IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Early Dementia Identification from Text and Speech Using ML and NLP

[1]Bharathi N, [2]Dileep D

[1]Assistant Professor, Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, SSAHE, Tumkur, Karnataka, India

[2] Research Scholar, Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, SSAHE,
Tumkur, Karnataka, India

Doi: https://doi.org/10.56975/ijcrt.v14i7.309919

Abstract: This work presents a system for early detection using text and speech data with machine learning (ML) and natural language processing (NLP) algorithms. A brain condition called dementia impairs memory. thinking, communication, and reasoning abilities. Early signs of dementia can be seen in language usage, such as repeated words, grammatical mistakes, hesitation while speaking, and reduced vocabulary. The proposed system uses a hybrid approach that combines speech analysis, text feature extraction, and ML models to improve detection accuracy compared to normal health applications that mainly depend on questionnaires or manual observation. The system analyzes speech features like pause duration, speech rate, pitch variation, and fluency. It also studies text features such as sentence complexity, vocabulary richness, and word frequency. The extracted features are processed using ML models including Logistic Regression, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM). The system also supports multiple Indian languages using NLP preprocessing and LLM-based translation methods. Experimental results show that the proposed approach achieves good accuracy, precision, recall, and F1-score in dementia prediction. The system provides a non-invasive, low-cost, and scalable solution for early dementia detection and can help healthcare professionals in timely diagnosis and patient care.

Keywords: Machine Learning, Natural Language Processing, Dementia Detection, Speech Analysis, Text Classification, Feature Extraction

I. INTRODUCTION

Millions of individuals worldwide, particularly older folks, are impacted by dementia, a serious and expanding public health issue. It is a collection of symptoms associated with the slow deterioration of cognitive processes such as logic and remembering, thinking, and communication rather than a single illness .As the condition progresses, it interferes with everyday activities, reduces independence, and creates emotional as well as financial stress for both households as well as medical systems. The number of dementia cases is predicted to increase due to the aging of the world's population, making early detection and treatment more crucial. In order to improve patient outcomes and reduce The development of dementia, early identification is essential. Early identification of cognitive deterioration allows medical professionals to implement supportive measures, lifestyle modifications, and treatment that may lessen the severity of symptoms. However, traditional diagnostic procedures frequently rely on neuropsychological testing, clinical assessments, and brain imaging modalities like CT and MRI scans. These methods are beneficial, yet they are typically expensive, time-consuming, and require specialized knowledge. Additionally, they frequently only identify the illness after a discernible impairment has already occurred. New opportunities for examining cognitive decrease using language-based data have been made possible by recent

advancements in NLP and ML.. Since language is closely connected to brain function, changes in cognition often appear in speech and writing patterns. People in the initial phases of dementia may show reduced vocabulary, repeated words or phrases, grammatical mistakes, shorter sentences, and more hesitation while speaking. These changes can be subtle in routine observation but can be captured effectively through computational analysis. Speech and text provide valuable information for recognizing these early symptoms. NLP methods that can be used to extract linguistic patterns from written language, while speech processing can reveal acoustic indicators such as pause duration, speech rate, pitch variation, and fluency. These characteristics can subsequently be used by ML models to ascertain if an individual exhibits symptoms of cognitive impairment. This makes the approach a non invasive, scalable, and economical substitute for conventional diagnostic techniques. The absence of practical and effective methods for early dementia detection utilizing common voice and text data is the primary issue this work attempts to address. Many existing methods require specialized equipment and expert involvement, which may not always be available. Thus, it is extremely beneficial to have an automated system that can examine language patterns and offer early signs of cognitive loss. This work aims to create a system that makes advantage of To detect dementia early, audio and text data are analyzed using ML and NLP. The system is designed to extract relevant features, train classification models, and predict whether a person is at risk. In this way, it can support healthcare professionals in making timely and informed decisions and contribute to better patient care.

II.SYSTEM DESIGN: The overall system is illustrated through a block diagram that present the data's journey from input to output. Model training, feature extraction, Preprocessing, and classification are some of the steps the system uses to handle audio and text. For precise dementia detection, each step transforms unprocessed data into relevant information.

