International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 02 | Feb - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

# Medbot: A Deep Learning-Based Medical Chatbot for Symptom Detection and Suggestion

**V. S. Tamboli [1], Bhalerao Yash Nilesh[2], Tupe Omkar Ashok[3], Dighe Gaurav Ramnath[3]**

*[1]Professor, Dept. of Computer Technology, P.Dr.V.V.P. Institute of Technology and Engineering, Loni, Maharashtra, India*

*[2,3,4] Final year Diploma Student, P.Dr.V.V.P. Institute of Technology and Engineering, Loni, Maharashtra, India*

---

**ABSTRACT -** To live a long and healthy life, which is attainable for many people on this planet, the healthcare industry is crucial. These people make use of medical facilities for treatment, routine examinations, and symptom-based disease diagnosis. By analyzing symptoms, these medical care facilities accurately diagnose a variety of ailments and provide appropriate therapy. But these facilities are under a lot of pressure, and doctors are overworked, because of the huge increase in the global population and the number of people. Hence, technology needs to be integrated into this method to make it more convenient for both patients and health care professionals. For this reason, a Chabot can serve patients well by automatically making accurate diagnoses. Thus, a highly efficient method for an automated Chabot that makes use of Decision Trees, Deep Belief Networks, Linear Regression, and K Nearest Neighbors Clustering. This strategy has been thoroughly tested using the methods described in this paper.

***Key Words***: K- nearest Neighbor classification, Regression analysis, Hidden Markov model, Decision Tree.

## 1.INTRODUCTION

In 2026, medbots—medical chatbots powered by artificial intelligence—are predicted to revolutionize healthcare by making basic medical advice more accessible, particularly for those without instantaneous access to physicians. These bots can alleviate the strain on hospital emergency rooms by doing preliminary triage around the clock using deep learning models that comprehend natural language, slang, and emotional indicators. While providing individualized, data-driven medical insights that lessen the likelihood of self-diagnosis mistakes, they aid in differentiating between less significant health issues and more serious problems.When creating a Medbot, it is important to follow all privacy requirements, including HIPAA and GDPR, have well-defined clinical goals, collaborate with stakeholders and healthcare experts, and use high-quality medical data sources like SNOMED CT and ICD-10. From a technical standpoint, Medbots employ hybrid deep learning architectures. These designs include, for language understanding, RNNs or LSTMs, and for symptom-based predictions, probabilistic multi-label classifiers. They also keep conversational context. For example, Explainable AI increases openness and user trust, while Retrieval-Augmented Generation makes sure that suggestions are based on current medical research. All things considered, Medbots will play an essential role in the digital healthcare systems of the future.

1] RetStroke, developed by Saeed Shurrab et al., is a clinically-informed system that can use retinal pictures to diagnose and predict strokes. The multimodal nature of RetStroke is its primary strength; it trains and infers with the patient's demographics, vital signs, and comorbidities to improve accuracy. RetStroke outperformed OCT and infrared, two examples of unimodal models trained solely on image data, in terms of performance growth. In addition, Retstroke outperforms RetFound, an existing foundation model with 300 million parameters that was pre-trained on a somewhat bigger OCT dataset in a unimodal situation [24]. the author looked into integrating the clinical data into the preexisting RetFound foundation model and shown that this modification results in a notable enhancement of performance. The clinical information used in the author's proposed framework is valuable, as these data show. The author also found that RetStroke was robust when tested on a variety of patient subgroups, including those defined by age, stroke subtype, and co-morbidities. In sum, RetStroke is a fresh and original contribution to oculomics that has the potential to deepen our understanding of a critical illness like stroke. Thanks to its innovative architecture, RetStroke offers a number of benefits that might be useful for AI researchers and physicians alike, opening up exciting new avenues for study. To begin with, RetStroke improves stroke prevention by filling a gap in the current clinical landscape by creating less expensive, faster, and more independent data modalities for stroke screening. On top of that, RetStroke does double duty by both predicting the likelihood of a stroke and identifying any long-term retinal damage. In addition to helping with risk prediction for stroke prevention, this provides valuable insights on the changes in the retina that can occur as a result of a stroke, which have an effect on eye health. From a technical standpoint, RetStroke uses ResNet-18, a lightweight and basic network, which uses less resources than RetFound. The use of simple fusion techniques in its design further simplifies the models. Lastly, RetStroke takes advantage of demonstrated performance-enhancing, easily-collectible, basic clinical qualities. Other pertinent use cases could benefit from these features. However, there are a few caveats to this study that do highlight some interesting avenues for future investigation. Concerns over the generalizability of RetStroke arise from the fact that it has only been tested on an internal private dataset, even if the author's technique is compared against external baselines. Nevertheless, healthcare data privacy and legal constraints make it difficult to get such data, therefore more work is needed to validate the model on external datasets and new baselines. Secondly, RetStroke's performance is impacted since the dataset used to construct the model is relatively small, consisting of just 183 cases. This limits the model's ability to learn features that are specific to strokes.

