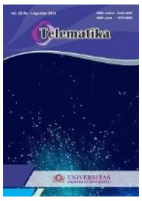

Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA “2” Kemenristek/BRIN, No. 85/M/KPT/2020

UNIVERSITAS AMIKOM PURWOKERTO

Improving Alzheimer's Disease Prediction Accuracy using Feature Selection, K Fold Cross Validation, and KNN Imputer Techniques

Kirso[1], Mila Desi Anasanti[2]

[1]Computer Science Master’s Study, Faculty of Information Technology
[2]Bart and London Genome Center
[2]Department of information Studies
[1]Nusa Mandiri University, Jakarta, Indonesia
[2]Queen Mary University of London, London, United Kingdom
[2]University College London, London, United Kingdom

ARTICLE INFO

History of the article:
Received October 14, 2024
Revised February 17, 2025
Accepted February 28, 2025

Keywords:
Alzheimer's Disease Prediction
KNN Imputer
Feature Selection Techniques
K-Fold Cross Validation
Machine Learning Algorithms

Correspondence:
E-mail:
mila.mld@nusamandiri.ac.id

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline and memory loss; it accounts for 60–70% of dementia cases. Early diagnosis remains challenging due to the subtlety of its symptoms. This study explores the effectiveness of ensemble methods, feature selection techniques, and imputation strategies in enhancing the accuracy of AD diagnosis. We applied an ensemble method with Chi-Square feature selection, achieving a high accuracy of 95.733% with 7 optimal features. The combination of classifiers, including Gradient Boosting (GB), Support Vector Machine (SVM), and Logistic Regression (LR), contributed to the high performance. Additionally, the use of KNN Imputer and K-Fold Cross Validation significantly improved accuracy, regardless of whether feature selection was employed. Notably, feature selection slightly reduced model complexity but resulted in a marginal decrease in accuracy. The study highlights the importance of these methods in achieving reliable AD predictions, though dataset dependency and potential biases from methodological choices are acknowledged. Future work may involve exploring alternative classifiers and validating findings across diverse datasets to enhance generalizability and address these limitations.

INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that leads to cognitive dysfunction, memory loss, and the accumulation of amyloid plaques and neurofibrillary tangles in the brain (Yıldız, et al., 2021) (Hughes, et al., 2020). It is one of the leading causes of dementia, affecting approximately 40 million people globally, and poses significant challenges to healthcare systems, particularly in low- and middle-income countries, which contribute to 60% of the total dementia cases (Sara, et al., 2022). In the United States alone, around 5.1 million people suffer from AD (Peavy, et al., 2020) (Kavitha, et al., 2022), with many not receiving adequate medical care. According to the World Health Organization (WHO), more than 55 million people globally will suffer from dementia by 2023, with low- and middle-income countries contributing to 60% of dementia cases. Alzheimer's disease accounts for 60-70% of dementia cases (Ebrahimi, et al., 2020) (Organization, 2023), and patients require continuous care as the disease progresses. AD manifests as memory loss, cognitive decline, and behavioral disturbances

http://dx.doi.org/10.35671/telematika.v18i1.3055

(Zhang, et al., 2023). In its later stages, AD can lead to life-threatening complications such as infections, malnutrition, and dehydration (Needham, 2022).

Advancements in research and the utilization of smart sensors and machine learning (ML) algorithms have introduced new possibilities for the diagnosis, monitoring, and prediction of AD (Gillani & Arslan, 2021) (Shiino, et al., 2021). ML models, particularly those trained on brain imaging data, have demonstrated promise in diagnosing AD and predicting its progression (Shiino, et al., 2021). Recent studies highlight the significance of ML in detecting Alzheimer's disease at an early stage, a critical factor for improving patient outcomes (Uddin, et al., 2023). For instance, Kavitha et al. (2022) utilized data from the Open Access Series of Imaging Studies (OASIS) to develop accurate ML models for predicting early-stage AD (Kavitha, et al., 2022).

Machine learning, particularly deep learning techniques such as Convolutional Neural Networks (CNNs), has shown significant potential in predicting AD by analyzing neuroimaging data (Beltrán, et al., 2020) (Marzban, et al., 2020). These models leverage features extracted from MRI scans and biomarkers, providing a data-driven approach to AD diagnosis. In one study, Random Forest classifiers achieved an accuracy of 86.92% in predicting AD from the OASIS dataset (Kavitha, et al., 2022). In contrast, deep learning models like BiLSTM have demonstrated even higher accuracy rates, reaching 95.59% (Dashtipour, et al., 2021).

While these studies have made important strides, several challenges remain, including the selection of appropriate classifiers, handling missing data, and determining optimal feature selection methods. Some researchers have employed imputation techniques, such as median imputation, to address missing data (Kavitha, et al., 2022), while others have used more advanced methods like KNN imputation. Additionally, feature selection methods like Chi-Square, Information Gain, and F-Score are commonly applied to improve model performance by reducing dimensionality and enhancing accuracy (Basheer, Bhatia, & Sakri, 2021).

This study aims to build on these previous findings by exploring the use of KNN imputation for handling missing data and investigating a variety of feature selection techniques, including Spearman correlation, Chi-Square, Information Gain, Pearson correlation, and F-Score. The study will also evaluate the performance of different classifiers, with a particular focus on ensemble methods. To better understand how various machine learning models have performed in Alzheimer's disease prediction, **Table 1** provides a comparative summary of several classifier models and their results from previous studies. By comparing the accuracy of classifiers with and without feature selection, this research seeks to contribute to the growing body of knowledge on improving machine learning-based approaches to AD diagnosis.

Table 1. Comparison of Classifier Models for Alzheimer’s Disease Prediction

Researchers' Study	*Dataset*	*Models*	*Classification accuracy*
Kavitha et al (Kavitha, et al., 2022)	OASIS Dataset	Random Forest	86,92% Accuracy in Random Forest Classifier
Basher et al. (Basheer, Bhatia, & Sakri, 2021)	OASIS Imaging Dataset	M-CapNet	92,3% Accuracy with M-CapNet
Malavika et al	OASIS (longitudinal MRI	KNN, Adaboost, SVM, Logistic Regression, Decision Tree, Random Forest	86,8% Accuracy in Random Forest Classifier

Researchers' Study	*Dataset*	*Models*	*Classification accuracy*
(Malavika, Rajathi, Vanitha, & Parameswari, 2020)	data)		
Dashtipour, et al. (2021)	OASIS Dataset	SVM, BiLSTM	82.24% (SVM), 95.59% (BiLSTM)

METHODS

The research methodology for data analysis typically involves several key stages to ensure the accuracy and reliability of the results. These stages include data collection, data preprocessing, feature selection, data splitting, model training and evaluation where model validation uses k-fold cross-validation, and prediction models using LR, RF, KNN, NB, SVM, DT, AB, GB, XGB, ANN, and ensemble classifiers, as well as performance evaluation. The steps taken to obtain the results of the analysis are illustrated in Figure 1.

