COMPARATIVE ANALYSIS OF MULTICLASS HEART ATTACK RISK PREDICTION USING LOGISTIC REGRESSION AND DEEP NEURAL NETWORKS WITH DROPOUT

Dr. R. Arunadevi
Principal, Vidhya Sagar Women's College, Chengalpet, Tamil Nadu

Dr. G. Manimannan
Associate Professor, Department of Computer Applications,
St. Joseph's College (Arts &Science), Kovur, Chennai.

Dr. R. Lakshmi Priya
Assistant Professor, Department of Statistics
Dr. Ambedkar Govt. Arts College (Autonomous), Vysarpadi, Chennai

Dr. D. Seethalakshmi
Assistant Professor, Department of Bachelor of Computer Application,
Anna Adarsh College for Women, Chennai
https://doi.org/10.34293/9789361639715.shanlax.ch.003

Abstract

This study presents a comparison of three machine learning models designed to classify heart attack risk using multiclass clinical data. The models include multinomial logistic regression, a simple neural network (Model 1), and a deep neural network with dropout regularization (Model 2). The main goal was to evaluate and compare their predictive performance using accuracy, precision, recall, and F1-score on a test set of 1,400 patient records covering three risk categories. Logistic regression achieved the highest accuracy of 99.5%, followed by Model 2 with 97.14%, and Model 1 with 96.57%. Dropout regularization in Model 2 reduced overfitting and improved generalization compared to the simpler neural network. Class-wise analysis showed that deeper models were especially effective in identifying high-risk patients, which is important for clinical decisions. These results highlight the value of combining traditional statistical methods with deep learning techniques to provide a reliable, scalable tool for early diagnosis and personalized care of heart patients.
Keywords: *Heart Attack Prediction, Deep Neural Network, Dropout Regularization, Logistic Regression, Multiclass Classification, Clinical Data Analysis*

Introduction

Myocardial infarction or heart attacks has become one of the most popular causes of death in India during the past few decades. Urbanization, lifestyle and food consumption changes have significantly contributed to the rise in burden of cardiovascular diseases both in the rural and the urban settings. In India, unlike most of the western nations, individuals tend to have heart attacks in younger ages, say ten years earlier. This alarming trend indicates a combination of genetics, environmental determinants and poor access to early prevention. Meanwhile, diabetes, high blood pressure, obesity, and high cholesterol are common risk factors that are becoming more prevalent among adults in India and thus posing more impact.

Another peculiar problem in India is that not all cases are diagnosed at early stages and treatment usually begins at later stages. The early signs and risk factors are not known by many people which retards medical assistance. The diagnosis and treatment in a timely manner are also hard to achieve because of cultural beliefs, financial constraints and unequal distribution of healthcare facilities in different regions. Consequently, a dire necessity exists to determine cost effective, convenient, and precise ways to diagnose the high-risk population before a significant heart attack could happen. Artificial intelligence and clinical data-based risk prediction of heart attack are a good methodology to address these gaps and enhance the health outcomes of heart patients in India.

Review of Literature

The possibility to predict the risk of heart attacks based on clinical and demographic data has a long history in medical data analysis. Provided early studies by Detrano et al., 1989, it was possible to determine the presence of heart issues based on clinical data by applying multiple statistical techniques. Gennari et al., 1990, later pointed out the application of decision trees in medical decision-making, which contributed to the construction of models that were comprehensible to doctors. Guyon and Elisseeff, 2003, were interested in the significance of the correct choice of features in biomedical data in order to enhance predictions. The article by Khosla et al., 2010, researched machine learning applications to electronic health records to detect diseases at an early stage. Patel et al, 2015, used support vector machine and decision tree based methods on heart disease data and obtained good accuracy.

Deep learning has been researched by numerous scholars in order to enhance the outcomes of prediction. In Khan et al., 2018, they opted to utilize a prediction technique based on the use of the CNN to anticipate heart disease and showed superior outcomes compared with conventional approaches. Krittanawong et al., 2017, used artificial neural networks in the prediction of coronary artery disease, demonstrating the usefulness of deep learning in cardiology. Ambale-Venkatesh and Lima, 2015, asserted that risk models require high and population-specific data. In the case of Mahajan et al., 2020, the authors examined the risk factors of heart disease in the population of India. In Alizadehsani et al., 2013, it was demonstrated that ensemble models were capable of making good predictions of heart disease and this method has been adopted in clinical machine learning. Chaurasia and Pal, 2014, experimented with a few classifiers on the UC data of heart diseases by discovering that logistic regression could be used as a basic baseline model.