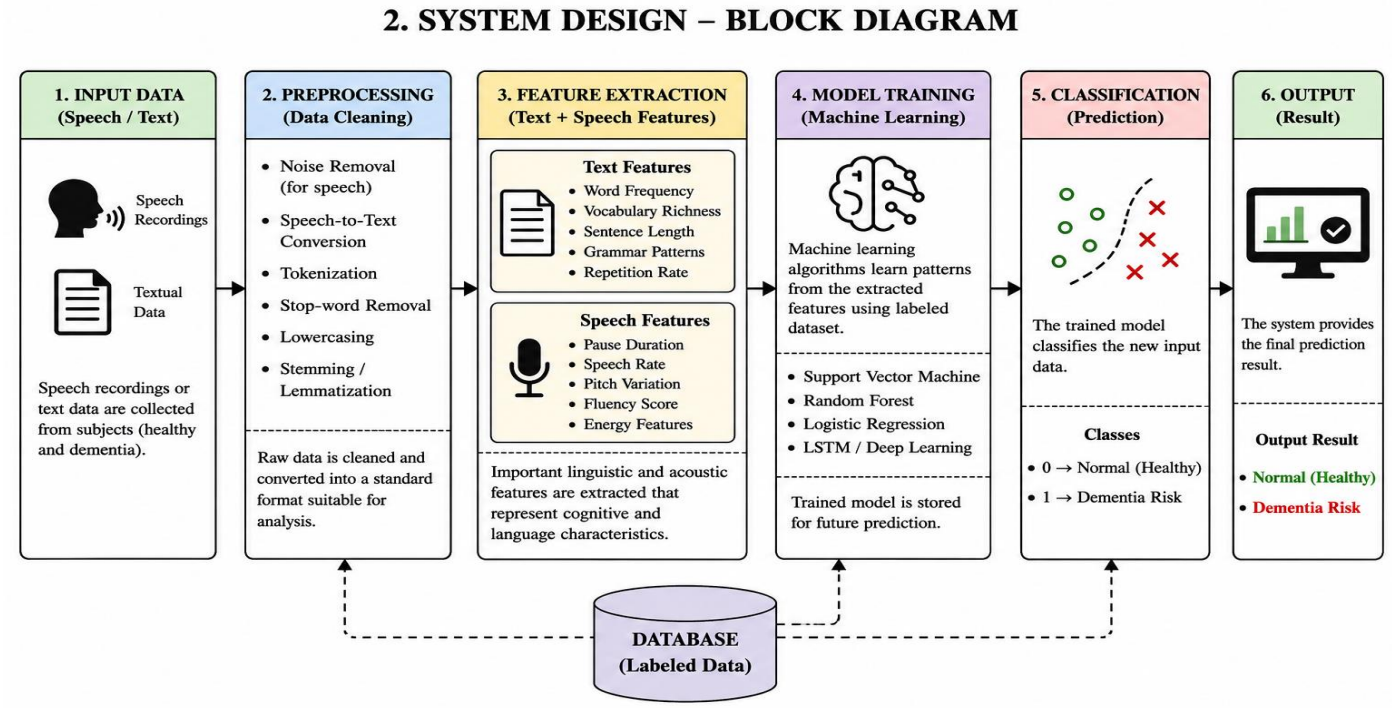


Fig. 1: Block Diagram of Early Dementia Detection System using Speech and Text

2.1 PSEUDOCODE

- Step 1: Read the input data (speech or text).
- Step 2: Preprocess the data to eliminate noise and unwanted elements.
- Step 3: Convert speech into text if required.
- Step 4: Extract linguistic and acoustic features from the data.
- Step 5: Train the ML model using extracted features.
- Step 6: Test the model with new input data.
- Step 7: Classify the data as Normal or Dementia.

Step 1: Data Collection: The first stage comprises obtaining data from multiple sources, including recorded conversations, clinical interviews. The dataset usually contains samples from both ill and healthy people diagnosed with early-stage dementia. A balanced dataset helps the model learn meaningful patterns and improve prediction accuracy.

Step 2: Preprocessing: Pre processing is required to clean and standardize the raw data before analysis.
• Remove noise from speech recordings.
• Convert using speech recognition to translate the speech into text techniques.
• Tokenize text into words and sentences.
• Remove stop words such as “is”, “the”, and “and”.
• Convert all text to lowercase for uniformity.
• Apply stemming or lemmatization to simplify words to their most basic form.
This stage improves the data's quality and helps guarantee the accuracy and significance Among the characteristics that are extracted.