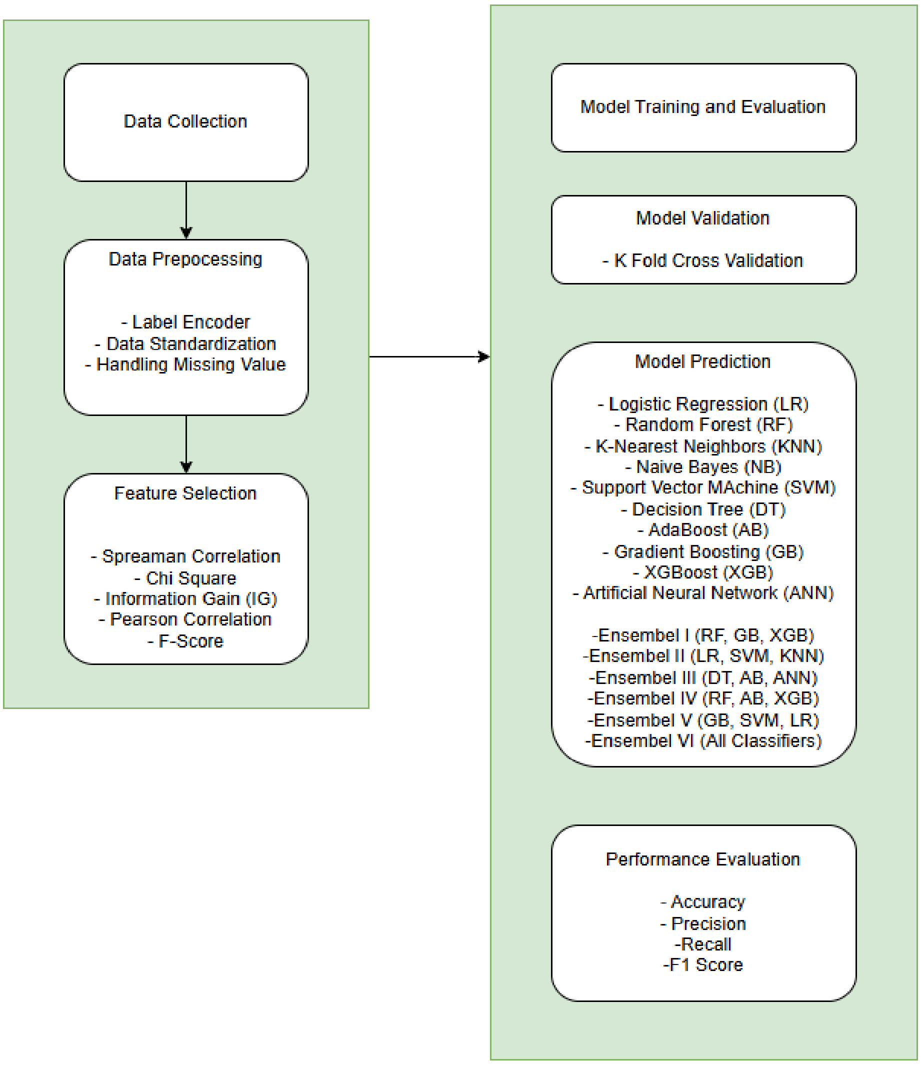

Figure 1. Research Method

1. Data Collection

Data collection is the initial step in gathering relevant data for analysis (Umar, et al., 2024). The dataset used is a longitudinal cross-section from the OASIS dataset, sourced from https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers. This dataset contains MRI data on conversions, serving as the primary data source. It includes data from 150 patients aged 60 to 96 years, with all subjects using their right hand. Among the 150 patients, 72 do not have dementia.

2. **Data preprocessing**

Data preprocessing is an essential step for preparing the data by cleaning, handling missing values, and standardizing data for analysis (Biswas & Rajan, 2021). The dataset contains 12 variables, listed in Table 1.

Table 1. OASIS Prepossessing Data

No	*Variable*	*Description*	*Data Type*
1	Group	Class	String
2	M/F	Gender	String
3	Age	Age	Integer
4	Educ	Years of education	Integer
5	SES	Socio Economic Status	Integer
6	MMSE	Mini-Mental State Examination	Integer
7	CDR	Clinical Dementia Rating	Integer
8	eTIV	Estimated Total Intracranial Volume	Integer
9	nWBV	Normalize Whole Brain Volume	Integer
10	ASF	Atlas Scaling Factor	Integer

In this study, label encoding was applied to the **Group** and **M/F** variables. The "Demented" and "Converted" values were replaced with 1, and the "Non-Demented" value was replaced with 0. Missing values in the **SES** and **MMSE** variables were handled using the KNN imputer.

3. **Feature selection**

Feature selection is an important step to identify the most relevant features for the model, which helps reduce dimensionality and improve model efficiency (Celard, et al., 2020). Feature selection methods used in this study include Spearman Correlation, Chi-Square, Information Gain, Pearson Correlation, and F-Score. These methods help streamline the analysis by focusing on the most impactful features, enhancing the efficiency of machine learning models.

4. **Model Training and Evaluation**

Model training involves building a predictive model using the training data, followed by evaluation of its performance on test data (Mnguni, 2021). Validation is conducted to assess the generalizability of the model, ensuring robustness (Istiqoh, Qodir, & Ahmad, 2022). Model prediction involves utilizing the trained model to make predictions on new data points (Zhang, et al., 2023). In this stage, machine learning classifiers are used for model training and evaluation. The models used are LR, RF, KNN, NB, SVM, DT, AB, GB, XGB, and ensemble classifiers.

5. **Model Validation**

K-fold cross-validation is a commonly used technique for model evaluation. It involves dividing the dataset into K subsets (folds), using each fold as a test set while the rest of the data serves as the training set (Paramita, 2022) (Chen, et al., 2023). This process is repeated K times, each time using a different fold for testing. This technique helps assess model performance, reducing the risk of overfitting and providing a better estimate of the model's generalizability (Ayinla & Oremei, 2024) (AlZu’b, Zraiqat, & Hendawi, 2022).