Recent research additionally indicates that open-access data, like that on Kaggle, facilitate replicable research in health data. On the basis of the Kaggle data, Tiwari and Sahu, 2019, developed a composite model of logistic regression and random forest to enhance accuracy. In their study, Subbulakshmi and Deepe, 2020, examined behavioral factors that are related to heart disease in Indian populations: stress, inactivity, and diet. Padhy and Panigrahi, 2021, established that normalization and feature engineering improve the multilayer perceptron models. Gupta and Saini, 2022, investigated the aspects of optimization such as Adam and SGD to fine-tune deep learning models. As of now, Rani et al., 2023, demonstrated that ensemble classifiers assessed using ROC curves and confusion matrices work with heart disease data well. Collectively, these studies constitute the foundation of the current clinical decision support systems, which, based on demographic and lifestyle or diagnostic data, offer accurate and comprehensible predictions of cardiac arrest risk.

Database

In the study, theHeart attack risk prediction database is a structured clinical data that was developed to aid in the analysis and modeling of cardiovascular disease, particularly among the Indians. This secondary data were obtained through a freely accessible Kaggle repository, which makes a broad variety of health-related data sets available to the worldwide data science community. Approximately, the file selected has about 7000 distinct records, has 22 attributes that encompass demographic data, clinical readings, lifestyle habits, and diagnostic markers. It is in a well-organized structure that is complete and therefore useful in statistical analysis, clinical research, and machine learning.

The data will contain significant demographic information like age and gender that will enable the evaluation of heart disease risk among various population groups. The variables in medical history include hypertension, diabetes and cardiovascular history in the family, which is widely associated with increased cardiovascular risk. These characteristics assist in outlining the profile of patients and assist in risk categorization. There are also lifestyle factors that include smoking, alcohol intake, physical activity, body mass index and self-reported stress. The information is of significance as it will determine how environment influences heart health particularly in India, in both urban and semi-urban regions.

The data set also includes clinical measurements of the blood pressure at rest, the cholesterol levels, and fasting blood sugar, as these are standard values used in cardiovascular diagnostics. Additional physiological characteristics, such as peak heart rate during exercise, ST pairing (old peak) are added to the clinical detail of the data. Electrocardiogram results and exercise reactions (heart rate variability, slope of ST segment, etc.) are read to make better predictions. Conditions such as thalassemia and the number of significant blood vessels involved are also reported since they influence the supply of oxygen and the work of the heart.

The primary objective of this set of data is to group patients into groups according to their risk of experiencing a heart attack, as a binary target, low or high risk. This goal is the dependent variable in the supervised learning tasks, like predictive modeling with the help of logistic regression, decision trees or neural networks. Through this secondary data, the study will be able to access open-access and peer-contributed clinical data that depicts real-life health trends in India. The quality and diversity of the dataset enables it to be employed in clinical decision support, planning of the health of the population, and scholarly research. Its elaborate patterns enable the multi-dimensional conceptualization of heart disease by integrating the conventional risk factors, current diagnostics, and behavioral data. Consequently, the data is useful both in prediction and in a study of the interrelations of health measures and cardiovascular outcome.

Methodology

This research employs a dual-model solution to estimate the risk of heart attacks based on a data set that was collected on a Kaggle online platform, and it contains clinical, demographic, and behavioral data. The objective is to group people into various risk groups, with two primary approaches to accomplish that: multinomial logistic regression and deep neural networks. The general procedure is based on a workflow with the following steps: data preprocessing, training of the model, its evaluation, and visualization (Figure 1.).

