

# A novel AI model for the extraction and prediction of Alzheimer disease from electronic health record

Sonam V. Maju, Gnana Prakasi Oliver Sirya Pushpam

Department of Computer Science and Engineering, CHRIST University, Bengaluru, India

## Article Info

### Article history:

Received May 4, 2024

Revised Sep 17, 2024

Accepted Sep 30, 2024

### Keywords:

Big data

Dark data

Electronic health record

Machine learning

Natural language processing

## ABSTRACT

Dark data is an emerging concept, with its existence, identification, and utilization being key areas of research. This study examines various aspects and impacts of dark data in the healthcare domain and designs a model to extract essential clinical parameters for Alzheimer's from electronic health records (EHR). The novelty of dark data lies in its significant impact across sectors. In healthcare, even the smallest data points are crucial for diagnosis, prediction, and treatment. Thus, identifying and extracting dark data from medical data corpora enhances decision-making. In this research, a natural language processing (NLP) model is employed to extract clinical information related to Alzheimer's disease, and a machine learning algorithm is used for prediction. Named entity recognition (NER) with SpaCy is utilized to extract clinical departments from doctors' descriptions stored in EHRs. This NER model is trained on custom data containing processed EHR text and associated entity annotations. The extracted clinical departments can then be used for future Alzheimer's diagnosis via support vector machine (SVM) algorithms. Results show improved accuracy with the use of extracted dark data, highlighting its importance in predicting Alzheimer's disease. This research also explores the presence of dark data in various domains and proposes a dark data extraction model for the clinical domain using NLP.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Sonam V. Maju

Department of Computer Science and Engineering, CHRIST University

Bengaluru, India

Email: sonam.maju@res.christuniversity.in

## 1. INTRODUCTION

Data is everything, and by the year 2020, there will be 2.5 quintillion bytes of data saved [1]-[3]. The International Data Cooperation projects that 175 zeta-bytes of fresh data will be created globally in 2025, based on Forbes' prediction that global data creation will reach 35 zettabytes (35 trillion gigabytes) in 2020. In a literal sense, David J. Hand offers an overview, uses and outcomes of dark data, and its applications [4]. The paper provides a concise explanation of dark data using real-world examples such as COVID-19, autism, and the challenger space shuttle disaster in 1986 and mentioned dark data being in a domain where the presence of dark data is less evident but frequent in a medical diagnosis. According to David Hand, statisticians are quite aware of certain varieties of dark data and a prominent example of dark data is that that statisticians are aware of is non-response in surveys, where respondents frequently indicate that they do not wish to be included in the group of respondents who have provided a certain response. The existence of dark data across industries including IoT, healthcare, finance, multimedia, etc. is depicted in Figure 1. Dark data may be found in many different sectors, but among the most crucial is the medical industry. Finding dark data from medical records can help prevent certain diseases and manage their severity. Further data will be

gathered from patients and healthcare professionals based on all the readings from the health monitoring sensors. IoT devices come in a variety of forms and can gather health data from patients. Any combination of clinical study conducted by academics, research on pathogens and other micro-organisms conducted by medical foundations, other case studies, high-drama ER treatments, and unexpected virus outbreaks can be included in the data collected from IoT devices. Determining the association and connection between illnesses is a crucial factor in the early identification of illnesses. Figure 1 illustrates the startling quantity of black data in healthcare, and it is obvious that removing this data from electronic health records electronic health records (HER) records will improve decision-making.