2] In order to classify cardiac illness from echo-images, Muhammad Raoof et al. investigated the possibilities of deep learning with models based on VGG16 and its derivatives. With an astounding accuracy of 92.18 percent, unaltered VGG16

effectively and swiftly detects cardiac problems from echocardiography pictures. Even if the author was able to achieve a lesser level of accuracy in the other trials, it was because they utilized extra layers and regularization procedures. The significance of hyperparameter tweaking is highlighted by these findings in the other trial, where a dropout rate of 0.4 and momentum of 0.99 produced an accuracy that surpassed all models that have performed up to this point (94.92%). So, it's possible to avoid overfitting and get your model to perform better with just the correct amount of regularization. There aren't many noteworthy things that the author has contributed to the field of medical imaging and heart disease detection through their study. This finding lends credence to the use of deep learning models, most notably VGG16, for the accurate diagnosis of cardiac conditions using echocardiography pictures. This paper presents the author's view on the significance of the hyperparameters momentum and dropout rate in the model's performance. In order to improve deep learning models for medical picture categorization, future research should follow these findings. Optimal hyperparameters for regularization are further investigated in this paper, with the goal of fine-tuning the VGG16 architecture and, ultimately, achieving a superior model configuration for cardiac disease classification. There are a lot of promising avenues for further research, even though this work does provide important findings. Given the breadth of deep learning architectures, it is necessary to explore more designs first. It is possible that ResNet or Inception, two architectures not used in this work, might converge more quickly or perform better than the VGG model. The availability of a bigger and more diverse dataset is the second requirement. The model will be more robust and able to generalize well to any patient in this fashion. It is also possible to artificially lengthen and diversify the data using certain data augmentation technologies. 3] The work of Md Zahin Muntaqim et al. was a giant leap forward in the fight against the urgent problem of eye disease classification, an essential part of healthcare around the world that needs quick identification to avoid blindness. Automated detection systems are essential, thus the author set out to solve their current shortcomings, such as poor feature representation, large computing overheads, and inadequate illness coverage, using a novel approach. Using the powerful capabilities of deep learning technologies, the author has developed a groundbreaking technique for detecting eye diseases. The author laid the groundwork for the development of their model by carefully analyzing the data and using several image augmentation techniques to make it resistant to rotations and translations. Efficient feature extraction and classification are made possible by the author's innovative lightweight three-stage deep learning architecture, which is defined by a careful combination of identity and convolutional blocks. The core of the author's methodology is the building of a deep learning model.

This paper's second portion provides an analysis of the previous studies that were considered as Litereature Survey. In Section 3, the course of action is described in full detail as Proposed methodology. Part 4 digs into the experimental evaluation, while Section 5 explores potential changes before wrapping up the article with a conclusion on the current proposal.

## 2. LITERATURE SURVEY

4] 57% recall, 93.4% binary accuracy, 55.3% F1-score, and 81% AUC were the highest results achieved by the MobileNetV1 model with geometric image augmentation, according to Muhammad Irtaza et al. (4]. When dealing with huge datasets, synthetic samples produced using GAN can assist with class imbalance and increasing data diversity for training. Although the author did use a DCGAN model to create synthetic X-ray pictures, they discovered that the MobileNetV1 model did not benefit from these images. F1-Score, AUC, Recall, and Precision were 54.1, 78.4, and 58.2 when synthetic samples were added, respectively. The failure of the DCGAN model to enhance the MobileNetV1 model's performance could be due to a number of factors. To begin, the NIH Chest X-ray 14 dataset contains a wide variety of lung disorders, but the DCGAN model may not have been able to adequately represent this diversity due to the limited sample sizes used for training each class. Second, the author may have seen that the DCGAN model was overfitting to the training data, which would have resulted in subpar classification accuracy. The author's tests provide strong evidence that deep learning models may accurately classify lung illnesses from chest X-ray pictures. The authors of this work hold the firm belief that their approach could one day aid in the early diagnosis of lung ailments by providing clinical decision assistance. Still, additional study is required to enhance these models' functionality, particularly on massive and unbalanced datasets. The author intends to use a bigger and more varied dataset of chest X-ray pictures to enhance the system's performance in future work.

5] A new approach to accurately detecting Parkinson's disease using handwritten records from the standard NewHandPD dataset was studied by Sura Mahmood Abdullah et al. In order to alleviate the strain of training time, the suggested architecture is built upon transfer learning models like ResNet, VGG19, and InceptionV3. An optimized feature vector is obtained by feeding the collective features from the TL models into the optimization process using a genetic algorithm. This allows for superior classication outcomes. During optimization, KNN serves as the objective function and accuracy is considered the tness value. Use of KNN, which is computationally less intensive, allows for the classication to be performed using the optimized features. Several analyses are conducted to study and evaluate the proposed model's performance. The proposed model outperformed other recently studied systems in terms of classication accuracy. Among other performance metrics, the loss is incredibly small and has high accuracy. The suggested model outperformed the alternatives in reliably identifying Parkinson's disease, according to the results of the experiments and the performance comparison study.