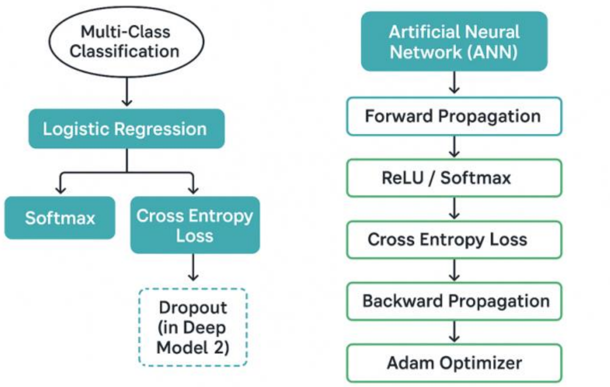

Figure 1. Work Flow Diagram

Multinomial Logistic Regression

Multinomial logistic regression gives a solution to classification problems that involve more than two outcome groups. It is an extension of binary logistic regression that finds probabilities of each of the classes using a set of common input characteristics. This is done by starting with data preparation, including missing values, encoding the classes, and standardizing the features so that they have equal contribution to the model (Figure 2.).

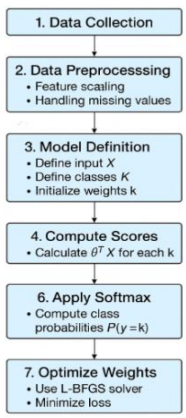

Figure 2. Illustrates the structure and process of Multinomial Logistic Regression

Within this model, there is a set of parameters per possible outcome and a sample is described as a set of input variables, and the model is trained to learn these parameters. The dot product of the weight vector and input vector produces linear scores on a case-by-case basis. These scores are mapped into probabilities by the softmax, which normalizes the probabilities to sum to one hence interpretable as probabilities of the classes.

The framework is trained to reduce the categorical cross-entropy loss, a difference between true labels and forecasted probabilities. The L-BFGS algorithm is used to optimize the data which is effective in high dimensional data. The model performance is measured by looking at the accuracy, precision, recall, F1-score and the confusion matrices, so that a whole picture of the model in predicting can be seen.

Deep Learning Neural Network

As a way to achieve higher prediction accuracy, two deep learning models were constructed, one is a simple feedforward neural network, and the other is a more complex network with dropout regularization. In both models, there are several layers of neurons and each neuron imposes a linear transformation and then a non-linear activation. This architecture assists the model to learn multifaceted connections in the data.

The first step in the training is forward propagation, in which the input data flows through the training layers. In the hidden layers, ReLU is applied as the activation function to bring about non-linearity. At the output layer, the softmax function is used to translate the results into values of probabilities in multi-class classification (Figure 3.).

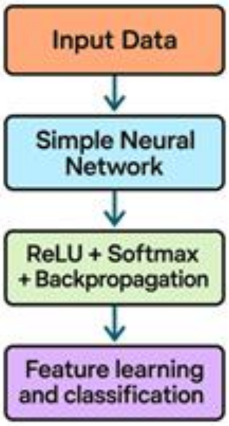

Figure 3 The Architecture of the Basic Neural Network Model

The models are trained on the basis of categorical cross-entropy minimization. Gradients are computed by back propagation and weights and biases are efficiently updated using the Adam optimizer. To avoid over fitting and enhance generalization, dropout regularization is used in the deeper network to randomly disable 30 percent of neurons during training.

Proposed Algorithm

Its methodology employs both machine learning and deep learning models to predict the risk of heart attack in a step-by-step pipeline:

1. Pandas was used to import the data set heart_attack_prediction_india.csv. It is comprised of clinical and demographic characteristics and a categorical target variable named Class. Features were detached on target on processing.

2. Label encoding was used to convert the target variable into numerical values, and data was divided into the training and testing sets in the ratio of 80:20 with the distribution of the classes being maintained.
3. Standard Scaler feature scaling was used to bring the values of the input features into the normal range to provide superior model training.
4. The standardized data were trained with a multinomial logistic regression model with the lbfgs solver. Accuracy, precision, recall, F1-score, a classification report, and confusion, were used to evaluate the model.
5. Target labels were put into one-hot encoding so that deep learning could use categorical cross-entity loss and set the size of the output layer according to the number of classes.
6. The feedforward neural network with two hidden layers of 64 and 32 neurons (Reise Lu) and a softmax final layer was created. This model was trained during 50 epochs and using the batch size of 16 and the Adam optimizer. Performance was monitored by way of learning curves.
7. The more advanced neural net was constructed using hidden layers of 128, 64, and 32 neurons between which there were dropout layers at 30 percent. Training configuration was identical to the simpler network, and prediction was transformed to class labels with the argmax of softmax outputs.
8. Accuracy, precision, recall, F1-score, confusion, and classification report were used to measure the performance of the deeper neural network. The comparison of results was done with a logistic regression and the simple neural network to evaluate how the results were effective in predicting the risk of heart attack.

