Advancing Dementia Prediction: Integrating Big Data Analytics and Machine Learning Models for Genomic Analysis

Ajay Hiremath
Post grad student at the
University of Nottingham
MSc in Data Science
Nottingham, UK
ajaychiremathgreprep@gmail.com

Prathamesh Mohan Dharamthok
Post grad student at the
University of Nottingham
MSc in Data Science
Nottingham, UK
prathamesh.d1998@gmail.com

Pratyush Joshi
Post grad student at the
University of Nottingham
MSc in Data Science
Nottingham, UK
pratyush.joshi1996@gmail.com

Muktesh Gawale
Post grad student at the
University of Nottingham
MSc in Data Science
Nottingham, UK
mukteshgawale@gmail.com

Raghav Anand
Post grad student at the
University of Nottingham
MSc in Data Science
Nottingham, UK
anand.raghav93@gmail.com

Dummy Harish Vaibhav Kashyap
Post grad student at the
University of Nottingham
MSc in Data Science
Nottingham, UK
vaibhavkashyap@gmail.com

Abstract— **Dementia with Lewy bodies (DLB) stands as a significant public health concern, representing the second most prevalent neurodegenerative dementia. Transcriptomic analyses have unveiled signatures indicative of synuclein aggregation, protein degradation, amyloid deposition, neuroinflammation, and mitochondrial dysfunction, yet the holistic understanding of DLB's transcriptomic profile remains nascent. Crucially, the absence of disease-modifying treatments or blood-based diagnostic biomarkers underscores the urgency for further research. This research paper endeavors to bridge this gap by proposing the application of data science methodologies, including genomic prediction and machine learning analyses of multi-omics data. By leveraging these advanced techniques, we aim to enhance our understanding of DLB's molecular underpinnings, expedite therapeutic interventions, and facilitate drug development efforts. Through a comprehensive review of the current genomic and transcriptomic landscape of DLB, coupled with an exploration of the potential of data science methodologies, this paper aims to provide insights that may lead to novel diagnostic and therapeutic strategies for DLB, ultimately mitigating the burden it imposes on individuals and society at large.**

Keywords—Data, Dementia, Lewy bodies, Data preprocessing, machine learning, big data analysis.

I. Introduction and Literature Review

Dementia refers to a decline in cognitive abilities that impairs daily functioning. It encompasses a range of symptoms such as memory loss, reasoning difficulties, and communication challenges. Symptoms may vary but commonly include memory problems, confusion, and mood changes. Timely diagnosis and effective management strategies are crucial for enhancing the well-being of individuals at risk and living with dementia and their caregivers. In recent years, the intersection of big data analytics and healthcare has paved the way for groundbreaking advancements in disease prediction and management. Leveraging the vast wealth of genomic data available, researchers are increasingly turning to machine learning models to unravel the intricate genetic underpinnings of dementia and predict individual susceptibility to the disease. Dementia, encompassing various subtypes such as Alzheimer's disease, vascular dementia, and dementia with Lewy bodies, poses a significant public health challenge globally. While dementia's causes are varied, recent findings highlight a significant genetic influence on disease susceptibility. We seek to develop predictive models capable of estimating an individual's likelihood of developing dementia.

This paper presents a comprehensive overview of our approach to addressing dementia prediction as a big data problem. We outline the methodology employed, including data preprocessing, feature selection, model training, and evaluation. Furthermore, we discuss the potential implications of our findings for clinical practice, public health policy, and future research directions. By leveraging the power of big data analytics and machine learning, we strive to advance our understanding of dementia and ultimately improve outcomes for affected individuals.

The aim of this study is to answer the following research questions.

1.What computational strategies and tools were employed to manage and process vast datasets in our study?

2.How did these approaches contribute to data handling efficiency, scalability, and the extraction of meaningful insights

3.What models were efficient in the prediction of the disease?