6] Classification of coronary artery disease using invasive coronary angiography pictures was investigated by Ariadna Jiménez-Partinen et al. Using different thresholds of lesion degree to classify as a "lesion," five state-of-the-art deep neural models were employed to differentiate between images with and without lesions. Four other kinds of studies, such as data augmentation and the removal of high-severe classes, were conducted once the dataset was partitioned into non-overlapping patches. The results demonstrated that lesions may be easily classified as either 99% or 100% (>90% F-measure, >95% AUC) with minimal data, but performance drastically

decreases when a lower degree is included in the positive class (65% F-measure, 80% AUC). By eliminating these outliers, the networks are able to achieve an AUC of 85% and an F-measure of 75% when data augmentation is used to detect severity levels of 70% and 50%, respectively. Beyond that, DenseNet-201 and NasNet-Mobile proved to be helpful in resolving the majority of the binary classification issues that were presented. This study has some caveats, one of which is that the dataset used only includes 42 individuals from a single institution. The author plans to use ICA photos from other patients at multiple hospitals to improve the current tests in future work. Another drawback is that, as seen in Table 2, there is an imbalance between the various lesion degree ranges. The experimental outcomes could be impacted by the absence of patches in each lesion degree range, even though they do not represent different classes. This is because the models employed for learning will perform better in the lesion degree ranges with more patches. We will continue to work on enhancing the overall performance of the classification system. Separate severity degree classification with more advanced preprocessing processes has the potential to increase homogeneity and, by extension, yield better results. Building one-of-a-kind deep networks from the ground up using architectures that zero down on certain local spatial attributes is another option. With an eye toward bringing these deep learning solutions to clinical settings, the given study seeks to enhance our knowledge of their limitations, requirements, and possible enhancements for ICA image processing. Lastly, the author proposes that tools that comprehend the requirements of deep learning solutions for ICA images could be useful for automating the interpretation of the images or calculating scales to aid in clinical decision-making in complicated CAD (such as the SYNTAX score), which would reduce the time and effort needed for CAD diagnosis and treatment.

7] According to Ifra Shaheen et al., diabetes is a slow-acting poison that can have a significant impact on people's lives if not caught early. To start, the author uses the ProWSyn method to level the playing field on the notoriously class-skewed DPD dataset. As a result, prejudices against the majority class were lessened. Author's paradigm incorporates two ensemble approaches: blending and hybridization. Through the integration of the highway and LeNet models, the author's hybrid model, Hi-Le, accomplishes remarkable outcomes, boasting an accuracy score of 94% and an F1-Score of 96%. The results demonstrate that the author's hybrid technique is effective, since the author's proposed model surpasses the performance of the independent DL models. By integrating the highway, TCN, and LeNet models, the author's blending model, HiTCLe, attains a 94% F1-score and 94% accuracy. Hi-Le and HiTCLe both have the ability to predict diabetes accurately and early on, which could save lives. Also, to make sure the author's models' performance metrics are reliable, the author's methodology includes a comprehensive review procedure that employs a 10-FCV strategy. In addition, the key mechanisms driving the model's predictions are shed light on by the enhanced interpretability of the author's study that follows from using SHAP. In addition, research has been conducted to assess the individual contributions of each component in the author's ensemble models, providing valuable insights into their efficacy and suggesting potential avenues for improvement. This paper makes a contribution to the field of early diabetes diagnosis by proposing unique ensemble procedures and employing innovative techniques for model evaluation and interpretation.

8] Early detection and proper treatment are the only ways to reduce the mortality rate of chronic renal failure, according to Kommuri Venkatrao et al. The healthcare industry is increasingly recognizing the value of categorization systems due to their ability to effectively categorize disease datasets. This study suggests a thorough and effective classification method for detecting renal disease based on objective clinical data. This diagnosis method relies on optimizing the DSCNN classifier with the dimensionality reduction strategy based on the Aquila optimization algorithm to identify the most essential risk factor characteristics linked with CKD. The Capsule Network method is utilized for feature extraction. The Sooty tern optimization algorithm (STOA) is used to optimize the parameters of the suggested classification technique, further improving performance. We test the efficiency of the produced model by looking at its recall, diagnostic accuracy, specificity, PPV, and FPR. Results showed that the suggested method outperformed state-of-the-art techniques on the repository CKD dataset. Using a clinical dataset, the suggested method achieves a classification accuracy of 99.18% and produces excellent results. So, it's safe to say that the suggested method outperforms and competes with the state-of-the-art methods in the literature, all while running in record speed and providing superior classification accuracy. Future studies will employ the clustering method to decrease the occurrence of inaccurate categorization and increase the efficacy of classification.

9] For the purpose of early dementia and Alzheimer's disease detection and diagnosis, HARSH VARDHAN BANSAL et al. offered a thorough multimodal framework. Clinical numerical indicators from the OASIS dataset and analysis of brain MRI images showing structural abnormalities are two separate but linked parts of the methodology. A sophisticated feature selection method based on firefly behavior is used to process the OASIS data. The chosen set is trained and classified using a 3 layer DNN model. In the MRI subdomain, features are extracted from pre-processed MRI images using ResNet-101. Following feature extraction, a lightweight support vector machine classifier is fed the data. An essential component, the decision-level fusion module unifies the two modalities' binary and multiclass results. Both the classification accuracy and the clarity of borderline cases are enhanced by this fusion method. The proposed work demonstrated a 4.7% higher F1-Score and a 1.5% higher accuracy when compared to state-of-the-art models like Dar et al. [23] (Accuracy: 0.891, F1-Score: 0.867), Khatun et al. [27] (Accuracy: 0.905, F1-Score: 0.888), and Rallabandi and Seetharaman [32] (Accuracy: 0.918, F1-Score: 0.901). This further affirms the efficacy of the dual-path learning and decision-level fusion strategy.

