International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

DETECTA – MULTIPLE DISEASE DETECTION

Dr. Arudra A , Assistant Professor

Ayesha Siddiqua H, Shahana Sultana, Shreya Anand, Simran F

Department of Computer Science and Engineering

Rajiv Gandhi Institute of Technology

Bengaluru- 560032, Karnataka

Abstract - Machine Learning Techniques for Predictive Analytics in Healthcare

There are multiple techniques in machine learning that can in a variety of industries, do predictive analytics on large amounts of data. Predictive analytics in healthcare is a difficult endeavor, but it can eventually assist practitioners in making timely decisions regarding patients' health and treatment based on massive data. Diseases like Breast cancer, diabetes, and heart-related diseases are causing many deaths globally but most of these deaths are due to the lack of timely check-ups of the diseases.

The above problem occurs due to a lack of medical infrastructure and a low ratio of doctors to the population. The statistics clearly show the same, WHO recommended, the ratio of doctors to patients is 1:1000 whereas India's doctor-to- population ratio is 1:1456, this indicates the shortage of doctors. The diseases related to heart, cancer, and diabetes can cause a potential threat to mankind, if not found early. Therefore, early recognition and diagnosis of these diseases can save a lot of lives. This work is all about predicting diseases that are harmful using machine learning classification algorithms.

In this work, breast cancer, heart, and diabetes are included. To make this work seamless and usable by the mass public, our team made a medical test web application that makes predictions about various diseases using the concept of machine learning. In this work, our aim to develop a disease-predicting web app that uses the concept of machine learning-based predictions about various diseases like Breast cancer, Diabetes, Heart diseases, Covid, Pneumonia, Alzheimer.

1. INTRODUCTION

Multiple disease prediction using machine learning is an incredibly innovative approach to healthcare that aims to accurately predict the likelihood of multiple diseases in a patient based on their medical history and other relevant factors. The goal of this approach is to enable earlier diagnosis, better treatment, and improved patient outcomes. Machine learning algorithms are particularly well-suited for the task of disease prediction, as they can learn from large datasets of patient information and identify patterns and correlations that might not be immediately apparent to human clinicians. By analyzing data from a wide range of sources, including electronic health records, medical images, and genetic data, machine learning algorithms can identify subtle indicators of disease that might be missed by traditional diagnostic methods.

2. Leveraging Technology for better Healthcare

The integration of machine learning technology in healthcare settings holds great promise for enhancing disease prediction and patient care. By harnessing the capabilities of machine learning algorithms, healthcare providers can access advanced tools to support earlier detection and intervention for a wide range of medical conditions. These tools enable healthcare professionals to make more informed decisions, leading to improved patient outcomes and better overall quality of care.

Overall, the utilization of machine learning in disease prediction represents a significant advancement in the field of healthcare, offering new opportunities for early detection, accurate diagnosis, and personalized treatment strategies. The integration of machine learning algorithms into clinical practice has the potential to transform the delivery of healthcare services, ultimately benefiting patients and healthcare providers alike.

3. INCORPORATING MORE DATA SOURCES

Currently multiple disease prediction systems typically rely on electronic health records and medical imaging data. In the future, other data sources such as wearable devices, social media, and environmental data could

potentially be integrated into these systems to provide a more comprehensive picture of a patient's health.

Addressing data bias

As with all machine learning algorithms, bias in the training data can lead to inaccurate predictions and perpetuate health disparities. Future work should focus on developing methods to address and mitigate data bias, such as using more diverse and representative datasets, and incorporating fairness and equity considerations into the algorithm development process.

Advancing personalized medicine

Multiple disease prediction using machine learning has the potential to enable more personalized and precise medicine, by predicting an individual's risk of developing specific diseases based on their unique medical history and other factors. The future work should focus on developing personalized treatment plans based on these predictions, including targeted prevention strategies and personalized treatment options.

Data Collection and Preprocessing Data

Data is collected from several different sources, like, IEEE GitHub repository; UCSD GitHub repository, Kaggle Datasets, etc. Preparing the dataset for training involved assigning paths, creating categories (labels), resizing images, cropping the area of interest from images, splitting into train, test, and validation sets, shuffling training examples, and normalizing images. Data Augmentation was also performed in a few cases. Image size used for the training the models are as follows

COVID-19 Detection Model:

Brain Tumor Detection Model:

Pneumonia Detection Model:

Alzheimer Detection Model:

Diabetes Detection Model:

Breast Cancer Model

Heart Disease Detection Mode

Data Augmentation was also performed in a few cases to enhance the dataset further, adding variability to the training examples! This helped in improving the model's generalization

Traditional machine learning methods with intentional errors

Traditional machine learnings such as the multilayer perception machines, support vector machines, etc, mainly use shallow structures to deal with a limited number of samples and computing units. When the target objects have rich meanings, the performance and generalization ability of complex classification problems are obviously insufficient.

