Enhanced Prediction of Heart Disease Using Machine Learning and Deep Learning

M. S. Guru Prasad[1](✉), D. K. Santhosh Kumar[2], M. S. Pratap[3], J. Kiran[3], S. Chandrappa[4], and Arnav Kotiyal[1]

[1] Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun, India
guru0927@gmail.com

[2] Department of Computer Science and Engineering, St Joseph Engineering College, Mangaluru, Karnataka, India

[3] School of Computing and Information Technology, Reva University, Bengaluru, India
pratapms.ms@reva.edu.in

[4] Department of Computer Science and Engineering, Jain (Deemed-to-be University), Bengaluru, India

Abstract. The provision of medical care is an essential component of human existence. As a result of the vast amount of psychiatric data included within the healthcare industry, machine learning models were utilized in order to efficiently deliver conclusions regarding heart disease prediction. The adoption of methods derived from machine learning enables the reliable classification of individuals according to whether or not they are healthy. The framework used in this study can understand the basics of effectively evaluating a patient's risk profile from features of clinical data. The aforementioned model was developed by utilizing both machine learning and deep learning in tandem with one another. Heart disease is widely acknowledged as one of the primary contributors to death rates across the globe. Large amounts of clinical data are stored in the many biomedical instruments and computer systems that are found in hospitals. Therefore, having a solid understanding of the data around heart disease is quite crucial if one wishes to increase the accuracy of their predictions. There have been a lot of experimental evaluations of the performance of models that have been developed using classification algorithms and relevant features that have been selected using a variety of different approaches to feature selection. The exploratory investigation used a dataset on heart illness to test four different classification strategies. These strategies were random forest, support vector machine, k-nearest neighbor, and convolutional neural network. The accuracy of machine learning algorithms utilized in the proposed work is Support Vector Machine 85.18%, Random Forest 92.5%, K-NN 74.07% and Convolutional Neural Network 85.18%

Keywords: Heart Disease · Deep Learning · Machine Learning · Random Forest · Support Vector Machine · K-Nearest Neighbor · Convolutional Neural Network

R. N. Shaw et al. (Eds.): ICACIS 2023, CCIS 1920, pp. 1–12, 2023.
https://doi.org/10.1007/978-3-031-45121-8_1

1 Introduction

The World Health Organization estimates that 17.9 million lives are lost annually due to cardiovascular disease [1]. It also predicts that the number of deaths from cardiovascular disease will grow to about 30 million by the year 2040 [2]. Heart disease can be attributed to a number of factors, including obesity, excessive cholesterol, a rise in triglyceride levels, hypertension, and more [3]. There are a number of common tests used by doctors to diagnose cardiovascular disease [4]. These include an ECG (echocardiogram), cardiac magnetic resonance imaging (MRI), a stress test (exercise stress test), and a nuclear cardiac stress test. Numerous computer technologies, such as those used to access patients' medical information and conduct research, can be utilized to accurately diagnose individuals and detect this disease in its early stages, before it has the chance to harm them [5]. Several different machine learning and deep learning models may be utilized in the direction of diagnose the condition as well as categories or forecast the consequences [6]. There are strategies for developing prediction models, as well as techniques for conducting in-depth analyses of patient data, which may be utilized to enhance the precision of such projections [7]. On the other hand, the expense of diagnosing and treating cardiovascular illness is so high that it is out of reach for the vast majority of people [8]. Using data mining techniques can help find signs of heart disease early and for less money. This will result in a reduction in the overall cost of diagnosis and treatment [9].

The use of machine learning techniques in the detection and classification of heart illness has been studied in prior studies [10]. Nevertheless, the focus of these studies is on the individual effects of particular machine learning approaches rather than on the optimization of these procedures utilising optimised methodologies [11]. In addition, very few researchers make an effort to apply hybrid optimization approaches for the purpose of improving the accuracy of machine learning classifications [12]. The majority of the research that has been proposed and published makes use of optimised approaches like Particle Swarm Optimization and Ant Colony Optimization in conjunction with a particular machine learning technique like SVM, KNN, or Random Forest [13].

