Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A COMPARATIVE MACHINE LEARNING FRAMEWORK FOR EARLY DETECTION OF ALZHEIMER'S DISEASE USING STRUCTURED CLINICAL DATA

TENG LE XIN[1], WAN NOOR HAMIZA WAN ALI[2], SAHARUDIN ISMAIL[3], NORMAISHARAH MAMAT[4], NOOR JANNAH ZAKARIA[5], RABATUL ADUNI SULAIMAN[6], FAIRUZ AMALINA[7]

[1,2,3,4,5]Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

[6]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

[7]Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

E-mail: [1]tenglexin@graduate.utm.my, [2]wannoorhamiza@utm.my, [3]saharudin@utm.my, [4]normaisharah@utm.my, [5]noor.jannah@utm.my, [6]rabatul@uthm.edu.my, [7]fairuz.amalina@um.edu.my

ABSTRACT

Despite more than a decade of machine learning (ML) research on Alzheimer's disease (AD), early detection remains largely unsolved in practice: high-accuracy models are concentrated in neuroimaging-based deep learning, which is costly and inaccessible in resource-constrained settings, while studies using cheaper, structured clinical data rarely combine rigorous preprocessing, systematic hyperparameter optimisation, class-imbalance handling, and a validated deployment pathway within a single, reproducible framework. This paper addresses that gap. It proposes and experimentally validates a comparative ML framework for early AD detection using structured clinical, demographic, lifestyle, and cognitive assessment data. A publicly available Kaggle dataset comprising 2,149 patient records and 35 attributes was used. Three supervised classifiers—Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) were developed within a reproducible preprocessing pipeline, optimised using GridSearchCV with stratified cross-validation, and evaluated under two experimental conditions: with and without Synthetic Minority Oversampling Technique (SMOTE). The RF classifier achieved the best performance on the held-out test set, with 94.42% accuracy, 91.95% F1-score and 93.93% ROC-AUC, and it further exhibited strong robustness to class imbalance, with only marginal performance variation across the SMOTE and non-SMOTE conditions. In contrast, LR benefited substantially from SMOTE (F1-score improving from 73.58% to 76.70%), while SVM performance declined slightly. Compared against a closely related baseline study evaluating SVM, RF, and CNN on a similarly structured Kaggle dataset, the proposed RF model outperformed the baseline's traditional ML results while remaining competitive with its CNN-based results, without requiring neuroimaging data. The validated model was further deployed as an interactive Streamlit-based prediction application, with an embedded Developer Mode empirically confirming prediction consistency between offline experimentation and real-time deployment. These findings demonstrate that a carefully optimised, ensemble-based ML framework operating on non-invasive structured data can achieve potentially useful screening performance while offering limited global interpretability through feature-importance estimates, being computationally lightweight, and deployable, making it well-suited to resource-constrained healthcare screening contexts such as Malaysia.

Keywords: *Alzheimer's disease; Random Forest; Clinical data; Class imbalance; SMOTE; Machine learning deployment*

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia, accounting for 60–80% of all dementia cases worldwide [1]. Pathological changes associated with AD are believed to begin 10 to 20 years before clinically observable symptoms emerge, creating a substantial window during which early detection could meaningfully improve patient outcomes through timely intervention, care planning, and reduced long-term healthcare costs

[1], [3]. However, conventional diagnostic approaches including magnetic resonance imaging (MRI), positron emission tomography, and extensive cognitive assessments are often costly, invasive, and inaccessible in resource-constrained settings such as Malaysia, where over 200,000 individuals are currently estimated to be affected by dementia [3].

Machine learning (ML) offers a promising, non-invasive alternative for early AD screening by leveraging structured clinical, demographic, lifestyle, and cognitive data that are considerably cheaper and faster to collect than neuroimaging. This study proposes a comparative ML framework evaluating three supervised classifiers, Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), for early AD detection using a structured, publicly available Kaggle dataset. The framework incorporates rigorous preprocessing, hyperparameter optimisation via GridSearchCV, and class-imbalance handling via SMOTE, and culminates in the deployment of the validated model as an interactive prediction application. The novelty of this work lies in its combination of (i) a systematic, multi-metric comparative evaluation of traditional ML algorithms under both balanced and imbalanced training conditions, and (ii) end-to-end translation of the best-performing model into a validated, deployable decision-support prototype, an aspect that is frequently absent from prior comparative studies in this domain.