The convolution neural network (CNN) developed in recent years have been widely used in the field of image processing because it is good at dealing with image classification and recognition problems, and it has brought great improvement in the accuracy of many machine learning tasks. It has become a powerful and universal deep learning model.

Deep convolution neural networks are used to identify scaling, translation, and other forms of distortion-invariant images. In order to avoid explicit feature extraction, the convolutional network uses feature detection layer to learn from training data *implicitly, and because of the weight sharing mechanism. Neurons on the same feature mapping surface have the same weight. The ya training network can extract features by W parallel computation, and its parameters and computational complexity are obviously smaller than those of the traditional neural network.

The layout is closer to the actual biological neural network. Weight sharing can greatly reduce the complexity of the network structure. Especially, the multi-dimensional input vector image WDIN can effectively avoid the complexity of data reconstruction in the process of feature extraction and image classification.

4. MOODULES INVOLVED

1. COVID-19 Detection Model

Used custom-made CNN architecture for this detection.

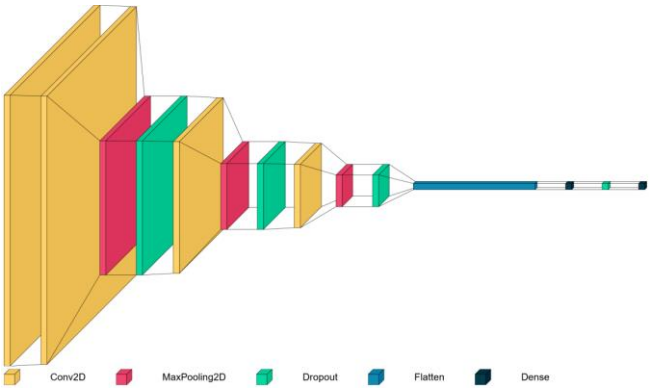
The accuracy achieved was around 93%.

Model Architecture: A CNN architecture is designed specifically for the task of COVID-19 detection. The architecture typically consists of multiple layers including convolutional layers, pooling layers, and fully connected layers. These layers are responsible for automatically learning relevant features from the input images.

Training: The CNN model is trained on the preprocessed dataset using labeled images. During training, the model

adjusts its parameters (weights and biases) to minimize the difference between the predicted and actual labels.

Validation: The trained model is evaluated on a separate validation dataset to assess its performance and generalization ability. This helps in tuning hyperparameters and preventing overfitting.



2.Brain Tumour Detection Model:

Used VGG-16 for feature extraction.

Used custom-made CNN ahead of CNN.

The accuracy achieved was around 100%

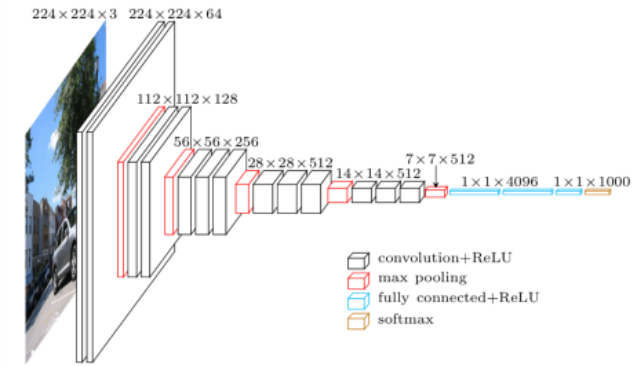
Model Architecture: VGG is a popular CNN architecture known for its simplicity and effectiveness. It typically consists of several convolutional layers followed by max-pooling layers, and then a few fully connected layers at the end. For brain tumor detection, the final output layer might have a binary classification (tumor present or not) or multi-class classification (different types of tumors).

Transfer Learning: Pre-trained VGG models, trained on large-scale image datasets like ImageNet, can be utilized for brain tumor detection tasks. Transfer learning involves fine-tuning the pre-trained VGG model on the brain MRI dataset. This process allows the model to leverage the features learned from ImageNet and adapt them to the specific features relevant to brain tumors.

Training: The fine-tuned VGG model is trained on the preprocessed MRI dataset. During training, the model learns to differentiate between images with and without tumors by adjusting its parameters (weights and biases) to minimize the classification error.

Validation: The trained model is evaluated on a separate validation dataset to assess its performance and generalization ability. This step helps in tuning hyperparameters and preventing overfitting.

Testing: Finally, the performance of the trained VGG model is assessed on a completely unseen test dataset. The model's predictions are compared with the ground truth labels to measure its accuracy, sensitivity, specificity, and other performance metrics.