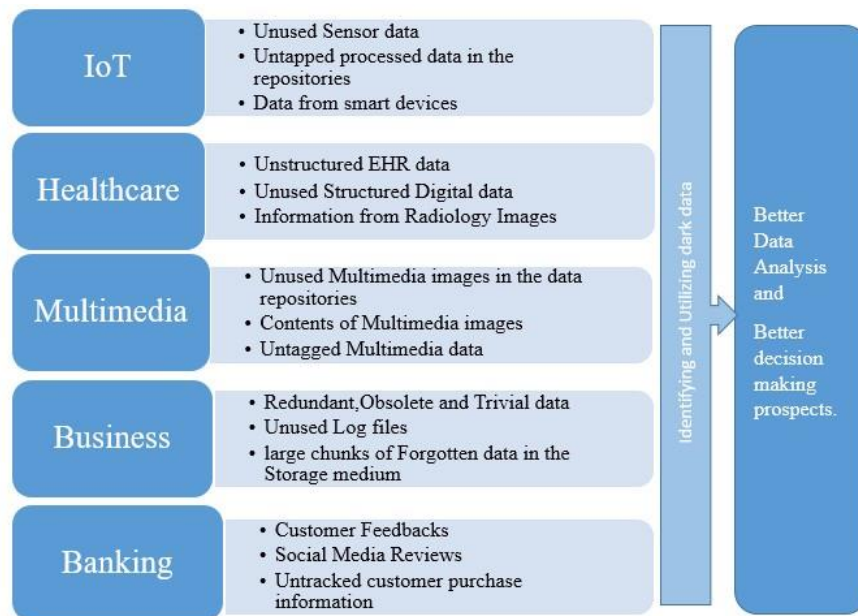


Figure 1. Impact of dark data in different sectors

The medical records go through data analysis, identification, extraction, and numerous other processes in order to obtain the black data from the EHR. States that dark data is any data that is large in quantity, inert, and necessitates the use of artificial intelligence techniques to mine and categorize it [5]. EHR are currently used by all clinical forums to track patient updates and previous treatment outcomes. EHR is being widely promoted by technological advancements and hospital information systems (HIS) [6] for pertinent medical analysis of clinical data. It is important to convert unstructured data to structured data because EHR include a vast amount of unstructured data and there is a risk of losing valuable information for disease analysis. Unstructured and semi-structured data can be transformed into structured data, opening a vast quantity of information for study, analysis, and learning. The outcomes can be utilized to create machine learning models and analyze data for the purpose of early disease diagnosis or prediction. For clinical and biomedical data [7], the authors present a versatile, production-grade named entity recognition (NER) system that can be used to identify the entities/keywords and may also be utilized to uncover underlying patterns and intent. Tarcar *et al.* [8] uses a mix of natural language processing techniques and an online annotation tool to maximize the performance of a customized NER model that was trained on a minimal amount of EHR training data. This solves the issue of data extraction from unstructured health records. Early on in Alzheimer's disease development, it is difficult to predict. Compared to later AD treatment, early AD treatment is more effective and causes less mild damage. Numerous studies offering various approaches have been published on disease prediction. The favourable correlation between neurological problems and a few other parameters such as medications, down syndrome, dementia, and cognitive impairment. As possible risk factors for seizures in Alzheimer's disease is covered in these studies. Based on a well-known large population study carried out in Rotterdam, [9]-[12] links dementia and diabetes and demonstrates the correlation of diseases. Research by [13], [14] examine and incorporate a variety of methodologies employed in various study genres, perhaps opening up new avenues for Alzheimer's disease research [15]. The items in seven-specified categories can be identified using the NER model that is proposed in this research. In this study, a named entity recognition model for natural language processing is put forth to extract pertinent data

from electronic health records. To locate and extract dark data that is available in electronic health records, the study also covers the phases of text processing, entity construction, and entity recognition. The examination of the results for the Alzheimer disease prediction is completed at this point.

The research paper starts with an Introduction to dark data, Alzheimer disease, the problem statement and existing works with a glance of methodology used in this work. Then the method section describes the NLP methods and prediction methods. The results and discussion include, the visualization of the results obtained through the method and a comparative study of prediction model. In the conclusion, an overall summary with result analysis and future work is discussed.

**2. METHOD**

The EHR data contains the details of patients from different departments like bariatrics, gynaecology, physiotherapy etc and also contains the data from discharge cell. Among these data, our first challenge is to extract Alzheimer related clinical entities from the EHR, as this research work is focussed only in Alzheimer disease. For this an NLP model with SpaCy and NER technique is used. Second challenge is to calculate the prediction accuracy of the retrieved data and is analysed using support vector machine (SVM) machine learning algorithm. The whole process can be divided into 3 stages as mentioned in Figure 2, data collection, extraction of selected departments from the clinical database using SpaCy NER, Accuracy test with the machine learning algorithm. The architecture explains the detailed workflow of our proposed AI model, from the dataset of 4999 patients, the model will extract only the required information which is necessary for the prediction model.

