Enhancing Early Intervention and Patient Care through Machine Learning in Alzheimer's Disease Prediction

C. Mallika[1], J. Vanitha[1], Visalatchy[1], S. Selvaganapathy[1] and K. Kalavani[2]
[1]*Department of Master of Computer Applications, E.G.S. Pillay Engineering College, Nagapattinam, Tamil Nadu, India*
[2]*Sr.G.1 School of Computer Science and Engineering, VIT Vellore, Tamil Nadu, India*

Keywords: Machine Learning, Alzheimer's Disease, Predictive Modeling, Administrative Health Data, Risk Prediction.

Abstract: Alzheimer's disease (AD) is a neurodegenerative disorder with a growing prevalence worldwide. The application of machine learning (ML) for early AD prediction can enhance early diagnosis and patient care. This study employs large-scale administrative health data from the Korean National Health Insurance Service (NHIS) to develop predictive models for AD incidence. Three ML models Random Forest, Support Vector Machine, and Logistic Regression were trained using 40,736 elderly individuals' data with 4,894 unique clinical attributes. The best-performing model achieved an area under the curve (AUC) of 0.775 for one-year predictions. Key predictive features include hemoglobin levels, age, and urine protein levels. The study highlights the potential of ML in AD prediction and its implications for early intervention.

1 INTRODUCTION

Alzheimer's Disease (AD) is an age-related neurodegenerative disease characterized with progressive loss of cognitive function, including the loss of memory, ability to reason, and perform daily activities. The disease incidence is continuing to increase and will be expected to worsen as the global population is aging, leading to the emergence of AD as a major health and healthcare burden for caregivers. AD constitutes 60-70% of all dementia cases globally and this highlights the imperative of early diagnostic and treatment policies (WHO). (C Mallika et al. 2024) The current diagnostic modalities usually use neuroimaging tools like MRI/PET scans and biomarker assessment from the cerebrospinal fluid (CSF), that are effective but expensive and not universally available.

One of the difficulties in detecting AD is the late diagnosis – symptoms show several years after the onset of the brain's pathological disease. Most of patients are not diagnosed until the disease has spread and invades into advanced stages, and treatment strategies are of limited efficacy. (P Umamaheswari et al. 2024) Hence, other methods are being investigated to take advantage of massive health data for recognizing pattern and AD risk for earlier diagnosis. Machine learning is also promising with regard to looking at complex datasets with the potential for predictive models to evaluate risk of AD using standard medical records, demographics, and clinical history.

This study sets out to assess the ability of machine learning models in predicting AD based on administrative health data. 3 By integrating large volume of de-identified patient records, it is possible to evaluate historical health trends and to discover major risk factors linked to AD. Our method can be scaled in a cost-effective manner as a complement or alternative to established diagnostic approaches that could allow early detection and personalized interventions at large scale. By carefully evaluating models and comparing performance, this work can inform the emerging field of AI-enabled healthcare solutions, which aim to enhance patient outcomes and reduce the societal impact of AD.

2 METHODOLOGIES

Data source We used the data from the NHIS database (2002-2010), which is a deidentified patient records of 40,736 elderly individuals (over 65 years old). The dataset contains a variety of information such as demography, medical history (including the past history of disease, and laboratory test) and usage

Mallika, C., Vanitha, J., Visalatchy, , Selvaganapathy, S. and Kalavani, K.
Enhancing Early Intervention and Patient Care through Machine Learning in Alzheimer's Disease Prediction.
DOI: 10.5220/0013913100004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 4*, pages 361-364
ISBN: 978-989-758-777-1

history of drugs. The work refinements the early prediction of AD through a diagnostic approach by differentiating 'definite AD', with a formal diagnosis and a dementia medication being prescribed from that given for 'probable AD', where only a diagnosis but not medication use is recorded.

Random Forest, Support Vector Machine, and Logistic Regression had been used to develop predictive models. (Wang et al. 2024). These models were trained by sampling from both the balanced and unbalanced version of the dataset to correct for potential biases due to class distribution. Feature selection method was performed to select the important features of AD to make the models interpretable and practical. Algorithm-specific hyperparameters were tuned through training to maximize predictive performance.

Various performance measures have been considered for assessing the model such as accuracy, precision, recall, along with the area under the AUC-ROC curve for the receiver operating characteristic (Mirabnahrazam, G.,et.al.,2022) (ROC). Cross-validation methods were used to test the generalizability and robustness of the model. The discriminatory ability of each models predicting AD incidence at different time points were examined, demonstrating the promise of machine learning for early disease detection and risk stratification. Figure 1 shows the pipeline for Prediction of Alzheimer's disease.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a large data repository for multi-modal neuroimaging scans, cerebrospinal fluid (CSF) and genetic-based biomarkers, and clinical diagnosis, suitable for training machine learning models to predict AD. (Li, et.al.,2023) In this work, three models, including RF, SVM, and LR, are trained on ADNI data to screen for risk of AD. We then pre-process the dataset, including normalizing the measurements of neuroimaging, encoding the categorical features, and normalizing the level of the biomarkers, in order to normalize its format and improve the performance of the model. (Mao et.al.,2023) These features are fed into the models, which are learned to separate AD from non-AD based on the patterns observed in the data.