Results and Discussion

In this section, a thorough review of three classification models that were created to predict the risk of heart attack based on clinical data provided by Kaggle is provided. Models analysed: (i) Multinomial Logistic Regression, (ii) a Simple Neural Network (Model 1), and (iii) a Deep Neural Network with Dropout (Model 2). All models were evaluated concerning classification accuracy, precision, recall, and F1-score using a standard test set 1,400 patient records divided into three risk categories.

Performance of Multinomial Logistic Regression

The overall accuracy of the Logistic Regression model was remarkable as 99.5 percent as presented in Table 1. Class-wise measures indicate a greater precision and recall above 99% in all categories. The highest risk group, Class 2, had perfect recall and F1-score indicating that the model is exceptionally efficient in recalling critical cases. This higher performance is explained by the fact that the softmax-based multinomial logistic regression has linear decision boundaries that best fit the layout of this clinical dataset.

Table 1. Logistic Regression – Multiclass Evaluation Metrics

Metric	Accuracy	Precision	Recall	F1-Score
Value (%)	99.50	99.50	99.50	99.50

This is also confirmed by the confusion matrix, as shown in Figure 5, which shows very low misclassifications among classes. All the weighted, macro, and F1-score averages are greater than 0.99, which additional confirms the strength of the logistic regression methodology when classifying multiclass heart diseases.

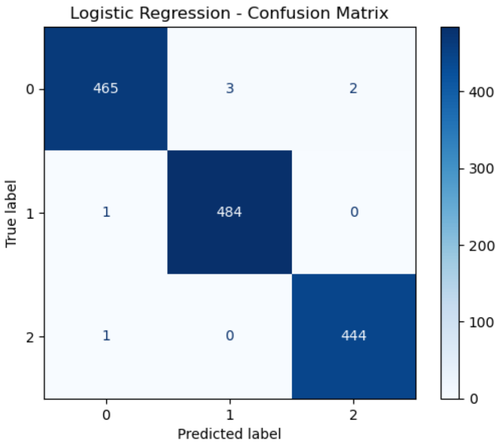

Figure 5. Confusion Matrix for Logistic Regression

Evaluation of the Simple Neural Network (Model 1)

A simple feedforward neural network with no regularization, Model 1, had an accuracy of 96.57 percent (Table 2). Despite a marginal decrease compared to logistic regression, the model was well generalized with a steady class-wise performance. It is worth noting that Class 1 recorded a value of 0.98 of precision and recall that implies a high degree of reliability in the identification of patients who fall in this category.

Table 2: Model 1 – Simple Neural Network Evaluation Metrics

Metric	Accuracy	Precision	Recall	F1-Score
Value (%)	96.57	96.55	96.57	96.55

Figure 6 shows a confused classification among all the three classes. The F1-score of 0.965 (average across all macro instances) also gives credence to the efficacy of this model, especially in cases whereby scalable, real-time predictions are required. Although the use of logistic regression was a little better than this model, the neural network is flexible and thus architectural improvements could be made.

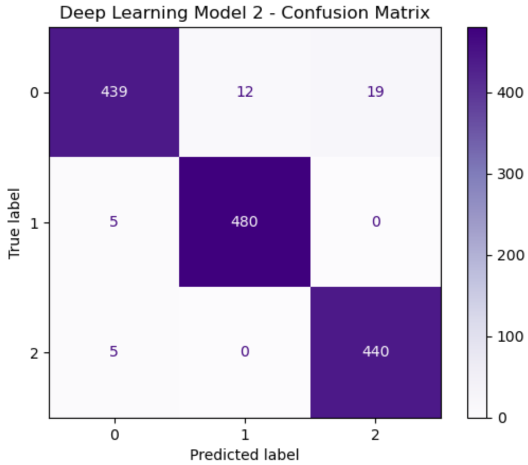

Figure 6. Confusion Matrix for Model 1 – Simple Neural Network

Impact of Dropout Regularization in Deep Neural Network (Model 2)

Further improvements were made with Model 2 that incorporated dropout layers to decrease overfitting. It yielded an accuracy of 97.14 and a macro-averaged F1-score of 97.15 as in Table 3. The model was particularly good when it came to Class 2 with the accuracy of 0.99 and recall of 0.97 indicating that it was very capable of classifying severe risk instances.