6. Model Prediction

Once validated, the trained models are used to make predictions on new data points (Oh, Tannenbaum, & Deasy, 2022). The prediction models used in this study include:

a. Logistic Regression (RF)

Logistic regression is a statistical model that is commonly used in binary classification tasks (Kost, Rheinbach, & Schaeben, 2019).

b. Decision Tree (DT)

Decision tree classifier is a versatile classifier that can handle both numerical and categorical data, offering clear interpretability(Abana, 2019). DT has also been utilized in explainable ML, explicitly showing how different features contribute to predictions (Cao, Sarlin, & Jung, 2020).

c. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a straightforward and effective classification technique widely used across various domains. However, it can be computationally expensive and slow when working with large datasets (Naveed, Madhloom, & Husain, 2021). KNN is capable of handling multiclass data without assuming any specific data distribution, making it flexible for evolving datasets and suitable for unstructured data. Despite its strengths, KNN suffers from slow prediction times on large datasets, is highly sensitive to data scale, and may perform poorly with high-dimensional data due to the increasing complexity in calculating distances.

d. Random Forest (RF)

Random Forest is a highly versatile ensemble learning method that has found extensive applications across various fields due to its robustness and high accuracy. According to (Barbara Pes, 2021), RF excels as a classifier, particularly when learning from high-dimensional or class-imbalanced datasets, demonstrating significant success in dealing with complex data structures and imbalanced classification tasks.

e. AdaBoost (AB)

AdaBoost, or Adaptive Boosting, is an ensemble technique that combines several weak classifiers to create a strong and accurate overall classifier. It is widely used across numerous domains due to its effectiveness and adaptability in handling a variety of data types and classification challenges. AdaBoost is valued for its ability to improve the performance of weak learners, especially in complex problems.

f. Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting (XGB) is an advanced gradient boosting algorithm that has become a popular choice in machine learning due to its efficiency and effectiveness. XGB builds a robust model by combining multiple weak classifiers, typically decision trees, to form a strong ensemble capable of handling both classification and regression tasks. The algorithm enhances model performance by applying a second-order Taylor expansion to the cost function, leveraging both first- and second-order derivatives for better optimization (Xu, Wu, & Chen, 2022).

g. Naive Bayes (NB)

Naive Bayes is a popular classification algorithm that uses Bayesian principles and assumes feature independence to make predictions. In a study by Winarti et al. (Winarti, et al., 2021), NB was compared with KNN for classifying Indonesian language articles, showcasing its application in natural language processing. This supervised algorithm relies on training data and prior knowledge for making accurate predictions.

h. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised algorithm used for both classification and regression tasks. In a study by Zhang et al. (Zhang, Lin, & Wang, 2021), SVM was employed for forecasting e-commerce transaction trends by integrating an enhanced Whale Optimization Algorithm, showcasing the algorithm's versatility in predictive analytics.

i. Ensemble

Ensemble voting methods, such as hard voting and soft voting, are essential for combining the predictions of multiple classifiers to enhance overall accuracy and robustness in machine learning. Hard voting is suitable for models predicting distinct class labels, while soft voting is preferred when models provide probabilities for each class. The study elucidates the differences between hard and soft voting, highlighting their applications based on the nature of classifier outputs (Πεππές, et al., 2021). Ensemble strategies like hard voting, soft voting, and model stacking have been successfully utilized in various domains. Akhtar et al. (Akhtar, et al., 2021) showcased the effectiveness of soft and hard voting in an ensemble model for enhanced identification of thyroid disorders, emphasizing the importance of ensemble techniques in healthcare applications. In conclusion, ensemble voting methods, particularly soft and hard voting, provide a robust mechanism to harness the strengths of multiple classifiers, thereby boosting predictive performance across diverse fields.

7. Performance Evaluation

To evaluate machine learning models, performance metrics such as accuracy, precision, recall, and F1 score are commonly used. Studies have demonstrated their effectiveness in various applications. Xu et al. (Xu, et al., 2022) utilized the F1 score to assess models predicting clinic attendance and HIV/STI testing uptake, while Dong et al. (Dong, et al., 2022) used accuracy and F1 score for diabetic kidney disease prediction. Ljubobratović et al. (Ljubobratović, et al., 2022) highlighted AUC, accuracy, F1 score, and kappa for evaluating peach maturity prediction, emphasizing AUC as a key metric. These studies underscore the importance of robust evaluation metrics in assessing model effectiveness across different domains.

RESULTS

In this research, there were 150 patients aged between 60 and 96 years with the right-hand dominance, comprising 213 females, accounting for 57.1%, and 160 males, accounting for 42.9%, as depicted in Figure 2.

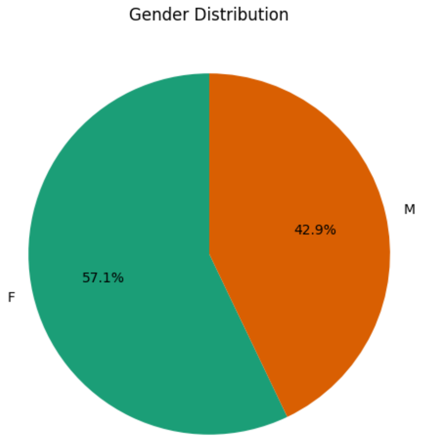

Figure.2. Gender Distribution

Additionally, in Figure 3 and Figure 4, there are categories for socio-economic status and educational level, where higher levels correspond to larger values.

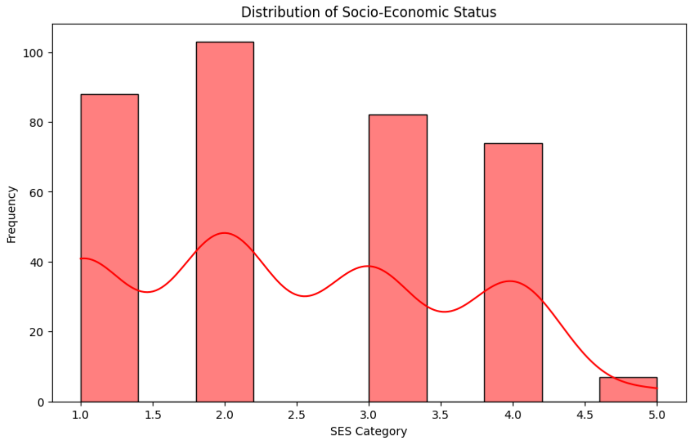

Figure 3. Distribution of Socio-Economic Status

Based on Figure 3, it is observed that the distribution of socio-economic status data is skewed to the left, indicating that the majority of samples have relatively low socio-economic status.

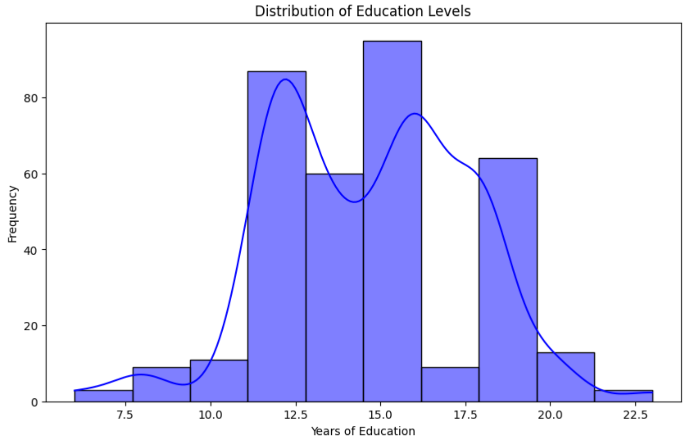

Figure 4. Distribution of Education Levels

Based on Figure 4, the result shows that the distribution of education levels is symmetrical but slightly skewed to the right. Therefore, the education level of the majority of research samples falls within the middle to high range.