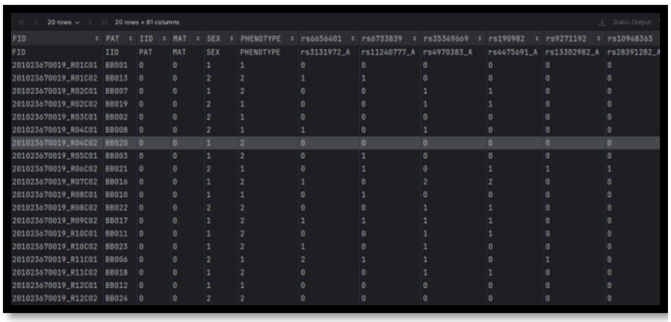
4.What features were selected to form the model along with what methods were used to extract those features?

II. Methodology

A. Data:
1. Data Format and Dimensions:
The dataset at our disposal comprises genetic data and demographic attributes of individuals. Initially stored in a RAW file format, the dataset presented formidable dimensions, comprising 535 rows and approximately 300,000 columns. This vast dataset posed inherent challenges in data handling and analysis, necessitating the

implementation of sophisticated processing techniques to extract meaningful insights.

2 Problems Faced: For effective data processing, we tried to import the entire dataset for initial analysis using PySpark. However, we had trouble loading and showing the complete dataset in an acceptable amount of time because of its enormous size. We experienced issues with Databricks, such as high processing times, which made the switch to a local machine approach necessary for increased productivity and efficiency. We attempted extracting a first subset of 50,000 rows to lessen the problem and improve the efficiency of exploratory data analysis. Unfortunately, handling and quickly displaying even this selection proved to be difficult, demonstrating the tremendous processing demands imposed by the sheer magnitude of the dataset.



20 rows ⌄ 20 rows × 81 columns Static Output

FID	PAT	IID	MAT	SEX	PHENOTYPE	rs6656401	rs6733839	rs35349669	rs190982	rs9271192	rs10948363
FID	IID	PAT	MAT	SEX	PHENOTYPE	rs3131972_A	rs11240777_A	rs4970383_A	rs4475691_A	rs13302982_A	rs28391282_A
201023670019_R01C01	BB001	0	0	1	1	0	0	0	0	0	0
201023670019_R01C02	BB013	0	0	2	2	1	1	0	0	0	0
201023670019_R02C01	BB007	0	0	1	2	0	0	1	1	0	0
201023670019_R02C02	BB019	0	0	2	1	0	0	1	1	0	0
201023670019_R03C01	BB002	0	0	2	1	0	0	0	0	0	0
201023670019_R04C01	BB008	0	0	2	1	1	0	1	0	0	0
201023670019_R04C02	BB020	0	0	1	2	0	0	0	0	0	0
201023670019_R05C01	BB003	0	0	1	2	0	1	0	0	0	0
201023670019_R06C02	BB021	0	0	2	1	0	1	0	1	1	1
201023670019_R07C02	BB016	0	0	1	2	1	0	2	2	0	0
201023670019_R08C01	BB010	0	0	1	1	0	1	0	0	0	0
201023670019_R08C02	BB022	0	0	2	2	0	0	1	1	0	0
201023670019_R09C02	BB017	0	0	1	2	1	1	1	1	0	0
201023670019_R10C01	BB011	0	0	1	2	0	0	1	1	0	0
201023670019_R10C02	BB023	0	0	1	2	1	0	1	0	0	0
201023670019_R11C01	BB006	0	0	2	1	2	1	1	0	1	0
201023670019_R11C02	BB018	0	0	1	2	0	0	1	1	0	0
201023670019_R12C01	BB012	0	0	1	1	0	0	0	0	0	0
201023670019_R12C02	BB024	0	0	2	2	0	0	0	0	0	0

B. Data Preprocessing and Cleaning:

1. Addressing Memory Issues: A distributed processing strategy was implemented to overcome the challenges posed by the dataset's size. The data was divided into 2977 distinct nodes, each containing approximately 100 columns. These nodes were orchestrated by a master node, resulting in a significant improvement in overall efficiency.