10] The authors Thavavel Vaiyapuri et al. present a novel approach to EDCD-DLDR that is powered by the Internet of Things. In order to facilitate the early detection of diabetes, the EDCD-DLDR strategy seeks to equip IoT devices with the ability to gather patient medical data and utilize the DL model. Acquiring data, normalizing data, selecting features, detecting diabetes, and tweaking hyperparameters are the five main steps in implementing the suggested work. Data capture utilizing the Internet of Things (IoT) is the first step in the EDCDDLDR approach, and Z-score normalization is used to standardize the acquired data. In order to reduce dimensionality, the EDCD-

DLDR method employs the ARO-FS model. Furthermore, the ABiGRU model can identify cases of diabetes. Improving the ABiGRU network's performance is the goal of the POA-based hyperparameter selection model. A small series of simulations demonstrate how well the EDCDDLDR technique performs on the Kaggle dataset. An improved accuracy value of 97.14% compared to previous methods was highlighted by the experimental validation of the EDCD-DLDR approach. Potential risks in scaling IoT data collecting operations owing to variability in sensor types and data quality are posed by the constraints of the EDCD-DLDR approach. In addition, ARO-FS can help with feature selection, however how well it works might differ depending on the dataset and feature space. While the ABiGRU technique is highly effective for sequential data, like diabetes identification, it could require a lot of tuning and processing power. Further optimization of ABiGRU model parameters, enhancement of feature selection techniques to accommodate different datasets, exploration of novel hyperparameter optimization methods beyond POA, and enhancement of IoT integration for robust data handling are all potential areas for future research. Statement on Data Availability Visit https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset, reference number [33], to see the publicly available data that supports the results of this study.

11] In order to screen for ILD utilizing complete HRCT images, Surendra Reddy Vinta et al. proposed a novel and effective approach based on a hybrid deep learning network. Raising the precision of the deep learning algorithms across the board has improved the method's efficacy. In the first stage of lung segmentation, the enhanced U-Net++ is used to remove the unwanted background from the HRCT pictures. This paves the way for the subsequent step, which involves deep feature extraction, to extract ILD. Utilizing RAPNet, the segmented pictures of the lungs were used. Consolidation, micro-nodules, ground glass, fibrosis, emphysema, and normal are the six types of ILDs that the improved MobileUNetV3 classified using deep learning features. Current models that rely on deep learning have been compared to the efficacy of the proposed approach. By a wide margin, the suggested method beats an identical full-image-based approach and prior techniques that account for five ILD types. The suggested method emphasizes the potential for improving overall efficiency by picking the best CNN method for a given job and raising accuracy across the board. Across many ILD disorder classes, the suggested technique attains an accuracy rate of 99.10%. One problem with the suggested method is that the combined attributes dimension is too broad. In the future, a feature reduction method might make it feasible to decrease this set of features. A CREDENTIAL They attest that the work in question is completely unique, has never been published before, and is not presently under consideration for publication by any other entity.

12] This study demonstrated the feasibility of using deep learning, namely the EfficientNet-B3 architecture, to predict cardiovascular risk from retinal fundus images. N. D. Bisna et al. proposed this method with the use of integrated clinical data. By producing very accurate predictions, the model proved it could be used as a non-invasive screening method. This method improves early risk stratification by detecting small retinal vascular abnormalities linked to cardiovascular disorders, which is an addition to existing diagnostic tools such as electrocardiograms (ECGs) and blood tests. Nevertheless, more enhancements utilizing techniques like cost-sensitive learning are necessary due to the presence of false negatives. Future research should prioritize validating the model's generalizability across different populations and clinical scenarios. It should also address potential biases in datasets and account for real-world variables. Validation and comparative study are necessary to assess the efficiency and cost-effectiveness of clinical integration in relation to current practices, notwithstanding the promising outcomes. The practical implementation of retinal imaging as a screening tool relies on robust algorithms capable of handling diverse patient demographics and image variability. The technology has the potential to be both successful and accessible. Research into the prevention and management of cardiovascular disease should focus on conducting comprehensive clinical trials to evaluate its impact on patient outcomes and explore its feasibility for seamless integration into existing healthcare systems.