The other features used are the Mini-Mental State Examination, Clinical Dementia Rating, Estimated Total Intracranial Volume, and Normalized Whole Brain Volume as shown in Figure 5, Figure 6, Figure 7, and Figure 8.

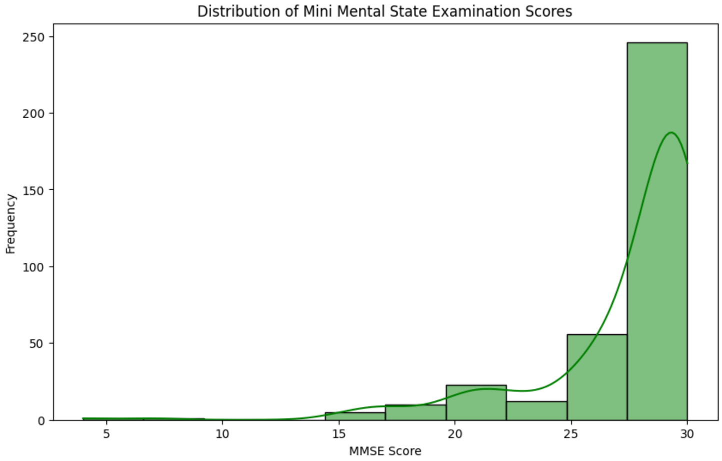

Figure 5. Distribution of MMSE

Based on Figure 5, the results indicate that the majority of research samples have MMSE scores ranging from 25 to 30, where the maximum score for MMSE is 30. This indicates that the research samples have good cognitive levels or cognitive functions, especially in terms of orientation, memory, calculation, abstract thinking, language, and visual-spatial abilities with a good level. However, a small portion of the samples has relatively low MMSE scores or good cognitive levels or cognitive functions, especially in terms of orientation, memory, calculation, abstract thinking, language, and visual-spatial abilities, with a less satisfactory level.

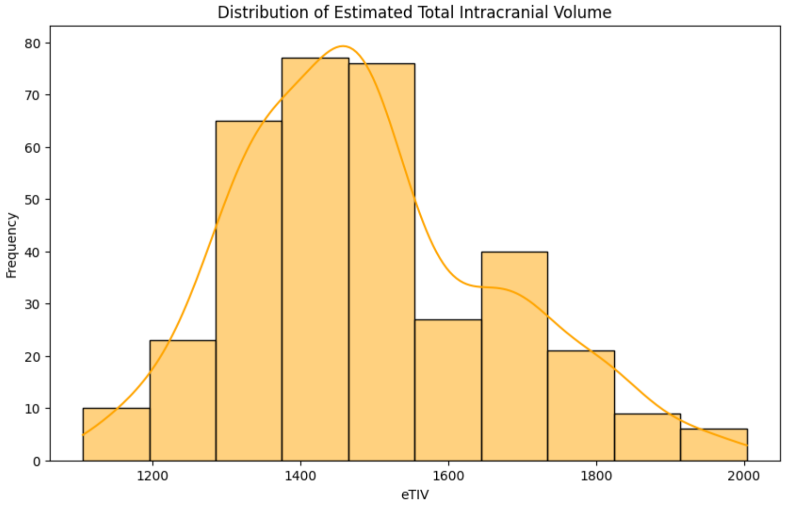

Figure 6. Distribution of eTIV

Based on Figure 6, the results show that the eTIV values for the majority of samples fall within the range of 1200 – 1600 cm^3. Referring to the normal eTIV values for adults, which range from 1200 – 2000 cm^3, it can be said that the majority of research samples have normal eTIV values.

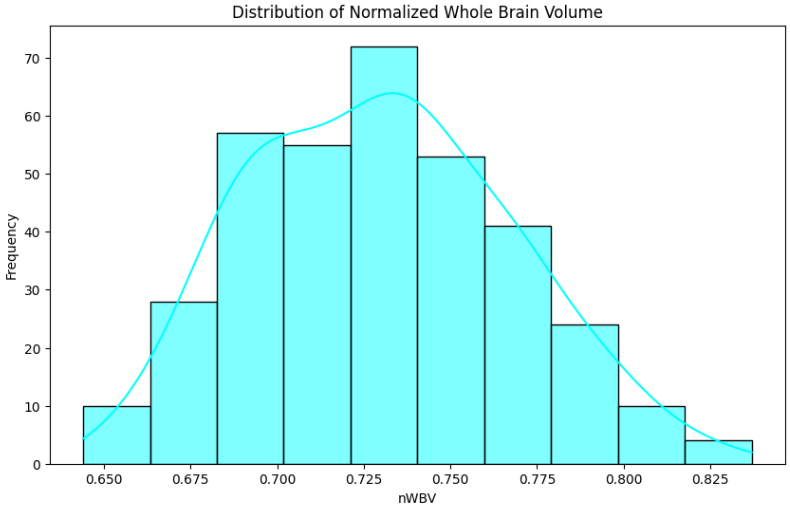

Figure 7. Distribution of nWBV

Based on Figure 7, all research samples have nWBV values below 100%. Some samples even have nWBV values ranging from 0.700 to 0.750 or 70% to 75% of their brain volume. If the Normalized Whole Brain Volume (nWBV) value is less than 100%, it indicates that the normalized brain volume is smaller than the estimated total intracranial volume. In this context, an nWBV value less than 100% may indicate a reduction in relative brain volume compared to an individual's intracranial size. This could be due to various factors, such as brain atrophy associated with aging or certain medical conditions that affect brain volume.

1. Spearman correlation-based feature selection compared to using no feature selection.

The application of Spearman correlation-based feature selection across various classifiers yielded results, as shown in Table 2.

Table 2. The accuracy values of classifiers with Spearman correlation feature selection versus without feature selection.

No	*Classifier*	*Kavitha Research*	*Features*					*Without Feature Selection*
			4	*5*	*6*	*7*	*8*	
1	LR	-	94,659	94,659	94,659	94,659	94,395	94,659
2	RF	86,920	92,226	93,848	94,388	95,192	95,462	95,733
3	KNN	-	94,118	94,388	92,525	91,451	91,977	92,809
4	NB	-	94,666	94,666	94,666	94,666	94,666	94,666
5	SVM	81,670	94,659	94,659	94,659	94,659	94,659	94,659
6	DT	80,460	90,356	90,085	91,693	92,781	92,511	91,415
7	AD	-	94,403	93,862	93,044	93,855	93,855	93,314
8	GB	-	92,496	92,767	93,030	95,192	95,192	95,192
9	XGB	85,920	91,970	93,300	93,037	95,192	95,192	95,192
10	ANN	-	94,659	94,659	94,659	94,659	93,848	94,388
11	Ensemble 1	-	92,226	93,307	94,104	95,469	95,469	95,733
12	Ensemble 2	-	94,659	94,659	94,659	94,659	94,659	94,659
13	Ensemble 3	-	90,896	90,619	92,240	93,585	92,767	92,504
14	Ensemble 4	-	92,504	93,841	94,644	95,462	95,192	94,659
15	Ensemble 5	-	94,659	94,659	94,659	94,659	94,659	95,733
16	Ensemble 6	85,120	94,659	94,659	94,659	94,659	94,659	94,659