2 Literature Survey

Masethe, H. D. et al. [1] proposed a model, it comprised both the selection of features and the verification of presence of duplicates in the data. The model utilized both machine learning and deep learning methods in order to provide an accurate forecast of heart illness. The machine learning strategy incorporated linear model selection, one of its components being the use of linear regression. The KNN classifier was used for the purpose of concentrating on the neighbour selection technique. After that, a tree-based technique known as the Decision Tree Classifier was utilised. Finally, the random forest classifier, one of the most commonly used ensemble algorithms, was employed. A Support Vector machine was utilised both for determining whether or not the data had a high dimensionality and for managing it. The sequential modelling approach was utilised for the Deep Learning model. This study's findings suggest that, machine learning algorithms fared better. In the past, many academics have argued that ML should be used even when the dataset is small, but this article shows that this is true.

Ramprakash, P et al. [2] uses Artificial Neural Networks and Deep Neural Network Algorithms in this model. Trainest-hold-out approval conspire is used in the suggested model for model approval. Train-test validation, which specifies that 80% of the information is used for preparation and 20% is used for testing, was implemented using the 80–20 rule. The sklearn package is used to slice up the data for use in both the testing and training phases. Out of a total of 303 samples, 242 examples were chosen and used to create the model, while the other 61 samples were used as testing information to evaluate the model's execution. Predictive models may be evaluated using a variety of metrics, including accuracy, sensitivity, specificity, and the Matthews correlation coefficient (MCC). If there are heart-related difficulties, this system returns a value of 1; otherwise, it returns a value of 0. This model's main drawback is that its precision is lower than that of other models. In the trials, it was found that the proposed method improved prognosis accuracy. Patients with heart disease will be able to be identified with the help of this research. When a patient is anticipated to have a favourable outcome, their reports and data can be examined in detail. In the future, a genetic algorithm may be applied to improve accuracy. In addition to a patient's family history of heart disease, this information may also be included in the dataset, which increases the model's accuracy.

Shah, D. et al. [3] implemented SVM, Decision Tree, Nave Bayes, and Random Forest algorithms in their work. These include deep neural networks and artificial intelligence (AI). The Weka Data Mining Tool is used to create this model. Clustering, classification, regression, visualisation, and feature selection are among the usual data mining activities that may be performed using the programme. It makes it simple to import data in the form of files, URLs, or databases. Computer-Structured Query Language (CSV). In the confusion matrix, true positive, accuracy, recall, and false negative are all analysed and shown in an easy-to-understand fashion. Each model has been analysed for its performance in terms of precision, recall, ROC (Receiver Operating Characteristic), and percent accuracy. Table IV displays the models' correctness in terms of performance. A ROC value of less than 0.80 is typically regarded as "GOOD," whereas a ROC value of less than 0.77 is seen as "FAIR". The model with the ROC value closest to 1 is regarded to be the most accurate.

Gavhane A et al. [4] proposed the decision tree and Ada-Boost algorithm model. Examples, diverse algorithms, and techniques for evaluation are some of the elements employed in a given piece of work. Using this strategy, you'll be taught by a teacher. For illness datasets and medical characteristics, the Kaggle warehouse is used. A person's risk of developing heart disease may be estimated using this collection of data. Training and testing sets are further subdivided. Attributes of the dataset utilised in this proposed research are 14. If "target," an independent variable, is correctly predicted, then a person is healthy or has heart disease.

3 Proposed Architecture

The term "architectural design" refers to the graphic representation of a collection of architectural ideas, including "principles," "elements," and "components. [14]" It is an imaginative process in which the aim is to provide a structure for the system that satisfies both the functional and non-functional requirements of the system [15]. Because it is an

artistic process, the activities that take place inside it vary widely depending on the kind of system that is being created, the educational background and professional experience of the system architect, and the needs of the system [16]. Figure 1 shows the proposed architecture.