13] An important gap in prior research that mostly focused on binary classification was filled by K. Shyamala et al., who introduced a data-driven deep learning system for stage classification of Parkinson's Disease (PD) utilizing voice data. Using Random Forest-based feature selection, KMeansSMOTE for class balance, and Isolation Forest for outlier identification, the PSS-Net model successfully identified the most essential speech aspects for PD staging. With an accuracy of 89%, the Optimized and Fused Neural Network (OptiFusionNet) showed promise for stage-specific categorization. The SHAP method also boosts physician self-assurance while providing insights about feature contributions. First, by moving beyond binary categorization to staging, it fills a critical knowledge gap in the area by offering stage-specific insights for therapies. The second justification is that the model is interpretable, which boosts confidence and practicality, and doctors can understand how certain speech characteristics affect the classification with the help of SHAP (SHapley Additive exPlanations). Finally, PD can be transformed by the scalable, non-invasive, and transparent paradigm, which provides a personalized treatment plan, early detection, and affordability. More substantial clinical use of this study's findings will necessitate more investigation into new architectures and the integration of multimodal data. Additionally, it proves that voice data is useful for PD staging. Expanding upon this work with more diverse and essential datasets could enhance the generalizability and robustness of the models in future study. Speech and other biomarkers like gait or neuroimaging data integrated with other modalities could improve classification accuracy and reliability. Improving the PSS-Net further can make it useful for a wider range of neurodegenerative diseases, not just Parkinson's. Graph Neural Networks and Transformer models are two examples of emerging deep learning architectures that might improve categorization and feature extraction. Working together with clinical professionals to acquire labels annotated by specialists could be part of future work that helps evaluate inter-rater variability. Finally, more varied and representative data is required for clinical deployment; the present model is trained on a curated speech dataset from the UCI library. Data collected from non-invasive modalities, such as handwriting analysis or wearable sensors, will be a part of future upgrades. For clinical trust and adoption, it is essential that outcomes are easy to understand and use. Although SHAP was utilized to identify important speech features in this study, future work will incorporate expert comments, evaluate model findings

against clinical knowledge, and use SHAP-guided feature reduction. One part of the deployment approach is using model compression techniques to make it easier to deploy on mobile or embedded systems by reducing computational load. To guarantee clinical usability over the long term, continuous learning and temporal modeling are essential. The gap between theory and practice could be filled by creating clinically applicable, real-time solutions for Parkinson's disease prediction and staging.

14] Problems with using multiclass datasets for quick diagnosis and categorization of neurological diseases like Alzheimer's were discovered by Erol Kina et al. In order to diagnose and treat the illness, an accurate automated approach is necessary. Using EfficientNet Squeeze Attention Blocks and transfer learning, this study presented a lightweight convolutional architecture that is both efficient and light. The model was trained on a multiclass dataset to detect cases of Alzheimer's disease by utilizing two dropout layers. It employed a number of lightweight layers, including as batch normalization, global 2D pooling, and an L2 regularizer. The author utilized an SMOTE sampling technique to equalize the classes because the Alzheimer illness dataset was extremely unbalanced. Results from experiments showed that the suggested method achieved 95.28 percent, 98.05 percent, and 98.13 percent training accuracies and 90.17%, 92.17%, and 95.19% validation accuracies for balanced training data, completely balanced data, and imbalanced data, respectively. With an average precision of 99.39 and a minimum of 90.46, the suggested method achieved an area under the curve (AUC) of 99.93 and 97.85, respectively. There was a solitary categorization mistake made by the GLIOMA brain tumor. In general, the sampling strategy improves the suggested model's performance. Because they are basic structures that provide controlled timing with reduced computing complexity, memory requirements, and overfitting, the trials prove that the suggested designs are appropriate. The suggested framework falls short in some situations, including the control and non-tumor instances, where it achieved poor results on all measures. Applying feature extraction and selection approaches, the author will focus on poor examples in the future.

## 3. METHODOLOGY

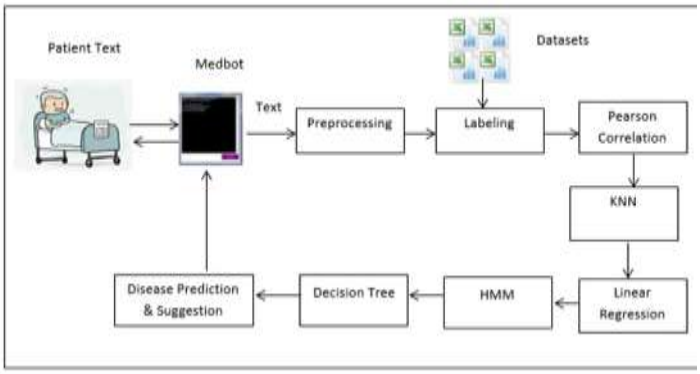



Figure 1: Proposed model System Overview

The proposed methodology for a medical chatbot that performs disease detection and suggestion using deep learning has been described in the system overview given in figure 1. The steps utilized for achieving the presented technique has been detailed in the steps given below.

*Step 1: Dataset collection, preprocessing and Labeling* – The system for the medbot comprising of disease prediction and suggestion requires an input of 3 datasets, consisting of kidney disease, Covid-19, and Heart Disease. The kidney disease dataset is downloaded from the URL - https://www.kaggle.com/mansoordaku/ckdisease, the Covid-19 dataset from URL - https://data.gov.il/dataset/covid-19, and heart disease from URL https://www.kaggle.com/ronitf/heart-disease-uci.

The datasets are extracted and provided to the presented technique as an input which initially performs the labeling procedure. This is performed by reading the datasets in the form of a double dimension list through the use of the JXL API. This process indexes the dataset which converts it into a labeled dataset. The user input is also grabbed using the interactive user interface that is designed in the Java programming language using the Swings Framework. This user interface prompts the user with the most common symptom of the 3 diseases. Once the user selects this disease the parameters related to the particular disease are selected from the patient as an input and preprocesses it before providing it to the system for further processing.

