples. [1] provides a thorough explanation of the ways in which computer-aided diagnosis can be used to enhance clinical treatment. [2] illustrates how glaucoma can be identified by classifying the photos that have been processed. A CNN-based technique for classifying lesions in a hemorrhagic stroke is demonstrated in [3]. In [4], a CNN-based automated method for lumbar spine disc detection is suggested. In [5], a CNN-based approach is put out to enhance Papanicolaou smear classification. In [5], an automatic model known as online transfer learning is put forth for the purpose of using ultrasound pictures to differentiate between benign and malignant thyroid nodules. There is no mention of using CNN in reference [6].

Literature Survey

To address the feature selection problem, J. P. Li et al. [2] introduced a unique fast conditional mutual information feature selection technique (FCMIM). In order to improve classification accuracy and shorten the time it takes for the classification system to execute, features selection techniques are employed. Additionally, the hyperparameter tuning and best practices for model assessment have been learned through the application of the leave one subject out cross-validation method. Support vector machines (SVMs) are used to boost the accuracy and rate of disease identification in order to enhance the performance of the suggested method. In order to detect heart failure in its early stages, G. G. N. Geweid et al. [3] developed a hybrid strategy combining dual SVM with nonparametric algorithm. This approach increased the accuracy and reliability of heart failure diagnosis and identification. Two SVM models are obtained by training SVM and its dual using the nonparametric approach. Sometimes the dual problem is easier to understand than the original since it