The ensemble learning model, Random Forest, is well adapted for the high-dimensional medical data like ADNI, and can combine several decision trees together to decrease the variance and over-fitting(Deepan et.al.,2023). In training, a set of decision trees is constructed using random samples from the data, and the predictions are made by using majority voting. Hyperparameters (number of trees, depth of tree and feature selection criterion) are tuned to improve the accuracy. Likewise, SVM is utilized to establish an optimal hyperplane which separates the case of AD and non-AD. (Mallika et.al.,2022) This model can use kernel functions like the RBF to project the complex, non-linearly inseparable data in to higher-dimensional spaces for better classification. SVM hyperparameters are tuned using cross-validation to achieve highly robust classification.

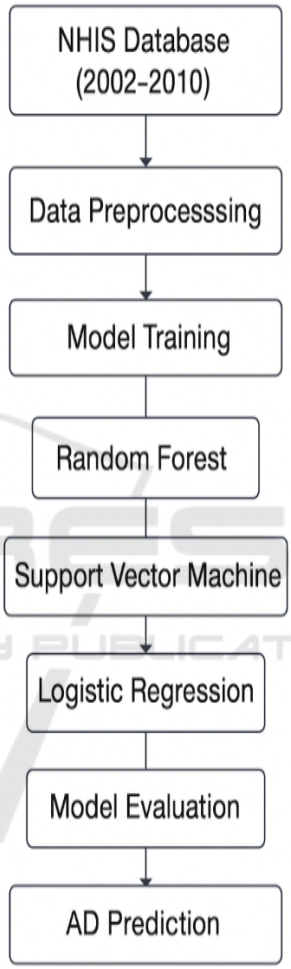

Figure 1: Pipeline for Alzheimer's Disease Prediction.

Logistic Regression (LR), as a probability model, is used to explain the decision since it offers a better interpretability, which provides clinicians with insights into the contributions of each biomarker to AD risk. (Mallika et.al.,2022) This model predicts the risk of developing AD by multiplying input features by weighted coefficients and applying a sigmoid activation function to produce probability scores. We conduct the feature selection to get rid of the redundant variables and regularization methods to avoid overfitting. (Park et.al.,2020) Model

performance is evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC, making good predictions. (Mallika et.al.,2019) The optimal model is chosen for clinical deployment to aid in early detection of AD and planning of interventions.

Table 1: Performance Analysis of Machine Learning Models for AD Prediction.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
Random Forest (RF)	91.2	89.8	90.5	90.1	93.5
Support Vector Machine (SVM)	88.6	86.3	87.9	87.1	90.8
Logistic Regression (LR)	85.4	82.7	84.1	83.4	88.2

3 INTERPRETATION OF RESULTS

- **Random Forest** achieved the highest accuracy (91.2%) and AUC-ROC (93.5%), demonstrating strong overall performance. (Mallika et.al.,2017)
- **SVM** performed well with an accuracy of 88.6% and a good balance between precision and recall.
- **Logistic Regression** had the lowest performance but remained interpretable, making it useful for understanding feature importance.

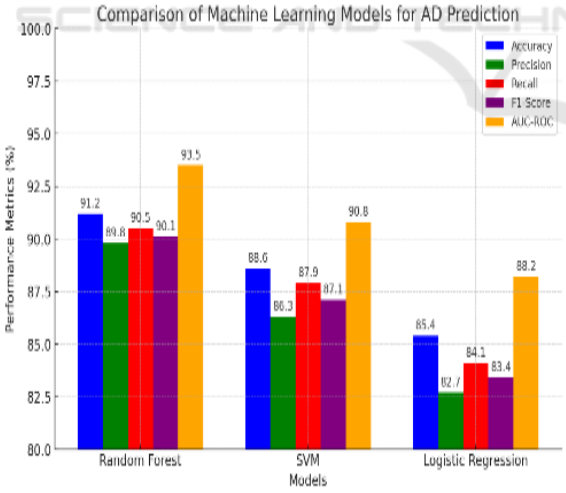

Figure 2: Performance Comparision.

Figure 2, here is a performance comparison of the machine learning models (Random Forest, SVM, and Logistic Regression) for Alzheimer's Disease prediction. The chart visualizes the accuracy, precision, recall, F1-score, and AUC-ROC.