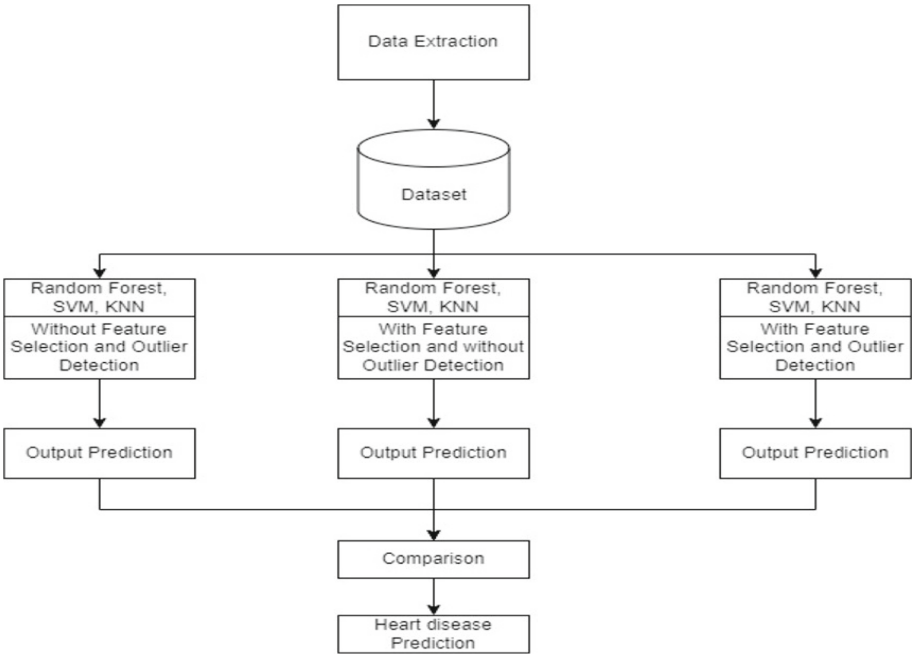

Fig. 1. Proposed Architecture

3.1 System Flow Diagram

A System flow diagram, often known as an SFD, is a diagram that displays the many types of data that will be input into and output from a system, in addition to showing where the information will come from, where it will go, and where it will be stored [17]. Flow diagram is also used as a synonym for flowchart and, occasionally, as a counterpoint to flowchart [18]. Flow diagrams are used to organize and arrange a complicated system, as well as to illustrate the underlying structure of the parts and how they interact [19]. The phrase "flow diagram" has multiple implications in theory and practice. The terms "flowchart" and "flow diagram" are frequently used interchangeably in the context of a process depiction The user will have to enter the required values, which will be taken as input. The data from the input is then retrieved which will be used for classification of the result (Fig. 2).

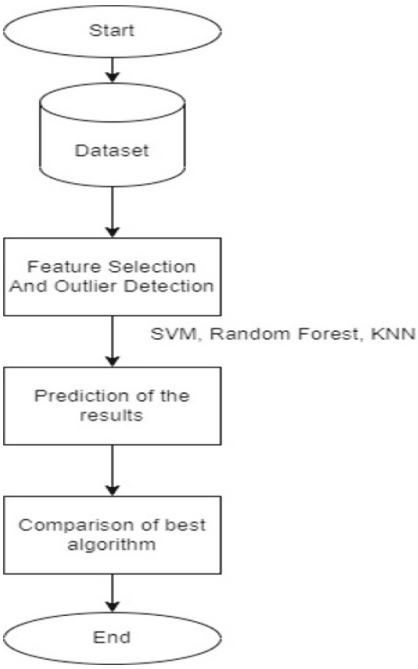

Fig. 2. System Flow Diagram of the proposed system

4 Results and Discussion

Kaggle provided the dataset that was utilised for the development of this suggested system. The databases make up this data collection, which was compiled in 1988 and dates back to that year. It has 76 properties, including the attribute that was predicted. However, all of the published tests only relate to employing a selection of 14 of those features. In this situation, the patient's heart disease status is the aim. It is integer valued with 0 = no disease and 1 = disease. The goal is included as one of the 1329 rows and 14 columns that make up the dataset. The thirteen characteristics of the dataset will serve as the data, while the target column will serve as the label. The dataset does not have any empty values (also known as null values).

The following are the results obtained using machine learning algorithms like SVM, Random Forest, KNN, and Deep Learning Algorithms like CNN. The proposed model predicts the accuracy as well as gives the graphical representation of the algorithms for a given ratio of test and training data so that comparison can be done and the best algorithm can be obtained for the heart disease dataset.