Step 3: Multilingual Language Processing: The proposed system supports multiple Indian languages to improve accessibility and usability among diverse users. Speech input in regional languages is first converted into text. Using automatic speech recognition techniques. The translated text is then processed using NLP preprocessing methods such as tokenization, stop-word removal, stemming ,and lemmatization. The multilingual functionality is implemented using LLM assisted prompting and translation-based preprocessing techniques, which convert regional language inputs into a standardized textual representation suitable for feature extraction and classification. This allows the system to analyze speech and text patterns across different languages while maintaining consistency in dementia detection performance. The multilingual capability improves the practicality of the system in real-world healthcare environments where patients may communicate in their native languages.
Step 4: Extraction of Features The processed data is converted into numerical values by feature extraction so that ML models can use them.
Text Features:
• Word frequency to measure repetition of words.
• Vocabulary richness to identify reduced language diversity.
• Sentence length to detect shorter or incomplete sentences.
• Grammar patterns to identify errors in sentence construction.

Speech Features:
• Pause duration to measure hesitation in speech.
• Speech rate to calculate words spoken per minute.
• Pitch variation to analyze tone changes.
• Fluency score to assess how smoothly the speech flows.
These characteristics aid in recognizing trends frequently observed in dementia sufferers.

Step 5: Model Training: At this point, ML models are trained using the extracted characteristics. To find patterns in labeled data, methods like LSTM and other deep learning models, Random Forest, Support Vector Machine, and Logistic Regression are employed. The retrieved features are utilized by the model to differentiate between dementia patients and healthy persons. Step 6: Classification and Output In the last phase, the trained model is employed to classify new input data. The system examines the traits that were extracted and predicts whether the individual is healthy or at risk of dementia.

The output is presented as:
• Class 0 → Healthy
• Class 1 → Dementia Risk This final result helps in early detection and supports medical decision-making.