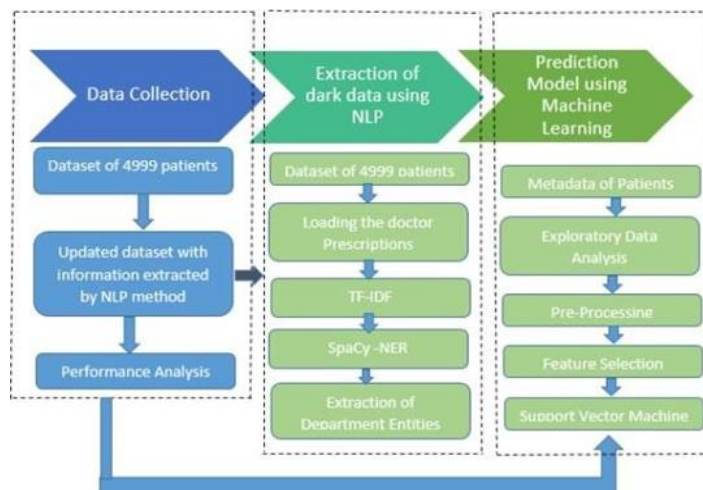


Figure 2. Architecture of prediction model

**2.1. Data collection**

Collecting clinical data and hospital data has become a challenge. The hospital dataset used in this research work is a publicly available dataset which has a good collection of transcribed medical reports from different clinical specialties. The main features of this datasets are, description, medical specialty, Sample name, transcription, and keywords. The description gives the abstract idea about the patient and the reason of hospital visit. The medical specialty gives the information about which clinical department the patient is consulting. Sample name gives the specific disease of the patient, and the transcription explains the patient details, medical history, allergies, symptoms, and assessment. There are forty-two medical specialties available in this dataset, a few of them are allergy/immunology, autopsy, bariatrics, cardiovascular/pulmonary, dermatology, endocrinology, neurology, and radiology. The most important part of this research work is to extract the medical specialties where the Alzheimer prediction parameters are available from description of the doctors about the patients.

**2.2. Extraction of selected departments using spaCy NER**

Extracting the required information from the dataset of 4,999 patients from 42 different clinical departments is a tedious task. As the focus of this research work is into Alzheimer disease, of the 42 departments, only the department which reflects the condition of Alzheimer need to be extracted. The

description column needs to be extracted and analyze the descriptions to identify the rows containing specific departments. Identifying the medical department entities from the ‘description’ parameters is done using NER. The importance of adding a new entity in this research work is shown in Figure 3, Figure 3(a) shows entity extraction without adding ‘specialty’ as a new entity with date, name, and quantity, and Figure 3(b) shows the information that will be extracted from the description with an added entity ‘specialty’ of the patients. The objective of our research work is to extract and identify clinical departments from the description of the patients. For the same a new entity needs to be added. The medical specialty feature is analyzed, and a list will be created with unique medical specialties along with the count of occurrences for each specialty.

<pre>Entities [('2d', 'CARDINAL')] Entities [('2d', 'CARDINAL')] Entities [('eea anastomosis', 'ORG'), ('many years', 'DATE')] Entities [] Entities [('2d', 'CARDINAL')] Entities [] Entities [('echocardiogram', 'DATE')] Entities [('25mm', 'QUANTITY')] Entities [] Entities [('cerebral angiogram', 'ORG'), ('moyamoya', 'ORG')] Entities [] Entities [('teeth 1 16 17 and 32', 'DATE'), ('teeth 1 16 17 and 32', 'DATE')] Entities [] Entities [('8', 'CARDINAL')]</pre>	<pre>Entities [('chiropractic', 'SPECIALTY')] Entities [('surgery', 'SPECIALTY')] Entities [('autopsy', 'SPECIALTY')] Entities [('surgery', 'SPECIALTY')] Entities [('surgery', 'SPECIALTY')] Entities [('surgery', 'SPECIALTY')] Entities [('surgery', 'SPECIALTY')] Entities [('surgery', 'SPECIALTY')] Entities [('chiropractic', 'SPECIALTY')] Entities [('surgery', 'SPECIALTY')] Entities [('surgery', 'SPECIALTY')]</pre>
(a)	(b)