Two primary methods were employed:

> *Column Dropping*: Columns with over 10 "NA" (not available) values were eliminated since they were considered to have an excessive amount of missing data. This prevented potential biases introduced by a high proportion of missing values.
>
> *Imputation*: For the remaining missing values, a strategic imputation technique was applied. Instead of removing rows with missing data, the fillna () method was utilized to replace these "NA" values with the value ‘3’.

2. This approach was chosen over dropna () to preserve valuable observations and avoid the model's performance. This refined preprocessing strategy enabled efficient handling of the large data set.

Feature Selection: To address the high-dimensional genetic data with approximately 300,000 columns, a feature selection strategy was used to identify the most influential genetic factors for predicting the target variable, PHENOTYPE (disease code). This involved a combination of feature techniques, encoding methods, and statistical testing.

3. Feature Engineering:
    a. Categorical variables were transformed into numerical representations using OneHotEncoding, which creates new binary features for each unique category within a categorical feature.
    b. String Indexer was utilized to assign unique integer indices to distinct categories, enhancing the model's ability to handle categorical data effectively.

4. Statistical Testing:
    a. A Chi-squared test was performed to evaluate the association between each genetic variable and the PHENOTYPE (disease code).
    b. This statistical test enabled ranking the features based on their significance in predicting the disease code.
    c. A subset of top-ranking genetic variables, identified by their "rs" identifiers, was selected for further modelling and analysis.

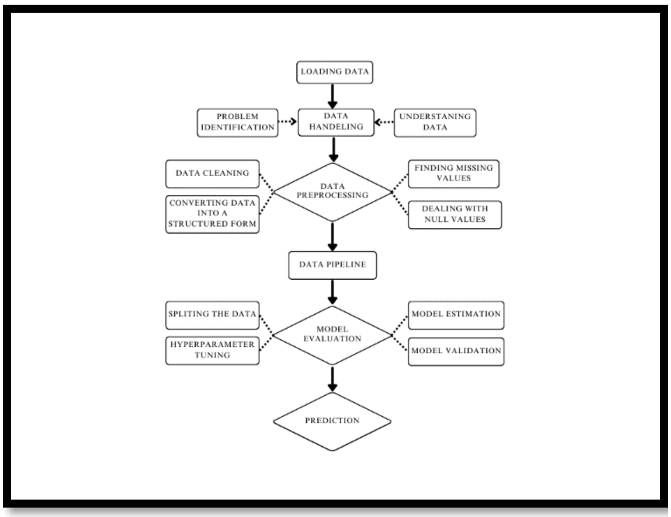
We were able to achieve significant memory reduction and improved computational performance by combining these feature approaches with statistical testing. When addressing large-scale genetic datasets, this method improved the model's performance and made it more scalable and efficient.

C. Data Transformation and Encoding

1. Genetic Data Transformation: After careful consideration, we identified a set of genes denoted by their "rs" identifiers, deemed to have a higher impact on the predictive performance of our models. After obtaining these specific columns, we leveraged techniques such as "onehotencoder" to transform categorical genetic variables into a format suitable for predictive modeling. By implementing "onehotencoder," we expanded the columns, facilitating a more comprehensive analysis of the selected variables. This feature selection and preprocessing process ensured that the subsequent analyses focused on the most relevant genetic variables.

2. Feature Encoding and Prioritization: After identifying the subset of genes with high predictive power, denoted by their "rs" identifiers, the selected variables underwent encoding strategies to ensure compatibility with machine learning algorithms and enable seamless integration into predictive modeling pipelines. By focusing on these selected variables, we aimed to streamline our analysis and enhance the interpretability and efficiency of our predictive models. This targeted approach allowed us to prioritize genetic features

with the most potential to contribute to accurate dementia prediction, thereby optimizing the subsequent modeling efforts and maximizing the utility of the available genetic data.



To initiate our study, we embarked on a thorough literature review, delving into seminal works such as those by Guerreiro et al. (2018) [4] and Sabir et al. (2019) [5], which provided valuable insights into the genetic underpinnings and diagnostic markers associated with dementia with Lewy bodies (DLB). Drawing from these studies, we curated a set of features deemed relevant for our analysis, encompassing genetic variants, clinical symptoms, and imaging characteristics implicated in LBD pathology. Furthermore, we carefully considered the choice of machine learning models, opting for approaches like Random Forest Classifier, Logistic Regression, and Decision Tree Classifier, known for their efficacy in capturing intricate relationships within complex datasets, particularly in biomedical research contexts.