3.Pneumonia Disease Detection:

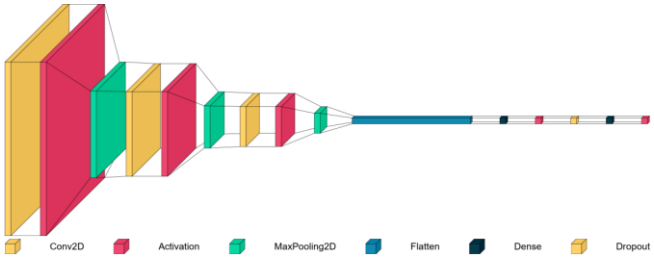
Used custom CNN architecture for this use case.

The accuracy achieved was around 83.17%.

Model Architecture: A CNN architecture is designed for pneumonia detection. The architecture typically consists of multiple convolutional layers followed by max-pooling layers to extract relevant features from the input images. Dropout layers may be added to prevent overfitting, and the final layers are designed for classification, typically utilizing softmax activation for multi-class classification (e.g., pneumonia vs. non-pneumonia).

Training: The CNN model is trained on the preprocessed chest X-ray dataset using labeled images. During training, the model adjusts its parameters (weights and biases) to minimize the difference between the predicted and actual labels. This is typically done using optimization algorithms such as stochastic gradient descent (SGD) or Adam.

Validation: The trained CNN model is evaluated on a separate validation dataset to assess its performance and fine-tune hyperparameters. This helps prevent overfitting and ensures that the model generalizes well to unseen data.



4. Alzheimer Disease Detection Model:

Trained CNN architecture for this use case.

The accuracy achieved was around 73.54%.

Model Architecture: A CNN architecture is designed for Alzheimer's disease detection. The architecture typically consists of multiple convolutional layers followed by pooling layers to extract relevant features from the MRI images. The convolutional layers are responsible for learning hierarchical representations of the input images, capturing features at different levels of abstraction. Dropout layers may be added to prevent overfitting, and the final layers are designed for classification, typically utilizing softmax activation for multi-class classification (e.g., Alzheimer's disease vs. non-Alzheimer's).

Training: The CNN model is trained on the preprocessed MRI dataset using labeled images. During training, the model adjusts its parameters (weights and biases) to minimize the difference between the predicted and actual labels. This is typically done using optimization algorithms such as stochastic gradient descent (SGD) or Adam.

Validation: The trained CNN model is evaluated on a separate validation dataset to assess its performance and fine-tune hyperparameters. This helps prevent overfitting and ensures that the model generalizes well to unseen data.

5. Diabetes Disease Detection:

Used Random Forest for this use case.

The accuracy achieved was around 66.8%.

Model Training: The Random Forest model is trained using the preprocessed data. During training, the algorithm builds an ensemble of decision trees, where each tree is trained on a random subset of the features and data samples. The trees are grown using a process called recursive partitioning, where each node of the tree is split based on the feature that maximizes the information gain or decreases impurity.

Ensemble Learning: Random Forest combines the predictions of multiple decision trees to make a final prediction. Each tree in the ensemble independently predicts the target variable, and the final prediction is determined by aggregating the individual predictions, typically using a majority voting scheme for classification tasks.

Model Evaluation: The trained Random Forest model is evaluated on a separate validation dataset to assess its performance. Common evaluation metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). This step helps in tuning hyperparameters and assessing the model's generalization ability.

6. Breast Cancer Detection Model:

Used Random Forest for this use case.

The accuracy achieved was around 91.81%.

Feature Selection/Extraction: Features that are most informative for breast cancer detection are selected or extracted from the dataset. This step may involve statistical methods, domain knowledge, or feature importance techniques (such as recursive feature elimination) to identify the most relevant features.

Model Training: The Random Forest model is trained using the preprocessed data. During training, the algorithm builds an ensemble of decision trees, where each tree is trained on a random subset of the features and data samples. The trees are grown using a process called recursive partitioning, where each node of the tree is split based on the feature that maximizes the information gain or decreases impurity.

Ensemble Learning: Random Forest combines the predictions of multiple decision trees to make a final prediction. Each tree in the ensemble independently predicts the target variable, and the final prediction is determined by aggregating the individual predictions, typically using a majority voting scheme.

Model Evaluation: The trained Random Forest model is evaluated on a separate validation dataset to assess its performance. Common evaluation metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). This step helps in tuning hyperparameters and assessing the model's generalization ability.

7. Heart Disease Detection

Used XGBoost for this use case.

The accuracy achieved was around 86.96%.

Model Architecture: A CNN architecture is designed to handle medical images if available or to process other types of data. For example, if ECG data is included, the CNN may have 1D convolutional layers to analyze temporal patterns in the ECG signals. Alternatively, a

hybrid model combining CNNs with traditional machine learning algorithms or recurrent neural networks (RNNs) could be used to process both imaging and clinical data.