Figure 3. Extraction of entities (a) data extracted before adding a new entity and (b) extraction after adding ‘specialty’ as new entity

Text pre-processing is done on the unique medical parameters, and this helps in consistency and ease of use in further processing. Later the named entities from the doctor’s description about the patient were recognized and SpaCy NLP pipeline is used for tokenizing the text, part-of-speech tagging, dependency parsing, named entity recognition, and other natural language processing tasks. Tokenization, as the name indicates, will slit the text into individual tokens and can be represented as shown (1).

$$S = t_1, t_2, \dots, t_n \quad (1)$$

Where  $t$  is the tokens and  $S$  is the set of all tokens. The entities and the respected entity labels will be extracted from the description, which is useful for understanding what types of named entities are present in the text data. Extraction of entities related to medical specialties is done using the previously set medical list, containing lowercased and stripped medical specialties and a custom function to process the description and extract entities related to medical specialties. To train a NER model using spaCy, the NER model was trained on custom training data represented by the train data list, which contains the processed ‘description’ texts along with their associated entity annotations. For custom NER, the encapsulation of annotations of text with corresponding entity labels can be represented as in (2).

$$A = (T_1, L_1), (T_2, L_2), \dots, t_n \quad (2)$$

Where  $T_1$  is the annotated text and  $L_1$  is the corresponding entity label. A blank spaCy language class model was created for the english language and this blank model does not have any pre-trained word vectors, and the NER component need to be trained from scratch. Each tuple’s annotations will iterate through the list of entities (start and end positions, and entity labels) and the label (entity type) of each entity is added to the NER component. Hence, in a higher dimensional space, word embedding can be represented as word vectors by converting words ( $w$ ) into word vectors ( $vw$ ) where, for a given embedding dimension  $d$ :

$$v_w = \epsilon R^d \quad (3)$$

The train data list is randomly shuffled ( $R$ ) before each training iteration to introduce variation during training. The examples need to be batched up in the train data list with varying sizes. This helps in efficient training with a dynamic batch size with the texts and annotations (entities) should be separated from the batch. The training for NER is performed using the provided training data and returns the trained NER model. The training process involves updating the model with batches of texts and annotations for a specified

number of iterations. This way, the model learns to recognize and predict entities within the 'description' texts. So, in fact the motive is to predict an entity label for each token in the description. The prediction  $P_i$  for the  $i$ th token can be represented as function  $F$ , which intakes the features and context of the token as input:

$$P_i = \mathcal{F}(v_t, context) \tag{4}$$

in order to minimize the loss function in the training of NER model can be represented as identifying the model parameters that minimizes the loss:

$$\theta^* = \arg \min_{\theta} \sum^n Loss(P_i, P_i) \tag{5}$$

where loss indicates the loss function comparing predicted labels  $P_i$  to  $L_i$ . Thus, the performance of the trained NER model can be evaluated on both the training and test data. It shows how well the model identifies entities in unseen test data after training. Only after adding specialty as one new entity, the algorithm was able to extract and identify clinical departments as specialty, which is shown in Figure 3(b). After adding 'specialty' as new entity. After identifying the new entities from each description, the next step is to validate the algorithm with all the doctor's description of patients and make sure the description which specifies a clinical department needs to be extracted. The validation results of the algorithm are shown in Figure 4, where the specialty entity is recognized and extracted if it's mentioned in the given description and gives null value if a clinical department is not mentioned in the given text.

```

autopsy asphyxia due to ligature strangulation
Entities [('autopsy', 'SPECIALTY')]

autopsy ligature strangulation and craniocerebral injuries
Entities [('autopsy', 'SPECIALTY')]
    
```

Figure 4. Validation of extracted clinical department

Once all the required departments for Alzheimer predictions are extracted from the electronic health record, the next hurdle is to identify the prediction parameters of Alzheimer disease. From The literature survey eleven prediction parameters were shortlisted and validated. The eleven prediction parameters are, Age, Gender, mini-mental state examination (MMSE), clinical dementia rating (CDR), normalized whole-brain volume (nWBV), estimated total intracranial volume (eTIV), atlas scaling factor (ASF), cholesterol, diabetics, blood pressure and body mass index.