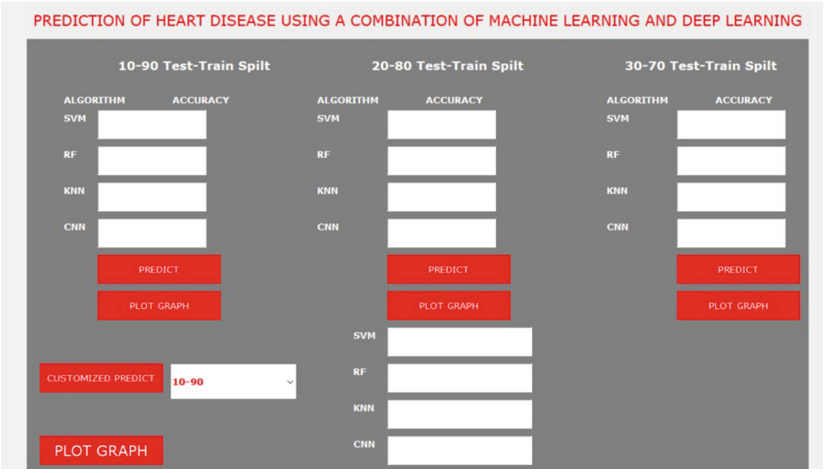

Fig. 3. User Interface of the proposed system

The above Fig. 3 shows the user interface of the proposed system where the user can see the predicted accuracy for various ratios of test and training data as well as a plot graph for the same. The customized predict allows the user to choose the required ratio of the split and also plot a graph for the same.

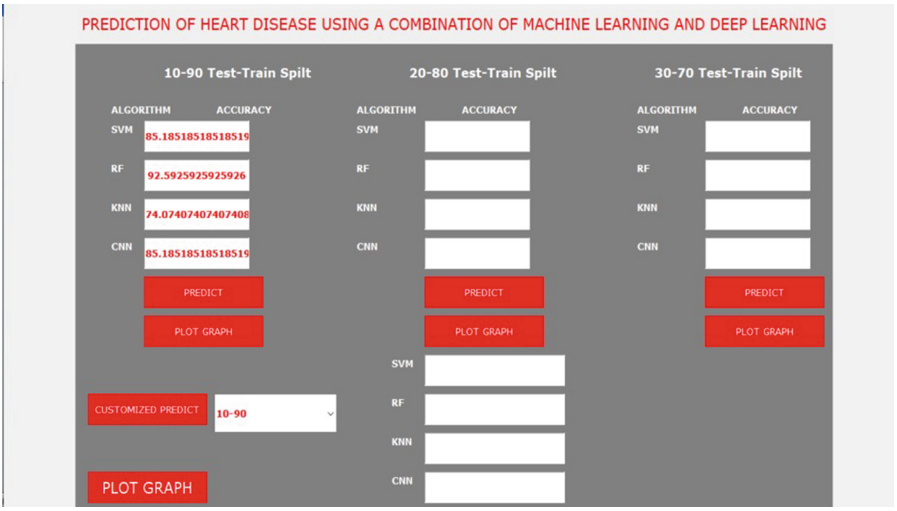

Fig. 4. Accuracy comparison of SVM, RF, KNN and CNN for 10% testing and 90% training data

The above Fig. 4 shows the predicted accuracy of SVM, RF, KNN and CNN for 10% testing and 90% training data.

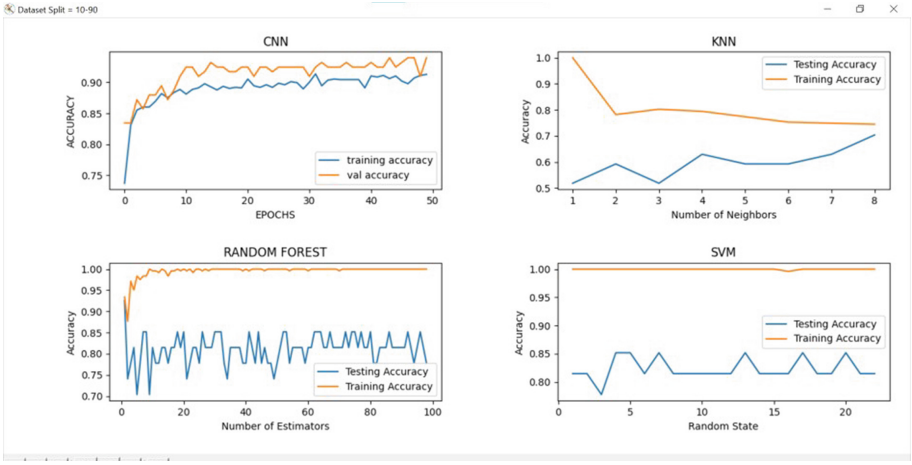

Fig. 5. Graph comparison of SVM, RF, KNN and CNN for 10% testing and 90% training data