*) 🟨 = Best Accuracy

Ensemble 1 = RF, GB, XGB
Ensemble 2 = LR, SVM, KNN
Ensemble 3 = DT, AD, ANN
Ensemble 4 = RF, AD, XGB
Ensemble 5 = GB, SVM, LR
Ensemble 6 = All classifiers

http://dx.doi.org/10.35671/telematika.v18i1.3055

In Table 2, the Naive Bayes classifier consistently achieved the highest accuracy of 94.666% when employing 4, 5, and 6 features. Using 7 and 8 features, Ensemble 1 maintained a consistent accuracy of 95.469%. However, without feature selection, RF, Ensemble 1, and Ensemble 5 obtained the highest accuracy of 95.733%, which is 0.264% higher than Ensemble 1 with seven features and 1.067% higher than Naive Bayes using only four features.

2. Chi Square-Based Feature Selection Compared To Using No Feature Selection

The application of Chi Square-based feature selection across various classifiers yielded results, as shown in Table 3.

Table 3. The accuracy values of classifiers with Chi-square feature selection versus without feature selection.

No	*Classifier*	*Kavitha Research*	*Features*					*Without Feature Selection*
			4	*5*	*6*	*7*	*8*	
1	LR	-	94,659	94,659	94,659	94,659	94,659	94,659
2	RF	86,920	94,381	94,659	95,192	94,929	94,922	95,733
3	KNN	-	94,659	94,125	91,992	91,451	92,006	92,809
4	NB	-	94,666	94,666	94,403	94,666	94,666	94,666
5	SVM	81,670	94,659	94,659	94,659	94,659	94,659	94,659
6	DT	80,460	92,240	90,370	91,977	92,781	93,037	91,415
7	AD	-	93,862	94,125	94,936	93,855	93,314	93,314
8	GB	-	93,065	94,395	95,206	95,192	94,922	95,192
9	XGB	85,920	93,578	93,578	95,462	95,192	95,192	95,192
10	ANN	-	94,659	94,659	94,659	94,659	94,388	94,388
11	Ensemble 1	-	95,192	94,388	95,199	95,469	95,192	**95,733**
12	Ensemble 2	-	94,659	94,659	94,659	94,659	94,659	94,659
13	Ensemble 3	-	92,767	91,444	93,585	93,585	93,300	92,504
14	Ensemble 4	-	94,659	94,659	94,659	94,659	94,659	94,659
15	Ensemble 5	-	94,644	93,848	95,199	95,733	95,733	95,733
16	Ensemble 6	85,120	94,659	94,659	94,659	94,659	94,659	94,659

*) 🟨 = Best Accuracy

Ensemble 1 = RF, GB, XGB
Ensemble 2 = LR, SVM, KNN
Ensemble 3 = DT, AD, ANN
Ensemble 4 = RF, AD, XGB
Ensemble 5 = GB, SVM, LR
Ensemble 6 = All classifiers

In Table 3, the utilization of 4 features by Ensemble 1 yielded the highest accuracy of 95.192%. Employing five features, the NB classifier achieved the highest accuracy of 94.666%. Using six features, the XGB classifier attained the highest accuracy of 95.462%. When employing 7 and 8 features, Ensemble 5 consistently achieved an accuracy of 95.733%. Conversely, without feature selection, the highest accuracy was achieved by RF, Ensemble 1, and Ensemble 5, all recording 95.733%, identical to Ensemble 5 with seven features. This accuracy was 0.271% higher than that of XGB with six features, 1.067% higher than NB with five features, and 0.541% higher than Ensemble 1 with only four features.

3. Information Gain-Based Feature Selection Compared To Using No Feature Selection

The application of Information gain-based feature selection across various classifiers yielded results as shown in Table 4.

Table 4. The accuracy values of classifiers with Information Gain feature selection versus those without feature selection.

No	*Classifier*	*Kavitha Research*	*Features*					*Without Feature Selection*
			4	*5*	*6*	*7*	*8*	
1	LR	-	94,659	94,659	94,659	94,659	94,395	94,659
2	RF	86,920	94,922	95,192	95,455	95,199	94,922	95,733
3	KNN	-	95,192	95,185	93,044	92,518	91,977	92,809

No	*Classifier*	*Kavitha Research*	*Features*					*Without Feature Selection*
			4	*5*	*6*	*7*	*8*	
4	NB	-	94,666	94,666	94,666	94,403	94,666	94,666
5	SVM	81,670	94,659	94,659	94,659	94,659	94,659	94,659
6	DT	80,460	90,619	91,152	91,700	92,511	92,240	91,415
7	AD	-	94,125	93,855	93,855	93,585	93,855	93,314
8	GB	-	95,192	95,192	94,118	94,659	95,192	95,192
9	XGB	85,920	95,192	95,192	95,185	94,922	95,192	95,192
10	ANN	-	94,659	94,659	94,659	94,659	94,118	94,388
11	Ensemble 1	-	95,462	95,462	95,185	95,199	95,733	**95,733**
12	Ensemble 2	-	94,659	94,659	94,659	94,659	94,659	94,659
13	Ensemble 3	-	92,226	91,686	92,760	93,307	93,044	92,504
14	Ensemble 4	-	94,929	94,929	94,659	94,659	94,659	94,659
15	Ensemble 5	-	95,462	95,462	94,922	95,199	95,199	95,733
16	Ensemble 6	85,120	94,659	94,659	94,659	94,659	94,659	94,659

*) 🟨 = Best Accuracy

Ensemble 1 = RF, GB, XGB
Ensemble 2 = LR, SVM, KNN
Ensemble 3 = DT, AD, ANN
Ensemble 4 = RF, AD, XGB
Ensemble 5 = GB, SVM, LR
Ensemble 6 = All classifiers

In Table 4, the use of 4 and 5 features by Ensemble 1 and 5 consistently resulted in the highest accuracy of 95.462%. Utilizing six features, the RF classifier achieved the highest accuracy of 95.455%. For seven features, RF, Ensemble 1, and 5 achieved the highest accuracy of 95.199%, while using eight features, Ensemble 1 consistently achieved an accuracy of 95.733%. Conversely, without feature selection, the highest accuracy was obtained by RF, Ensemble 1, and 5, all achieving 95.733%. This accuracy was identical to Ensemble 1 with eight features, 0.534% higher than RF, Ensemble 1, and 5 with seven features, 0.278% higher than RF with six features, and 0.271% higher than Ensemble 1 and 5 with only four features.

4. Pearson Correlation-Based Feature Selection Compared To Using No Feature Selection

The application of Pearson Correlation-based feature selection across various classifiers yielded results as shown in Table 5.

Table 5. The accuracy values of classifiers with Pearson Correlation feature selection versus those without feature selection.