*Step 2: Pearson Correlation* – The outputs achieved in the previous step have been taken as an input into this step for the purpose of achieving the correlation between the two entities. The two entities in this approach are the attributes of the dataset and the user input. The correlation between these two entities is calculated using the Pearson Correlation approach. The Pearson Correlation allows for the realization of the correlation coefficient which is achieved for each of the rows of the dataset. This results in a correlation list which is then provided to the next step for the clustering. The equation for the Pearson correlation has been given in the equation 1 below.

$$r = \frac{\sum(x_i-\bar{y})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \text{ ------- (1)}$$

*Step 3: K Nearest Neighbor Clustering* – The labeled list and the user input is provided as an input to this step of the system for performing the clustering operation. The 3 datasets, namely, kidney disease, Covid-19 and heart disease, are in the form of a double dimension list. The K -Nearest Neighbors are used to perform the segregation of input data into clusters which are useful in determining the semantic groups. The clusters are obtained through the use of the following steps.

*Distance Evaluation* – The Euclidean Distance is used to determine the distance in the selected attributes of the input double dimension list in comparison with the user input for the same. The distance evaluated for the selected attributes is appended to the end of the particular row as the row distance $R_D$. This is performed for all the rows in the list and the respective row distances are appended accurately through the evaluation using the equation 1 given below. These row distances are also subjected for the average row distance evaluation which is then stored appropriately.

ED=√( 〖∑(ATi-ATj) 〗 ^2) _____ (1)

Where,

ED=Euclidian Distance

ATi=Attribute at index i

ATj= Attribute at index j

*Centroid Estimation* – The output from the previous step of distance calculation is provided as an input here. The list containing the distances added to the end of the rows is used for the purpose of centroid estimation. For achieving the centroid, the list is first sorted into the ascending order of the row distances. This sorted list is subjected to data point selection randomly. These data points are nothing but row distances that are k in number. These row distances are then used to determine the boundaries through the use of the Average row distance acquired previously.

The selected row distances from the data points and the average row distance is then used to form the minimum and maximum values by addition of both the values and subtraction of the same respectively. These boundaries are highly useful for the formation of the clusters in the next step.

*Cluster Formation* – The k boundaries attained in the previous step are used as a major aspect in this step for the purpose of cluster formation. The row distances in the double dimension list are subjected to scrutiny based on the boundaries attained in the previous step. The clusters are then formed by the row distances that abide by these boundaries which are then stored as a cluster list and transferred to the next step of the system. The entire process can be illustrated through the algorithm 1 given below.

---

ALGORITHM 1: KNN Classified Cluster Formation

---

//Input : Sorted Distance List $SD_{L,}$
//Output:Cluster List $K_{CL}$
1: ***Start***
2: $I_L = \emptyset$ [Inner Layer] $O_L = \emptyset$ [Outer Layer], $K_{CL}=\emptyset$
3: MIN= 0 , MAX=$SDL_{SIZE-1}$
4: K= ( MAX-MIN ) /2
5: K=MIN+K
6: ***for*** i=0 to Size of $SD_L$
7: R = $SD_{L\,[i]}$
8: ***if***(i<=K), ***then***
9: $I_{L=}\ I_{L+R}$
10: ***else***
11: $O_{L\,=}\ O_{L\,+R}$
12: ***end for***
13: $K_{CL[0]} = I_L$
14: $K_{CL[1]} = O_L$
15: return $K_{CL}$
16: ***Stop***

---

*Step 4: Linear Regression* – The linear regression procedure performs the regression between the user input and the cluster attributes that are formed in the previous step. These values are provided as input to this step for the regression analysis through the use of Linear Regression. The regression analysis performed through the linear regression determines the change between two different variables and quantifies it. The lists are the x [] and y[ ] out of which the x [] is the independent list and the y []is the dependent list. The equation for the same is given in equation 1 below.

Y=Mx + B _____ (2)

The regression is measured through the equation given above for which the values of the slope given as m and the value of the intercept given as b are unknown. These values are achieved by the evaluation of the equation 3 and 4 given below. The values of x [ ] in these are the user attributes and the value Y [ ] is the clusters values achieved in the previous step. These values are added to the equation to achieve the required values of m and b.

$$M = \frac{N\sum(xy)-\sum x\sum y}{N\sum(x^2)-(\sum x)^2} \quad \text{_____ (3)}$$

$$B = \frac{\sum y - M\sum x}{N} \quad \text{_____ (4)}$$

Where:
x = Independent variable
y = Dependent variable
M = Slope or Gradient
B = the Y Intercept
N= Size of the array
Y=Intercept value

The attained values of m and b from the above calculations are then used in the equation 2 above to attain the dependent variable values. Here an independent value from X[ ] is used in equation 2, and its obtained values are averaged to get the mean regression value for each of the rows in the cluster. The regression of the values of x [] and y[ ] allow for a greater understanding of the relationship between the two variables. These are the regression values from the linear regression analysis which are then aggregated in the form of a list which is given to the next step as an input.

*Step 5: Hidden Markov Model* – The Hidden Markov Model is highly effective and useful model for the purpose of identification and detection based problems. The Hidden Markov Model utilizes time series for the purpose of detection of the hidden or the unobservable states. For providing the input to this step of the methodology the clusters achieved in the previous steps are being utilized as an input to this step of the procedure.