The above Fig. 5 depicts the graphical representation of SVM, RF, KNN, and CNN for 10% testing and 90% training data. The Y axis is the accuracy of the respective algorithm, and the graph is plotted comparing the actual and predicted accuracy.

The following Figs. 6 and 8 show the predicted accuracy of SVM, RF, KNN, and CNN for 20% testing and 80% training data and 20% testing and 80% training accuracy, respectively, and Figs. 7 and 9 give the graphical representations for the respective ratio of the data.

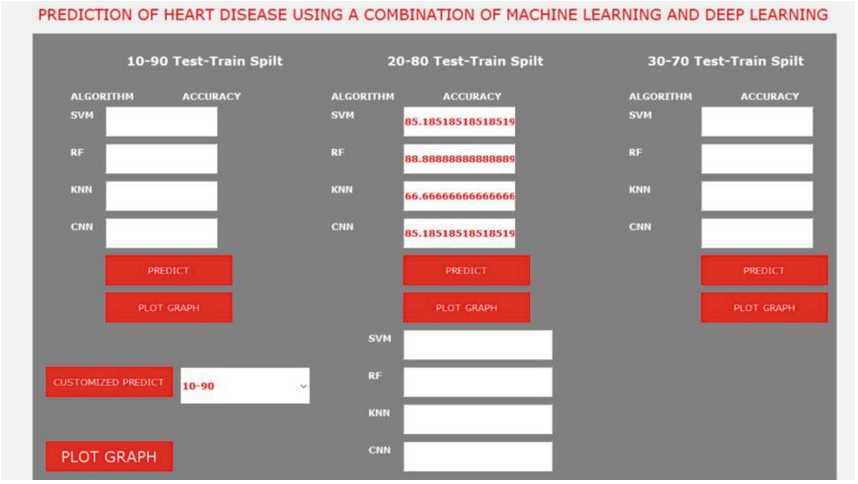

Fig. 6. Accuracy comparison of SVM, RF, KNN and CNN for 20% testing and 80% training data

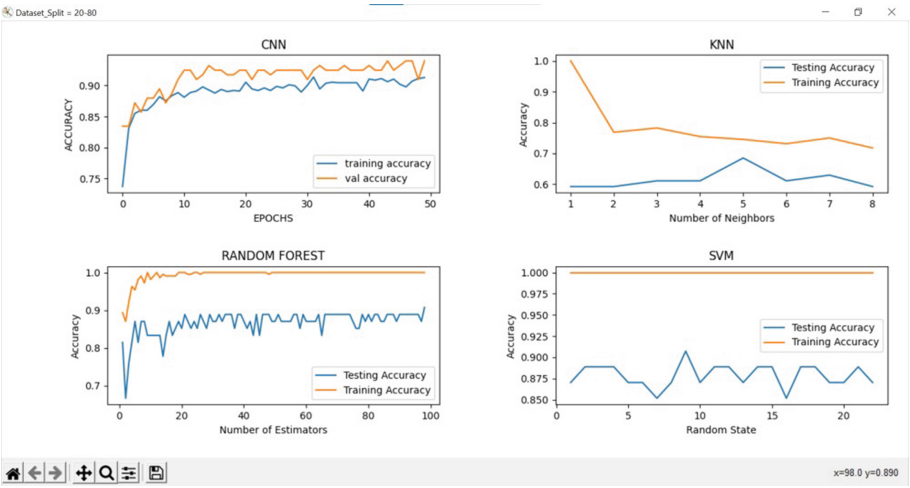

Fig. 7. Graph comparison of SVM, RF, KNN and CNN for 10% testing and 90% training data