No	*Classifier*	*Kavitha Research*	*Features*					*Without Feature Selection*
			4	*5*	*6*	*7*	*8*	
1	LR	-	94,659	94,659	94,659	94,659	94,395	94,659
2	RF	86,920	91,686	93,592	94,118	95,733	95,192	95,733
3	KNN	-	94,118	94,388	92,525	91,451	91,977	92,809
4	NB	-	94,666	94,666	94,666	94,666	94,666	94,666
5	SVM	81,670	94,659	94,659	94,659	94,659	94,659	94,659
6	DT	80,460	90,356	90,085	91,159	92,233	92,511	91,415
7	AD	-	94,403	93,862	93,044	93,855	93,855	93,314
8	GB	-	92,496	92,767	93,030	95,192	95,192	95,192
9	XGB	85,920	91,970	93,300	93,037	95,192	95,192	95,192
10	ANN	-	94,659	94,659	94,659	94,118	93,848	94,388
11	Ensemble 1	-	91,963	93,307	94,104	95,469	95,469	**95,733**
12	Ensemble 2	-	94,659	94,659	94,659	94,659	94,659	94,659
13	Ensemble 3	-	90,896	90,889	91,430	92,226	93,044	92,504
14	Ensemble 4	-	94,659	94,659	94,659	94,659	94,659	94,659
15	Ensemble 5	-	92,504	93,841	94,644	95,462	95,462	95,733
16	Ensemble 6	85,120	94,659	94,659	94,659	94,659	94,659	94,659

*) 🟨 = Best Accuracy

Ensemble 1 = RF, GB, XGB
Ensemble 2 = LR, SVM, KNN
Ensemble 3 = DT, AD, ANN
Ensemble 4 = RF, AD, XGB
Ensemble 5 = GB, SVM, LR
Ensemble 6 = All classifiers

In Table 5, using 4, 5, and 6 features, the NB classifier consistently achieved the highest accuracy of 94.666%. With seven features, the RF classifier obtained the highest accuracy of 95.733%. Utilizing eight features, Ensemble 1 achieved the highest accuracy of 95.469%. Conversely, without feature selection, the highest accuracy was achieved by RF, Ensemble 1, and Ensemble 5, all recording 95.733%, which equaled the accuracy of RF with seven features. This accuracy was 0.264% higher than Ensemble 1 with eight features and 1.067% higher than NB with only four features.

5. F Score-Based Feature Selection Compared To Using No Feature Selection

The application of F Score-based feature selection across various classifiers yielded results, as shown in Table 6.

Table 6. The accuracy values of classifiers with F Score feature selection versus without feature selection.

No	Classifier	Kavitha Research	Features					Without Feature Selection
			4	5	6	7	8	
1	LR	-	94,659	94,659	94,659	94,659	94,395	94,659
2	RF	86,920	91,423	94,118	94,915	95,469	95,462	95,733
3	KNN	-	94,118	94,388	92,525	91,451	91,977	92,809
4	NB	-	94,666	94,666	94,666	94,666	94,666	94,666
5	SVM	81,670	94,659	94,659	94,659	94,659	94,659	94,659
6	DT	80,460	90,626	90,356	91,693	92,781	91,166	91,415
7	AD	-	94,403	93,862	93,044	93,855	93,855	93,314
8	GB	-	92,496	92,767	93,030	94,922	95,192	95,192
9	XGB	85,920	91,970	93,300	93,037	95,192	95,192	95,192
10	ANN	-	94,659	94,659	94,659	94,659	93,848	94,388
11	Ensemble 1	-	92,226	93,848	94,104	95,469	95,469	**95,733**
12	Ensemble 2	-	94,659	94,659	94,659	94,659	94,659	94,659
13	Ensemble 3	-	90,626	90,619	91,970	92,504	93,030	92,504
14	Ensemble 4	-	94,659	94,659	94,659	94,659	94,659	94,659
15	Ensemble 5	-	92,233	93,300	94,644	95,462	95,199	95,733
16	Ensemble 6	85,120	94,659	94,659	94,659	94,659	94,659	94,659

*) 🟨 = Best Accuracy

Ensemble 1 = RF, GB, XGB
Ensemble 2 = LR, SVM, KNN
Ensemble 3 = DT, AD, ANN
Ensemble 4 = RF, AD, XGB
Ensemble 5 = GB, SVM, LR
Ensemble 6 = All classifiers

In Table 6, using 4 and 5 features, the NB classifier consistently achieved the highest accuracy of 94.666%. With six features, the RF classifier obtained the highest accuracy of 94.915%. Employing seven features, both RF and Ensemble 1 achieved the highest accuracy of 95.469%. Using eight features, Ensemble 1 also achieved the highest accuracy of 95.469%. Conversely, without feature selection, the highest accuracy was obtained by RF, Ensemble 1, and Ensemble 5, all recording 95.733%. This accuracy was 0.264% higher than RF and Ensemble 1 with seven features, 0.814% higher than RF with six features, and 1.067% higher than NB with only four features.

DISCUSSION

This study achieved the highest accuracy of 95.733% using KNN Imputer and a combination of classifiers without feature selection, surpassing the accuracies reported in previous studies: 86.92% with Random Forest and missing value removal [5], 92.3% with McapNet and advanced deep learning techniques [17], and 86.8% with Random Forest [20]. This superiority was attained through the utilization of ensemble methods and systematic evaluation of feature selection techniques, demonstrating an advantage over approaches that simply remove missing values or employ single classifiers.

An analysis of the nWBV data distribution revealed that all research samples exhibited below-normal values. The implementation of KNN Imputer and K-Fold Cross Validation played a crucial role in enhancing accuracy, regardless of feature selection. While prior studies achieved a maximum accuracy of 86.92%, the integration of KNN Imputer, K-Fold Cross Validation, and feature selection increased accuracy by 8.813%. Using Information Gain for feature selection reduced the number of features from 9 to 4, optimizing computational efficiency and interpretability. Although accuracy slightly decreased from 95.733% to 95.462% (a reduction of 0.271%), this trade-off is minimal compared to the advantages of mitigating overfitting, enhancing model generalizability, and prioritizing the most relevant features for robust data analysis. These findings affirm that well-chosen imputation techniques and feature selection strategies significantly improve model performance, resulting in more reliable and precise predictions for Alzheimer's disease diagnosis.

However, several limitations should be acknowledged in this study:

1. **Dataset Composition**: The dataset may lack diversity in terms of demographic representation, including race, ethnicity, and geographical distribution. This limitation could impact the model's generalizability to broader populations.
2. **Computational Complexity**: The ensemble methods used, while improving accuracy, introduce computational overhead, which may affect scalability and real-time application feasibility.
3. **Real-World Applicability**: The model's performance in clinical settings remains uncertain due to potential discrepancies between the dataset and real-world medical cases. External validation with independent datasets is necessary to confirm its robustness.
4. **Feature Selection Bias**: The reliance on Information Gain for feature selection may introduce bias, potentially overlooking other important predictors. Further research should explore alternative feature selection techniques.
5. **Evaluation Metrics and Methodological Assumptions**: The chosen metrics and methodologies could influence result interpretation and comparability with other studies. Future research should assess model consistency across different evaluation frameworks.