**2.3. Prediction model**

As defined, this research focuses on the early prediction of Alzheimer's disease from a dataset of 4,999 data points. Generally, for early prediction of Alzheimer's, features like Mini-Mental State Examination (MMSE), normalized whole- brain volume (nWBV), estimated total intracranial volume (eTIV), and Atlas scaling factor (ASF) are used from the EHR records. However, the EHR data of any health record includes many other features; for example, the diabetes, Blood Pressure, Cholesterol -related parameters may be recorded but not utilized for calculating the results, and thereby becoming dark data. In this research, along with the Alzheimer's features, the relation of Alzheimer's with other diseases is identified, and those parameters are also included in training the prediction model to increase the accuracy of the model and for better predictions. The above-mentioned parameters are extracted from electronic health records, pre-processed, and divided into training and testing data, and later prediction analysis of Alzheimer disease is done through SVM algorithm. SVM is used in identifying a hyperplane that maximizes the margin between Alzheimer and non -Alzheimer patients. The maximum distance between the nearest data points of these two different classes can be done using the mathematical equation of hyperplane:

$$w \cdot x + b = 0 \tag{6}$$

where,  $w$  is the weight perpendicular to the hyperplane,  $x$  is the input feature vector and  $b$  is the bias term. SVM aims to maximize the distance between the hyperplane and nearest data point is called margin. The algorithm needs to maximize this margin while minimizing the classification errors. The distance from a data point  $x_i$  to the hyperplane can be calculated by the following formula, distance:

$$\text{Distance, } D = \frac{|w \cdot x_i + b|}{\|w\|} \text{ and margin can be calculated by, } \frac{2}{\|w\|}$$



Mathematically the above-mentioned formulas will work perfectly for hyperplane and maximizing the distance of the margin but practically the data might not be easily separable. So, a regularization parameter C can be used as a trade of between maximizing the margin and minimizing the classification errors. The appropriate value of C can be done using hyper parameter tuning using techniques like cross validation. Here, k fold cross validation method is used with 5 as the value of k, later the scores are calculated with mean cross validation accuracy. The model needs to train and evaluate with different C values to determine which value performs best on the data points. Another important parameter of SVM is the kernel which subtly map the data points to a higher-dimensional space, allowing SVMs to understand the nonlinear relationships in the data. The suitable choice of the kernel bandwidth or parameter is critical for obtaining good performance with the RBF kernel. The RBF kernel and gamma parameter can be expressed mathematically as follows:

$$K(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$$

where x and x' are the input feature vectors,  $\|x - x'\|$  is the representation of Euclidean distance between the vectors,  $\sigma$  is the kernel bandwidth and finally  $\frac{1}{2\sigma^2}$  is the  $\gamma$  parameter. Thus, the SVM model is designed after iterating the values for each parameter, like C, gamma, and kernel values and then the performance is analysed. The accuracy performance of the curated SVM model on the dataset is analyzed and depicted in Table 1 with best accuracy on cross validation is 98%, test accuracy with best parameters is 97% and Recall is 95%.

### 3. RESULTS AND DISCUSSION

The results of this research work can be explained in two ways, extracting Alzheimer related parameters from EHR records prediction of Alzheimer disease using the extracted parameters. The analysis of patient entry in each clinical department is done and is visualized using word cloud, which is shown in Figure 5. After the basic analysis, the first challenge of the research work was to extract the required departments from doctor's description through Natural Language processing and the results of extraction are plotted in Figure 6 where Figure 6(a) shows the list of all the medical departments extracted from the description and the number of occurrences. From all the entities extracted, Figure 6(b) shows the entities utilized for the prediction of Alzheimer disease.



