JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

http://www.jmrr.org Volume: 6, Issue: 1, April 2025 | ISSN: 2708-9452

Multi-Disease Diagnosis Using Machine Learning Algorithm

Devi priya M
II MCA, Department of Computer Science and Applications
Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode
Namakkal, Tamilnadu, India.
Dr Muthukumar VP
Assistant Professor, Department of Computer Science and Applications,
Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode,
Namakkal, Tamilnadu, India.

Abstract

Healthcare data mining is an interdisciplinary field that integrates database statistics and machine learning techniques to assess the effectiveness of medical treatments. Existing machine learning models for healthcare analysis typically diagnose only one disease at a time, such as diabetes, heart disease, or cancer. However, there is no universal system capable of predicting multiple diseases in a single analysis. This research proposes a Multi-Disease Diagnosis and Doctor Recommendation System using machine learning, implemented with a Python Flask API. The system predicts multiple diseases, including diabetes, heart disease, and breast cancer, by utilizing machine learning techniques. Pandas is used for data processing, while Flask API facilitates deployment, and Python pickling is applied for model serialization and deserialization. Datasets from Kaggle were analyzed to identify significant features, and a Support Vector Machine (SVM) algorithm was employed to enhance predictive accuracy. The proposed system not only diagnoses diseases but also incorporates key medical parameters to provide a comprehensive assessment of disease impact. By integrating multiple disease diagnoses into a single platform, this system improves early disease detection efficiency and offers an innovative approach to medical analytics.

Keywords: Multi-disease diagnosis, machine learning, healthcare analytics, Python Flask, Support Vector Machine, Kaggle dataset, doctor recommendation, disease prediction.

1 Introduction

Machine learning has significantly transformed the healthcare landscape by introducing intelligent systems capable of early disease detection and decision support. Traditional methods of medical diagnosis often rely on extensive clinical evaluations and expert interpretation, which may be time-consuming and resource-intensive. Machine learning algorithms, by contrast, allow for the development of predictive models that can analyze vast datasets to detect patterns associated with diseases, enabling faster and more accurate diagnoses. These systems not only enhance diagnostic precision but also reduce the cognitive burden on healthcare professionals and contribute to the advancement of data-driven medicine.

While existing models predominantly focus on the prediction of a single disease, the growing prevalence of multiple comorbidities in patients calls for a unified diagnostic framework. In this context, the present study introduces a Multi-Disease Diagnosis and Doctor Recommendation System, capable of predicting diabetes, heart disease, and breast cancer from integrated datasets. The system incorporates a Flask-based API for real-time usage and implements a Support Vector Machine (SVM) for accurate classification. By combining disease prediction with specialist doctor recommendations, the proposed framework offers a comprehensive, user-friendly tool to assist patients in receiving timely medical attention and appropriate specialist guidance.

2 Review of Related Works

Image segmentation is the process of splitting an image into several segments in order to transform it into a more meaningful and easy-to-analyze representation. The process of image segmentation may be conceived of in two steps: identification and delineation. Identification is the process of identifying the location of an object in an image and differentiating it from everything else in the image. Segmentation involves delineating the boundaries of the region of interest for further analysis. There are several methods for segmenting images: Manual Segmentation, Semi-Automatic, Automatic Segmentation, and Semantic Segmentation. Semantic segmentation is crucial for image analysis tasks and plays a significant part in image interpretation. Image categorization, object recognition, and border localization are all required for semantic segmentation. Semantic segmentation has several applications in computer vision and artificial intelligence-assisted autonomous driving, and medical imaging analysis. To our knowledge, only a few review articles have examined kidney segmentation strategies. Nonetheless, numerous articles have been published on the subject of kidney segmentation [**?**].