The various attributes and collection of values relating to these attributes along with the regression list is taken as an input and added to a double dimension array list along with the corresponding user information entered by the patient. These values are being used for the purpose of extracting the probability of the instance of a particular disease out of the three diseases being identified in the presented approach.

In this type of the approach the probability is realized through the effective calculation of the time taken for the attributes to transform from one state to another. This procedure for the calculation of the probability values is performed recursively for each of the rows of all the attributes taken as an input. The termination equation is given by the equation 5 given below.

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \text{ --- (5)}$$

Where,

$P(O|\lambda)$ is the observational probability of the sequence O based on the λ (Hidden Markov Model) for a summation of all the variables in time T.

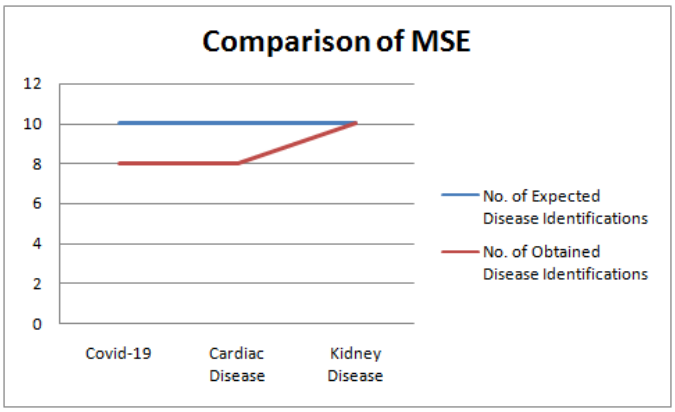
The achieved values of the probabilities are utilized by adding all of these values into a list that is then sorted in the descending order of the probability scores. This list is effectively shortening by selection of 50% of the data size and then provided to the next step of the procedure.

*Step 6: Decision Tree* – The probability values generated from the Hidden Markov Model stage of the technique has taken as an input in this step of the approach. This step involved the categorization of probability values using the decision tree methodology. This categorization is accomplished through the application of if-then rules, which properly select the right result for disease diagnosis and recommendation realization. The categorization can also help to limit the amount of false positives in the system. The system's ideas are utilized to populate the medbot user interface.

## 4. RESULTS AND DISCUSSIONS

The proposed methodology for medbot a medical chatbot for disease identification and recommendation by deep learning has been demonstrated in the Java programming language. To make the technique more effectively executed, the NetBeans IDE has indeed been utilized. The graphical user interface was created using the Swings Framework. The system was implemented on a developer computer with an Intel Core i5 CPU, 8 GB of RAM, and 1 TB of hard disk space as its setup.

The experimental evaluation of the approach is an essential necessity to determine the performance of the presented system.




This determines the amount of error achieved by the approach as well as stipulates if the deep learning methodologies have been accurately implemented in the system properly. For the purpose of evaluation, the RMSE evaluation mechanism is being utilized. The system provides appropriate detection of the error achieved by the methodology for the detection or identification of the disease.

### Performance Evaluation through Root Mean Square Approach

Several tests were conducted to establish the error produced by the suggested approach, the process for disease detection employing Convolutional neural networks. The decreased accuracy reached by the approaches attributable to the Disease Detection component's predisposition for error may be utilized to define the predefined threshold.

The Root Mean Square Error, or RMSE, is used to calculate the error caused by the specified approach. The presence of any type of imprecision in the suggested strategy for disease identification through HMM indicates the identification accuracy of the proposed methodology. The RMSE approach makes it easier to calculate errors between two continuously connected parameters. The metrics tested in this technique are the expected disease identification and the achieved disease identification. Equation 6 is used to compute the error estimates.

Where,

_ (6)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}}$$

∑ - Summation

$(x_1 - x_2)^2$ - Differences Squared for the summation in between the expected disease identification and the achieved disease identification counts

n - Number of Trails

These two properties were measured on 3 distinct diseases based on the user inputs on the medical chatbot standalone application. The outcomes of these assessments are depicted in table 1 below.

| Disease | No. of Expected Disease Identifications | No. of Obtained Disease Identifications | MSE |
|---|---|---|---|
| Covid-19 | 10 | 8 | 4 |
| Cardiac Disease | 10 | 8 | 4 |
| Kidney Disease | 10 | 10 | 0 |

Table 1: Mean Square Error measurement

Figure 3: Comparison of MSE in No. of expected disease identification and the No. of achieved disease identification

The experimental analysis of the concept's outputs for 10 trials for each disease have made it simpler to comprehend the error rate visually, as illustrated in figure 3. The graph shows the system's error rate in guessing the Disease that the patient is suffering from.