2.2 MACHINE LEARNING MODELS

Models of ML for classification are trained using the retrieved features

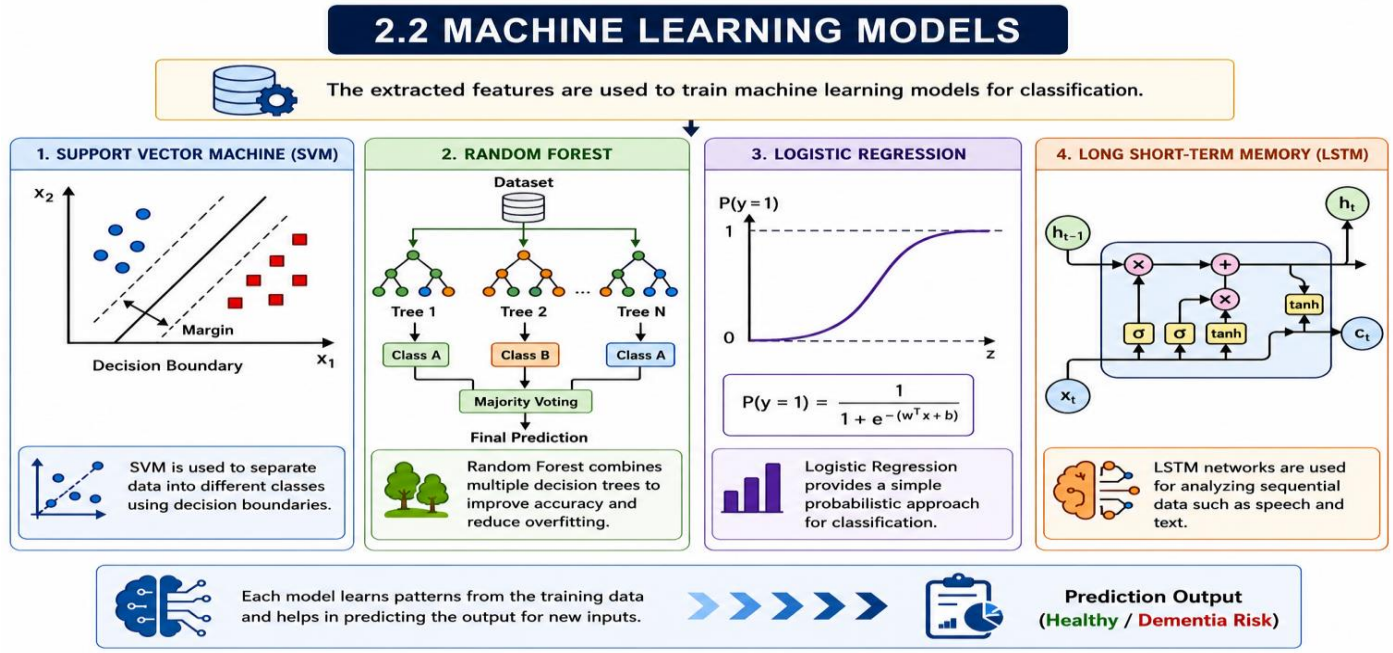


Fig. 2: Machine Learning Models Used for Early Dementia Detection

.

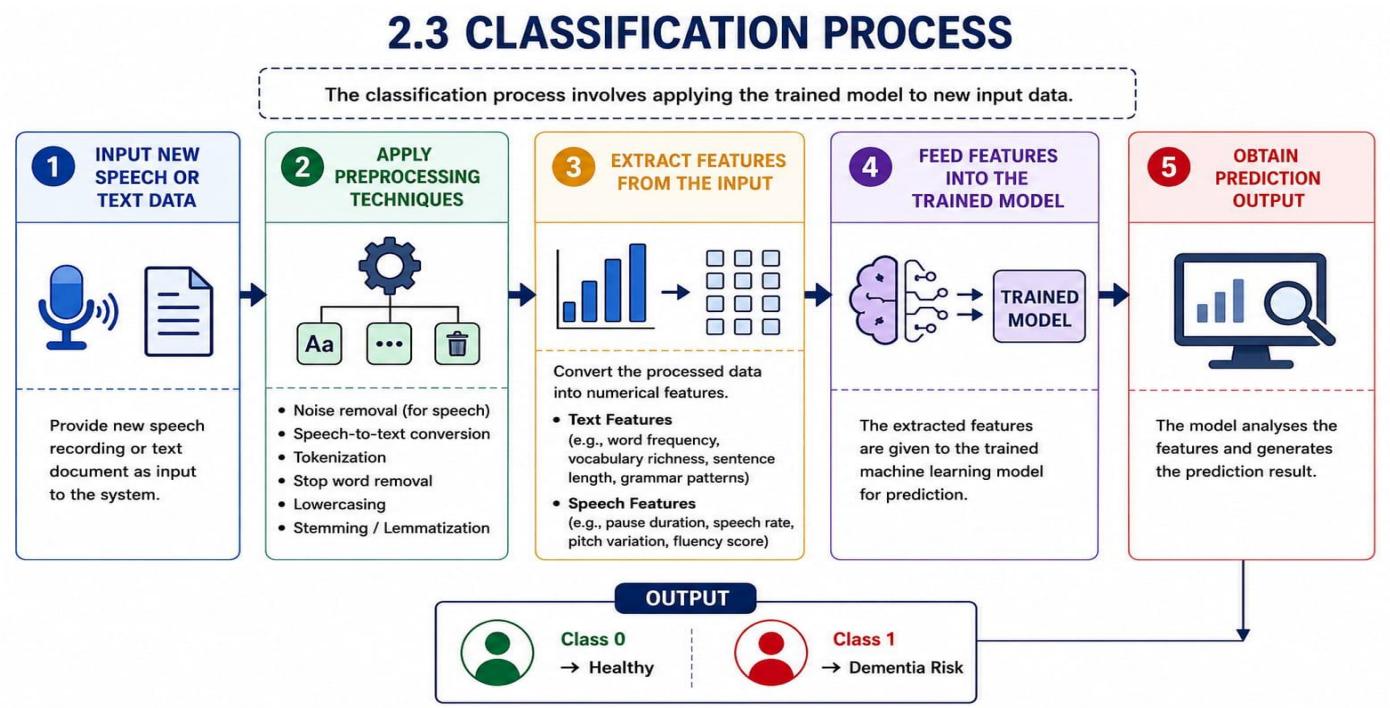
• Support Vector Machine (SVM) uses decision boundaries to divide data into many classifications.
• A number of decision trees are merged by Random Forest to reduce over fitting and improve accuracy.
• Logistic regression offers a simple probabilistic method for classification.
• Long Short-Term Memory (LSTM) networks can be used to interpret sequential data, such as speech and text.