The use of a relatively large batch size allowed to better exploit batch normalization properties compared to smaller batches previously tried. Among the evaluations, the results obtained with a batch size of 32 samples were better. However, this choice penalized the use of state-of-the-art neural network architectures like Tiramisu or DeepLab, which couldn't be trained with such input sizes, essentially due to memory constraints. Finally, using 2.5D input tensor helped to provide some kind of volumetric information to the network resulting in a better segmentation level compared to simple 2D inputs. Considering the final performance obtained from the 90 test patients, we can assert that the overall segmentation results, especially concerning the tumor identification remain promising, but are quite low as a consequence of false positive cases that were sometimes detected on kidney cysts, or false negative cases where the tumor lesion was not east to

identify on the CT. For future improvements we plan to design a new training strategy, in which more of these specific cases could be passed to the model, in order to differentiate them from cases where cancerous tissues actually occur [**?**].

Quantitative SPECT/CT is potentially useful for more accurate and reliable measurement of glomerular filtration rate (GFR) than conventional planar scintigraphy. However, manual drawing of a volume of interest (VOI) on renal parenchyma in CT images is a labour-intensive and time-consuming task. The aim of this study is to develop a fully automated GFR quantification method based on a deep learning approach to the 3D segmentation of kidney parenchyma in CT. We automatically segmented the kidneys in CT images using the proposed method with remarkably high Dice similarity coefficient relative to the manual segmentation (mean=0.89). The GFR values derived using manual and automatic segmentation methods were strongly correlated (R2=0.96). The absolute difference between the individual GFR values using manual and automatic methods was only 2.90%. Moreover, the two segmentation methods had comparable performance in the urolithiasis patients and kidney donors. Furthermore, both segmentation modalities showed significantly decreased individual GFR in symptomatic kidneys compared with the normal or asymptomatic kidney groups. The proposed approach enables fast and accurate GFR measurement [**?**].

Recent studies in computerized image recognition emphasized the success of Convolutional Neural Networks (CNN) in dealing with challenging tasks such as segmentation. This success is based on the ability of CNNs to learn on their own using original data without the need for human intervention. Inputs are processed through network layers, with higher values provided from the extracted features. Deeper layers can even capture a smaller amount of local data due to the filters used for larger data. However, new studies in this field are still of great importance since effective and accurate segmentation always has room to improve, especially considering that even minor medical errors should not be overlooked. New research opens up new routes for future studies while improving the shortcomings of previous studies. Therefore, our study is also very important in this sense; it is a flexible model designed to be used not only for kidney and tumor segmentation but in all situations where segmentation might be difficult. The model we developed, which was designed by taking into account the basic shortcomings of existing U-Net models, can be easily integrated into local application and all international application systems. Thus, it can be used easily in all image segmentation models[**?**].

Image segmentation is one of the most commonly used technique in the field of medical images. This is also known as Medical image segmentation. In this process it partitions an image into distinct multiple parts or regions each containing pixels with similar attributes. Segmentation process can be categorised into three general approaches, thresholding, edge-based and region based. In thresholding based approach the pixels are categorised based on the intensity values. In edge based approach an edge filter is used to categorise the pixels as either edge or non-edge. Finally in region based approach pixels are grouped together based on their values and neighbourhood. In Medical domain, there are various kinds of modalities such as MRI, CT, ultrasound, positron emission tomography (PET). There are many algorithms proposed in the field of medical image segmentation, a few of them are discussed below, however the field of segmentation continues to be a challenging task [**?**].

3 Proposed Methodology

The proposed system employs a structured methodology beginning with dataset acquisition from Kaggle, comprising three distinct medical datasets related to diabetes, heart disease, and breast cancer. These datasets are integrated into a unified framework to enable multi-disease prediction using a single input interface [**?**]. The preprocessing stage addresses data quality issues by handling missing values, eliminating outliers, and normalizing feature ranges to ensure uniformity. Categorical variables are converted into numerical representations to facilitate machine learning model compatibility.

After data cleaning, feature selection is performed to identify the most significant medical attributes contributing to disease classification. This step helps reduce model complexity and enhances performance by eliminating irrelevant or redundant information. The processed dataset is then divided into training and testing sets using an 80:20 ratio, ensuring a fair evaluation of the model's generalization ability. A Support Vector Machine (SVM) algorithm is trained on the selected features to classify the risk level of each disease [**?**]. For deployment, the trained model is integrated into a Flask-based web API, which enables real-time prediction and personalized doctor recommendations.