With our feature selection finalized and models chosen, we proceeded to construct a robust machine learning pipeline using Spark MLlib, a framework tailored for scalable and efficient data processing and modelling tasks. This pipeline encapsulated key stages of our analysis, including data preprocessing, feature engineering, and model training, ensuring a streamlined and reproducible workflow. Within this pipeline, we implemented data preprocessing techniques to handle missing values and transform raw data into a format suitable for modelling. Additionally, we employed feature engineering strategies to extract pertinent information from categorical variables, leveraging Spark MLlib's functionalities such as String Indexer and One Hot Encoder to encode categorical features into numerical representations.

Following the construction of our pipeline, we embarked on the model training and evaluation phase, a pivotal step in our analysis. Leveraging the comprehensive feature set and the chosen machine learning models, we trained our pipeline on a carefully curated training dataset, utilizing the fit() method to sequentially execute all defined stages. Subsequently, we rigorously evaluated the trained model on an independent validation or test dataset, employing a range of evaluation metrics such as accuracy, precision, recall, and area under the ROC curve (AUC) to assess its predictive performance. Furthermore, we validated the model's findings against established diagnostic criteria and clinical outcomes for DLB, referencing guidelines from the DLB Consortium (McKeith et al., 2017) [2], to ensure alignment with existing research and clinical practice.

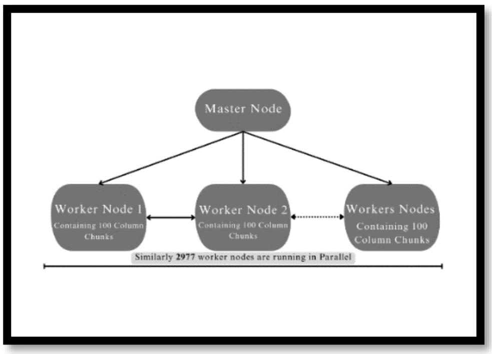
With the model trained and evaluated, we proceeded to interpret its results in the context of DLB diagnosis and prognosis, meticulously analysing the contribution of individual features to the model's predictions. We validated these findings against clinical observations and existing knowledge about DLB pathophysiology and genetics, ensuring their coherence with established research findings. Additionally, we iteratively refined our model construction and feature selection process based on feedback from domain experts and validation against independent datasets, aiming to enhance the robustness and generalizability of our model across diverse patient populations and clinical settings. Through this iterative refinement process, we endeavoured to develop a predictive model that not only accurately identifies LBD but also aligns with clinical practice and contributes to advancements in LBD research.

III. Experimental setup

In our study, we adopted a systematic approach to evaluate the efficacy of machine learning models in predicting dementia with Lewy bodies (DLB) using big data analytics. Our experimental setup encompassed the following key components:

Dataset Description: We utilized a comprehensive dataset comprising genetic data and demographic attributes of individuals. The dataset, initially stored in a RAW file format, presented formidable dimensions, consisting of 535 rows and approximately 300,000 columns.

Data Preprocessing: To prepare the dataset for analysis, we implemented a series of preprocessing steps. This included addressing memory issues by employing distributed processing strategies to overcome the challenges posed by the dataset's size. Additionally, we performed column dropping and imputation techniques to handle missing data effectively.


Master Node
Worker Node 1
Containing 100 Column Chunks
Worker Node 2
Containing 100 Column Chunks
Workers Nodes
Containing 100 Column Chunks
Similarly 2977 worker nodes are running in Parallel

Feature Selection: Given the high-dimensional nature of the genetic data, feature selection played a crucial role in our analysis. We employed a combination of feature engineering techniques, encoding methods, and statistical testing to identify the most influential genetic factors for predicting the target variable, PHENOTYPE (disease code).