From the training set, each model picks up patterns that aid in forecasting results for fresh inputs.

2.3 CLASSIFICATION PROCESS

The classification process involves applying the trained model to new input data.
• Input new speech or text data.
• Apply preprocessing techniques.
• Extract features from the input.
• Feed features into the trained model.
• Obtain prediction output.

Output:
• Class 0 → Healthy
• Class 1 → Dementia Risk



2.4 PERFORMANCE EVALUATION

Standard classification criteria that assess the model's accuracy in identifying speech-related cognitive deterioration and text data are utilized to gauge the effectiveness of the recommended early dementia detection method. Since the system is intended for medical prediction, both correct detections and possible errors must be carefully assessed.

2.4.1. Accuracy: By calculating the proportion of precise projections among all forecasts, accuracy assesses the system's overall performance.
• It includes both correctly identified healthy individuals and dementia patients.
• A model that works well on the dataset is indicated by a higher accuracy.
• However, in dementia detection, accuracy alone is insufficient since there might be dire repercussions if a dementia case is missed.

2.4.2. Precision: Precision quantifies the percentage of individuals that are accurately categorized as "dementia risk."
• It concentrates on reducing false alarms (healthy people wrongly predicted as dementia).
• High precision guarantees that there is no needless fear or incorrect diagnosis caused by the system.
• This is important in real-world healthcare applications where incorrect labeling can affect patient confidence.

2.4.3. Recall (Sensitivity) Recall measures how effectively the system detects actual dementia cases.
• It calculates how many real dementia patients are correctly identified. High recall is extremely important in this application because missing a dementia case (false negative) can delay diagnosis and treatment.
• A good model should aim for high recall to ensure early detection.

2.4.4. F1-Score: Precision and recall are balanced by the F1-score.
• It is helpful when both identifying actual cases and preventing false positives are crucial.
• In dementia detection, F1-score gives a deeper comprehension of overall model reliability.
• A higher F1-score suggests that the model consistently achieves good recall and precision.

2.4.5. Confusion Matrix (Application-Based): The confusion matrix helps in understanding the model's predictions in detail
.• True Positive (TP): Correctly detected dementia patients based on speech/text patterns
• True Negative (TN): Correctly identified healthy individuals
• False Positive (FP): Healthy individuals wrongly classified as dementia risk
• False Negative (FN): Dementia patients wrongly classified as healthy (most critical error)

In this application, minimizing False Negatives is very important because undetected dementia can delay medical intervention.