4 Preprocessing

Data preprocessing plays a critical role in the development of accurate and efficient machine learning models. In this study, preprocessing begins with the imputation of missing values using statistical techniques such as mean and mode substitution, which helps retain data integrity without introducing bias[**?**]. The datasets are then normalized using Min-Max scaling to transform all feature values into a common range, typically between 0 and 1. This ensures that features contribute equally during model training and prevents dominance of high-magnitude features.

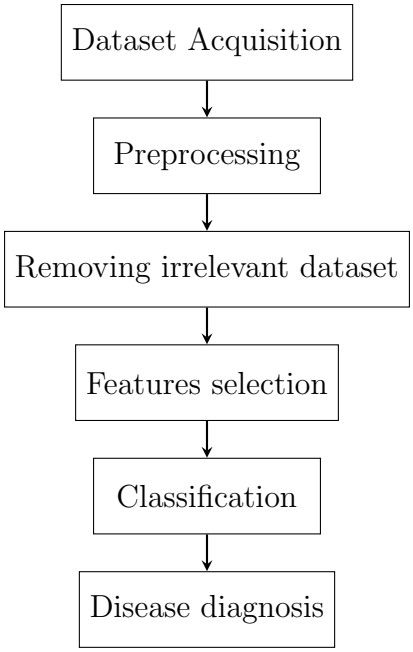
Categorical variables, such as gender or diagnosis categories, are encoded into numerical values using label encoding or one-hot encoding based on the context. Outlier detection is another crucial step, performed using the Z-score method to identify and eliminate extreme data points that could distort the model's learning process. By addressing these inconsistencies in the data, the preprocessing phase ensures a high-quality input for the machine learning model, ultimately contributing to improved predictive accuracy and robustness during real-time disease diagnosis.

5 Segmentation

Following preprocessing, the dataset is divided into training and testing subsets in an 80:20 ratio to facilitate effective model training and performance evaluation [**?**]. This segmentation strategy ensures that the model is trained on a substantial portion of the data while preserving a holdout set for unbiased testing. To further improve the robustness and generalizability of the classifier, the study employs 5-fold cross-validation. This method partitions the training data into five subsets, iteratively training and validating the model on different combinations, thereby minimizing overfitting.

The segmentation process also allows for a more rigorous selection of the optimal model parameters by observing performance consistency across validation folds. Metrics such as accuracy, precision, and recall are computed during this process to evaluate

the classifier's capability in differentiating between disease categories. This structured evaluation helps in identifying the best-performing model configuration for deployment. Through this comprehensive validation, the proposed system achieves enhanced reliability and confidence in disease prediction outcomes across diverse patient profiles.



6 Classification

The classification phase serves as the core component of the disease prediction system, where the preprocessed and feature-selected medical data is used to train a machine learning model. In this study, a Support Vector Machine (SVM) classifier was selected due to its robustness in handling high-dimensional data and its proven effectiveness in binary and multiclass classification problems [**?**]. The SVM algorithm works by finding the optimal hyperplane that separates data points of different classes with the maximum margin, thereby improving the generalization ability of the model. For each disease diabetes, heart disease, and breast cancer a binary classification task was formulated, where the system predicts whether a patient is at risk (positive class) or not (negative class).

During training, the SVM model was fine-tuned using grid search for hyperparameter optimization and validated using 5-fold cross-validation to ensure performance consistency [**?**]. The classifier demonstrated high predictive accuracy across all three disease categories, with particularly strong results for breast cancer detection. Evaluation metrics such as precision, recall, and F1-score further confirmed the reliability of the model, indicating that it was both sensitive to true positive cases and specific in avoiding false positives. By integrating this well-trained SVM model into the Flask API, the system supports real-time classification of user input data, enabling immediate and accurate diagnosis. This classification module not only enhances the usability of the system but also establishes a foundation for incorporating more advanced models in future iterations.