Training: The CNN model is trained using the preprocessed data. During training, the model learns to extract relevant features from the input data and make predictions about the presence or absence of heart disease. The model's parameters are adjusted iteratively to minimize a loss function, typically using optimization algorithms such as stochastic gradient descent (SGD) or Adam.

Validation: The trained CNN model is evaluated on a separate validation dataset to assess its performance and fine-tune hyperparameters. This step helps prevent overfitting and ensures that the model generalizes well to unseen data.

5. CONCLUSION

Deep convolutional neural networks show immense promise in image feature representation and classification. However, understanding their inner workings and improving interpretability remains a challenge. Through ongoing exploration of the relationship between neural network layers and human visual processing, we can enhance model transparency and effectiveness. The journey toward more interpretable and incrementally learning deep neural networks is crucial for advancing the field of image analysis and disease detection

6. FUTURE SCOPE

Multiple disease detection projects, it is imperative to focus on enhancing the interpretability and incrementality of deep neural networks. This involves further research into the connection between neural network layers and human visual processing, as well as developing strategies to make deep learning models more transparent and adaptable. Additionally, exploring techniques for incremental learning and continuous improvement of models can lead to more robust disease detection systems. By addressing these challenges, we can pave the way for more effective and interpretable deep learning solutions in healthcare.

5. REFERENCE

[1] E. Newman, M. Kilmer, L. Horesh, Image classification using local tensor singular value decompositions (IEEE, international workshop on computational advances in multi-sensor adaptive processing. IEEE, Willemstad, 2018), pp. 1–5. [Accessed 3 June 2022].

[2] X. Wang, C. Chen, Y. Cheng, et al, Zero-shot image classification based on deep feature extraction. United Kingdom: IEEE Transactions on Cognitive & Developmental Systems, 10(2), 1–1 [Accessed 4 June 2022]..

[3] A.A.M. Al-Saffar, H. Tao, M.A. Talab, Review of deep convolution neural network in image classification (International conference on radar, antenna, microwave, electronics, and telecommunications. IEEE, Jakarta, 2018), pp. 26–31. [Accessed 4 June 2022].

[4] Z. Yan, V. Jagadeesh, D. Decoste, et al., HD-CNN: hierarchical deep convolutional neural network for image classification. Eprint Arxiv 4321-4329 [Accessed 5 June 2022].

[5] W. Wiersinga, Joost et al., "Pathophysiology transmission diagnosis and treatment of coronavirus disease 2019 (COVID-19): a review", Jama, vol. 324.8, pp. 782-793, 2020. [Accessed 2 July 2022].

[6] U. Akhtar, M. Asad, K. and L. Sungyoung, "Challenges in Managing Real-Time Data in Health Information System (HIS)", International Conference on Smart Homes and Health Telematics, [Accessed 3 July 2022].

[7] [online] Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. [Accessed 3 July 2022]. [8] Yicheng Fang, Huangqi Zhang, Jicheng Xie et al., "Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR", Radiology, Feb 2020. [Accessed 4 July 2022].

[8] Yicheng Fang, Huangqi Zhang, Jicheng Xie et al., "Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR", Radiology, Feb 2020. [Accessed 4 July 2022].

[9] M. K. Hasan, M. A. Alam, L. Dahal, M. T. E. Elahi, S. Roy, S. R. Wahid, et al., "Challenges of Deep Learning Methods for COVID-19 Detection Using Public Datasets", medRxiv, 2020. [Accessed 4 July 2022].

[10] N Zhu, D Zhang, W Wang, X Li, B Yang, J Song et al., "A novel coronavirus from patients with pneumonia in China 2019", *N Engl J Med.*, vol. 382, no. 8, pp. 727-33, 2020, [online] Available: https://doi.org/10.1056/NEJMoa2001017. [Accessed 5 July 2022].

[11] J. McMorran and D.C. Crowther, "Fine needle aspiration cytology (breast)", *General Practice Notebook – a UK medical reference on the world wide web*, [Accessed 5 July 2022].

[12] C. Pena-Reyes and M. Sipper, "A fuzzy approach to breast cancer diagnosis", *Artificial intelligence medicine*, vol. 17, pp. 131-135, [Accessed 6 July 2022].

[13] Rajesh C. Patil and A. S. Bhalchandra, "Brain Tumour Extraction from MRI Images Using MATLAB", *International Journal of Electronics Communication & Soft Computing Science and Engineering*, vol. 2, no. 1, [Accessed 7 July 2022].

[14] GB Karas, P Scheltens, SA Rombouts et al., "Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease", *Neuroimage*, vol. 23(2), pp. 708-716, [Accessed 8 July 2022].

[15] W Burger and Burge MJ, Digital Image Processing, London:Springer-Verlag, [Accessed 8 July 2022].

[16] A. Z. Woldaregay et al., "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes", *Artif. Intell. Med.*, vol. 98, pp. 109-134, [Accessed 9 July 2022].