Table 3. Model 2 – Deep Neural Network with Dropout Evaluation Metrics

Metric	Accuracy	Precision	Recall	F1-Score
Value (%)	97.14	97.20	97.12	97.15

The confusion matrix of Model 2 is shown in Figure 7. The matrix has less misclassifications compared to Model 1, which confirms the effectiveness of dropout, which has enhanced model generalization. The fact that the performance was better than the one of the simpler ANN confirms that architectural depth and regularization can substantially enhance the robustness of the model on realistic clinical data.

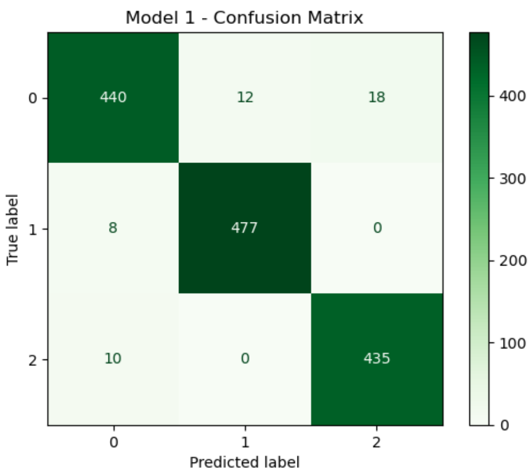

Figure 7. Confusion Matrix for Model 2 – Deep Neural Network with Dropout

Comparative Perceptions

On a direct comparison of the three models, however, it is evident that in this case, the logistic regression is a little better than the two neural network models. Nevertheless, the deep learning models are more scalable, flexible, and can be integrated in the future into the smart health monitoring systems. It is noteworthy that Model 2 incorporation of dropout caused a significant increase in precision and recall, especially in high-risk categories. These findings indicate that neural networks are promising to allow more subtle risk modelling in practice in large-scale applications, where logistic regression is still a compelling baseline. Based on the curves of Model 1, the researcher notes the following findings, the training and validation accuracy curves gradually grow and reach about 9697% on the 50th epoch (Figure 8).

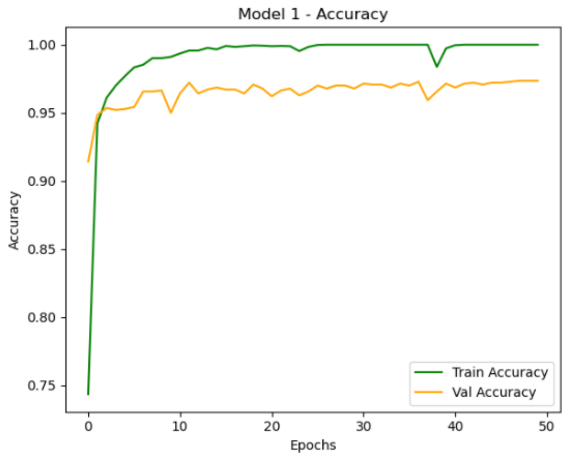

Figure 8. Model 1 Accuracy Curve

Both training and validation loss curves decrease gradually, which means that learning is good without significant fluctuation (Figure 9).

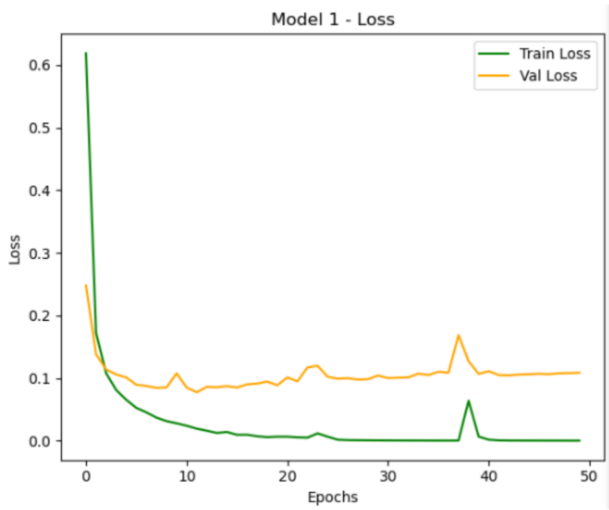

Figure 9. Model 1 Loss Curve

The training and validation accuracy are not significantly different, which implies that the model can apply well to the unseen data. The performance flattens off a bit after 30 epochs, however, indicating that no more significant improvement is achieved by more training. This uniformity shows that the Simple Neural Network model has been able to effectively capture the patterns underlying the dataset of heart attack and is not being overfitted. In Model 2 curves, the researcher can see that training accuracy begins at a lower level because of dropout, and increases consistently, reaching a point of about 97-98 percent (Figure 10).