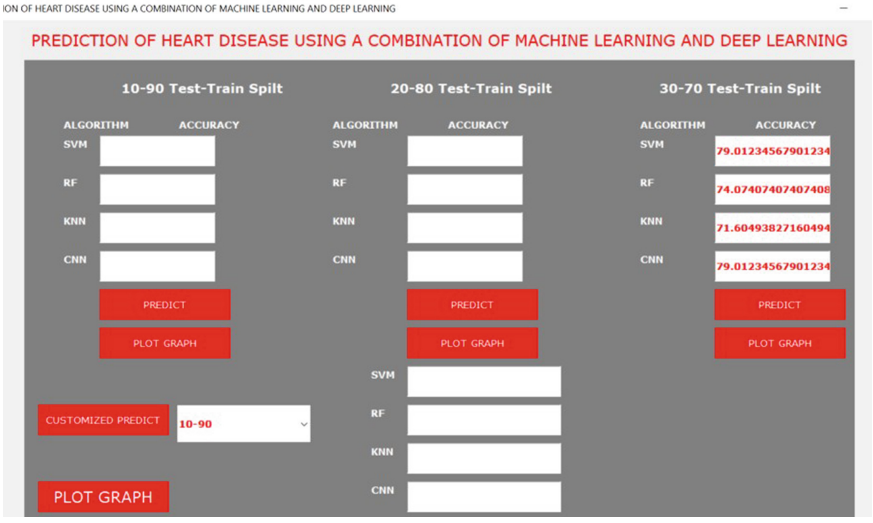

Fig. 8. Accuracy comparison of SVM, RF, KNN and CNN for 30% testing and 70% training data

Figure 10 shows the customised prediction accuracy of SVM, RF, KNN, and CNN algorithms. The snapshot contains the accuracy for a ratio of 60% test and 40% train data. However, the dropdown button "Customized Predict" allows the user to choose the desired ratio accordingly (Fig. 11).

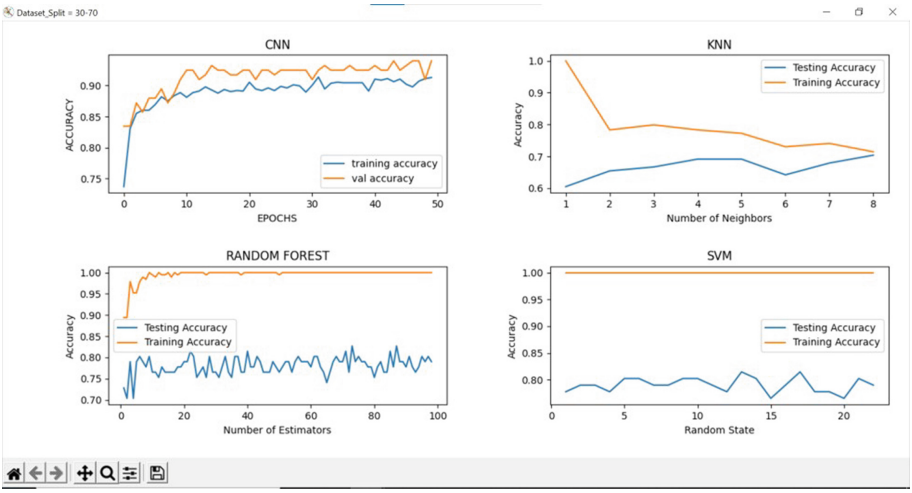

Fig. 9. Graph comparison of SVM, RF, KNN and CNN for 30% testing and 70% training data

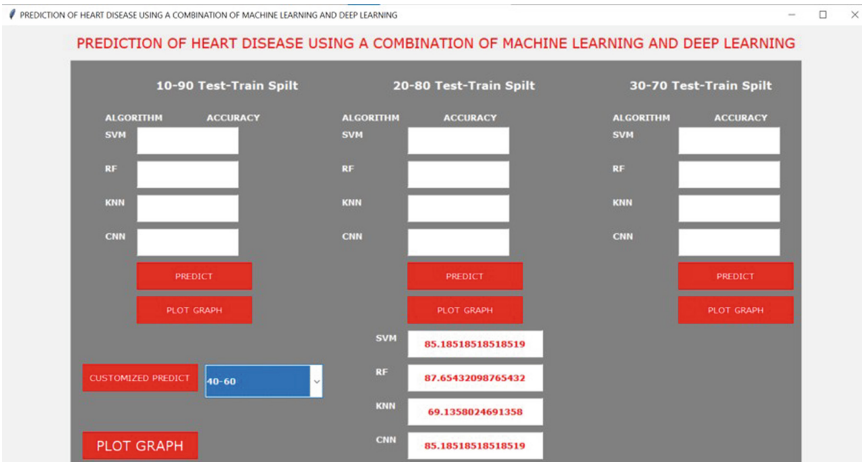

Fig. 10. Accuracy comparison of SVM, RF, KNN and CNN for customized 60% testing and 40% training data where the user can customize other values accordingly