Addressing these limitations will be crucial for developing a more comprehensive and adaptable predictive model. Future studies should validate findings across diverse datasets, consider alternative modelling approaches, and evaluate real-world clinical integration.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, this study achieved an impressive accuracy of 95.733% in diagnosing Alzheimer's disease using KNN Imputer and ensemble methods, surpassing previous benchmarks. The application of KNN Imputer and K-Fold Cross Validation significantly improved model accuracy. Feature selection through Information Gain further enhanced model performance by reducing complexity and improving interpretability, with only a minor reduction in accuracy. These findings underline the importance of choosing the right imputation and feature selection techniques to reduce overfitting, enhance model generalization, and optimize data-driven decision-making. Future studies should focus on validating these findings with diverse datasets, exploring additional feature selection methods, and assessing the model’s integration into real-world clinical settings.

REFERENCES

Abana, E. (2019). A Decision Tree Approach for Predicting Student Grades in Research Project using Weka. *IJACSA*(DOI: 10.14569/ijacsa.2019.0100739).

Akhtar, T., Gilani, S., Mushtaq, Z., Arif, S., Jamil, M., & Ayazet al., Y. (2021). Effective voting ensemble of homogenous ensembling with multiple attribute-selection approaches for improved identification of thyroid disorder. *Electronics, vol. 10, no. 23*(https://doi.org/10.3390/electronics10233026), 3026.

AlZu'b, S., Zraiqat, A., & Hendawi, S. (2022). Sustainable Development: A Semantics-aware Trends for Movies Recommendation System using Modern NLP. *International Journal of Advances in Soft Computing and Its Applications, 14(3)*(https://doi.org/10.15849/ijasca.221128.11), 154-173.

Ayinla, B., & Oremei, C. (2024). Development of Lr_multi- Cross-validation Model for Prediction of an Imbalanced Dataset in Flood Susceptible Area. (https://doi.org/10.21203/rs.3.rs-3826233/v1).

Barbara Pes. (2021). Learning from High-Dimensional and Class-Imbalanced Datasets Using Random Forests. *Information, 12(8)*(https://doi.org/10.3390/info12080286).

Basheer, S., Bhatia, S., & Sakri, S. (2021). Computational Modeling of Dementia Prediction Using Deep Neural Network: Analysis on OASIS Dataset. *IEEE Access , Volume: 9*(https://ieeexplore.ieee.org/document/9380278).

Beltrán, J., Wahba, B., Hose, N., Shasha, D., Kline, R., , ... , . (2020). Inexpensive, non-invasive biomarkers predict Alzheimer transition using machine learning analysis of the Alzheimer's Disease Neuroimaging (ADNI) database. *PLoS ONE, 15(7)*(https://doi.org/10.1371/journal.pone.0235663), e0235663.

Biswas, S., & Rajan, H. (2021). Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Softw*(https://doi.org/10.1145/3468264.3468536), 981–993.

Cao, H., Sarlin, R., & Jung, A. (2020). Learning Explainable Decision Rules via Maximum Satisfiability. *IEEE Access, Vol 8*(DOI: 10.1109/access.2020.3041040), 218180-218185.

Celard, P., Vieira, A., Iglesias, E., Borrajo, L., , & . (2020). LDA filter: A Latent Dirichlet Allocation preprocess method for Weka4. *PLoS ONE*(https://doi.org/10.1371/journal.pone.0241701).

Chen, C., Shi, X., Ye, X., Yang, L., , & . (2023). Intrusion detection model based on genetic algorithm optimization extreme learning machine of K-fold stratified cross-validation. *International Conference on Signal Processing and Communication Technology (SPCT 2022)*(https://doi.org/10.1117/12.2673803).

Chuan, Y., Zhao, C., He, Z., Wu, L., , & . (2021). The Success of AdaBoost and Its Application in Portfolio Management. *arXiv*(https://doi.org/10.48550/arXiv.2103.12345).

Dashtipour, K., Taylor, W., Ansari,, S., Zahid,, A., Gogate, M., Ahmad, J., . . . Abbai, Q. (2021). Detecting Alzheimer's disease using machine learning methods. *EAI*(https://hal.science/hal-03381752/document), HAL Id: hal-03381752.

Dong, Z., Wang, Q., Ke, Y., Zhang, W., Hong, Q., Liu, C., & et al. (2022). Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records. *J Transl Med, Vol.20*(DOI:10.1186/s12967-022-03339-1), Article number: 143.

Ebrahimi, K., Jourkesh, M., Sadigh-Eteghad, S., Stannard, S., Earnest, C., Ramsbottom, R., & et al. (2020). Effects of Physical Activity on Brain Energy Biomarkers in Alzheimer's Diseases. *Diseases, 8(2)*(https://doi.org/10.3390/diseases8020018), 18.

Ge, H., Ma, F., Li, Z., Tan, Z., Du, C., & . (2021). Improved accuracy of phenological detection in rice breeding by using ensemble models of machine learning based on uav-rgb imagery. *Remote Sensing, vol. 13, no. 14*(https://doi.org/10.3390/rs13142678), p. 2678.

Gillani, N., & Arslan, T. (2021). Intelligent Sensing Technologies for the Diagnosis, Monitoring and Therapy of Alzheimer's Disease: A Systematic Review. *Sensors, 21(12), 4249*(https://doi.org/10.3390/s21124249).

Hughes, C., Choi, M., Yi, J., Kim, S., Drews, A., George-Hyslop, P., & et al. (2020). Beta amyloid aggregates induce sensitised TLR4 signalling causing long-term potentiation deficit and rat neuronal cell death. *Communications Biology, 3*(https://doi.org/10.1038/s42003-020-0792-9), 79.

Istiqoh, A., Qodir, Z., & Ahmad, Z. (2022). Narrative Policy Framework: Presidential Threshold Policy Toward the 2024 Election. *J. Bina Praja, Volume 14 No 3*(DOI: 10.21787/jbp.14.2022.505-516), 505-516.

Kavitha, C., Mani, V., Srividhya, S., Khalaf, O., Romero, C., & . (2022). Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. *Front. Public Health, Volume 10*(https://doi.org/10.3389/fpubh.2022.853294).

Kost, S., Rheinbach, O., & Schaeben, H. (2019). Logistic regression for potential modeling. *Proc Appl Math and Mech*(https://doi.org/10.1002/pamm.201900039).

Ljubobratović, D., Vuković, M., Bakarić, M., Jemrić, T., Matetić, M., & . (2022). Assessment of Various Machine Learning Models for Peach Maturity Prediction Using Non-Destructive Sensor Data. *Sensors, 22(15)*(DOI:10.3390/s22155791), 5791.