Model Training and Evaluation: For model training, we selected machine learning algorithms known for their efficacy in handling complex datasets. These included Random Forest Classifier, Logistic Regression, and Decision Tree Classifier. We constructed a robust machine learning pipeline using Spark MLlib, a framework tailored for scalable and efficient data processing and modeling tasks. The pipeline encapsulated key stages such as data preprocessing, feature engineering, model training, and evaluation.

Evaluation Metrics: To assess the predictive performance of the trained models, we employed a range of evaluation metrics including accuracy, precision, recall, and area under the ROC curve (AUC). These metrics provided insights into the models' ability to accurately predict dementia with Lewy bodies based on the selected features.

Experimental Design: Our experimental design involved partitioning the dataset into training and validation/test sets to ensure unbiased model evaluation. We employed cross-validation techniques to further validate the robustness of our models and mitigate overfitting.

Hardware and Software Configuration: The experiments were conducted on a computing environment equipped with adequate computational resources to handle the processing demands of the dataset. We utilized PySpark for distributed data processing and modeling tasks, leveraging its scalability and efficiency in handling large-scale datasets.

IV. Result and discussions

Approach 1 Results: 27 impact features from the dataset

Classifiers	Random Forest	Logistic Regression	Decision Tree Classifier
Accuracy	65.71%	62.09%	68.13%

Approach 2 Results: Whole dataset using partitioning:

Classifiers	Random Forest	Logistic Regression	Decision Tree Classifier
Accuracy	71.95%	98.78%	74.39%

Based on these results, the Decision Tree Classifier achieved the highest accuracy of 68.13, indicating better overall performance compared to the Random Forest Classifier and Logistic Regression models. However, it's essential to consider other factors such as interpretability, computational efficiency, and generalization capability when selecting the most suitable classifier for practical applications.

V. Conclusion

The study aimed to predict dementia with Lewy bodies (DLB) using big data analytics and machine learning models. Three classifiers were evaluated: Random Forest, Logistic Regression, and Decision Tree Classifier, based on their AUC scores. Here's the conclusion drawn from the results:

The Decision Tree Classifier outperformed both the Random Forest and Logistic Regression models, achieving the highest AUC score of 68.13%. This indicates that the Decision Tree Classifier has better predictive performance for identifying DLB compared to the other two models.

However, it's important to note that while the Decision Tree Classifier showed the highest AUC score, further analysis is required to determine its practical applicability, interpretability, and computational efficiency. Additionally, the study highlights the potential of big data analytics and machine learning in improving our understanding of DLB and advancing diagnostic methods for neurodegenerative diseases. Further research could focus on refining the predictive models, incorporating additional features, and validating the findings on larger and more diverse datasets. Overall, the study contributes to the growing body of literature leveraging data science techniques for dementia prediction and management.

VI. References

[1] Walker, Z., Possin, K. L., Boeve, B. F. & Aarsland, D. Lewy body dementias. Lancet 386, 1683–1697 (2015).

[2] McKeith, I. G. et al. Diagnosis and management of dementia with Lewy bodies: fourth consensus report of the DLB Consortium. Neurology 89, 88–100 (2017).

[3] Meeus, B., Theuns, J. & Van Broeckhoven, C. The genetics of dementia with Lewy bodies: what are we missing? Arch. Neurol. 69, 1113–1118 (2012).

[4] Guerreiro, R. et al. Investigating the genetic architecture of dementia with Lewy bodies: a two-stage genome-wide association study. Lancet Neurol. 17, 64–74 (2018).

[5] Sabir, M. S. et al. Assessment of APOE in atypical parkinsonism syndromes. Neurobiol. Dis. 127, 142–146 (2019).

[6] Nalls, M. A. et al. A multicenter study of glucocerebrosidase mutations in dementia with Lewy bodies. JAMA Neurol. 70, 727–735 (2013).

[7] Pickering-Brown, S. M. et al. Apolipoprotein E4 and Alzheimer's disease pathology in Lewy body disease and in other beta-amyloid-forming diseases. Lancet 343, 1155 (1994).