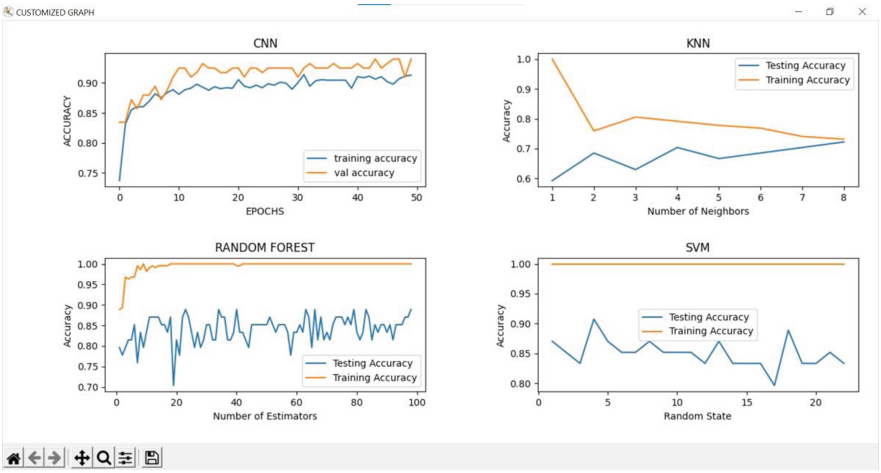

Fig. 11. Graph comparison of SVM, RF, KNN and CNN for customized testing and training data where the user can customize other values accordingly

5 Conclusion

Using the accuracy of the algorithms, this system seeks to provide a useful framework for comparing them. The SVM, Random Forest, KNN, and CNN, a deep learning method, can all be predicted with reasonable accuracy by the system. We feed the system various ratios of training and testing data to see how accurate it performs. We then visualise the difference between the system's actual performance and what we projected it would be. An algorithm based on performance-based algorithms for heart disease datasets and a certain ratio of data has been presented. We discovered that machine learning techniques such as the Random Forest algorithm outperformed all others in our study. In the past, many academics have argued that ML should be used even when the dataset is small, but this article shows that this is true. Machine learning and deep learning data graphs are used to make comparisons. Outliers must be identified and isolated using the Isolation Forest approach. This technique is used to locate outliers in datasets with Gaussian distributions, which were also discovered throughout the analysis process. The problem here is that the dataset has a small sample size. Deep learning and machine learning may both benefit greatly from huge datasets. Deep learning may be used in conjunction with several additional improvements and a larger dataset to achieve more promising outcomes.

References

1. Masethe, H.D., Masethe, M.A.: Prediction of heart disease using classification algorithms. In: Proceedings of the world Congress on Engineering and Computer Science, vol. 2, no. 1, pp. 25–29 (2014)
2. Ramprakash, P., Sarumathi, R., Mowriya, R., Nithyavishnupriya, S.: Heart disease prediction using deep neural network. In: 2020 International Conference on Inventive Computation Technologies (ICICT), pp. 666–670. IEEE (2020)

