A Deep Dive into Disease Prediction using Artificial Neural Networks: Insights from Data Analysis

Anil   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*— In order to diagnose and cure diseases at an early stage, medical diagnosis is a crucial component of healthcare. Machine learning models have gained popularity in the area of medical diagnostics due to the expansion of medical data availability. The goal of this study is to create an Artificial Neural Network (ANN) model for symptom-based disease prediction. Hyperparameter tuning is performed on the created model, and it is then contrasted with models like Random Forest, Logistic Regression, and K Nearest Neighbours. With an accuracy rating of 99.87%, the results show that the constructed ANN model performs better than other models. The model's capacity to precisely distinguish between positive and negative situations is shown by the ROC-AUC curve.

Keywords—Disease prognosis, ANN, Hyperparamter tuning, Logistic Regression, KNN

# Introduction

## Disease prediction refers to the use of statistical, computational, or machine learning techniques to predict the likelihood of an individual developing a particular disease or medical condition. Predicting diseases is crucial for a number of reasons, including the reduction of healthcare costs, early disease detection, disease prevention in individuals at risk, and optimal resource allocation. In terms of identifying people who are at a high risk for disease and enabling early intervention and treatment, disease prediction models have demonstrated promising results. Infection expectation models can likewise be utilized to enhance medical care asset portion by recognizing people who are at high gamble of fostering a sickness and giving designated intercessions and therapies to forestall or deal with the illness. The healthcare system's overall efficiency and costs could both be improved by this. For a wide range of diseases, including cancer, heart disease and many others, disease prediction models have been developed in recent years. Early prevention and treatment have been made possible by these models, which have demonstrated promising results in identifying people who are at high risk for disease.

A promising strategy for healthcare disease prediction is machine learning (ML). Predicting an individual's likelihood of developing a disease or medical condition is made possible by ML algorithms by analysing large datasets of medical records, genetic data, and other health-related information. The use of artificial neural networks (ANNs), a type of deep learning algorithm, in healthcare disease prediction research is on the rise. ANN is a powerful tool for disease prediction because it can identify complex patterns and relationships in the data that humans may find difficult to recognize. However, ANN models can be difficult to interpret, making it difficult to comprehend the underlying factors that influence disease risk and to explain the model's predictions to patients and clinicians. Although ANN-based disease prediction models have the potential to improve healthcare outcomes and enable personalized medicine, more research is required to address their limitations and guarantee their safe and effective application in clinical settings.

This study aims to build an Artificial Neural Network (ANN) model that would be taking features as input and predicting what disease has occurred based on those features. The research has followed all the necessary steps to build an efficient model that would give good results when compared to present solutions.

# Related work

Segmentation and diagnosis with machine learning algorithms are common in the biomedical field, but little is known about how they are used to classify skin diseases. A study [1] extracted RGB colour features and GLCM texture features from images of plaque psoriasis, lichen planus, and chronic eczema to address this issue. Four well-known machine learning algorithms were used to test various feature combinations to compare their classification abilities. Straight Discriminant Investigation and Backing Vector Machine showed the most elevated precision out of the four calculations tried. This emphasizes the noteworthiness of selecting the suitable machine learning calculations and include extraction strategies for skin illness classification.

Machine learning and pattern recognition have the potential to enhance the objectivity and accuracy of disease diagnosis processes. In the biomedical field, machine learning is a good way to make automatic algorithms from high-dimensional and multimodal data. This study [2] paper thinks about various AI calculations for the determination of illnesses like coronary illness, diabetes, liver sickness, dengue, and hepatitis. When it comes to the detection of heart disease, SVM has an improved accuracy of 94.60%, Naive Bayes has the highest classification accuracy of 95%, and FT has a correctness of 97.10% when it comes to the diagnosis of liver disease. Using the RS theory, dengue disease was detected with 100% accuracy, and a feed forward neural network correctly classified hepatitis with 98% accuracy. The review features the benefits and impediments of these calculations and accentuates the significance of choosing the right AI calculation for the examination of infections and dynamic cycles.

This study [3] looked at academic conference and journal papers that used data mining techniques to classify and diagnose diseases using public medical datasets and were published between 2007 and 2019The review found a huge expansion in the use of information mining methods in disease classification. In any case, there was restricted spotlight on creating strategies utilizing steady renditions of information mining procedures. The research expects to give helpful data to help analysts in creating clinical support systems with ideas on cutting edge advancement strategies.

The use of data mining methods in machine learning to extract useful information from medical databases for disease prediction is the subject of this paper [4]. An example of 4920 patient records with 41 illnesses was dissected utilizing 95 chose independent factors firmly connected with the sicknesses. For the purpose of disease prediction, the study makes use of the Random Forest classifier , Decision Tree classifier and Naive Bayes classifier algorithms. The comparison study of the algorithms' findings is presented.