presents an alternative viewpoint. This feature compares the outputs of the SVM model with the dual SVM model in order to identify heart failure disease in ECG signals. An reliable hybrid feature selection method was proposed by R. Alzubi et al. [4] for identifying the most informative SNPs and choosing the best SNP subset. The foundation of the suggested approach is the combination of two methods: the SVM recursive feature elimination and the Conditional Mutual Information Maximisation (CMIM) method. A prediction method based on machine learning was presented by Haq et al. [5]. The support vector machine (SVM) was employed as a predictive model for the prediction of Parkinson's disease (PD) in the construction of the suggested system. For accurate target categorization of people with Parkinson's disease (PD) and healthy individuals, the L1-norm SVM of features selection was employed to choose relevant and highly connected characteristics. Based on a feature weight value, the L1-norm SVM created a new subset of features from the PD dataset. The K-fold cross-validation approach was applied to validate the suggested system. For the purpose of evaluating the performance of the model, additional metrics were generated, including execution time, F1 score, accuracy, sensitivity, specificity, and precision. This paper included the PD dataset. The best subset of the chosen features was obtained with the ideal accuracy, which may have been caused by different contributions from the PD characteristics. Using Decision Trees (DT) and several deep learning models, a contemporary generic strategy has been developed for the identification and classification of wheat illnesses by W. Haider et al. [6]. Additionally, once the outputs of both algorithms were validated by subject matter experts, the accuracy of the CNN increased by 4.3%, improving the decision trees' accuracy by 28.5%, and the outcome was decision rules for wheat illnesses in a knowledge-based system. A unified convolutional neural network (CNN) representing the matrix-based convolutional neural network (M-bCNN) was proposed by Z. Lin et al. [7]. Its distinguishing feature is the convolutional kernel matrix, which is made up of parallel convolutional layers merged with DropConnect, an exponential linear unit, local response normalisation, and other techniques to overcome vanishing gradient and over-fitting. Compared with the widely used plain networks, it may effectively increase the data streams, neurons, and connection channels of the model with a bearable addition of parameters. As a result, it will increase the number of non-linear mappings and improve representational ability while maintaining a manageable parameter growth. Because the subcategories of winter wheat leaf diseases share a lot of characteristics, these photos were used as experimental samples. In order to effectively depict the texture and form aspects of plant diseases in an image-based leaf monitoring system applied in a mobile-cloud context, P. Siricharoen et al. [8] presented a feature extraction and selection process. Along with the suggested and improved shape features based on diseased regions, we also add mean and peak indices of histogram-of-shape as disease property representations. A variety of textural and region-based features are collected from prior works. To discover the best possible features based on many factors, including the chance of feature error, correlation, targeted-class relevancy, and the separability quality of a feature, a total of 260 color-based characteristics and 163 shape attributes are searched. The automatic detection and diagnosis of illnesses affecting maize leaves, as presented by X. Zhang et al. [9], is much appreciated. The upgraded deep learning-based GoogLeNet and Cifar10 models are suggested for leaf disease recognition in this paper, with the goal of reducing the number of network parameters and increasing the identification accuracy of maize leaf diseases. By varying the pooling combinations, adding dropout operations and rectified linear unit functions, decreasing the number of classifiers, and modifying the parameters, two improved models are generated that are used to train and test nine different types of maize leaf images. Furthermore, compared to the VGG and AlexNet structures, the enhanced models have a substantially lower number of parameters. The GoogLeNet model has a top - 1 average identification accuracy of 98.9% during the recognition of eight types of maize leaf diseases, while the Cifar10 model achieves an average accuracy of 98.8%. The enhanced techniques may increase maize leaf disease accuracy and decrease convergence iterations, both of which can boost the effectiveness of model training and recognition. A innovative approach to improve the accuracy of cardiovascular disease prediction by employing machine learning algorithms to identify key traits was developed by S. Mohan et al. [10]. Several feature combinations and well-known categorization methods are used to introduce the prediction model. Using the hybrid random forest with a linear model (HRFLM) as our prediction model for heart disease, we achieve an improved performance level with an accuracy level of 88.7%. A machine intelligence framework for the diagnosis of heart disease, called MIFH, was created by A. Gupta et al. [11]. In order to extract and identify features from the UCI heart disease Cleveland dataset and train machine learning predictive models, MIFH uses factor analysis of mixed data (FAMD). The holdout validation approach is used to validate the MIFH framework. A hybrid grid search algorithm (HGSA) that can simultaneously optimise the two models was proposed by L. Ali et al. [12]. Six assessment metrics are used to assess the efficacy of the suggested method: area under the curve (AUC), ROC charts, accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC). The experimental findings verify that the suggested approach enhances a traditional SVM model's performance by 3.3%. Furthermore, the suggested approach performs better than the ten previously suggested approaches, which produced accuracy values between 57.85% and 91.83%. A disease prediction model (DPM) based on an individual's risk factors data is proposed by N. L. Fitriyani et al. [13] to provide an early prognosis for type 2 diabetes and hypertension. The suggested DPM consists of an ensemble methodology to predict diseases, a synthetic minority oversampling technique tomek link (SMOTETomek) to balance data distribution, and an isolation forest (iForest) based outlier identification method to eliminate outlier data. To enable the practical application of the suggested DPM, a mobile application

is built in conjunction with the suggested methodology. The created mobile application collects risk factor information and transmits it to a distant server, enabling the suggested DPM to diagnose a person's current state. Once unforeseen health issues (such as type 2 diabetes and/or hypertension) emerge at an early stage, individuals can take appropriate and prompt action to reduce and prevent their risks. This is accomplished by sending the prediction result back to the mobile application. A radial basis function neural network (RBFN)-based approach to predict eight cardiac arrhythmias was proposed by J. P. Kelwade et al. [14]. The heart rate time series used in this method are taken from the MIT-BIH arrhythmia database. Every arrhythmia's heart rate time series is used to identify its linear and nonlinear characteristics. Thirty percent of each feature dataset is utilised to forecast eight heart illnesses, with the remaining seventy percent being used to train RBFN. By using this method, the overall prediction accuracy is 96.33%, which is higher than the approaches found in the literature. When the full connection model presented by Xiaoting Yin et al. [15] is compared to the CNN training data, the accuracy and generalizability of the CNN model are demonstrated. The effectiveness benefits of the CNN model across several product categories are demonstrated by using the non-deep learning model as the baseline for comparison.