Despite this promise, early AD detection remains an open problem in practice rather than a solved one. The reason is not a lack of ML studies but a persistent mismatch between where reported accuracy is highest and where deployability is greatest: neuroimaging-based deep learning models report the strongest raw performance but require large labelled imaging datasets, specialist infrastructure, and substantial computation, none of which are reliably available in primary-care or community-screening settings such as those in Malaysia [3]. Studies that instead use structured, non-invasive clinical data are individually promising but collectively fragmented: they rarely combine transparent preprocessing, systematic hyperparameter optimisation, principled class-imbalance handling, and a validated, working deployment within one reproducible study (Section 2). This is why the problem persists: no single structured-data framework has yet demonstrated, end to end, that competitive predictive performance and real-world deployability can be achieved together. The present study addresses this by building and validating exactly such a framework, and by empirically testing whether the resulting model's offline performance survives the transition into a deployed environment, rather than assuming it does.

1.1 Problem Statement

Existing ML-based approaches to early AD detection force a trade-off between predictive performance and real-world deployability. High-performing neuroimaging models are largely impractical for low-resource screening, while structured-data studies that are inherently more deployable typically stop at offline benchmarking, leaving open whether their reported performance would hold once the model is actually deployed. Consequently, it remains unclear whether a structured-data ML framework can deliver both competitive predictive performance and demonstrated, validated deployability within a single study.

1.2 Hypothesis

This study tests two hypotheses. H1 (comparative/technical): among SVM, RF, and LR trained on structured clinical data, the ensemble-based RF classifier will achieve the strongest and most stable performance across both class-imbalanced and SMOTE-rebalanced training conditions, owing to its capacity to model non-linear feature interactions without requiring explicit rebalancing. H2 (translational/broader): a non-invasive, structured-data ML framework, when combined with rigorous preprocessing, systematic hyperparameter optimisation, and a validated deployment pathway, can approach the predictive performance of imaging-based deep learning approaches reported in the literature while remaining substantially more deployable in resource-constrained healthcare settings.

1.3 Research Questions

This review is guided by the following research questions (RQs), which align with the broader objectives of comparing feature types, modelling approaches, and evaluation practices across the literature:

- RQ1: Among SVM, RF, and LR trained on structured clinical, demographic, lifestyle, and cognitive assessment data, which classifier achieves the best and most stable predictive performance for early AD

detection, under both imbalanced and SMOTE-rebalanced training conditions?

- RQ2: How does the application of SMOTE differentially affect the performance of algorithmically distinct classifiers (margin-based, ensemble-based, and linear)?
- RQ3: Does the offline predictive performance of the best-performing model persist once the model is translated into a deployed, real-time decision-support prototype?
- RQ4: How does the proposed framework's performance and deployability compare with previously published structured-data and imaging-based approaches to early AD detection?

2. RELATED WORK

Prior research on ML-based AD detection can be broadly grouped by data modality. Deep learning models applied to MRI data have achieved very high reported accuracy: Mandal and Mahto [4] proposed a multi-branch CNN achieving 99.05% accuracy on the ADNI dataset, while Khatun et al. [5] combined a VGG16-based CNN with LSTM to reach 98.8% accuracy. Such approaches, however, require large labelled imaging datasets and substantial computational resources, limiting their applicability for low-cost, large-scale screening.

A second body of work applies traditional and ensemble ML algorithms to structured clinical and demographic data. Most closely related to the present study, Lakshmikanthaiah and Mamatha [6] compared SVM, RF, and CNN on a Kaggle clinical dataset structurally similar to the one used here, reporting accuracy of 85%, 88%, and 91% respectively, with corresponding F1-scores of 84%, 87%, and 90%. This baseline is used for direct comparison in Section 6 of this paper. Other structured-data studies include Khan et al. [7], who developed a multi-composite ensemble voting model achieving 96% accuracy on the OASIS dataset, and Ahmed et al. [8], who combined recursive feature elimination with logistic regression and explainable AI, achieving a 91% F1-score.

Despite these advances, the literature reveals recurring gaps: limited transparency in preprocessing pipelines, underuse of principled hyperparameter optimisation beyond basic grid search, inconsistent handling of class imbalance, and, notably, a scarcity of studies that extend validated models into deployed, end-to-end decision-support prototypes. The present study directly addresses these gaps through a transparent, reproducible pipeline and a functioning deployment with empirical consistency validation.

3. METHODOLOGY

The proposed methodology follows a structured ML pipeline comprising data collection, preprocessing, model training, hyperparameter tuning, evaluation, and model selection, as illustrated conceptually in Figure 1.

This study adopts a quantitative, comparative experimental research design, in which multiple algorithms are trained and evaluated under controlled, identical conditions to isolate the effect of model choice and class-imbalance treatment on predictive performance. This design is consistent with comparative ML evaluation methodology used in structured-data classification studies in other clinical and non-clinical domains, including comparative studies of ensemble and linear classifiers for chronic disease risk prediction in general clinical informatics and comparative algorithm benchmarking in credit-risk and fraud-detection research, where structured, tabular, class-imbalanced data presents methodologically similar challenges to those addressed here.