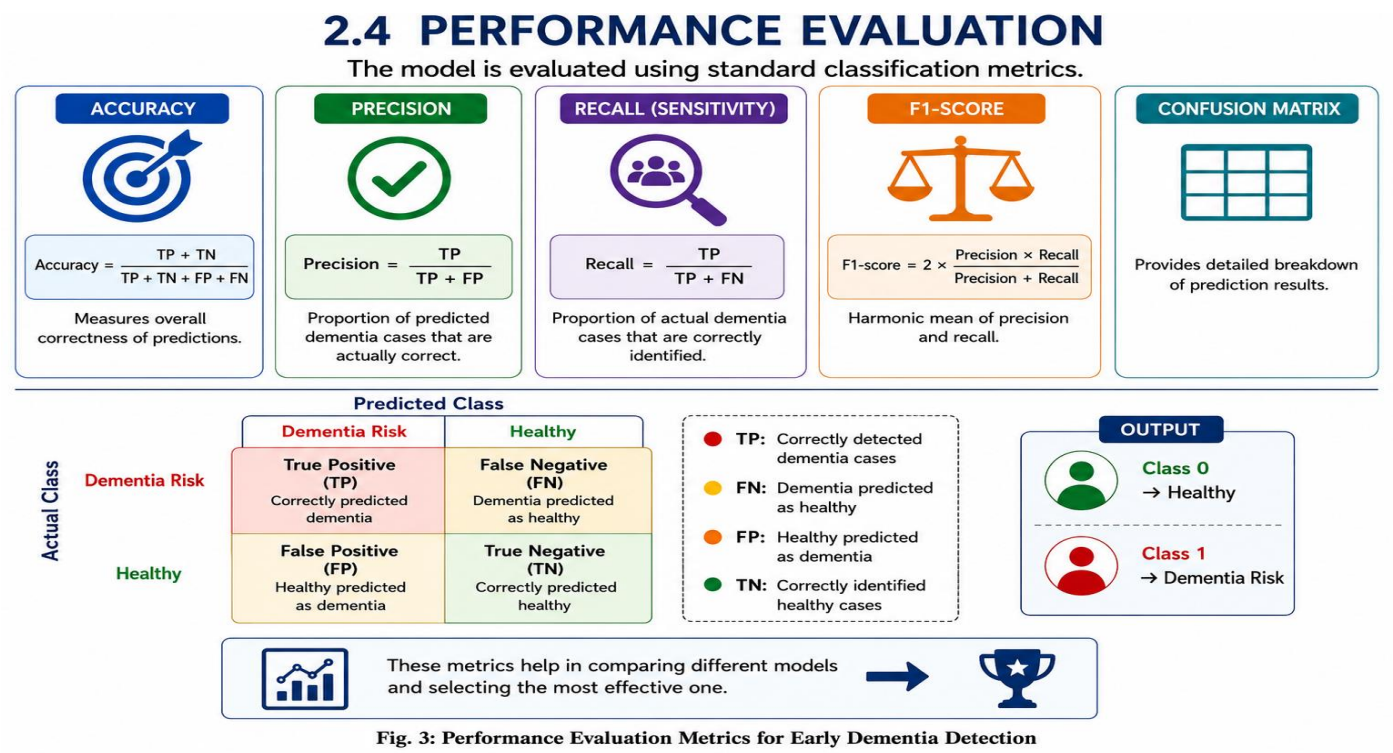


Fig. 3: Performance Evaluation Metrics for Early Dementia Detection

2.5 RESULTS AND PERFORMANCE ANALYSIS: The proposed dementia detection system was evaluated using speech and text datasets collected from healthy individuals and patients showing early symptoms of dementia. The extracted linguistic and acoustic features were used to train and test different ML models. The dataset was divided into training and testing sets using an 80:20 ratio. Performance was measured using Accuracy, Precision, Recall, and F1-Score. The results show that the LSTM model achieved the highest performance because of its ability to capture sequential speech and language patterns effectively. Random Forest also produced strong classification accuracy with reduced over fitting. The proposed hybrid ML and NLP approach performs better than traditional questionnaire-based or generic health monitoring systems because it analyzes both linguistic and acoustic behavioral patterns automatically. The confusion matrix analysis also indicates that the proposed system successfully minimizes false negatives, which is highly important in dementia detection since missed diagnosis may delay medical treatment.

Conclusion of Evaluation: These assessment metrics offer a comprehensive view of the system's efficacy in early dementia identification. The most dependable model can be selected by taking into account F1 score, recall, accuracy, and precision all at once. For accurately and quickly identifying dementia cases from speech and text data, recall and F1-score in particular are essential.

2.6 CLINICAL VALIDATION AND SAFETY: To ensure reliability and safety, the generated predictions and feature interpretations were compared with established dementia-related linguistic indicators available in clinical research literature. The extracted speech and text characteristics such as hesitation, reduced vocabulary richness, and fluency reduction are consistent with commonly observed symptoms of cognitive decline. The system is designed as a supportive screening tool rather than a replacement for professional medical diagnosis. Final clinical decisions should always be reviewed and confirmed by healthcare professionals, neurologists, or trained medical experts. Future improvements may include collaboration with clinical specialists and nutritionists to further validate prediction outputs and improve healthcare reliability.

III. FUTURE WORK:

The proposed early dementia detection system can be enhanced and expanded in several ways to improve its accuracy, usability, and real-world effectiveness.