M., S., & G., T. (2023). Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA). *Materialstoday: Proseeding, Volume:1 Part 2*(https://www.sciencedirect.com/science/article/abs/pii/S2214785321020757), 182-190.

Malavika, G., Rajathi, N., Vanitha, V., & Parameswari, P. (2020). Alzheimer Disease Forecasting using Machine Learning Algorithm. *Biosc.Biotech.Res.Comm, Special Issue Vol 13 No 11*(https://bbrc.in/wp-content/uploads/2021/01/Galley-Proof-004.pdf), 15-19.

Marzban, E., Eldeib, A., Yassine, I., Kadah, Y., , & . (2020). Alzheimer’s disease diagnosis from diffusion tensor images using convolutional neural networks. *PLoS ONE, vol. 15, no. 3*(https://doi.org/10.1371/journal.pone.0230409), e0230409.

Mnguni, L. (2021). Strategies for the Development and Application of Research Frameworks in Sciences Education Research. *JESR, Vol. 11 No. 6 (2021): November 2021*(https://doi.org/10.36941/jesr-2021-0123).

Naveed, N., Madhloom, H., & Husain, M. (2021). Breast Cancer Diagnosis Using Wrapper-Based Feature Selection and Artificial Neural Network. *acs, Vol 17 No.3*(https://doi.org/10.23743/acs-2021-18), 19–30.

Needham, R. (2022). *Alzheimer's Disease: A Caregiver's Guide with Answers to Questions and a Path to Spiritual Healing.* Columbus, OH: Gatekeeper Press.

Oh, J., Tannenbaum, A., & Deasy, J. (2022). Automatic identification of drug-induced liver injury literature using natural language processing and machine learning methods. (https://doi.org/10.1101/2022.08.10.503489).

Organization, W. H. (2023, March 15). *Dementia*. (www.who.int) Retrieved May 03, 2024, from https://www.who.int/news-room/fact-sheets/detail/dementia

Paramita, A. S. (2022). Implementation of the K-Nearest Neighbor Algorithm for the Classification of Student Thesis Subjects. *Journal of Applied Data Sciences, vol. 3, no. 3*(https://doi.org/10.47738/jads.v3i3.66), 128-136.

Patel, M., Ta, J., & Chou, F.-S. (2021). Non-Linear Algorithms in Supervised Classical Machine Learning. *Neonatology Today, 16(7)*(https://doi.org/10.51362/neonatology.today/202171674043), 40-43.

Peavy, G., Jenkins, C., Little, E., Gigliotti, C., Calcetas, A., Edland, S., & et al. (2020). Community Memory Screening as a Strategy for Recruiting Older Adults into Alzheimer’s Disease Research. *Preprint, Version 2*(https://doi.org/10.21203/rs.2.19958/v2).

Pino, R., Mendoza, R., & Sambayan, R. (2021). A Baybayin word recognition system. *PeerJ Computer Science, 7:e596*(https://doi.org/10.7717/peerj-cs.596).

Sai, P., Rajalakshmi, T., & Snekhalatha, U. (2021). Non-invasive thyroid detection based on electroglottogram signal using machine learning classifiers. *Proc Inst Mech Eng H, 235(10)*(https://doi.org/10.1177/09544119211028070), 1128-1145.

Sara, D., Sami, A., Khan Md., H., Asif, K., Mirjam, J., & A S M , F. (2022). Dementia Prediction Using Machine Learning. *CENTERIS- International Conference on Enterprise Information System/ ProjMAn- International Conference on Project Management/ HCist-International Conference on Health and SOcial Care Information System and Technologies 2022. -*.

Shiino, A., Shirakashi, Y., Ishida, M., Tanigaki, K., Japanese Alzheimer’s Disease Neuroimaging Initiati, & . (2021). Machine learning of brain structural biomarkers for Alzheimer's disease (AD) diagnosis, prediction of disease progression, and amyloid beta deposition in the Japanese population. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, Volume 13, Issue 1* (https://doi.org/10.1002/dad2.12246).

Uddin, K., Alam, M., Jannat-E-Anawar, Uddin, M., Aryal , S., & . (2023). A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease. *Biomedical Materials & Devices, Volume 1*(DOI: 10.1007/s44174-023-00078-9), 882-898.

Umar, M., Zhanfang, C., Shuaib, K., Liu, Y., , & . (2024). Effects of Feature Selection and Normalization on Network Intrusion Detection. *figshare. Preprint.*(https://doi.org/10.36227/techrxiv.12480425.v3).

Winarti, T., Indriyawati, H., Vydia, V., Christanto, F., , & . (2021). Performance comparison between naive bayes and k- nearest neighbor algorithm for the classification of Indonesian language articles. *IJ-AI, Vol 10 No 2*(http://doi.org/10.11591/ijai.v10.i2.pp452-457), 452-457.

Xu, X., K Fairley, C., Chow, E., Lee, D., Zhang, L., & Ong, J. (2022). Using machine learning approaches to predict timely clinic attendance and the uptake of HIV/STI testing post clinic reminder messages. *Sci Rep, 12(1)*(DOI:10.1038/s41598-022-12033-7), Article number: 8757.

Xu, Y., Wu, G., & Chen, Y. (2022). Predicting Patients' Satisfaction With Doctors in Online Medical Communities. *Journal of Organizational and End User Computing (JOEUC), 34(4)*(http://doi.org/10.4018/JOEUC.287571), 1-17.

Yıldız, Z., Eren, N., Orçun, A., Gökyiğit, F., Turgay, F., & Celebi, L. (2021). Serum apelin-13 levels and total oxidant/antioxidant status of patients with Alzheimer's disease. *Aging Medicine, 4*(DOI: 10.1002/agm2.12173), 201-205.

Zhang, L., Sindakis, S., Dhaulta, N., Asongu, S., , & . (2023). Economic Crisis Management during the Covid-19 Pandemic: The Role of Entrepreneurship for Improving the Nigerian Mono-Economy. *Journal of the Knowledge Economy, Version 1*(https://doi.org/10.21203/rs.3.rs-1438381/v1).

Zhang, R., Zeng, M., Zhang, X., Yang, Z., Lv, N., & et al. (2023). Therapeutic Candidates for Alzheimer's Disease: Saponins. *International Journal of Molecular Sciences, 24(13)*(https://doi.org/10.3390/ijms241310505), 10505.

Zhang, S., Lin, H.-C., & Wang, X. (2021). Forecast of E-Commerce Transactions Trend Using Integration of Enhanced Whale Optimization Algorithm and Support Vector Machine. *Computational Intelligence and Neuroscience*(https://doi.org/10.1155/2021/9931521), Article ID 9931521.

Πεππές, N., Daskalakis, E., Alexakis, T., Adamopoulou, E., Demestichas, K., & . (2021). Performance of machine learning-based multi-model voting ensemble methods for network threat detection in agriculture 4.0. *Sensors, vol. 21, no. 22*(https://doi.org/10.3390/s212274753), 7475.