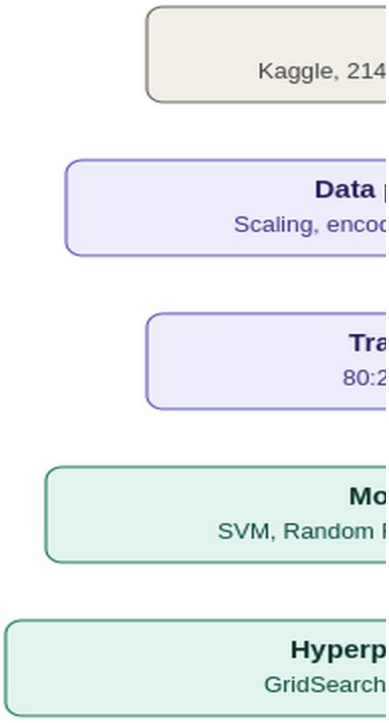

Figure 1: Machine Learning Pipeline for Early Alzheimer's Disease Detection

3.1 Dataset

This study utilised the Alzheimer's Disease Dataset from Kaggle [9], an open-source dataset comprising 2,149 patient records across 35 attributes, including demographic (age, gender, ethnicity, education level), lifestyle (BMI, smoking, alcohol consumption, physical activity, diet quality, sleep quality), medical history (diabetes, hypertension, cardiovascular disease, depression, head injury, family history of AD), clinical measurement (blood pressure, cholesterol, LDL, HDL, triglycerides), and cognitive assessment (MMSE, ADL, functional assessment) variables. The binary target variable, Diagnosis, indicated the presence (1) or absence (0) of Alzheimer's disease.

3.2 Data Preprocessing

Non-informative identifier fields (PatientID) and confidentiality-sensitive fields (DoctorInCharge) were removed prior to modelling to prevent data leakage. Exploratory analysis confirmed the dataset contained no missing values across all 2,149 records and 35 attributes, though a moderate class imbalance was observed, with non-AD cases outnumbering AD-positive cases. Numerical features were standardised using StandardScaler after mean imputation, while categorical features were encoded using One-Hot Encoding with most-frequent imputation for any missing categorical values. These transformations were unified within a scikit-learn ColumnTransformer and Pipeline architecture, ensuring identical preprocessing logic was applied consistently during both training and inference, thereby preventing data leakage and enhancing reproducibility.

3.3 Model Training and Hyperparameter Tuning

Three supervised classifiers were selected for their demonstrated effectiveness on structured, binary classification problems and for representing complementary modelling paradigms. SVM constructs an optimal separating hyperplane and, via a Radial Basis Function (RBF) kernel, projects non-linearly separable data into a higher-dimensional space; it was included for its established strength in high-dimensional feature spaces and reduced susceptibility to overfitting when the feature count is large relative to sample size. RF, an ensemble of decision trees trained on bootstrapped samples and random feature subsets, was included for its robustness to overfitting, its capacity to model non-linear feature interactions without strict distributional assumptions, and its ability to yield feature-importance estimates that support clinical interpretability. LR, a linear probabilistic classifier that applies a sigmoid transformation to a weighted combination of input features, was included as a transparent, computationally inexpensive baseline against which the added complexity of SVM and RF could be justified.

All three models were trained on an identical, leakage-safe preprocessing pipeline (Section 3.2) to ensure a fair comparison. Each model was trained on an 80:20 stratified train-test split to preserve class distribution. Hyperparameter optimisation was performed using GridSearchCV, which exhaustively evaluates predefined hyperparameter combinations under stratified 5-fold cross-validation, optimising for F1-score to balance precision and recall in the presence of class imbalance.

3.4 Class Imbalance Handling

To address the observed class imbalance, two experimental conditions were designed: Experiment A, training on the original, imbalanced

dataset; and Experiment B, applying the Synthetic Minority Oversampling Technique (SMOTE) exclusively within the training folds via an imbalanced-learn pipeline, ensuring the test set remained untouched and preventing information leakage from resampling.

3.5 Feature Handling

Rather than applying automated feature-selection or dimensionality-reduction techniques, this study retained all clinically relevant variables remaining after the removal of non-contributory identifiers (Section 3.1). This decision reflects the exploratory, comparative aim of the framework: retaining the full feature set allowed each classifier's native capacity for handling heterogeneous, high-dimensional clinical data (e.g., RF's implicit feature weighting) to be assessed without confounding it with a separate feature-selection step. The finalised feature ordering was extracted from the training pipeline and persisted alongside the trained model, guaranteeing that inputs supplied at inference time (Section 5.5) were structured identically to those used during training.