Predicting the Onset of Diabetes with Machine Learning Methods: The prevalence rate of diabetes in Taiwan is increasing annually, with an estimated 2.18 million people affected. The study [5] focused on 15,000 women between the ages of 20 and 80 who had visited a medical center from 2018 to 2022, with or without a diabetes diagnosis. Eight characteristics were analyzed, including age, insulin level, and body mass index. Various neural network models were trained using Microsoft Machine Learning Studio, and the two-class boosted decision tree was found to be the best predictor with an area under the curve of 0.991. Holistic treatments and prevention are necessary to avoid complications and medical resource consumption.

An approach to improve the performance and accuracy of three distinct classifiers for predicting early-stage breast cancer is proposed in this paper [6]. Two benchmark datasets are used to validate and analyze the classifiers, which are Naive Bayes (NB), Decision Tree (J48) & Sequential Minimal Optimization (SMO). Resampling the data to lessen the impact of classes that aren't evenly distributed addresses the issue. The accuracy, true positive, false positive, Roc curve, and the classifier's efficiency are all measured. In WBC, SMO gave best result where as in Breast Cancer dataset, J48 outperformed others, indicating that using a resample filter improves the classifier's performance.

Using KNN and CNN algorithms, a disease prediction system based on symptoms is proposed in this paper [7]. For precise prediction, the system takes into account daily routines and information from checkups. Compared to KNN, CNN's disease prediction accuracy is 84.5 %. Additionally, the system indicates a lower or higher risk for general disease. The difficulties of making an accurate diagnosis based on symptoms can be overcome by making use of data mining and machine learning in disease prediction.

This paper [8] proposes a general disease prediction system based on patient symptoms, using the Random Forest algorithm for accurate disease prediction. For precise predictions, the system takes into account routines and checkup data. After making a prediction, the system is able to determine whether there is a lower or higher risk of general disease. The Random Forest algorithm is more accurate and efficient than other algorithms, according to the findings of experiments, but it takes more time and memory. For accurate analysis and early patient care, data mining helps uncover hidden patterns in medical data.

The classification efficacy of various artificial neural network (ANN) models in osteoporosis, an orthopedic condition, is the subject of this study [9]. Probabilistic neural networks (PNNs) and multilayer perceptrons (MLPs) were utilized, with PNNs having spread values ranging from 0.1 to 50 and four or two neurons in the output layer.

In this [10] review, we presents a characterization approach utilizing Multi-facet Perceptron (MLP)with Back-Proliferation learning calculation and a component choice calculation alongside biomedical test values to analyze coronary illness. For diagnosis, a number of tests are required of patients. In many instances, not all of the tests help to accurately diagnose a disease. Our task is to reduce the number of attributes used to classify the presence of heart disease. The original classification of heart disease involves 13 attributes. By utilizing Information Gain, we can reduce the number of attributes that must be obtained from patients. The diagnosis of patients is categorized using artificial neural networks. Thirteen ascribes are diminished to 8 credits. The training data set's accuracy difference between 13 and 8 features is 1.1%, while the validation data set's accuracy difference is 0.82 percent.

# Methodology

As the project aimed at building a god classifier that could predict what kind of disease can happen based on the symptoms present, we followed proper systematic process of any data science project. All the steps are elaborated below:

## Data Preprocessing and EDA

The study uses this

(<https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning> ) Kaggle dataset as the base dataset. The dataset is well documented and proper guide on the data is given over the above link.

The dataset contains 2 CSV files, one for training and another for testing. It contains 133 columns, where 132 columns are the symptoms and the last ‘prognosis’ is the target column. The target column contains 42 different classes, i.e 42 different diseases. The independent columns includes various day-to-day common symptoms like itching, skin rashes, shivering, chills etc.

After data gathering, exploratory data analysis was done. And, it was observed the data has an equal distribution for each of the class, i.e, every disease has same number of samples, which meant the dataset was balanced. The dataset was then checked for missing values and outliers. But, the data didn’t have any missing values, nor outliers, so imputation and outlier handling was skipped. Then, we tried to draw insights using the correlation matrix. And it was observed, most of the columns didn’t have much correlation with any other column. The correlation matrix is visualized as below:

A picture containing graphical user interface

Description automatically generated

Figure 1: Correlation matrix

Then, we checked distribution of entities in each of the categorical columns to know if the column has any role in prediction or not. We found a lot of redundant columns. The redundancy was found such that if any column has almost all entry as same, then it might have less contribution in causing the disease. Also, if two or more columns are exactly identical then, then for this dataset only of the column is enough to represent all of the identical columns. So, based on these criteria 45 redundant columns were found out and removed from the training data, as they are just overheads. Data distribution for each of the column was plotted. Since, there are a lot of features a few plots are presented in this report as below:

Shape

Description automatically generated

Figure 2: Output distribution for individual columns Explanatory data analysis

## Model Building and Training

Model architecture building was the next step. So, after the basic EDA and preprocessing, an Artificial neural network architecture was built. The model we used is a sequential ANN model, having 14 Dense layer, and dropout layers in between to handle overfitting. After we built the basic architecture, the model was trained for 50 epochs and validation set was used for validating our model after each epoch. The validation size was 20 %. With this model we achieved an accuracy 93.78% on the validation set. The training and validation accuracies and losses are plot as below:

Chart, line chart

Description automatically generated

Figure 4: Accuracies plot

Chart, line chart

Description automatically generated

Figure 5: Losses Plot

## Hyperparameter Tuning using GridSearchCV

After this, the model was tuned to get the best hyperparameters. We used Gridsearch CV to get the optimal set of hyperparameters. And the gridsearch cv provided us with the best set of hyperparameters. The optimal set was found to be rmsprop, with a batch size of 32 and epochs of 50. The best training accuracy was 98.03 % and the best validation set accuracy came out to be 99.87%. The results are displayed as below:

Chart, line chart

Description automatically generated

Figure 6: Accuracy Plot

Chart

Description automatically generated

Figure 7: Losses plot

## We can see the accuracy graph is quite smooth this time after gridsearch CV also giving higher accuracy than previous ones. Same applies to the losses plot.

## Test data Analysis

The final model that we built with the hyperparameters we got from gridsearch Cv, was tested on the test data. For each of the disease class our model was giving good results. There were only true positives(except for a minor deviation) in model prediction which was visualized in the confusion matrix as follow:

A picture containing chart

Description automatically generated

Figure 8: Confusion matrix on test data

Chart

Description automatically generated

Figure 9: Confusion matrix on Validation data

For each of the disease we plotted the ROC curve and it was observed the plot indicates that our model is giving 100% accuracy. The perfect classifier graph can be identified from the ROC curves. Here for the classification we used One-Vs-Rest approach while plotting the ROC curve. Some of the ROC curves is displayed as follow:

Chart, line chart

Description automatically generated Figure 10: ROC Curve -1 Chart, line chart

Description automatically generated Figure 11: ROC Curve -2

## Comparative analysis with other models

After we have built the final ANN model and got an accuracy score of 99.87 %. We now compared the performance of some other models that are used for multiclass classification. Here, for comparison we took Random Forest, Logistics Regression and K Nearest Neighbors as comparative models. Although almost all of the model gave quite decent accuracy, RandomForest giving an accuracy of 85.7 %, Logistic Regression 97.61%, and K Nearest Neighbors 97.61%. The various confusion matrix for prediction of different model is displayed as follow:

A picture containing chart

Description automatically generated

Figure 12: Random Forest

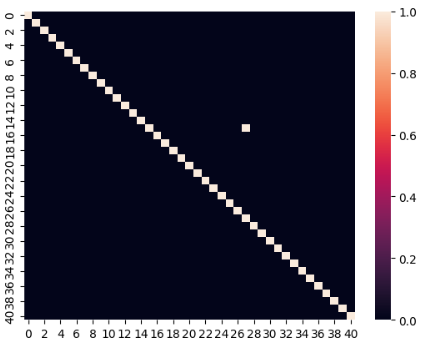


Figure 13: Logistic Regression

Shape

Description automatically generated

Figure 14: KNearest Neighbors

# Results and Findings

It is a noteworthy accomplishment in the field of medical diagnosis that the illness prediction ANN model was constructed with such great accuracy. The model has the potential to save lives by enabling early detection and prompt treatment due to its accuracy of 99.87% in predicting whether the disease will exist or not. The model's architecture and parameters were optimised through the use of hyperparameter tuning, which has proven beneficial in raising the model's accuracy.

Additionally, the ANN model's superiority has been demonstrated by comparisons with models like Random Forest, Logistic Regression, and K Nearest Neighbours. The ANN model has shown its promise as the go-to model for disease prediction, with the other models producing accuracy ratings of 85.7%, 97.61%, and 97.61%, respectively. The ROC-AUC curve's perfect classifier attribute adds to the model's accuracy and capacity to differentiate between positive and negative situations. Overall, the ANN model's performance is impressive and has the potential to make a substantial impact on the medical industry.

# Discussion and Future works

There are still a number of ways to enhance the performance of the constructed ANN model, despite the fact that it has demonstrated outstanding accuracy for disease prediction. Increasing the amount of data utilised to train the model is one strategy. The sample size used today might not be large enough to account for all the complexity and unpredictability in the data. More data can be added to a model to increase generalisation and decrease overfitting. Additionally, incorporating diverse datasets from various sources can aid in strengthening and improving the model.

Investigating how various feature selection methods affect the performance of the model is another strategy. The model can be made simpler and more effective by choosing the most useful and pertinent aspects. The performance of the model can be enhanced by using feature selection techniques like PCA and Lasso regularisation to find the dataset's most important variables.

The creation of more precise and reliable disease prediction models in the future has the potential to completely change the practise of medicine. More complex models, such deep learning models, can be created to increase the accuracy of disease prediction as data becomes more readily available. Additionally, combining data from various sources, including imaging, genomic, and electronic health records, can help create more detailed and individualised disease prediction models. Overall, the subject of disease prediction still has a lot of space for growth, and future research should concentrate on creating more precise and reliable models.

# Conclusion

The developed ANN model has proven to be remarkably accurate at predicting the existence or absence of various diseases based on their symptoms. The efficiency of the model's design and parameters has been confirmed by the hyperparameter tuning procedure and comparison with other models. However, by expanding the dataset and applying feature selection methods, further advancements can be made. The accuracy and performance of the model show promise for the future of medical diagnosis, with the potential to save lives through early detection and prompt treatment. Predicting various diseases based on their symptoms has the potential to greatly enhance patient outcomes and support individualised healthcare approaches.

# References

[1] P. R. Hegde, M. M. Shenoy, and B. H. Shekar, “Comparison of Machine Learning Algorithms for Skin Disease Classification Using Color and Texture Features,” *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2018, doi: <https://doi.org/10.1109/icacci.2018.8554512>.

[2] M. Fatima and M. Pasha, “Survey of Machine Learning Algorithms for Disease Diagnostic,” *Journal of Intelligent Learning Systems and Applications*, vol. 09, no. 01, pp. 1–16, 2017, doi: <https://doi.org/10.4236/jilsa.2017.91001>.

[3] M. Nilashi *et al.*, “Disease Diagnosis Using Machine Learning Techniques: A Review and Classification,” *Journal of Soft Computing and Decision Support Systems*, vol. 7, no. 1, pp. 19–30, Mar. 2020, Available: <http://www.jscdss.com/index.php/files/article/view/224>

[4] S. Grampurohit and C. Sagarnal, “Disease Prediction using Machine Learning Algorithms,” *2020 International Conference for Emerging Technology (INCET)*, Jun. 2020, doi: <https://doi.org/10.1109/incet49848.2020.9154130>.

[5] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, “Predicting the Onset of Diabetes with Machine Learning Methods,” *Journal of Personalized Medicine*, vol. 13, no. 3, p. 406, Feb. 2023, doi: <https://doi.org/10.3390/jpm13030406>.

[6] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, “Analysis of Breast Cancer Detection Using Different Machine Learning Techniques,” *Data Mining and Big Data*, pp. 108–117, 2020, doi: <https://doi.org/10.1007/978-981-15-7205-0_10>.

[7] D. Dahiwade, G. Patle, and E. Meshram, “Designing Disease Prediction Model Using Machine Learning Approach,” *IEEE Xplore*, Mar. 01, 2019. <https://ieeexplore.ieee.org/document/8819782>

[8] D. Devi, P. Sreevalli, Keerthana, Prathyusha, and M. Asia, “Prediction Of Diseases Using Random Forest Classification Algorithm,” 2020. <https://www.semanticscholar.org/paper/PREDICTION-OF-DISEASES-USING-RANDOM-FOREST-Devi-Sreevalli/2fbc6fb252d95f4b7423279154109a0c78cfd362> (accessed Mar. 26, 2023).

[9] D. H. Mantzaris, G. C. Anastassopoulos, and D. K. Lymberopoulos, “Medical disease prediction using Artificial Neural Networks,” *IEEE Xplore*, Oct. 01, 2008. <https://ieeexplore.ieee.org/abstract/document/4696782/> (accessed Apr. 04, 2022).

[10] A. Khemphila and V. Boonjing, “Heart Disease Classification Using Neural Network and Feature Selection,” *2011 21st International Conference on Systems Engineering*, 2017. <https://www.semanticscholar.org/paper/Heart-Disease-Classification-Using-Neural-Network-Khemphila-Boonjing/7780ba5f1550e8e19084ff07948ff91427fccaf9> (accessed Mar. 26, 2023).