## 5. CONCLUSION AND FUTURESCOPE

The healthcare system has been productively improved thanks to the introduction of a medical chatbot that can diagnose diseases and make helpful recommendations based on the specific case. The system takes the user-supplied symptoms as input. Through the user inputs given to the chatbot, these symptoms are efficiently preprocessed to generate a lightweight query that can be easily processed to achieve the diagnosis. A number of datasets containing diseases and symptoms are subsequently used to label the query. The disease range can be efficiently restricted by the system thanks to this labeling. K Nearest Neighbors is used to create suitable clusters after sending the tagged text via Pearson Correlation, which generates the correlation. After the clusters have been collected, the following stage of the process uses Linear Regression to extract the input symptoms' regression. The following phase uses the regression list to evaluate the disease and generate predictions using the Hidden Markov Model. In order to accurately categorize the diseases predicted and suggested by the Hidden Markov Model, the Decision Tree Module employs if-then rules. The user is thereafter notified of the outcomes via the graphical user interface. An efficient method for evaluating the approach's efficacy is root-mean-squared error (RMSE). By comparing the resultant error with the dominant illness detection mechanism, the proposed strategy is shown to be superior.

Future improvements to this method can include making this strategy available to healthcare providers and patients through a mobile app.

## REFERENCES

1. S. Shurrab, A. Nepal, T. J. Lee-St. John, N. G. Ghazi, B. Piechowski-Jozwiak and F. E. Shamout, "Multimodal Deep Learning for Stroke Prediction and Detection using Retinal Imaging and Clinical Data," 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Copenhagen, Denmark, 2025, pp. 1-7, doi: 10.1109/EMBC58623.2025.11253814.

2. M. Raoof, M. Mahtab, S. Masood Bhatti, M. Rashid and A. Jaffar, "Heart Disease Classification From Echocardiogram Images Using Deep Learning," in IEEE Access, vol. 13, pp. 8011-8022, 2025, doi: 10.1109/ACCESS.2024.3524732.

3. M. Zahin Muntaqim et al., "Eye Disease Detection Enhancement Using a Multi-Stage Deep Learning Approach," in IEEE Access, vol. 12, pp. 191393-191407, 2024, doi: 10.1109/ACCESS.2024.3476412.

4. R. S. Gargees, "Multi-Class Flat Classification of Lung Diseases Utilizing Deep Learning," 2022 IEEE IAS Global Conference on Emerging Technologies (GlobConET), Arad, Romania, 2022, pp. 804-809, doi: 10.1109/GlobConET53749.2022.9872480.

5. S. M. Abdullah et al., "Deep Transfer Learning Based Parkinson's Disease Detection Using Optimized Feature Selection," in IEEE Access, vol. 11, pp. 3511-3524, 2023, doi: 10.1109/ACCESS.2023.3233969.

6. A. Jiménez-Partinen, K. Thurnhofer-Hemsi, J. Rodríguez-Capitán, A. I. Molina-Ramos and E. J. Palomo, "Coronary Artery Disease Classification With Different Lesion Degree Ranges Based on Deep Learning," in IEEE Access, vol. 12, pp. 69229-69239, 2024, doi: 10.1109/ACCESS.2024.3401465.

7. I. Shaheen, N. Javaid, N. Alrajeh, Y. Asim and S. Aslam, "Hi-Le and HiTCLe: Ensemble Learning Approaches for Early Diabetes Detection Using Deep Learning and Explainable Artificial Intelligence," in IEEE Access, vol. 12, pp. 66516-66538, 2024, doi: 10.1109/ACCESS.2024.3398198.

8. K. Venkatrao and S. Kareemulla, "HDLNET: A Hybrid Deep Learning Network Model With Intelligent IOT for Detection and Classification of Chronic Kidney Disease," in IEEE Access, vol. 11, pp. 99638-99652, 2023, doi: 10.1109/ACCESS.2023.3312183.

9. H. V. Bansal, P. Gupta and V. Juneja, "A Multimodal Deep Learning Framework Using ResNet-101 and Firefly-Based Feature Selection for Early Diagnosis of Dementia and Alzheimer's Disease," in IEEE Access, vol. 13, pp. 184709-184721, 2025, doi: 10.1109/ACCESS.2025.3621157.

10. T. Vaiyapuriet al., "IoT-Enabled Early Detection of Diabetes Diseases Using Deep Learning and Dimensionality Reduction Techniques," in IEEE Access, vol. 12, pp. 143016-143028, 2024, doi: 10.1109/ACCESS.2024.3455751.

11. S. R. Vinta, B. Lakshmi, M. A. Safali and G. S. C. Kumar, "Segmentation and Classification of Interstitial Lung Diseases Based on Hybrid Deep Learning Network Model," in IEEE Access, vol. 12, pp. 50444-50458, 2024, doi: 10.1109/ACCESS.2024.3383144.

12. N. D. Bisna, P. Sona and A. James, "Retinal Image Analysis for Heart Disease Risk Prediction: A Deep Learning Approach," in IEEE Access, vol. 13, pp. 76388-76399, 2025, doi: 10.1109/ACCESS.2025.3562433.

13. K. Shyamala and T. M. Navamani, "Design of an Optimized Feature Driven Severity Stage Classifier for Parkinson's Disease Prediction Using Deep Learning," in IEEE Access, vol. 13, pp. 142140-142160, 2025, doi: 10.1109/ACCESS.2025.3597851.

14. E. Kina, "TLEABLCNN: Brain and Alzheimer's Disease Detection Using Attention-Based Explainable Deep Learning and SMOTE Using Imbalanced Brain MRI," in IEEE Access, vol. 13, pp. 27670-27683, 2025, doi: 10.1109/ACCESS.2025.3539550