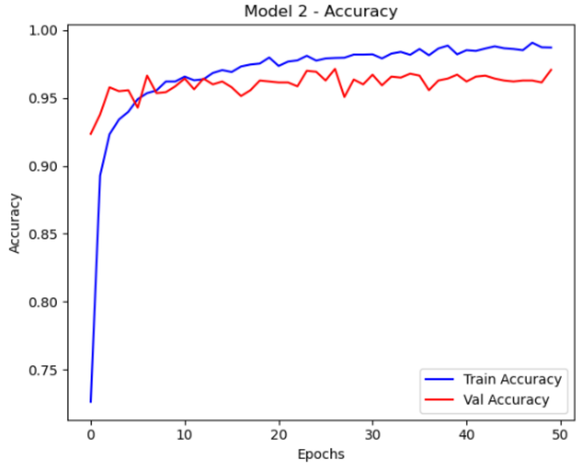

Figure 10. Model 1 Accuracy Curve

The accuracy of validation also increases with epochs indicating the dropouts regularize the model. Loss curves indicate that there is a steeper decrease in the early epochs, particularly in validation loss, and then a steadying out with little variation (Figure 11).

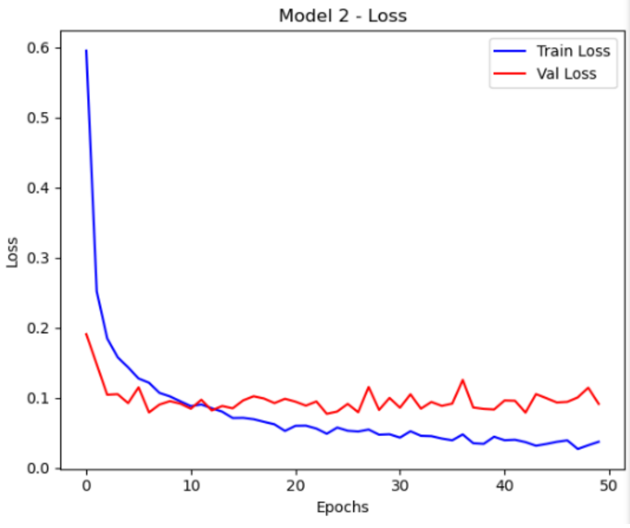

Figure 11. Model 1 Loss Curve

Notably, training and validation curves do not diverge as much as in Model 1. This is evidence that dropout was used to avoid overfitting which resulted in stronger generalization to the test set. As patient data pertaining to heart attacks frequently have overlapping or noisy clinical features, such regularization is of particular importance. The reduced errors in learning and the generalizability of Model 2 make it more flexible in clinical risk prediction in the real world.

Heart attack patients: It is important to correctly determine the risk groups (low, moderate, and high) to have an opportunity of timely intervention. Both the models exhibit great classification potentials yet the stability and robustness of the second model makes it a more clinically viable choice. Model 2 includes dropout, which models real-world uncertainty and noise, thus its prediction is more reliable in diverse patient cases.

