Elena Esparza Baña

IA MODELS FOR DISEASE DETECTION

Abstract: In this paper, a comparison of several models, including decision tree, random forest, algorithms, as well as more complex architectures such as MLP, DNN, and CNN models, is presented. The methodology and techniques employed are thoroughly explained. Finally, a conclusion is provided, offering a critical perspective on the models and their practical utility. Additionally, both binary and multi-class classification tasks are analyzed and discussed.

INTRODUCTION

Artificial Intelligence applications are expanding rapidly due to advancements in these technologies. The primary area which scientists, engineers and everyday people are interested in is the Health Industry. Various AI models have been developed and applied for disease detection, prediction of illnesses based on health records, identifying relationships between symptoms and health factors. All of these can help patients and health professionals in improving patients quality of life and having better disease prevention and management.
Some research has been done in the use of AI models for disease detection such as Ghiasi et al., 2020 where a survey on coronary artery disease (CAD) classification is carried out using a classification and regression tree (CART). Moreover, the Ulla et al. (2022) research develops several models to make comparisons in the performance of diabetes detection concluding that k-nearest neighbor (KNN) surpassed the other models with extraordinary high accuracy , sensitivity, specificity, and ROC/AUC score"

METHODOLOGY AND ARCHITECTURE OF THE MODELS

DATASETS

The dataset employed for heart disease, diabetes and stroke detection is titled 'Dataset1.csv' which contains over 250000 samples of patients data. It comprises 22 attributes including characteristics like high blood pressure or BMI, alongside records indicating the presence of these diseases. Most of the features are categorical like 'stroke' and 'diabetes', and others like 'BMI', 'PhysHlth' are continuous values. Analysis preprocessing techniques were applied to prepare the dataset for the model. Target encoding was used on 'Age' and one-hot encoding on 'Diabetes'. Additionally, normalization of continuous characteristics was applied to ensure consistent scaling. The class distribution is represented in the chart below. Given the substantial imbalance between positive and negative classes, Synthetic Minority Over-Sampling Technique (SMOT) and undersampling were applied on the training dataset for better performance of the models.

	DIABETES	STROKE	HEART DISEASE
0	0.842	0.959429	0.905814
1	0.158	0.040571	0.094186

Apart from applying SMOTE, stratification is used in the splitting of the data into training and test sets, such as in the sets fed into the decision tree, random forest models for binary classification. Stratification prevents biased training and misleading evaluation ensuring the both training and testing maintain the same proportions of classes.

The second dataset, named 'PneumoniaMNIST' is a collection of pediatric chest X-Ray images where some of them present pneumonia disease, labeled as 1, while the ones from healthy patients are labeled as 0, no presence of the disease. The dimensions of the dataset accumulate 5332 grayscale images. It is split into training and validation set at a 9:1 ratio. In the following chart, details of both datasets are shown and a representation of the classes proportions within the training set is given:

Set	Samples	Healthy	Pneumonia
Training	4708	3494	1214
Test	624	390	234

The model used for the detection of pneumonia is a convolutional neural network model. Therefore, the values of the images had been normalized and converted into np.arrays to be fitted into the it, being this one of the preprocessing techniques used on the dataset. Moreover, the SMOTE method has been applied to the training set for a better performance of the model.

BINARY CLASSIFICATION

- DECISION TREE

The first model developed is a decision tree classifier. This algorithm splits the data step by step, from the root to the leaves, where predictions are made. To control its complexity and avoid overfitting, the maximum depth per leaf is set to 5, and at least 10 samples are required to make a split. These limits help the model focus on important patterns and prevent it from learning noise in the data. Additionally, 'Class_weight' is set due to the imbalance presented. The dataset is split 70% for training and 30% for testing, which helps the model generalize better by learning from the majority of the data while being evaluated on unseen samples (Ghiasi et al., 2020)

A second tree classifier was implemented, using entropy instead of Gini impurity to evaluate splits. While Gini impurity measures the likelihood of incorrect classification, entropy focuses on minimizing uncertainty in the splits. On The other hand, entropy measures the uncertainty of the target variable after a split which the tree seeks to minimize too (Aviny, et at.,2023).

- RANDOM FOREST BI-CLASSIFIER

A Random Forest model was designed for disease detection, consisting of 200 decision trees with a maximum depth of 10. Each tree splits the dataset at each node using the entropy criterion, when there are at least 10 samples. The model also adjusts class weights to handle class imbalance. Undersampling of the dataset has been applied to reduce the number of samples from the majority class. However, in the code submitted, the contrast of the performance of the model between doing undersampling of the data or training the model with the training dataset without reducing it can be analyzed.

- MLP

A multilayer perceptron is an artificial neural network composed of multiple layers of neurons for binary classification. It learns complex relationships by passing the input through the layers adjusting the weights using backpropagation during the training. As an optimizer it uses stochastic gradient descent. The architecture includes an input layer with 12 neurons to match the number of features selected for the problem. Then, the hidden layers, the first one, about 200 neurons and the second one about 100 neurons and a ReLU activation function is applied. Finally, the output layer consists of a single neuron and the activation function is the sigmoid.