Figure 5. Clinical department analysis

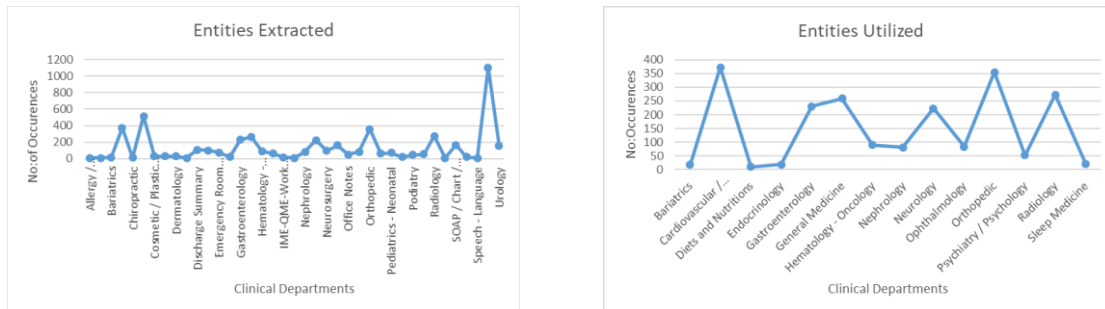


Figure 6. Extraction of clinical departments (a) the extracted medical entities and (b) utilized medical entities

After, extraction of entities, Alzheimer prediction parameters were identified, and a feature ranking is done among the mentioned eleven parameters. The feature importance of each parameter is presented in Figure 7. By the identification and addition of non-conventional parameters like diabetes, cholesterol and blood pressure for Alzheimer prediction, there is a huge difference in the accuracy and can bring better performance in the early prediction [16]. The feature importance can be done using regression models, the hybrid model of Lasso, Ridge and linear regression is used for calculating the feature ranking and the results are shown in Figure 7. Secondly, the results obtained after using SVM as the prediction algorithm is shown in Table 1, which clearly shows the importance of considering Glyhb aka diabetics as a very important parameter in diagnosing Alzheimer's. The accuracy obtained by considered only the conventional Alzheimer parameters was .7785 and is showing an improved performance when a new parameter was added, glyhb it will be .9892. The comparative study of the results of all existing methods are shown in Table 2. Eventhough the proposed model shows promising results, integrating quantum machine learning for refining the results is a scope for future research.

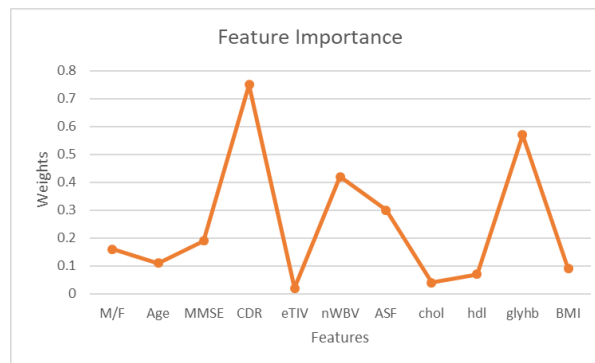


Figure 7. Feature ranking

Table 1. Performance metrics

Performance metrics	Results
Best accuracy on cross validation	98%
Best parameter for c	100
Best parameter for gamma	1
Test accuracy with best parameters	97%
Test recall with best parameters	95%

Table 2. Comparative analysis

Reference	Methods	Accuracy
Computational modelling of dementia prediction using deep neural network: analysis on OASIS Dataset [17]	Deep neural networks	92%
MRI deep learning-based solution for Alzheimer's disease prediction [18]	Deep learning and image processing technique	88%
Early-Stage Alzheimer's disease prediction using machine learning models [19]	Machine Learning Models	83%
Deep learning approach for early detection of Alzheimer's disease [20]	Convolutional neural networks	93%
An efficient deep neural network binary classifier for Alzheimer's disease classification [20]	Deep neural network binary classifier	85%
Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional auto encoders [21]	Deep learning using convolutional auto encoders	80%
Alzheimer's disease prediction using machine learning techniques and principal component analysis [22]	Machine learning models	75%
Explainable AI-based Alzheimer's prediction and management using multimodal data.[23]	Comparative Study of Alzheimer disease prediction with different machine learning models	Random forest (98%)
Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing [24]	NLP Methods	Differ based on category (ranges from 54% -96%)
Learning implicit sentiments in Alzheimer's disease recognition with contextual attention features [25]	NLP	91.6%