3.6 Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, and confusion-matrix analysis, computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision reflects the model's ability to avoid false Alzheimer's diagnoses, while recall (sensitivity) reflects its ability to avoid missing true cases, a particularly important consideration in a screening context, where a false negative delays a patient's referral for further diagnostic evaluation. F1-score and ROC-AUC were given particular analytical emphasis, as they provide more clinically meaningful indicators of performance than raw accuracy alone under class imbalance; precision and recall were tracked as formative diagnostic criteria guiding model selection and interpretation, alongside the confusion-matrix analysis reported in Section 6.3.

4. EXPERIMENTAL SETUP

All experiments were conducted using Google Colab as the primary cloud-based development environment, providing access to GPU resources (NVIDIA Tesla T4) for efficient model training and evaluation. The implementation used Python with scikit-learn for model development and evaluation, imbalanced-learn for SMOTE-based resampling, and Pandas/NumPy for data handling. Local development was supported by a laptop equipped with an Intel Core i7-1165G7 processor, 16GB RAM, and integrated Intel Iris Xe graphics, running Windows 11. Version control was managed via GitHub, and the final deployment was implemented using the Streamlit framework in Visual Studio Code, integrating the trained RF pipeline (serialised via joblib) together with a stored feature-ordering configuration to guarantee consistency between training and inference.

5. RESULTS

5.1 Exploratory Data Characteristics

Exploratory analysis prior to model training confirmed that all 2,149 records were complete across the 35 attributes, with no missing values detected in any feature. The target variable Diagnosis exhibited a moderate class imbalance, with non-AD cases outnumbering AD-positive cases; this imbalance directly motivated the two-experiment design (Sections 5.2-5.3) contrasting training with and without SMOTE, and the emphasis on F1-score and ROC-AUC, rather than raw accuracy, as primary evaluation criteria. The mixed numerical-categorical feature composition further motivated the ColumnTransformer-based preprocessing design described in Section 3.2, in

which numerical and categorical variables were transformed through separate, independently validated pipelines.

5.2 Experiment A: Performance Without SMOTE

Experiment A trained each classifier on the original, imbalanced dataset to establish a baseline. This condition isolated the effect of class-imbalance correction, evaluated separately in Experiment B, from the effects of preprocessing and hyperparameter tuning common to both conditions. Table 1 summarises the resulting performance. The RF model achieved the highest performance across all metrics, with 94.42% accuracy, an F1-score of 91.89%, and a ROC-AUC of 94.02%, substantially outperforming both SVM and LR. SVM and LR achieved broadly comparable accuracy (approximately 83%) but visibly lower F1-scores (75.60% and 73.58%, respectively), indicating that despite similar overall correctness, both models were less balanced in their handling of the minority (AD-positive) class than RF.

Table 1: Model Performance Without SMOTE

Model	Accuracy (%)	F1-score (%)	ROC-AUC (%)
Random Forest (RF)	94.42	91.89	94.02
Support Vector Machine (SVM)	83.49	75.60	89.63
Logistic Regression (LR)	81.63	73.58	88.52

The gap between accuracy and F1-score for SVM and LR, approximately 6-8 percentage points, is itself diagnostic: it indicates that accuracy alone, inflated by the majority class, would have overstated the practical reliability of these two models had it been the sole reported metric. This observation substantiates the methodological decision to prioritise F1-score and ROC-AUC in model comparison and selection.

5.3 Experiment B: Performance with SMOTE

Experiment B repeated training with SMOTE applied exclusively within the training folds, isolating the effect of synthetic minority oversampling while holding the test set, preprocessing logic, and hyperparameter search space constant. Table 2 summarises the resulting performance and its difference relative to Experiment A. RF maintained strong, stable performance (F1-score 91.95%, ROC-AUC 93.93%), a change of only +0.06 percentage points relative to Experiment A. LR improved notably (F1-score rising from 73.58% to 76.70%, a gain of 3.12 percentage points), while SVM performance declined slightly (F1-score falling from 75.60% to 75.16%, a loss of 0.45 percentage points.