3. Shah, D., Patel, S., Bharti, S.K.: Heart disease prediction using machine learning techniques. SN Comput. Sci. **1**(6), 1–6 (2020)
4. Gavhane, A., Kokkula, G., Pandya, I., Devadkar, K.: Prediction of heart disease using machine learning. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1275–1278. IEEE (2018)
5. Gujjar, J.P., Kumar, H.P., Prasad, M.G.: Advanced NLP frame-work for text processing. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, pp. 1–3 (2023). https://doi.org/10.1109/ISCON57294.2023.10112058
6. Guru, P.M.S., Praveen, G.J., Dodmane, R., Sardar, T.H., Ashwitha, A., Yeole, A.N.: Brain tumor identification and classification using a novel extraction method based on adapted alexnet architecture. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, pp. 1–5 (2023). https://doi.org/10.1109/ISCON57294.2023.10112075
7. Kumar, M.A., Pai, A.H., Agarwal, J., Christa, S., Prasad, G.M.S., Saifi, S.: Deep learning model to defend against covert channel attacks in the SDN networks. In: 2023 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), Ernakulam, India, pp. 1−5 (2023). https://doi.org/10.1109/ACCTHPA57160.2023.10083336
8. Kirubasri, G., Sankar, S., Guru Prasad, M.S., et al.: LQETA-RP: link quality based energy and trust aware routing protocol for wireless multimedia sensor networks. Int. J. Syst. Assur. Eng. Manag. (2023). https://doi.org/10.1007/s13198-023-01873-9
9. Guru Prasad, M.S., Naveen Kumar, H.N., Raju, K., et al.: Glaucoma detection using clustering and segmentation of the optic disc region from retinal fundus images. SN Comput. Sci. **4**, 192 (2023). https://doi.org/10.1007/s42979-022-01592-1
10. Kumar, H.N.N., Kumar, S.A., Prasad, G.M.S., Shah, M.A.: Automatic facial expression recognition combining texture and shape features from prominent facial regions. IET Image Process. **17**, 1111–1125 (2023). https://doi.org/10.1049/ipr2.12700
11. Rajawat, A.S., et.al.: Depression detection for elderly people using AI robotic systems leveraging the Nelder–Mead method. In: Artificial Intelligence for Future Generation Robotics, pp. 55–70. Elsevier (2021). ISBN 9780323854986, https://doi.org/10.1016/B978-0-323-85498-6.00006-X.Chakraborty
12. Chakraborty, A., Chatterjee, S., Majumder, K., Shaw, R.N., Ghosh, A.: A comparative study of myocardial infarction detection from ECG data using machine learning. In: Bianchini, M., Piuri, V., Das, S., Shaw, R.N. (eds.) Advanced Computing and Intelligent Technologies. LNNS, vol. 218, pp. 257–267. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-2164-2_21
13. Chandrappa, S., Chandra Shekar, P., Chaya, P., et al.: Machine learning algorithms for identifying fake currencies. SN Comput. Sci. **4**, 368 (2023). https://doi.org/10.1007/s42979-023-01812-2
14. Anand Kumar, M., Abirami, N., Guru Prasad, M.S., Mohankumar, M.: Stroke disease prediction based on ECG signals using deep learning techniques. In: 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, pp. 453–458 (2022). https://doi.org/10.1109/CISES54857.2022.9844403
15. Prasad, G., Jain, A.K., Jain, P., Nagesh, H.R.: A novel approach to optimize the performance of hadoop frameworks for sentiment analysis. Int. J. Open Source Softw. Process. (IJOSSP) **10**(4), 44–59 (2019)
16. Prasad, M.G., Pratap, M.S., Jain, P., Gujjar, J.P., Kumar, M.A., Kukreti, A.: RDI-SD: an efficient rice disease identification based on apache spark and deep learning technique. In: 2022 International Conference on Artificial Intelligence and Data Engineering (AIDE), Karkala, India, pp. 277–282 (2022). https://doi.org/10.1109/AIDE57180.2022.10060157

17. Pai, A., Anandkumar, M., Prasad, G., Agarwal, J., Christa, S.: Designing a secure audio/text based captcha using neural network. In: 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, pp. 510–514 (2023). https://doi.org/10.1109/Confluence56041.2023.10048791
18. Agarwal, J., Christa, S., Pai, A., Kumar, M.A., Prasad, G.: Machine learning application for news text classification. In: 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, pp. 463–466 (2023). https://doi.org/10.1109/Confluence56041.2023.10048856
19. Patel, V., Guru Prasad, M.S., Aditya Pai, H., Kumar, A.S., Praveen Gujjar, J., Naveen Kumar, H.N.: Real-time face mask detector. In: 2023 IEEE 3rd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET), Mysuru, India, pp. 1–5 (2023). https://doi.org/10.1109/TEMSMET56707.2023.10150182