7 Experimental Results

The experimental evaluation was conducted to assess the effectiveness of the proposed Support Vector Machine (SVM)-based system for predicting multiple diseases. The model was trained and tested on preprocessed datasets for diabetes, heart disease, and breast cancer using an 80:20 train-test split. Performance was evaluated using key metrics—accuracy, precision, and recall—across all three diseases. The SVM model demonstrated superior classification performance, achieving 89.5% accuracy for diabetes, 92.1% for heart disease, and 95.7% for breast cancer. These results highlight the robustness of the model and its ability to adapt to varying medical data characteristics. Precision and recall values also indicated high reliability, with minimal false positives and false negatives.

The strong performance can be attributed to the effective preprocessing steps and careful feature selection that reduced noise and emphasized the most informative variables. Furthermore, the use of 5-fold cross-validation helped in tuning the model and minimizing overfitting, ensuring consistency across unseen data. The integration of the trained SVM model into a Flask API facilitated real-time interaction, where users could input their medical data and receive instantaneous disease predictions. In addition, the system was able to recommend specialized doctors based on the identified disease, enhancing its practical applicability. This dual functionality—accurate disease detection and relevant medical recommendation—positions the proposed system as a comprehensive decision support tool for healthcare applications.

8 Conclusion

This study presents a Multi-Disease Diagnosis and Doctor Recommendation System using machine learning. The system successfully predicts diabetes, heart disease, and breast cancer with high accuracy using an SVM algorithm. It integrates data preprocessing, feature selection, and model optimization techniques to enhance prediction reliability. By utilizing a Flask API, the system is deployed as a real-time diagnostic tool, making disease detection more accessible. Additionally, doctor recommendations provide users with guidance for seeking medical assistance. Future improvements include incorporating more diseases, refining feature selection methods, and enhancing doctor recommendation algorithms.

References

[1] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," Mater. Today Proc., vol. 80, pp. 3682–3685, 2023.

[2] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," Healthcare, vol. 10, no. 3, p. 541, Mar. 2022.

[3] S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning machine (ELM) technique for heart disease diagnosis," in Proc. IEEE Canada Int. Humanitarian Technol. Conf. (IHTC), 2015, pp. 1–3.

[4] P. Sajda, "Machine learning for detection and diagnosis of disease," Annu. Rev. Biomed. Eng., vol. 8, no. 1, pp. 537–565, 2006.

[5] R. C. Deo, "Machine learning in medicine," Circulation, vol. 132, no. 20, pp. 1920–1930, 2015.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Belmont, CA, USA: Wadsworth and Brooks, 1984.

[7] A. Parashar, A. Gupta, and A. Gupta, "Machine learning techniques for diabetes prediction," Int. J. Emerg. Technol. Adv. Eng., vol. 4, no. 3, pp. 672–675, 2014.

[8] J. A. Paniagua, J. D. Molina-Antonio, F. Lopez-Martinez, et al., "Heart disease prediction using random forests," J. Med. Syst., vol. 43, no. 10, p. 329, 2019.

[9] A. Yaganteeswarudu, "Multi disease prediction model by using machine learning and Flask API," in Proc. 5th Int. Conf. Commun. Electron. Syst. (ICCES), 2020, pp. 1242–1246.

[10] T. Bhanuteja, K. V. N. Kumar, K. S. Poornachand, C. Ashish, and P. Anudeep, "Symptoms based multiple disease prediction model using machine learning approach," Int. J. Innov. Technol. Explor. Eng. (IJITEE), vol. 10, no. 3, pp. 2278–3075, 2021.

[11] I. Mohit, K. S. Kumar, U. A. K. Reddy, and B. S. Kumar, "An approach to detect multiple diseases using machine learning algorithm," J. Phys.: Conf. Ser., vol. 2089, no. 1, p. 012009, Nov. 2021.

Cite this article:

Devi priya M & Dr Muthukumar VP, "Multi-Disease Diagnosis Using Machine Learning Algorithm", Journal of Multidimensional Research and Review (JMRR), Vol.6, Iss.2, pp.14-20, 2025