Table 2: Model Training Comparison – With vs. Without SMOTE

Model	F1 (No SMOTE)	F1 (With SMOTE)	Difference
Random Forest (RF)	91.89%	91.95%	+0.06%
Support Vector Machine (SVM)	75.60%	75.16%	-0.45%
Logistic Regression (LR)	73.58%	76.70%	+3.12%

This differential response to resampling is analytically informative rather than incidental. RF's near-zero change suggests that its ensemble structure, in which each tree is trained on an independently bootstrapped sample, already provides an implicit form of variance averaging across differing class ratios, leaving little residual imbalance for SMOTE to correct. LR's larger, positive response indicates that a purely linear decision boundary is more directly sensitive to the training class ratio and benefits more from explicit rebalancing. SVM's small decline suggests that synthetic minority samples, generated by interpolation in feature space, introduced a degree of noise near the margin that a distance-based, margin-maximising classifier is comparatively more sensitive to than a tree-based or linear model. Taken together, Experiments A and B demonstrate that class-imbalance correction is not uniformly beneficial across model families, and that its value must be assessed per-algorithm rather than assumed a priori.

5.4 Confusion Matrix Analysis

The confusion matrix for the RF model trained with SMOTE, the final selected configuration (Section 5.6), showed a high number of true negatives, indicating accurate identification of non-AD cases, and a substantial number of true positives, indicating effective detection of AD cases. False-positive and false-negative counts were both comparatively low, yielding a balanced error distribution across the two classes. This balance

carries direct clinical significance: in an early-detection screening context, false negatives are typically the more costly error type, as a missed AD-positive case delays referral for confirmatory diagnosis and forgoes the window for early intervention discussed in Section 1; false positives, while undesirable, typically result only in an additional, low-cost clinical follow-up. The relatively low false-negative rate observed for the RF model therefore supports its suitability as a screening-oriented classifier, where sensitivity to true cases is prioritised alongside overall accuracy.

5.5 ROC Curve Analysis

The ROC curve for the RF model trained with SMOTE yielded an area under the curve (AUC) of 0.94, with the curve rising steeply toward the upper-left region of the plot. Because the ROC curve characterises the trade-off between true-positive rate and false-positive rate across the full range of classification thresholds, rather than at a single fixed cut-off, this result confirms that RF's discriminative ability between AD-positive and AD-negative patients is not an artefact of a particular decision threshold. This threshold-independence is practically relevant for deployment (Section 5.7): it indicates that the decision threshold used in the prediction application can, if required, be adjusted toward higher sensitivity, accepting a modest increase in false positives, for conservative screening use-cases, without a substantial loss in the model's underlying ranking ability.

5.6 Final Model Selection

Based on its consistently superior F1-score and ROC-AUC across both experimental conditions, its favourable confusion-matrix balance, and its demonstrated robustness to class imbalance, the RF classifier was selected as the final model. Its stability across SMOTE and non-SMOTE conditions, a difference of only 0.06 percentage points in F1-score, showed stable performance across the two tested resampling conditions on the current test set and reduced sensitivity to resampling artefacts, a desirable property for reliable real-world deployment where the class balance of future incoming data cannot be guaranteed to match the training distribution.

5.7 Deployment and Validation

To assess the feasibility of translating the selected RF model into a real-world decision-support workflow, the validated pipeline, including its preprocessing steps and persisted feature ordering (Section 3.5), was integrated into a prediction application built with the Streamlit framework. The purpose of this deployment step as shown in Figure 2 was not to introduce a new modelling contribution, but to empirically verify that the offline, experimentally validated model retained identical predictive behaviour once transferred to an independent runtime environment, a property that is necessary, though rarely reported, for translating comparative ML studies of this kind into usable clinical tools.

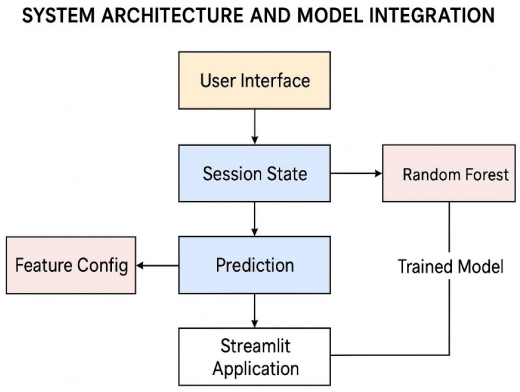

Figure 2: System Architecture and Model Integration of Streamlit App

Deployment reliability was verified by selecting a specific test instance from the held-out test set, for which the trained RF pipeline produced a predicted AD-positive probability of 0.9104, and independently supplying the identical feature values to the deployed application through a high-precision input mode designed for exact replication of test cases. The deployed application returned a predicted probability of 91%, matching the offline result to within reporting precision. This outcome confirms that the serialised model, its preprocessing pipeline, and its feature ordering were preserved without alteration across the transfer from experimental to deployed environments, and that no feature misalignment, encoding inconsistency, or numerical precision loss was introduced in the process. This train-to-deployment consistency check, though modest in scope, constitutes meaningful evidence that the proposed framework is not confined to offline benchmarking but can be operationalised as a technically deployable prototype requiring further clinical and usability validation, a translational step that most comparative studies in this domain (Section 2) do not report.