• **Implementation of Advanced Models:** To further enhance the system's performance, future research can concentrate on implementing advanced deep learning models, like transformer based architectures like

BERT and GPT. These models are better at comprehending context and semantic meaning, which can enhance speech and text data processing and result in more precise early dementia identification.

• **Real-Time Monitoring Systems:** The system may be expanded into a real-time monitoring tool that continuously analyzes speech patterns during conversations. This would enable early and ongoing detection of cognitive decline without requiring manual input, making it more practical for everyday use.

• **Multilingual Support:** Currently, many systems are limited to a single language. More people will be able to use the system if it is expanded to handle multiple languages. This requires adapting NLP models to handle different linguistic structures and speech variations.

• **Integration with Mobile and Healthcare Applications:** The system can be integrated into mobile applications or healthcare platforms, allowing users to easily record speech and receive analysis results. This integration can also allow doctors to monitor patients remotely and maintain digital health records for improved diagnosis and treatment planning.

IV. CONCLUSION

This work describes a ML and NLP based system to analyze speech and text to be able to detect dementia early. The suggested method focuses on spotting minute voice and speech abnormalities that emerge in the early phases of cognitive decline.The system can distinguish between individuals in good health and those who are at risk of dementia by examining characteristics including word usage, sentence structure, pause length, and speech fluency. The application of a multi-stage pipeline, it involves the extraction of features, data preparation,, model training, and classification, ensures that unprocessed data is effectively transformed into meaningful insights. By using a range of machine learning models, the system may find complex patterns in the data and improve prediction accuracy. Evaluation using metrics like as accuracy, precision, recall, and F1 score demonstrates how reliable and effective the recommended method is. All things considered, the technology offers a scalable, economical, and non-invasive method for early dementia screening. By facilitating prompt diagnosis and action, it can be a useful support tool for medical personnel, ultimately enhancing patient care and quality of life. This study introduced a hybrid machine learning and natural language processing method for early dementia detection through text and speech analysis.

To find early signs of cognitive deterioration, the suggested method examines both linguistic and auditory characteristics, such as word richness, sentence structure, pause length, speech fluency, and pitch fluctuation. The integration of multiple ML models such as Random Forest, SVM, Logistic Regression, and LSTM improves prediction capability and enables accurate classification of dementia risk. Experimental evaluation demonstrated promising performance, with LSTM achieving the highest accuracy and F1-score among the tested models. In contrast to conventional diagnosis methods that could include costly medical imaging and professional assessment, the suggested system offers a non-invasive, scalable, and economical option for initial dementia screening. The system's practical usability in a variety of healthcare settings is further enhanced by the bilingual support. Larger multilingual datasets, real-time monitoring apps coupled with healthcare platforms, and transformer-based deep learning models like BERT and GPT can all be used to improve the system in the future. Healthcare practitioners could benefit greatly from the proposed effort in terms of early diagnosis, prompt intervention, and better patient care.

REFERENCES

[1] J. R. Searle, “Speech acts and cognitive function in dementia,” Cognitive Neuropsychology, no. 4, vol. 12, pp. 345–362, 2015 [2] D. Weiner, P. Coley, and S. De Jager, “Language and communication deficits in Alzheimer’s disease,” Journal of Neurolinguistics, vol. 25, no. 6, pp. 520 535, 2012.

[3] F. Eyben, M. Wöllmer, and B. Schuller, “OpenSMILE: The Munich versatile and fast open source audio feature extractor,” in Proc. ACM Multimedia, 2010, pp. 1459–1462.

[4] S. Hochreiter and J. Schmidhuber, “Long short term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Pearson, 2023.

[6] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[7] K. Fraser, F. Rudzicz, and G. Hirst, “Detecting late life depression in cases of Alzheimer's through analysis of speech and language,” Computer Language and Speech, vol. 29, no. 1, pp. 172–184, 2015.

[8] K. Fraser, F. Rudzicz, and G. Hirst, “Linguistic Alzheimer's disease characteristics in narrative speaking,” Journal of Alzheimer’s Disease, vol. 49, no. 2, pp. 407–422, 2016.

[9] J. Alhanai, R. Au, and J. Glass, "Spoken language biomarkers for detecting cognitive impairment," IEEE Transactions on Audio, Speech, and Language Processing, vol. 25, no. 11, pp. 2081–2093, 2017.