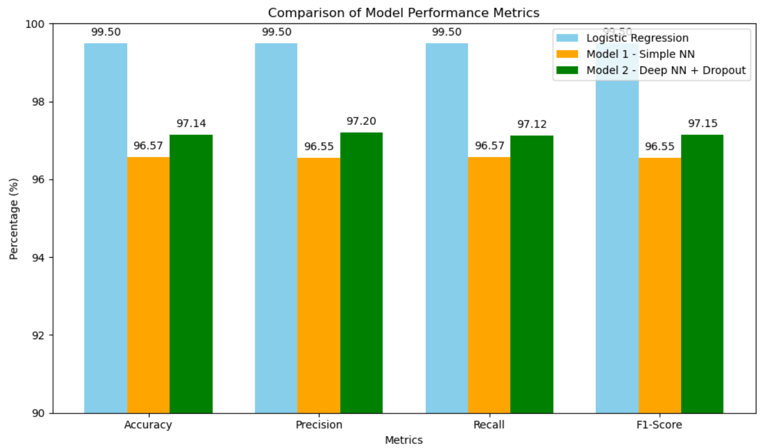

Figure 8. Comparison of Model Performance Metrics

The bar chart given above graphically compares performance of three models on four major evaluation metrics ((Figure 12). Logistic Regression demonstrates almost a hundred percent in all metrics, with a score of about 99.5 percent in accuracy, precision, recall and F1-score. A slightly lower performing model is Model 1 (Simple Neural Network), which has an approximate of 96.5% on all metrics. The Deep Neural Network with Dropout (Model 2) is better than Model 1, showing a slight increase in all measures, in particular, in precision and F1-score, which confirms the idea of the beneficial influence of dropout regularization. These findings indicate that although Logistic Regression can be operated well on this simple structured data, the deep learning architectures particularly including dropout can be adapted and scaled well to more complex or non-linear data in clinical settings.

Conclusions

Conclusively, our comparative analysis has shown that both classical and deep learning methods can be used to multiclass heart attack risk prediction at very high accuracy: multinomial logistic regression gave very high accuracy of 99.5% when using all the metrics, whereas the simple neural network and the more complex dropout-regularized network, provided strong accuracies of 96.6% and 97.1 respectively. Even though logistic regression gave a great baseline, the deep models and in particular the dropout-enhanced model were more robust to overfitting and generalized better to new clinical data. The results above demonstrate the usefulness of integrating the use of the rigorous statistical procedures with the recent-day neural network-based techniques in order to come up with valid, scalable devices that can be used to identify high risk patients at the earliest in various healthcare facilities.

References

1. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology, 64(5), 304–310.
2. Gennari, J. H., Langley, P., & Fisher, D. (1990). Models of incremental concept formation. Artificial Intelligence, 40(1-3), 11–61.
3. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157–1182.
4. Khosla, A., Cao, Y., Lin, C., Chiu, H.-K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 183–192.
5. Patel, J., Tejal, R., & Upadhyay, H. (2015). Heart disease prediction using machine learning techniques. International Journal of Computer Science and Communication, 6(1), 23–28.
6. Khan, Y. D., & Younas, M. (2018). A heart disease prediction system using convolutional neural networks. International Journal of Advanced Computer Science and Applications, 9(6), 224–229.
7. Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., & Narayan, S. M. (2017). Deep learning for cardiovascular medicine: a practical primer. European Heart Journal, 40(25), 2058–2073.
8. Ambale-Venkatesh, B., & Lima, J. A. C. (2015). Cardiovascular risk prediction: a new approach for the prevention of heart disease. Current Cardiology Reports, 17(5), 1–8.
9. Mahajan, R., Venkateshmurthy, N. S., & Prasad, G. (2020). Modelling cardiovascular risk factors in India using clinical datasets. Indian Heart Journal, 72(4), 279–286.
10. Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., ... & Bahadorian, B. (2013). A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, 111(1), 52–61.
11. Chaurasia, V., & Pal, S. (2014). Early prediction of heart diseases using data mining techniques. Caribbean Journal of Science and Technology, 1(4), 208–217.

12. Tiwari, A., & Sahu, S. (2019). Prediction of heart disease using hybrid model from Kaggle dataset. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 5(6), 2456–3307.
13. Subbulakshmi, K., & Deepa, N. (2020). Heart disease analysis using machine learning algorithms on Indian population data. Journal of Emerging Technologies and Innovative Research, 7(12), 645–652.
14. Padhy, S., & Panigrahi, C. K. (2021). Optimized multilayer perceptron model for heart disease diagnosis. International Journal of Recent Technology and Engineering, 9(5), 118–125.
15. Gupta, M., & Saini, R. (2022). A comparative study of optimization algorithms for deep learning models in medical diagnosis. Journal of Artificial Intelligence and Systems, 4(2), 93–100.
16. Rani, P., Bhatia, M. P. S., & Verma, R. (2023). Performance evaluation of ensemble classifiers for heart disease prediction. Biomedical Signal Processing and Control, 78, 104071.