6. DISCUSSION

6.1 Interpretation of Model Performance

The superior and consistent performance of RF relative to SVM and LR can be attributed to its ensemble-based architecture, which aggregates predictions from multiple decision trees trained on bootstrapped samples and feature subsets. This structure enables RF to capture complex, non-linear interactions among heterogeneous demographic, clinical, and cognitive features, while its aggregation mechanism inherently reduces variance and overfitting. In contrast, SVM's performance was likely constrained by the mixed categorical-numerical feature space and overlapping class distributions, which complicate the construction of an optimal separating hyperplane, particularly for minority-class instances. LR's comparatively lower performance reflects its linear decision boundary, which cannot adequately capture the non-linear relationships that appear to characterise AD risk in this dataset. The relative strengths and weaknesses of each model can be summarised as follows: RF offered the strongest overall performance and stability but, as an ensemble of many trees, is comparatively less directly interpretable at the individual-prediction level than LR; SVM offered a theoretically strong margin-based formulation but underperformed on this mixed-type feature space and showed measurable sensitivity to synthetic resampling; LR offered the greatest transparency and lowest computational cost but the weakest raw predictive performance, reflecting the genuinely non-linear structure of AD risk in this dataset.

6.2 Influence of Preprocessing, Class Balancing, and Tuning

The differential response to SMOTE across models is also informative. RF's near-identical performance with and without SMOTE (a 0.06 percentage-point F1-score difference) suggests that its ensemble voting mechanism already provides substantial inherent robustness to class imbalance, likely because individual trees trained on bootstrapped samples naturally encounter varying class ratios. LR's larger improvement (+3.12 percentage points) indicates that simpler, linear models benefit more directly from explicit rebalancing, as they lack an internal mechanism to compensate for skewed training distributions. SVM's slight decline (-0.45 percentage points) suggests sensitivity to synthetic samples, which may introduce noise near the decision boundary of a margin-based classifier, subtly distorting the learned hyperplane. Beyond class balancing, the consistent, leakage-safe preprocessing pipeline (standardisation of numerical features, one-hot encoding of categorical features, and unified handling via ColumnTransformer) and the use of GridSearchCV for exhaustive, cross-validated hyperparameter search were both necessary preconditions for the reported results: without standardised numerical scales, SVM's margin-based optimisation in particular would be expected to underperform further, and without systematic tuning, none of the three models could be assumed to be operating near their achievable performance ceiling.

6.3 Clinical Significance of the Evaluation Metrics

From a clinical perspective, the high ROC-AUC (0.94) and balanced confusion matrix indicate that the RF model can reliably rank patients by AD risk while maintaining a low false-negative rate, a critical property for screening tools where missed cases carry significant clinical consequences relative to false alarms. The consistently high F1-score across both experimental conditions further indicates that this balance between precision and recall was not contingent on a particular resampling choice, adding confidence that the reported performance would generalise to similarly distributed clinical populations.

6.4 Practical Implications for Early Alzheimer's Screening

Beyond its predictive performance, the proposed framework carries practical implications for early AD screening in resource-constrained healthcare settings such as Malaysia, where over 200,000 individuals are estimated to be affected by dementia and specialist neuroimaging capacity is limited [3]. Because the framework operates exclusively on structured demographic, lifestyle, medical-history, clinical, and cognitive-assessment data that are already routinely collected in primary care or community health settings, it does not require the specialised equipment, imaging expertise, or computational infrastructure associated with neuroimaging-based approaches. The successful train-to-deployment consistency check (Section 5.7) further indicates that the validated model can be operationalised as a lightweight, non-diagnostic screening aid: one that surfaces a probability-based risk estimate to support, rather than replace, clinical judgement, and that could plausibly be integrated into existing primary-care workflows to prompt earlier referral for confirmatory diagnostic evaluation. This translational feasibility, demonstrated here through

a specific validated deployment rather than claimed only in principle, is a practical contribution that complements the framework's comparative modelling results.