An intelligent learning model

Every day, the healthcare sector must make a wide range of difficult decisions, from allocating resources and developing policies to diagnosing patients and designing their treatments. Intelligent learning models must be used since these decisions are complex and high-stakes. They can supplement human judgement, increase decision accuracy, and boost overall healthcare results. Intelligent learning models analyse enormous volumes of healthcare data by utilising developments in machine learning (ML), artificial intelligence (AI), and data analytics. These models are intended to help healthcare practitioners make well-informed decisions by seeing trends, forecasting results, and offering useful insights. The delivery of healthcare might be completely transformed by incorporating these technologies into systems, making it more effective, efficient, and personalised.

Data from many sources, such as electronic health records (EHRs), lab findings, medical imaging, and patient-generated health data, are poured into healthcare systems. Healthcare professionals find it difficult to manually extract relevant insights from this data due to its great volume and complexity. Real-time processing and analysis of this data by intelligent learning models can reveal hidden patterns and correlations that human analysis might overlook. Making decisions is also made more difficult by the variety of diseases and the variation in how patients react to treatments. Conventional methods frequently depend on broad recommendations that could not take unique patient characteristics into consideration. Patient-specific data can be incorporated into intelligent learning models, enabling more focused and customised therapies that enhance patient outcomes.

Preprocessing of the Dataset

In this step, the noise was removed by using various algorithms. However, many outliers must be treated carefully, and the dataset's distribution must be corrected. There were two methods employed. Unimpressive results were obtained when data was applied straight to deep learning algorithms without feature selection or outliers. However, the outcomes are pretty encouraging once the overfitting issue is resolved by employing the dataset's normal distribution, and the outlier is then detected using Isolation Forest. Various charting methods were employed to assess the data's distribution, skewness, and outlier detection. All of these preprocessing methods are crucial when transferring the data for classification or prediction.

Normal Distribution

The majority of data points in a normal distribution cluster at the middle of the range, while the remaining data points taper off symmetrically towards either extreme. A normal distribution is a sort of continuous probability distribution. The distribution's mean is another name for the midpoint of the range. A Gaussian distribution or probability bell curve are other names for the normal distribution. It shows that values close to the mean occur more frequently than those further from the mean and is symmetric about the mean.

The formula for the normal distribution is expressed below.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Here, x is value of the variable; f(x) represents the probability density function; μ (mu) is the mean; and σ (sigma) is the standard deviation.

Radial basis function networks (RBFNs)

Artificial neural networks have particular examples, such as radial basis function networks (RBFNs). Artificial neural networks (ANNNs) are modelled after biological systems insofar as they replicate the internal organisation of the brain, complete with neurons and their connections. The system's input comes from external stimuli (i.e. the network). Weighting the connections between the nodes (i.e., the neurons) inside the network controls how the signals spread. Training is the procedure that modifies the weights of artificial neural networks. It is possible for the training process to be automatic (also known as unsupervised training or training without

a teacher) or to require teacher reaction (also known as supervised training or training with a teacher). In that each neuron in an RBFN consists of a radial basis function (RBF), there is only one hidden layer of neurons, one output node, and a two-stage training technique available, RBFNs are special examples of artificial neural networks.

An RBF $\emptyset_c$ is a function with a symmetric output around a centre μ_c, i.e.

$$\emptyset_c(\bar{x}) = \emptyset(\| \bar{x} - \overline{\mu_c} \|)$$

Where || || denotes the vector norm.
One symmetric function that is often used as an RBF is the Gaussian function

$$\emptyset(r) = e^{-\frac{r^2}{\sigma^2}}$$

With the Euclidean norm. As is well-known, Gaussian functions are characterized by a width or scale parameter σ. RBFs are also in many cases dependent on a scale parameter called the width of the function.

Dataset Description

The proposed approach is applied on three different disease datasets collected from various sources.

The first dataset, the Heart 2 dataset, has two classes, 303 instances, and 14 attributes. Classes describe the patient's cardiac disease status. They range in integer value from 0 (no heart problems) to 4. The Kaggle repository is where this dataset was obtained. The Haberman's Survival dataset is the second one. This dataset's data is used to assess whether a patient had a positive five-year survival rate from breast cancer or a negative five-year survival rate due to death. The data originates from a 1958–1970 study carried out at the University of Chicago Billings Hospital. There are two classes, thirty-six instances, and three characteristics. The dataset used by Haberman was obtained from the Keel repository. The Pima Indians Diabetes dataset includes data that can be used to categorize or forecast whether or not women of Pima Indian heritage who are younger than 21 have diabetes, based on whether they tested positive or negative. This data collection has two classes, 768 instances, and eight properties that were pulled from the Keel repository.

Experimental Results

The confusion matrix is used to analyze the model's performance. This will detail how well the categorization models perform with the provided test data. This will list the known values for the test data. This matrix has an overall number of predictions and two properties, such as original and forecasted values.

1936
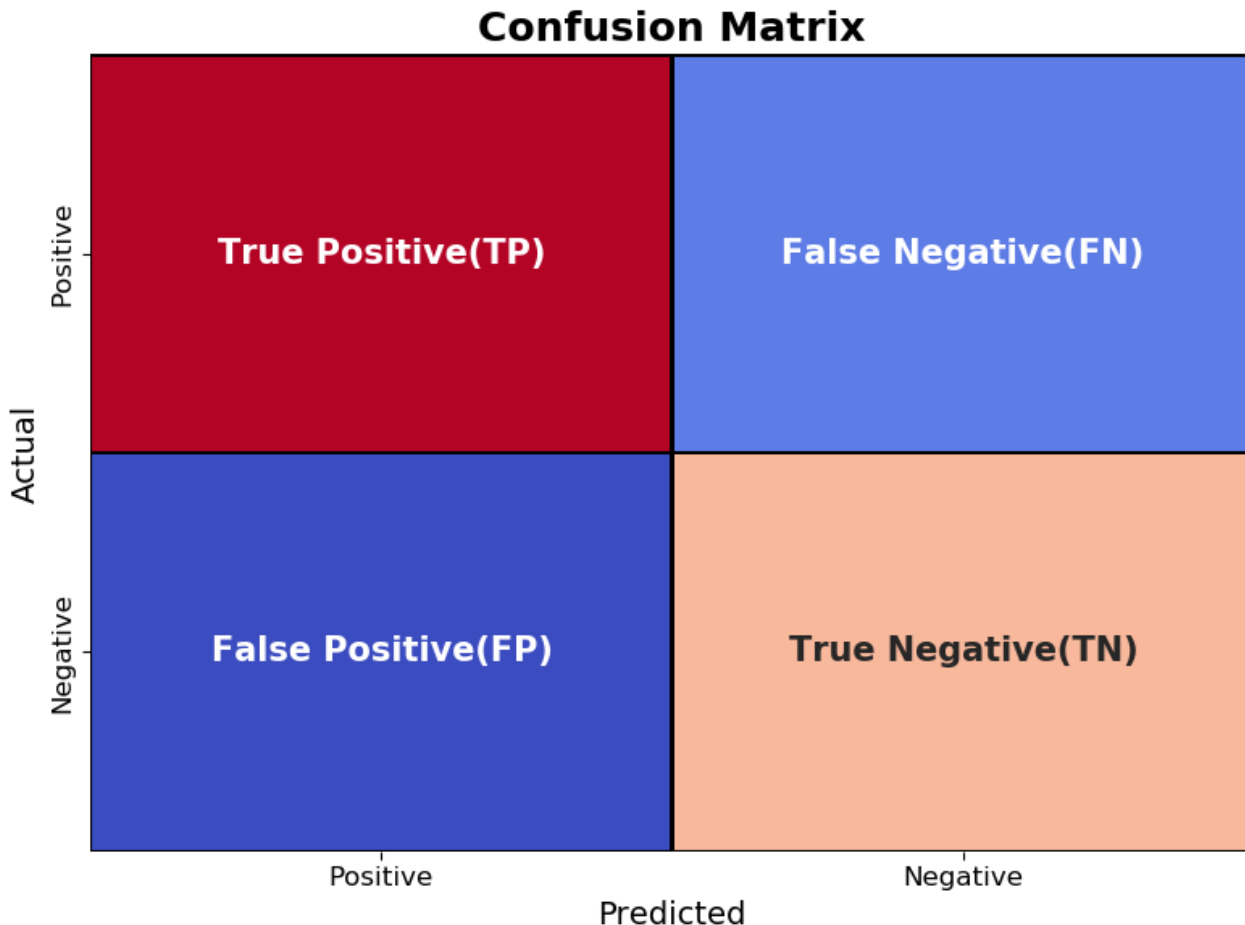
$$Precision = \frac{TP}{TP + FP}$$

$$F1\ Measure = 2 \times \frac{precision * recall}{precision + recall}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{No.\ of\ TP + No.\ of\ FN}$$

Table 1: Datasets and description

Dataset	Instances	Attributes	Online Source
Heart-2 Dataset	303	14	Kaggle
Haberman's Survival dataset	306	3	Keel repository
Pima Indians Diabetes dataset	768	8	Kaggle

Table 2: Comparative performance between Existing and Proposed Algorithms for Heart 2 dataset

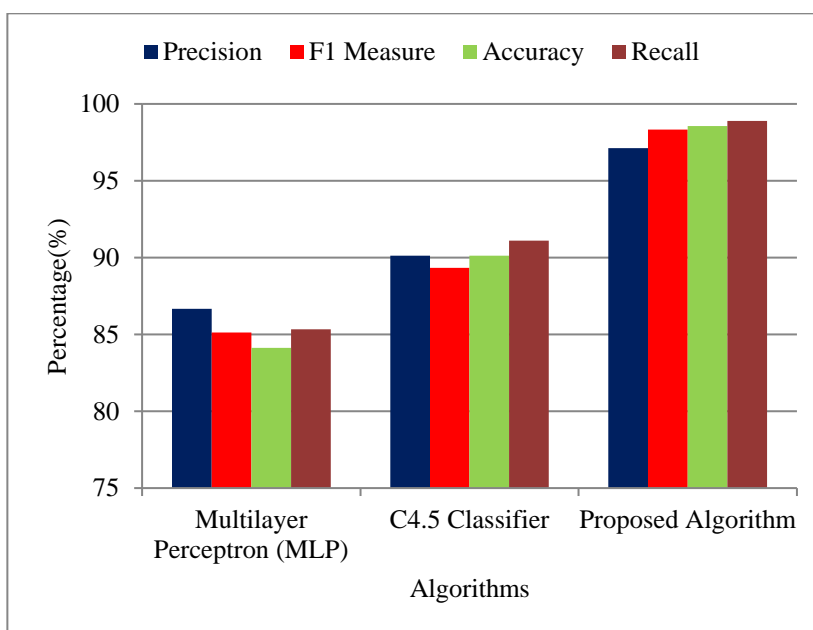| Algorithm	Precision	F1 Measure	Accuracy	Recall
Multilayer Perceptron (MLP)	86.67	85.12	84.12	85.34
C4.5 Classifier	90.12	89.34	90.12	91.1
Proposed Algorithm	97.12	98.34	98.56	98.89

Figure 2: Comparative performance between Existing and Proposed Algorithms for Heart 2 dataset
Table 3: Comparative performance between Existing and Proposed Algorithms for Haberman's Survival dataset

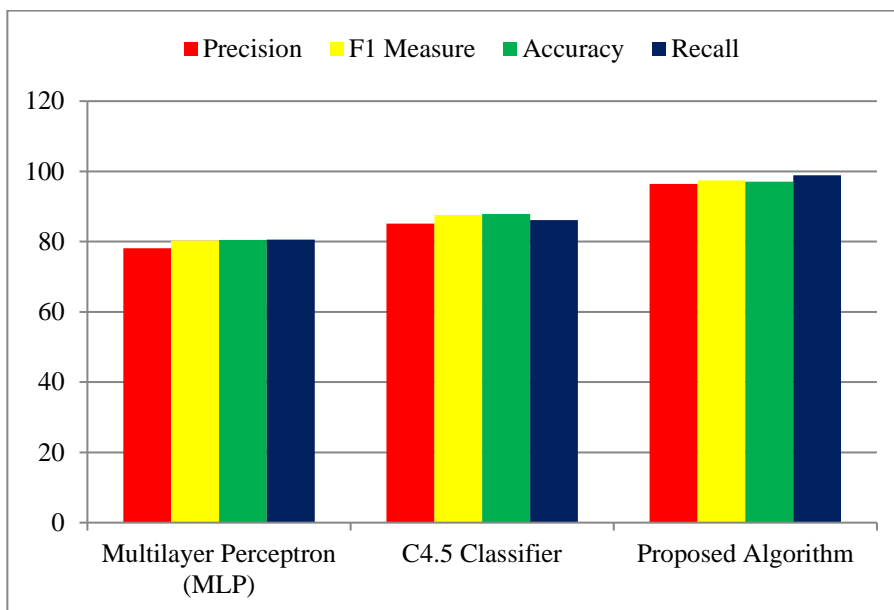| Algorithm	Precision	F1 Measure	Accuracy	Recall
Multilayer Perceptron (MLP)	78.12	80.34	80.45	80.56
C4.5 Classifier	85.12	87.56	87.9	86.12
Proposed Algorithm	96.45	97.45	97.12	98.89

Figure 3: Comparative performance between Existing and Proposed Algorithms for Haberman's Survival dataset

Table 4: Comparative performance between Existing and Proposed Algorithms for Pima Indians Diabetes dataset

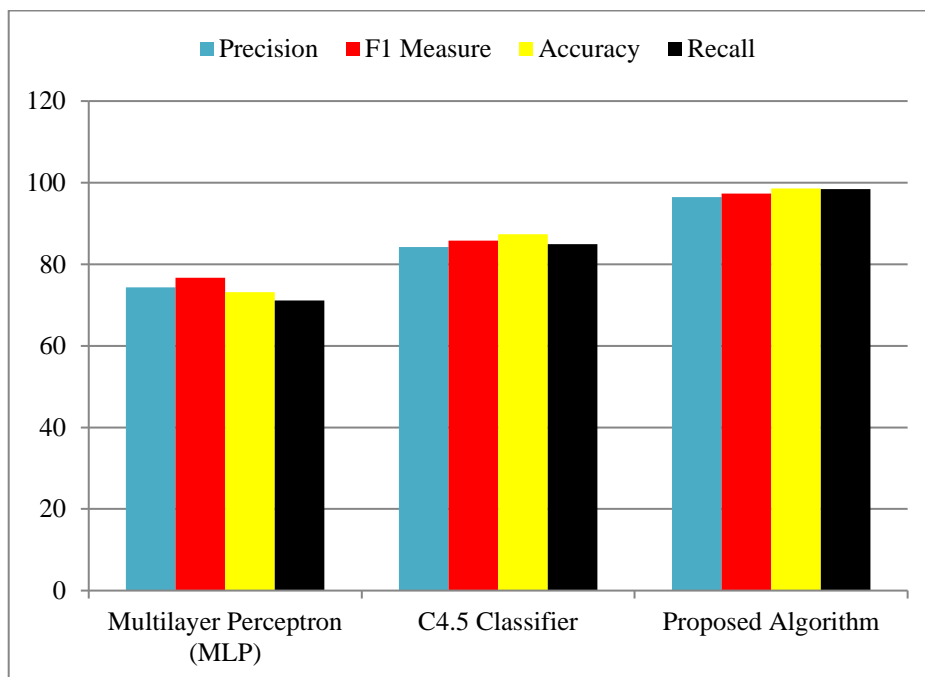| Algorithm	Precision	F1 Measure	Accuracy	Recall
Multilayer Perceptron (MLP)	74.34	76.67	73.12	71.09
C4.5 Classifier	84.23	85.78	87.34	84.89
Proposed Algorithm	96.45	97.34	98.56	98.45

Figure 4: Comparative performance between Existing and Proposed Algorithms for Pima Indians Diabetes dataset

Conclusion

The proposed approach in this paper provides reliable results when compared to current models. Confusion matrix, precision, specificity, sensitivity, and F1 score are the methodologies that are compared. When data preprocessing is used, the proposed method classifier performed better in the machine learning technique for the 13 features that were present in the dataset. Reducing the computational time is beneficial when implementing a model. Additionally, it was discovered that the dataset needed to be normalized in order to prevent overfitting of the training model and to ensure that the accuracy obtained is adequate for evaluating the model on real-world data problems, which can differ significantly from the dataset used for training. When

we compared the accuracy with other researchers, it rose with the algorithm we used in the ANN design. More datasets can be used to enable deep learning with a variety of additional optimizations, potentially yielding more encouraging outcomes. To raise the evaluation findings once more, machine learning and a number of other optimization strategies might be applied. The data can be normalized in several ways, and the outcomes can be contrasted. Additionally, there may be more ways to combine specific multimedia with ML and DL models trained on cardiac diseases for the convenience of physicians and patients.

References

[1] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[2] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[3] G. G. N. Geweid and M. A. Abdallah, "A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique," IEEE Access, vol. 7, pp. 149595–149611, 2019.

[4] R. Alzubi, N. Ramzan, H. Alzoubi, and A. Amira, "A hybrid feature selection method for complex diseases SNPs," IEEE Access, vol. 6, pp. 1292–1301, 2018.

[5] A. U. Haq, J. P. Li, M. H. Memon, J. Khan, A. Malik, T. Ahmad, A. Ali, S. Nazir, I. Ahad, and M. Shahid, "Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings," IEEE Access, vol. 7, pp. 37718–37734, 2019.

[6] W. Haider, A. -U. Rehman, N. M. Durrani and S. U. Rehman, "A Generic Approach for Wheat Disease Classification and Verification Using Expert Opinion for Knowledge-Based Decisions," in IEEE Access, vol. 9, pp. 31104-31129, 2021, doi: 10.1109/ACCESS.2021.3058582.

[7] Z. Lin, S. Mu, F. Huang, K. A. Mateen, M. Wang, W. Gao, and J. Jia, "A unified matrix-based convolutional neural network for fine-grained image classification of wheat leaf diseases," IEEE Access, vol. 7, pp. 11570–11590, 2019, doi: 10.1109/ACCESS.2019.2891739.

[8] P. Siricharoen, B. Scotney, P. Morrow, and G. Parr, "Texture and shape attribute selection for plant disease monitoring in a mobile cloud-based environment," in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2016, p. 489.

[9] X. Zhang, Y. Qiao, F. Meng, C. Fan, and M. Zhang, "Identification of maize leaf diseases using improved deep convolutional neural networks," IEEE Access, vol. 6, pp. 30370–30377, 2018.

[10] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019.

[11] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," IEEE Access, vol. 8, pp. 14659–14674, 2020, doi: 10.1109/ACCESS.2019. 2962755.

[12] L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," IEEE Access, vol. 7, pp. 54007–54014, 2019.

[13] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension," IEEE Access, vol. 7, pp. 144777–144789, 2019.

[14] J. P. Kelwade and S. S. Salankar, "Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series," in Proc. IEEE 1st Int. Conf. Control, Meas. Instrum. (CMI), Jan. 2016, pp. 454–458.

[15] Xiaoting Yin, Xiaosha Tao, "Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning", Scientific Programming, vol. 2021, Article ID 2179692, 9 pages, 2021.