MULTICLASS CLASSIFICATION

- DEEP NEURAL NETWORK

A neural network with multiple layers of neurons consisting of an input, hidden and output layer forming a sequential model. 12 neurons in the input layer, for each of the features. The hidden part of the model includes two layers, with 12 and 8 neurons, respectively, both using the ReLU activation function. The output layer consists of 4 neurons, one for each class, with a softmax activation function applied to generate output probabilities.
The compilation of the model is done by applying categorical cross-entropy for multi classification. The training lasts 150 epochs with a batch of 30. The training is visualized by plotting accuracy and loss across epochs

BINARY CLASSIFICATION FOR PNEUMONIA DETECTION PROCESSING OF IMAGES

- CONVOLUTIONAL NEURAL NETWORK MODEL

CNN, a series of convolutional layers are implemented to process the images as a binary classification model for detection of pneumonia disease. The first layer of the model uses 32 filters of size 3x3 and applies the ReLU activation function. The input shape of the model is the pixel size. Then, a max-pooling layer that applies downsampling on the image reduces its spatial dimension by half. The second convolutional layer uses 64 filters of size 3x3, followed by another max-pooling layer to enhance feature retention. A third convolutional layer with 64 filters of the same size completes the feature extraction stage. These convolutional layers are crucial for extracting key image features such as edges and textures, which the model uses to classify the images. After the convolutional layers, there is the flatten layer that transforms the 3d tensor into a unidimensional vector to be fed into the dense layer.
The dense layer is a fully connected layer with 64 units that uses the ReLU activation function to learn feature combinations extracted by the preceding layers. It is followed by another dense layer of a single unit corresponding to the model's output. It applies the sigmoid activation function to obtain the probability classifying it.
The model is trained using Adam optimizer, allowing it to adapt the learning rate for each parameter individually, and binary cross entropy loss, to quantify its performance, most common for binary classification, which calculates how far the predicted probability is from the actual label (Maity et al. 2024)

ANALYZE AND DISCUSSION

There are 5 AI models in total for binary and multiclass classification. Respecting binary classification for the detection of stroke, diabetes and Heart disease, the same selected features had been set to train each of the models: DT,with gini impurity and entropy criterion, RF, entropy, and MLP network. The influence of these factors are mention in 2004 INTERHEART study [13] cited in Mohammad M. Ghiasi (2020) including smoking, hypertension, daily consumption of vegetables

and fruits, sex and age. This explains the consistent feature selection. While having a variety of models, having different sets of features would have brought a more complex understanding of the features that best suited each model and disease.

The performance of the DT and RF models, despite their simplicity, on the training datasets without applying undersampling or oversampling was highly accurate. In some of the experiments, the precision and recall percentages for both classes-particularly for the majority class 0- were significantly high. The MLP model performed as well with the majority class achieving 73 007 of true positives and precision of 0.96. However, the precision for class1 initially was zero.After applying undersampling the precision increased to 0.08 and the recall to 0.75 .

Regarding the DNN model for multi-classification, its performance improves with the application of preprocessing techniques like SMOTE, one-hot encoding and normalization of the features. The classes were assigned as in the following figure where the performance of the model on each class can be seen.

Accuracy: 58.78%

	precision	recall	f1-score	support
0	0.92	0.65	0.76	58178
1	0.12	0.34	0.18	3097
2	0.20	0.37	0.26	5930
3	0.24	0.41	0.30	8899

For the evaluation of the CNN model for processing images the accuracy is calculated using the classification report function from sklearn to obtain the percentage of correct predictions. With the initial dataset, this metric was worse than the SMOTE training set. As shown in the figure, 88% of the predictions are correct. The F1-score is 81% for the class 0, healthy and 91% for class 1, presence of pneumonia. The model could be improved by adding the layers or filters and adjusting the learning rate.

Classification Report:

	precision	recall	f1-score	support
Class 0	0.98	0.68	0.81	234
Class 1	0.84	0.99	0.91	390
accuracy			0.88	624
macro avg	0.91	0.84	0.86	624
weighted avg	0.89	0.88	0.87	624

CONCLUSION

In conclusion, the AI models were implemented for binary and multi-class classification, targeting stroke, diabetes, and heart disease detection using a consistent set of relevant features such as smoking, hypertension, and dietary habits. DT and RF models performed well without resampling, while MLP improved minority class precision and recall after undersampling. DNN performance improved through preprocessing techniques like SMOTE and normalization. The CNN model achieved 88% accuracy, with F1-scores of 81% for healthy cases and 91% for pneumonia. Further improvements could be made by adjusting its architecture and learning rate.

REFERENCES

S. Maity, A. Thakur, M. Faujdar, A. Kumar, and A. Singh, “Medical Image Classification: Binary Classification of X-Ray Images to Classify Fractures in Bones,” in *Lecture Notes in Networks and Systems.*, Springer, Singapore, 2024, p. pp 331–341. Accessed: Jan. 13, 2025. [Online]. Available: https://doi.org/10.1007/978-981-97-7360-2_29

M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, “Decision tree-based diagnosis of coronary artery disease: CART model,” Computer Methods and Programs in Biomedicine, vol. 192, p. 105400, Aug. 2020, doi: https://doi.org/10.1016/j.cmpb.2020.105400.

Aviny, H. R., Ghasemi, M., Fazlazad, M. (2023). 'Cardiovascular Disease Diagnosis Using the Combination of Principal Component Analysis Algorithm and Regression Tree', Transactions on Machine Intelligence, 6(2), pp. 114-125. doi: 10.47176/TMI.2023.114

Ullah, Z. et al. (2022) 'Detecting high-risk factors and early diagnosis of diabetes using machine learning methods', *Computational Intelligence and Neuroscience*, 2022, pp. 1–10. doi: https://doi.org/10.1155/2022/2557795.