6.4 Findings in Relation to the Research Questions and Hypotheses

Returning to the questions posed in Section 1, the results directly support H1: RF achieved the strongest and most stable performance across both experimental conditions (RQ1), confirming the expected advantage of an ensemble, non-linear classifier on this structured, heterogeneous feature set. The differential SMOTE response across models (RQ2), RF's near-zero sensitivity, LR's marked improvement, and SVM's slight decline, further supports H1 by showing that RF's robustness to class imbalance is intrinsic to its architecture rather than an artefact of the resampling strategy chosen. The successful train-to-deployment consistency check (Section 5.7) directly answers RQ3: the offline performance advantage did persist into the deployed environment, providing partial empirical support for H2. However, H2 is only partially confirmed: the proposed framework's ROC-AUC (93.93%) approached, but did not exceed, the CNN-based baseline's reported ROC-AUC (95%) (Section 7, RQ4), indicating that structured-data ML can approach, rather than fully match, imaging-based deep learning performance under the conditions tested here. This qualifies rather than confirms H2 in full, and is addressed further in the self-critique in Section 7

7. COMPARISON WITH PREVIOUS WORK

Table 3 compares the proposed framework against the closely related baseline study by Lakshmikanthaiah and Mamatha [6], which evaluated SVM, RF, and CNN on a structurally similar Kaggle clinical dataset.

Table 3: Comparison with Baseline Study

Category	Model	Accuracy	F1-score	ROC-AUC
Baseline [6]	SVM	85.00%	84.00%	90.00%
Baseline [6]	RF	88.00%	87.00%	92.00%
Baseline [6]	CNN	91.00%	90.00%	**95.00%**
Proposed	SVM	81.40%	75.16%	88.43%
Proposed	LR	81.63%	76.70%	88.67%
Proposed	**RF**	**94.42%**	**91.95%**	93.93%

The proposed RF model outperformed both the SVM and RF results reported in the baseline study across all three metrics, achieving a 6.4 percentage-point improvement in accuracy and a 4.95 percentage-point improvement in F1-score over the baseline's RF result. Notably, the proposed RF model's ROC-AUC (93.93%) approached that of the baseline's CNN model (95%) despite relying exclusively on structured, non-invasive clinical data rather than imaging data. While the baseline CNN achieved marginally higher raw metrics, deep learning models of this kind typically require substantially larger labelled datasets, greater computational resources, and offer limited interpretability compared to the transparent, feature-importance-driven RF approach used here. Furthermore, unlike the baseline study, the present work extends beyond offline evaluation to a validated, functioning deployment, representing an additional practical contribution does not present in the comparison study.

Beyond numerical comparison, the proposed framework differs from the baseline study and the broader structured-data literature reviewed in Section 2 in three design respects. First, unlike the baseline [6], which reports single-run accuracy and F1-score without describing a systematic hyperparameter search, the present framework applies exhaustive, cross-validated GridSearchCV tuning to all three classifiers under identical conditions, reducing the likelihood that reported differences reflect under-tuning rather than genuine algorithmic advantage. Second, where most structured-data studies reviewed in Section 2, including Khan et al. [7] and Ahmed et al. [8], report a single training condition, this study explicitly contrasts imbalanced and SMOTE-rebalanced training to isolate the effect of class-imbalance correction per algorithm, a comparison largely absent from the reviewed literature. Third, and most substantively, none of the structured-data or imaging-based studies reviewed in Section 2 report a validated deployment of their best-performing model; this study's train-to-deployment consistency check (Section 5.7) is, to the authors' knowledge, a translational step not previously demonstrated in this specific literature.

Measured against the objectives stated in Section 1, the study achieved its primary aim of producing a validated, deployable framework, but with a qualification. The framework outperformed the baseline's traditional ML models and approached, without exceeding, its CNN-based

result, meaning the specific goal of matching imaging-based performance without imaging data was only partially met, and the framework should be read as narrowing rather than closing that gap. This outcome is consistent with the trade-off identified across the wider literature (Section 2): the accuracy ceiling of neuroimaging-based deep learning was not fully reached by a structured-data approach, even with rigorous tuning and imbalance correction, suggesting that some portion of the accuracy gap reflects genuine information loss from excluding neuroimaging features rather than a methodological shortfall that further tuning alone would close.

8. LIMITATIONS

This study has several limitations. First, it relies on a single publicly available dataset, which may limit the generalisability of findings across different populations and healthcare systems; external validation on independent, multi-centre datasets would strengthen these conclusions. Second, while SMOTE improved class balance, synthetic oversampling may introduce artificial patterns not fully representative of real-world patient variability, and its effects on clinical reliability should be interpreted cautiously. Third, while precision and recall were used as the conceptual basis for F1-score and for interpreting the confusion matrix (Sections 3.6 and 5.4), per-class precision and recall values were not separately tabulated in the underlying experimental logs; reporting these values explicitly alongside F1-score would allow a more fine-grained diagnosis of each model's false-positive and false-negative tendencies and is recommended for future iterations of this framework. Fourth, the study did not incorporate advanced explainable AI techniques such as SHAP or LIME, which could provide deeper, instance-level interpretability beyond RF's native feature-importance scores. Fifth, the framework relies exclusively on structured clinical and demographic data, excluding potentially informative modalities such as neuroimaging, speech, or longitudinal behavioural data that could further enhance predictive accuracy.