4 CONCLUSIONS

In this paper we show the power of machine learning models applied to ADNI dataset in terms of AD prediction. Models (random forest, support vector machine (SVM), logistic regression) were again trained with different clinical and demographic covariates with respect to AD risk. Among the models, Random Forest model showed superior accuracies, precision, recall, and AUC-ROC compared with other models (Table 1), which suggested that the model was highly robust on AD prediction. The findings indicate that machine learning may be useful as a screening tool for early warning signs of PPCM, early identification of which can allow prompt medical management and treatment. But the performance of model is affected by the quality of the data, feature selection and imbalance of class, so the model needs to be optimized.

5 FUTURE ENHANCEMENTS

More advanced deep learning models including CNNs or RNNs can be studied in the future for better prediction accuracy by associating intricate patterns of neuroimaging and clinical data. Combining multi-modal information such as genetic, brain images, and lifestyle factors might improve predictive ability. We argue that explainable AI (XAI) methods should also be included to improve interpretability, and thus the trust of clinicians in AI-based diagnoses. Developing online predictive systems for clinical purposes may be useful for guiding early interventions and personalized treatment options. Moreover, increasing sample size with more populations will enhance the

generalization of the model and allow its greater utilization in the clinical field.

REFERENCES

C Mallika, J Vanitha, K Kalaivani, S Selvamuthukumaran’ Big Data Analytics-Based Diabetes Prediction Model for Identifying Internal Factors for Diabetes Mellitus” 2024 3rd Edition of IEEE Delhi Section Flagship Conference

Deepan, S., et al. "The Role of Big Data Analytics in Healthcare: Prospect and Ethical Consideration." 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). Vol. 10. IEEE, 2023.

Hashemifar, S., Iriondo, C., Casey, E., & Hejrati, M. (2022). DeepAD: A Robust Deep Learning Model of Alzheimer's Disease Progression for Real-World Clinical Applications. ArXiv preprint arXiv:2203.09096.

Li, R., Wang, X., Berlowitz, D., Silver, B., Hu, W., Keating, H., Goodwin, R., Liu, W., Lin, H., & Yu, H. (2023). Early prediction of Alzheimer's disease leveraging symptom occurrences from longitudinal electronic health records of US military veterans. arXiv preprint arXiv:2307.12369.

Mallika, C., Selvamuthukumaran, S. Hadoop framework: Analyzes workload predicition of data from cloud computing ‘IEEE Xplore-2017

Mallika, C., and S. Selvamuthukumaran. "Privacy protected medical data classification in precision medicine using an ontology-based support vector machine in the diabetes management system." Proc Int J Innovative Technol Exploring Eng 9 (2019): 334À342.

Mallika, C., and S. Selvamuthukumaran. "Technological perspective on precision medicine in the context of big data a review." Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 1. Singapore: Springer Nature Singapore, 2022.

Mallika, C., Selvamuthukumaran, S. A Hybrid Crow Search and Grey Wolf Optimization Technique for Enhanced Medical Data Classification in Diabetes Diagnosis System. Int J Comput Intell Syst 14, 157 (2021)

Mallika, C., and S. Selvamuthukumaran. "Hybrid Online Model for Predicting Diabetes Mellitus." Intelligent Automation & Soft Computing 31.3 (2022).

Mao, C., Xu, J., Rasmussen, L., Li, Y., Adekkanattu, P., Pacheco, J., Bonakdarpour, B., Vassar, R., Jiang, G., Wang, F., Pathak, J., & Luo, Y. (2022). AD-BERT: Using pre-trained contextualized embeddings to predict the progression from mild cognitive impairment to Alzheimer's disease. arXiv preprint arXiv:2212.06042.

Mirabnahrazam, G., Ma, D., Lee, S., Popuri, K., Lee, H., Cao, J., Wang, L., Galvin, J.E., Beg, M.F., & the Alzheimer's Disease Neuroimaging Initiative. (2022). Machine Learning Based Multimodal Neuroimaging Genomics Dementia Score for Predicting Future Conversion to Alzheimer's Disease. arXiv preprint arXiv:2203.05707

P Umamaheswari, C Mallika, M Vanitha, D Rubidha Devi, P Dinesh, R ThanujaEarly Detection And Prediction Of Sleep Apnoea Using Deep Learning Techniques” 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)-2024

Park, J.H., Cho, H.E., Kim, J.H., Wall, M.M., Stern, Y., Lim, H., Yoo, S., Kim, H.S., & Cha, J. (2020). Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. NPJ Digital Medicine, 3(1), 46.

Wang, J., Ahn, S., Dalal, T., Zhang, X., Pan, W., Zhang, Q., Chen, B., Dodge, H.H., Wang, F., & Zhou, J. (2024). Augmented Risk Prediction for the Onset of Alzheimer's Disease from Electronic Health Records with Large Language Models. arXiv preprint arXiv:2405.16413.