#### 4. CONCLUSION

Extracting the relevant data for analysis and processing can be very challenging. Despite numerous solutions available for data extraction, these methods must adapt to the specific characteristics of the problem at hand. While the recent studies work on conventional Alzheimer parameters, our method emphasizes on more clinical parameters, hidden as dark data. There could be correlations between Alzheimer's and other neurological conditions like bipolar disorder that need to be identified. Based on results, the proposed methodology successfully extracted the necessary departments from doctors' prescriptions using SpaCy NER, and the metadata were used as prediction parameters achieving an impressive accuracy of 97%. In the future, more diseases and symptoms will be identified through clinical research by incorporating quantum machine learning algorithms and converting the classical data into quantum data

#### REFERENCES




- [1] B. Marr, "How much data do we create every day? The mind-blowing stats everyone should read," *Forbes*. [Online]. Available: <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read>
- [2] M. Hobart, "The 'dark data' conundrum," *Computer Fraud and Security*, vol. 2020, no. 7, pp. 13–16, Jan. 2020, doi: 10.1016/S1361-3723(20)30075-0.
- [3] L. Clissa, M. Lassnig, and L. Rinaldi, "How big is big data? a comprehensive survey of data production, storage, and streaming in science and industry," *Frontiers in Big Data*, vol. 6, Oct. 2023, doi: 10.3389/fdata.2023.1271639.
- [4] D. J. Hand, "Dark data," *Significance*, vol. 17, no. 3, pp. 42–44, Jun. 2020, doi: 10.1111/1740-9713.01406.
- [5] H. S. Kim, "Dark data in real-world evidence: challenges, implications, and the imperative of data literacy in medical research," *Journal of Korean Medical Science*, vol. 39, no. 9, 2024, doi: 10.3346/jkms.2024.39.e92.
- [6] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: A review," *Journal of Healthcare Engineering*, vol. 2018, pp. 1–9, 2018, doi: 10.1155/2018/4302425.
- [7] V. Kocaman and D. Talby, "Accurate clinical and biomedical named entity recognition at scale," *Software Impacts*, vol. 13, p. 100373, Aug. 2022, doi: 10.1016/j.simpa.2022.100373.
- [8] A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, "A clinical named entity recognition model using pretrained word embedding and deep neural networks," *Decision Analytics Journal*, vol. 10, p. 100426, Mar. 2024, doi: 10.1016/j.dajour.2024.100426.
- [9] K. Munot, N. Mehta, S. Mishra, and B. Khanna, "Importance of dark data and its applications," in *2019 IEEE International Conference on System, Computation, Automation and Networking, ICSCAN 2019*, IEEE, Mar. 2019, pp. 1–6, doi: 10.1109/ICSCAN.2019.8878789.
- [10] A. Javeed, A. L. Dallora, J. S. Berglund, A. Ali, L. Ali, and P. Anderberg, "Machine learning for dementia prediction: a systematic review and future research directions," *Journal of Medical Systems*, vol. 47, no. 1, p. 17, Feb. 2023, doi: 10.1007/s10916-023-01906-7.
- [11] M. Asadollahi, M. Atazadeh, and M. Noroozian, "Seizure in Alzheimer's disease: an underestimated phenomenon," *American Journal of Alzheimer's Disease and other Dementias*, vol. 34, no. 2, pp. 81–88, Mar. 2019, doi: 10.1177/1533317518813551.
- [12] V. Viswan, N. Shaffi, M. Mahmud, K. Subramanian, and F. Hajamohideen, "Explainable artificial intelligence in Alzheimer's disease classification: a systematic review," *Cognitive Computation*, vol. 16, no. 1, pp. 1–44, Jan. 2024, doi: 10.1007/s12559-023-10192-x.
- [13] Z. Sun and C. Tao, "Named entity recognition and normalization for Alzheimer's disease eligibility criteria," in *Proceedings - 2023 IEEE 11th International Conference on Healthcare Informatics, ICHI 2023*, IEEE, Jun. 2023, pp. 558–564, doi: 10.1109/ICHI57859.2023.00100.
- [14] A. Ott *et al.*, "Prevalence of Alzheimer's disease and vascular dementia: association with education. The Rotterdam study," *Bmj*, vol. 310, no. 6985, p. 970, Apr. 1995, doi: 10.1136/bmj.310.6985.970.
- [15] A. K. Tarcar, A. Tiwari, D. Rao, V. N. Dhaimodker, P. Rebelo, and R. Desai, "Healthcare NER models using language model pretraining," in *CEUR Workshop Proceedings*, New York, NY, USA: ACM, Jan. 2020, pp. 12–18, doi: 10.1145/3336191.3371879.
- [16] S. V. Maju and G. Prakasi Oliver Sirya Pushpam, "A novel two-tier feature selection model for Alzheimer's disease prediction," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 33, no. 1, p. 227, Jan. 2024, doi: 10.11591/ijeecs.v33.i1.pp227-235.
- [17] S. Basheer, S. Bhatia, and S. B. Sakri, "Computational modeling of dementia prediction using deep neural network: analysis on OASIS dataset," *IEEE Access*, vol. 9, pp. 42449–42462, 2021, doi: 10.1109/ACCESS.2021.3066213.
- [18] C. L. Saratxaga *et al.*, "Mri deep learning-based solution for Alzheimer's disease prediction," *Journal of Personalized Medicine*, vol. 11, no. 9, p. 902, Sep. 2021, doi: 10.3390/jpm11090902.
- [19] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. Tavera Romero, "Early-Stage Alzheimer's disease prediction using machine learning models," *Frontiers in Public Health*, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.853294.
- [20] H. A. Helaly, M. Badawy, and A. Y. Haikal, "Deep learning approach for early detection of Alzheimer's disease," *Cognitive Computation*, vol. 14, no. 5, pp. 1711–1727, Sep. 2022, doi: 10.1007/s12559-021-09946-2.
- [21] R. Prajapati, U. Khatri, and G. R. Kwon, "An Efficient Deep Neural Network Binary Classifier for Alzheimer's disease classification," in *3rd International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2021*, IEEE, Apr. 2021, pp. 231–234, doi: 10.1109/ICAIIIC51459.2021.9415212.
- [22] F. J. Martinez-Murcia, A. Ortiz, J. M. Gorriz, J. Ramirez, and D. Castillo-Barnes, "Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 17–26, Jan. 2020, doi: 10.1109/JBHI.2019.2914970.
- [23] M. Sudharsan and G. Thailambal, "Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA)," *Materials Today: Proceedings*, vol. 81, no. 2, pp. 182–190, 2021, doi: 10.1016/j.matpr.2021.03.061.
- [24] S. Jahan *et al.*, "Explainable AI-based Alzheimer's prediction and management using multimodal data," *PLoS ONE*, vol. 18, no. 11 November, p. e0294253, Nov. 2023, doi: 10.1371/journal.pone.0294253.






- [25] I. Y. Oh, S. E. Schindler, N. Ghoshal, A. M. Lai, P. R. O. Payne, and A. Gupta, "Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing," *JAMIA Open*, vol. 6, no. 1, Jan. 2023, doi: 10.1093/jamiaopen/ooad014.
- [26] N. Liu, Z. Yuan, Y. Chen, C. Liu, and L. Wang, "Learning implicit sentiments in Alzheimer's disease recognition with contextual attention features," *Frontiers in Aging Neuroscience*, vol. 15, May 2023, doi: 10.3389/fnagi.2023.1122799.

## BIOGRAPHIES OF AUTHORS



**Sonam V. Maju**    received the B.Tech. degree in information technology from Mahatma Gandhi University, India in 2013 and the M.Tech. degree in computer science and engineering from Mahatma Gandhi University, India in 2015. She is currently pursuing the Ph.D. degree in computer science and engineering at CHRIST University, Bangalore, India. Her research interest includes the machine learning, natural language processing and clinical research. She has published one Indian patent and several other publications on Clinical Research. She can be contacted at email: sonam.maju@res.christuniversity.in.



**Gnana Prakasi Oliver Sirya Pushpam**    is an associate professor at the CHRIST University, department of computer science and engineering, Bangalore, India. She received her B.Tech. degree in information technology and M.E. degree in software engineering from Anna University. Her research interests include mobile Ad hoc Networks, IoT, software engineering and machine learning. Major research focus in MAC and network layers and published research papers in reputed journals. Currently focusing on analyzing the prediction algorithms using recurrent neural network with clinical data. She has more than 15 publications in refereed national and international journals. She can be contacted at email: gnana.prakasi@christuniversity.in.