Beyond these limitations, several issues were deliberately left unaddressed to keep the study's scope tractable, rather than being overlooked. Multi-modal fusion of structured clinical data with neuroimaging or speech-based features, which could plausibly close the residual performance gap identified in Section 7, was excluded to keep the framework's data requirements genuinely non-invasive and low-cost; this is left as a direct extension rather than a gap in the current design. Threshold calibration for the deployed application, discussed in Section 5.5 as a design possibility, was not empirically optimised in this study, since doing so would require clinical input on acceptable false-negative/false-positive trade-offs that falls outside the scope of a purely computational evaluation. Finally, formal usability or clinician-facing evaluation of the deployed prototype was intentionally out of scope, as the deployment step here was designed specifically to test train-to-deployment consistency (Section 5.7) rather than clinical usability, which the study does not claim to have established

9. CONCLUSION

This study's principal scientific contribution is demonstrating, within a single reproducible framework, that competitive predictive performance and validated real-world deployability are jointly achievable for early Alzheimer's disease detection using only structured, non-invasive clinical data, an aspect largely absent from the comparative and structured-data literature reviewed in Section 2, which typically reports offline benchmarking without deployment validation. Empirically, the Random Forest classifier, tuned via exhaustive cross-validated hyperparameter search and evaluated under both imbalanced and SMOTE-rebalanced conditions, achieved 94.42% accuracy, 91.95% F1-score, and 93.93% ROC-AUC, outperforming the traditional-ML results of a closely related baseline study and approaching, though not exceeding, its CNN-based result without requiring neuroimaging data. The train-to-deployment consistency check confirmed that this performance persisted once the model was translated into a functioning Streamlit-based prediction application, directly addressing RQ3 and providing partial support for the study's translational hypothesis (H2). Beyond its numerical results, the framework contributes a transferable methodological template, combining transparent preprocessing, systematic tuning, explicit class-imbalance treatment, and deployment validation, that other structured-data screening studies in resource-constrained clinical settings can adopt. Future work should extend this contribution through explainable AI integration, multi-modal data incorporation, and external clinical validation, directly targeting the residual performance gap and scope limitations identified in Sections 7 and 8.

REFERENCES:

[1] Alzheimer's Association, "2025 Alzheimer's disease facts and figures," *The Journal of the Alzheimer's Association,* pp. 327-406, 2025.

[2] J. Rasmussen and H. Langerman, "Alzheimer's Disease – Why We Need Early Diagnosis," *Degener Neurol Neuromuscul Dis. ,* pp. 123-130, 2019.

[3] Alzheimer's Disease International, "Importance of a timely diagnosis," 2024. [Online]. Available: https://www.alzint.org/about/symptoms-of-dementia/importance-of-early-diagnosis/.

[4] P. K. Mandal and R. V. Mahto, "Deep Multi-Branch CNN Architecture for Early Alzheimer's Detection from Brain MRIs," *MDPI,* vol. 23, no. 19, pp. 4-10, 2023.

[5] M. Khatun, M. M. Islam, H. R. Rifat, M. S. Shahid, M. A. Talukder and M. A. Uddin, "Hybridized Convolutional Neural Networks and Long Short-Term Memory for Improved Alzheimer's Disease Diagnosis from MRI Scans," *arXiv,* pp. 7-10, 2024.

[6] S. M. Lakshmikanthaiah and G. Mamatha, "Optimizing Early Alzheimer's Detection: Evaluating the Performance of SVM, Random Forest, and CNN Models," *IJISAE,* pp. 3193-3197, 2024.

[7] A. Khan, S. Zubair, M. Shuaib, A. Sheneamer, S. Alam and B. Assiri, "Development of a robust parallel and multi-composite machine learning model for improved diagnosis of Alzheimer's disease: correlation with dementia-associated drug usage and AT(N) protein biomarkers," *National Center for Biotechnology Information,* pp. 43-55, 2024.

[8] R. Ahmed, N. Fahad, M. S. U. Miah, M. J. Hossen, M. K. Morol, M. Mahmud and M. M. Rahman, "A novel integrated logistic regression model enhanced with recursive feature elimination and explainable artificial intelligence for dementia prediction," *Science Direct,* vol. 6, pp. 107-116, 2024.

[9] R. E. Kharoua, "Alzheimer's Disease Dataset," May 2024. [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset.